الجـــــــــــمهـــورية الجـــــــزائريـــة الديمقـــــراطيـــة الشـــــــــعبيـــة

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التعليـــــــــــــــــــم العـــــــــــالي والبحـــث العـــــــــــلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جـــــــامعة غـــــردايـــــــة

Université de Ghardaia

كـــــــليـــــة العلـــــــوم والتكنـــــولوجيـــــــا

Faculté des Sciences et de Technologie

قسم الرياضيات والإعلام الآلي

Département des Mathématiques et Informatique

MÉMOIRE

Présenté pour l'obtention du **diplôme** de **MASTER**

En : Informatique

Spécialité : Systèmes Intelligents pour l'Extraction de Connaissances

Par : Aissa REFFIS

Sujet

Réalisation d'un outil d'aide à la modération

M. Slimane BELLAOUAR	Univ. Ghardaia	Président
M. Massoud BATKA	Univ. Ghardaia	Examinateur
M. Abdelkader OULD MAHRAZ	Univ. Ghardaia	Directeur de mémoire

Année Universitaire 2018/2019

DÉDICACE

Je dédie ce travail

À mes parents

À ma famille

À mes collègues

A tous ceux qui m'ont toujours encouragé

REMERCIEMENTS

Tout d'abord, je tiens à remercier le bon Dieu le tout puissant de m'avoir donné la force et le courage de mener à bien ce modeste travail, également je remercie infiniment mes parents, qui mon encouragé

Je tiens à remercier tous ceux et celle qui ont contribué à finaliser ce travail.

Mes remerciements vont à M.OULD MAHRAZ Abdelkader mon encadreur pour m'avoir guidé pour la réalisation de ce projet.

Enfin, je tiens à remercier tous ceux qui m'ont aidé et assisté durant mes études

ABSTRACT

Modern information and communication technologies have extremely changed the content of the Web. Social networks are one of the most profound changes that have happened. Because of its, the user has become today the first producer of content after having been a passive consumer of content in the past.

This new type of content production has many flaws and problems that raise many worries (defamation, extortion, hate speech, ... etc).

The methods that can be used to address these issues can be summarized in user identification, and moderation. Although user identification is a simple process, moderation is a complex, expensive and time-consuming process.

The purpose of this thesis is to propose a tool that helps moderate content. To do this we collect the different techniques cited in the literary can used to verify the quality of textual content generated by the user.

We chose the text classification to elaborate our approach which is based notably on the textual similarities calculation.

Keywords

Moderation, social networks, web content, text classification, textual similarities.

ملخص

لقد غيرت تقنيات المعلومات والاتصالات الحديثة محتوى الويب بشكل كبير. الشبكات الاجتماعية هي واحدة من أعمق التغييرات التي حدثت. فبفضلها أصبح المستخدم اليوم المنتج الأول للمحتوى بعد أن كان مستهلكًا سلبيًا للمحتوى في الماضي.

يحتوي هذا النوع الجديد من إنتاج المحتوى على العديد من العيوب والمشاكل التي تثير الكثير من (المخاوف (التشهير والابتزاز وخطاب الكراهية ...).

يمكن تلخيص الطرق التي يمكن استخدامها لمعالجة هذه المشكلات في تسجيل المستخدم وإدارة المحتوى.

على الرغم من أن تسجيل المستخدم هو عملية بسيطة، إلا أن الإشراف عملية معقدة ومكلفة وتستغرق وقتًا طويلاً .

الغرض من هذه المذكرة هو اقتراح أداة تساعد في إدارة المحتوى. للقيام بذلك، نقوم بدراسة التقنيات المختلفة التي يمكن استخدامها للتحقق من جودة المحتوى النصي الذي تم إنشاؤه بواسطة المستخدم اخترنا تصنيف النص كأساس لمقاربتنا المقترحة، والتي تستند بشكل خاص على حساب التشابه بين النصوص.

كلمات مفتاحيه

إدارة المحتوى، الشبكات الاجتماعية، محتوى الويب، تصنيف النصوص، التشابه بين النصوص.

RÉSUMÉ

Les technologies modernes de l'information et de la communication ont considérablement modifié le contenu Web. Les réseaux sociaux sont l'un des changements les plus profonds qui se soient produits. Grâce à eux, l'internaute est devenu aujourd'hui le premier producteur de contenu après avoir été un consommateur passif de contenu dans le passé.

Ce nouveau type de production de contenu présente de nombreux défauts et problèmes qui suscitent de nombreuses préoccupations (diffamation, l'extorsion, discours de haine, ...etc).

Les méthodes pouvant être utilisées pour traiter ces problèmes peuvent être résumées en identification de l'utilisateur, et la modération. Bien que l'identification des utilisateurs soit un processus simple, la modération est un processus complexe, coûteux et long.

Le but de ce mémoire est de proposer un outil qui aide à la modération de contenus. Pour ce faire on recueille les différentes technique citées dans la littérateurs permettant de vérifier la qualité du contenu textuel généré par l'utilisateur.

On a choisies la classification de texte pour élaborer notre approche qui est basé notamment sur le calcule de similarité des textes.

Mots clé

Modération, réseaux sociaux, contenu Web, classification de textes, similarité des textes.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION GÉNÉRALE

Dans les débuts d'Internet, l'internaute était considéré comme un consommateur passif de contenu, mais il est actuellement le premier producteur de contenu et dispose d'une plus grande capacité de publication de contenu. Cette évolution est principalement due au grand développement quotidien d'Internet, en particulier ceux liés aux réseaux sociaux, qui sont grâce au développement avancé de la technologie Web 2.0, peuvent être les services Internet les plus populaires de nos jours.

Les réseaux sociaux prennent en charge la création et le partage de contenu généré par l'utilisateur, y compris les technologies Web et les téléphones mobiles, susceptibles de faciliter le dialogue interactif entre organisations, communautés et individus. Les réseaux sociaux prennent différentes formes : magazines, forums Internet, blogs de différents types : culturel, social, éducatif, marketing ..., en plus du wiki, et du podcast.

Les internautes génèrent de grandes quantités de contenu. Ce contenu est différent, allant des commentaires de blogs aux sondages en ligne, en passant par les histoires des gens, les dernières nouvelles ...

Le problème est que dans les sites Web traditionnels, la qualité du contenu peut être surveillée, ce qui est une opération relativement simple.

Mais sur les sites de réseaux sociaux, les autorités du site Web qui publie le contenu généré par l'utilisateur est responsable des contributions des utilisateurs et tente de valider le contenu afin d'éviter les problèmes juridiques pouvant découler de ce contenu.

Les méthodes pouvant être utilisées pour traiter les problèmes ci-dessus peuvent être résumées par l'identification de l'utilisateur, la modération. Bien que l'identification des utilisateurs soit un processus simple, la modération est un processus complexe, coûteux et long.

Dans notre mémoire on étudie les techniques permettant de vérifier la qualité du contenu généré par l'utilisateur, en particulière le contenu textuel, et proposer une approche pour ce là.

Ce mémoire est composé de quatre chapitres, le premier chapitre contient les notions et les concepts de base du Web tel que le modèle client-serveur et l'HTML, le deuxième chapitre est consacré à la modération du contenu, ses définitions et ses concepts, et le troisième chapitre regroupe les connaissances nécessaires pour cerner toutes les facettes du sujet telles, le web mining, le web scraping, ainsi que l'apprentissage automatique et les concepts relatifs, et aussi la représentation et la classification du texte.

Le dernier chapitre propose une approche basé sur le calcul de similarité pour la modération des commentaires dans un blog ou un site web et à la fin une conclusion générale.

CHAPITRE 1

LE WEB, DEFINITIONS ET CONCEPTS FONDAMENTAUX

1.1 Introduction

L'internet est un réseau d'ordinateurs à l'échelle planétaire. Ces ordinateurs (Servers, PCs, Smartphones, Objets connectés) interconnectés à travers un protocole qui leur permet de dialoguer (le protocole TCP/IP). Actuellement des services et des normes se sont mis en place concernant Internet pour faciliter les échanges. On peut citer le service FTP pour l'échange de fichiers, le service Mail pour les courriers électroniques. Les derniers services sont les services d'information comme le Web, GOPHER (qui permet de consulter des documents) et WAIS (base de données).

Le Web est aussi un système d'information hypermédia sur Internet qui a donné à l'Internet une forme de cohésion en établissant des liens entre les masses d'informations autrefois dispersées. Il a donné aussi aux utilisateurs de l'internet un outil efficace pour accéder aux documents d'une façon très simple grâce à des interfaces clients conviviales.

1.2 Définition du web

Le Web fait aujourd'hui partie intégrante de notre vie quotidienne, il améliore et facilite notre vie. Le Web facilite la recherche d'information et la communication et minimise les coûts d'envoi des messages. Le Web est un réseau maillé à la base de la page web et dont la structure repose sur la notion de liens. Plus précisément, le Web est un "système d'information multimédia, une page web peut contenir : du Texte (html ...), des images (JPEG, PNG, ...), de la vidéo/audio (mp3, mov, wav ...).. etc. .

1.3 Les caracteristiques du web

Il existe de nombreuses caractéristiques qui rendent le Web si intéressant :

Interactif : L'environnement hypermédia favorise l'interaction. Naviguer dans différents pages d'un site web peut garder l'internaute actif. .

Archivé : la plupart des fichiers permanents des documents et les sessions interactives en ligne sont emmagasinés et disponibles à des fins de recherche. .

Dynamique : il est en évolution chaque jour et les implications de différents contenus mise en ligne sont énormes.

Ouvert : dépend des protocoles et des normes largement acceptés. Actuellement, presque toutes les plates-formes informatiques prennent en considération l'internet. .

Cela facilite pour l'utilisateur le téléchargement de contenu accessible dans le monde entier.

Distribué : les informations sont archivées sur des milliers de serveurs de réseau à travers le monde entier. Cela signifie que quel que soit le lieu où se trouve l'utilisateur, il peut accéder aux mêmes informations.

Anonymat : Les internautes sont cachés derrière des écrans et peuvent rester anonymes [2].

1.4 Le modele client/serveur

L'environnement client-serveur désigne un modèle de communication réseau entre plusieurs programmes, l'un est qualifié de client qui envoie des requêtes. L'autre qualifié de serveur qui attend les requêtes des clients et y répond. Ce modèle diffère de TELNET (le service qui permet une connexion interactive à distance entre deux ordinateurs) ou FTP (service de transfert de fichiers) où le client est inactif. Le World Wide Web fonctionne principalement sur le modelé client/serveur. .

Un serveur web est un programme (résident) qui s'exécute sur un ordinateur dans le but de répondre aux requêtes du logiciel client web exécuté sur d'autres ordinateurs. Ces requêtes peuvent être le transfert d'un fichier ou le résultat de l'exécution d'un programme indépendant du logiciel client : Appel d'API (application programming interface), cette nouveauté est très originale. Parmi ces programmes, on peut citer Webstar sur les Macs et APACHE sur les stations SUN et les PCs.

Un document est une unité fournie par un serveur en réponse à une demande de client. Nous utilisons également le terme ressource. Une ressource peut être un fichier hypertexte, un fichier Text, XML/JSON ou le contenu d'un répertoire obtenu par FTP, un menu GOPHER ou une base de données WAIS.

Un client web est un programme qui permet à un utilisateur de soumettre des requêtes à un serveur web et d'afficher les résultats. Ce programme est également capable d'interagir avec d'autres types de serveurs : FTP, GOPHER, WAIS en utilisant leur propre protocole. Le client web peut être appelé visualiseur ou navigateur (browser en anglais), par exemple Google

Chrome, Mozilla Firefox, Opera… etc

En général, un service est une abstraction des ressources informatiques et le client ne doit pas nécessairement se préoccuper de la façon dont le serveur exécute lorsqu'il répond à la requête et fournit la réponse.

Le client doit seulement comprendre la réponse en fonction du protocole utilisé entre lui et le serveur qui est dans ce cas le protocole de transfert hypertext ou HTTP (HyperText Transfer Protocol) qui est l'ensemble des règles de transfert de fichiers (texte, images graphiques, fichiers audio, vidéo et autres fichiers multimédia) sur le World Wide Web [3] .

Dès qu'un utilisateur Web ouvre son navigateur Web, il utilise indirectement http, qu'est un protocole d'application qui s'exécute sur la suite de protocoles TCP / IP.

Les concepts HTTP incluent (comme l'indique la partie hypertexte du nom) l'idée que les fichiers peuvent contenir des références à d'autres fichiers dont la sélection entraînera des demandes de transfert supplémentaires. Le navigateur Web est un client HTTP qui envoie des demandes aux serveurs. Lorsque l'utilisateur du navigateur entre les demandes de fichier en "ouvrant" un fichier Web (en tapant un URL ou un localisateur de ressources uniformes) ou en cliquant sur un lien hypertexte, le navigateur crée une requête HTTP et l'envoie à l'adresse de protocole Internet (adresse IP) indiqué par l'URL. Le serveur de destination reçoit la demande et renvoie le ou les fichiers demandés associés à la demande.

En conclusion, on a d'un côté le client qui effectue des requêtes en direction du serveur, et de l'autre côté on a le serveur qui exécute ces requêtes et renvoie le résultat au client. Le client et le serveur sont, en pratique, deux logiciels différents communiquant avec le même protocole à travers le réseau. Le rôle d'un programme client est :

- Compiler les messages de l'utilisateur en messages correspondant au protocole Exchange utilisé avec le serveur.
- Contacter le serveur demandé et de lui transmettre la requête désignée.
- Attendre la réponse de la part serveur.
- Mettre en forme la réponse obtenue de serveur et de la présenter d'une façon convenable à l'utilisateur car en fait le serveur renvoie un fichier brut au client, et sa mise en page et de sa mise en forme pour l'utilisateur, sont des tâches de client [4].

1.5 HTML (HyperText Markup Language)

Le langage HTML (l'abréviation de "Hypertext Markup Language" en anglais, en français "langage hypertexte de balisage") est derrière presque tout ce que nous voyons et faisons sur le Web, que ce soit une recherche sur une information, vérifier son compte CCP, ou suivre l'actualité sur un site de nouvelles, ou tout simplement consulter les avis sur sa page Facebook…, on utilise un navigateur, cela signifier qu'on utilise HTML.

Ce langage de balisage sous-jacent du World Wide Web est le langage de présentation de contenu Web. HTML a été inventé par Tim Berners-Lee en 1989. Cette norme n'a cessé d'évoluer et de se développer depuis la version initiale, la plus récente version est HTML5, développée par le World Wide Web Consortium (W3C), et (WHATWG) le groupe de travail sur la technologie des applications hypertextes Web ("WHATWG " est une collaboration non officielle des différents développeurs de navigateurs web ayant pour but le développement de nouvelles technologies destinées à faciliter l'écriture et le déploiement d'applications à travers le Web).

Bien que chaque révision de HTML ait créé de nouvelles fonctionnalités et restructuré d'anciennes, la grammaire de base des documents HTML n'a guère changé au fil des ans et reste relativement stable dans un avenir proche, ce qui en fait l'un des standards les plus importants pour travailler avec et sur le Web.

HTML sert pour baliser le contenu d'une page, c'est à dire pour le structurer et lui donner du sens. Il indique aux navigateurs tel contenu est une image ou une vidéo, et que tel texte doit être considéré comme un paragraphe, un titre, ou un lien.

FIGURE 1.1 – Vue d'un simple document html sur un navigateur [1] .

Les fichiers HTML sont écrits en tant que fichiers de texte brut (Le texte brut, ou texte pur, ou texte simple, "plain text" en anglais est une notion liée à la représentation du texte utilisée entre les dispositifs électroniques. D'où un fichier de texte brut dont le contenu représente uniquement une suite de caractères ; il utilise nécessairement une forme particulière de codage, le code ASCII par exemple.).

Ces fichiers prennent généralement l'extension" .html" (".htm" sur les systèmes d'exploitation qui ne supporte pas plus de 3 caractères de suffixe), ils peuvent être édités et modifiés avec n'importe quel éditeur de texte (par exemple : Bloc-notes, ou Notepad sous Windows ; SimpleText sous Macintosh ; pico sous Linux ; ...).

Les fichiers HTML se composent de deux choses :

- Le contenu : C'est ce que l'internaute voit sur une page web.
- Les balises (Tags en anglais) : chaque balise est écrite entre deux chevrons (angle brackets en anglais).

En d'autres termes les balises sont délimitées par les deux symboles "supérieur à" et "inférieur à", "<. . . > ". À l'intérieur de ces chevrons une séquence de caractères (des parties du texte).

En effet les balises permettent de spécifier quelles parties de document doivent être affichées en tant que titres, les parties qui doivent être organisées sous forme de tableau, les parties contenant des liens, ... et bien d'autres nombreux formulaires et différents formats.

Le navigateur n'affiche pas ces balises, mais affiche leur contenu. Il existe deux types de balises :

— Le premier type : les balises dont le contenu est affiché directement sur la page Web, par exemple <img /> et <input />.
— Le Deuxième type : les balises qui décrit le texte qu'elle contient à l'intérieur ; exemple <p> texte à l'intérieur de la balise </ p>

Cette balise peut contenir une autre balise en tant que sous-élément de cette balise, par exemple <p>texte <i> italique </ i> </ p>.

Les balises indiquent aux navigateurs (et plus particulièrement aux analyseurs syntaxiques) la structure de la page et la fonction de ses différentes parties.

Étant donné que les navigateurs (browsers en anglais) varient d'un à l'autre, il est préférable au concepteur de pré- visualiser sa page avec tous les navigateurs existants et de s'assurer qu'elle s'apparaîtra correctement dans tous les navigateurs, car l'internaute qui visitera la page, va utiliser un navigateur de son choix.

Alors ce qu'on voit dans le navigateur n'est pas le document HTML lui-même, mais une interprétation de celui-ci.

On prend l'exemple suivant :

Les figures 1.1 et 1.2 montrent le même document HTML «OurFirstHTML.html».

La figure 1.1 affiche une version interprétée du fichier comme nous avons l'habitude de voir.

La figure 1.2 montre le code source du document.

FIGURE 1.2 – Le code source d'un simple document html [1].

1.5.1 Les balises HTML

Le code à l'intérieur d'un fichier HTML consiste en lignes de texte entouré de balises. Ces balises indiquent où une mise en forme doit être appliquée, comment cette mise en forme doit apparaître, quelles images doivent être placées à certains endroits, etc

Il y'a même quelques balises permettent de combiner des différents médias tels que des photos, des vidéos ou de la musique entre le texte de la page.

Supposons qu'on veut incliner un mot en particulier, comme par exemple le mot " L'informatique " dans la phrase suivante : ***L'informatique* est un domaine extrêmement vaste et dynamique.**

En HTML, il n'y a pas de bouton sur lequel on clique pour mettre le texte sélectionné en italique, comme dans les logiciels de traitements de texte.

Donc, on doit baliser (taguer) le mot que on souhaite mettre en italique.

Le code pour activer l'italique est <i>, et le code pour désactiver l'italique est </ i> d'où notre code HTML sera comme ceci :

<i> L'informatique</ i> est un domaine extrêmement vaste et dynamique.

On remarque la barre oblique (en anglais slash) "/" dans la balise de fermeture (</ i>). Cette barre oblique différencie une balise d'ouverture d'une balise de fermeture [1].

1.5.2 La syntaxe des balises HTML

Les balises HTML respectent une syntaxe simple, mais rigide :

- Un chevron ouvrant (inférieur à," < ").
- Le nom de cette balise.
- Des attributs (l'attribut est optionnel, il se compose d'un espace, suivi du nom de l'attribut, d'un signe égal = et d'une valeur entre doubles quottes "").
- Un chevron fermant (supérieur à, " > ").

Chaque balise fermante doit avoir le même nom que sa balise ouvrante.

Exemples :

- <title>First HTML</title>
- Link to Homepage

Notons que les balises HTML ne sont pas sensibles à la casse, elles peuvent être saisies en minuscule, en majuscule ou même une combinaison des deux. En d'autres termes, la balise <article> peut être écrite ainsi : <Article>, <ArTicle> ou autrement.

La rigidité syntaxique pour la fermeture des balises est utile afin de pouvoir imbriquer différents éléments les uns dans les autres.

1.5.3 Qu'est-ce qu'un élément HTML ?

Un élément HTML, est le bloc de construction de base du document HTML. Il se compose d'une balise ouvrante, d'un contenu textuel et d'une balise fermante.

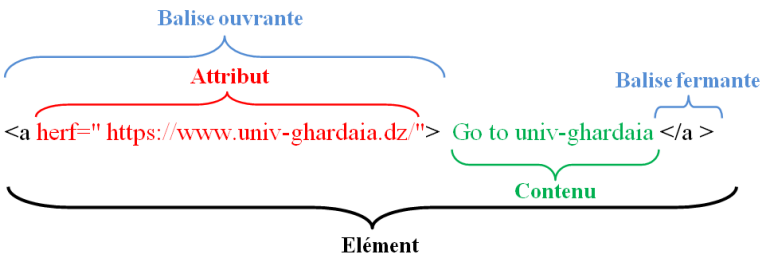


FIGURE 1.3 – Les composants d'un élément html

1.5.4 Les types des balises HTML

Il existe deux types de balise :

— **Balise paire :**

En général, les balises fonctionnent par paires, où la première balise s'appelée "la balise ouvrante" et la seconde s'appelée "la balise fermante".

Ces balises contiennent de texte entre eux, auquel l'effet de cette balise sera appliqué.

Exemple :

<b> UN TEXTE EN GRAS </ b>

Ici <b> est la balise ouvrante, </ b> est la balise fermante et « UN TEXTE EN GRAS »est le texte entre lequel apparaîtra comme UN TEXTE EN GRAS sur l'écran du navigateur.

D'autres exemples de balise paire :

—
—
—
— <h1> </h1> ,...etc

— **Balise impaire :**

Une balise impaire, appelé aussi balises auto-fermantes ou balises vides, est une balise unique ne nécessitant pas de balise associée, elle ne contient pas de texte.

La balise <img> est le plus simple exemple de cela . Le navigateur remplace la balise <img> par l'image indiquée, ignorant le texte de l'élément. Pour cette raison, <img> n'a pas de balise fermante.

Ces balises peuvent être écrites comme <> ou </> les deux fonctionnent de la même manière, le choix du style vous appartient.

Exemple :

—
 ou
— <hr> ou <hr />

1.5.5 Certaines balises parmi les plus fréquemment utilisées

- **<! DOCTYPE> :** pour définir le type de document.
- **<html> :** définit un document HTML.
- **<head> :** fournit des informations générales (métadonnées) sur le document.
- **< title > :** définit un titre pour le document.
- **<body> :** pour définir le corps du document.
- **<h1> à <h6> :** pour définir les différents niveaux de titres pour structurer un (gros) contenu.
- **<a> :** pour créer des liens vers des ressources externes.
- **<p> :** pour définir un paragraphe de texte.
- **
 :** insère un saut de ligne (un retour chariot).
- **<hr> :** définit un changement thématique dans le contenu.
- **<table> :** pour définir une table.
- **<img> :** pour intégrer une image dans un document.
- ** et :** elles permettent simplement de séparer des sections d'un document.
- **<ul>, <ol> et <li> :** pour créer des listes. permet de définir une liste non-ordonnée, <ol> de définir une liste ordonnée, et <li> permettra de définir un élément de la liste.
- **<link> :** définit la relation entre un document et une ressource externe (la plupart du temps utilisée pour créer un lien vers des feuilles de style).
- **<script> :** est utilisé pour intégrer ou faire référence à un script exécutable. Cela fait généralement référence à du code JavaScript.
- **<! –...–> :** pour insérer des commentaires, ils sont écrits avec du texte contenu entre < !– et –>.

1.5.6 Structure globale d'un document HTML

Le document HTML a une structure de base définie par un ensemble de balises spéciales car les éléments définis dans ces balises ne doivent pas apparaître plus d'une fois dans le document (sauf l'élément <title> qui est exclu de cette règle). Le navigateur gérera les cas où ces éléments ne sont pas fournis (ce qui est déconseillé, et non recommandé de faire).

<html> Cet élément indique le langage utilisé pour le document, et entre les balises de cet élément, plusieurs balises sont ouvertes et refermées.

L'élément <html> est l'élément racine du document qui se divise en deux branches, <head> et <body> (chaque document HTML ne contient qu'une seule racine. C'est où tous les autres éléments du document doivent être placés).

<head> Cet élément définit la tête du document, c'est un conteneur pour les métadonnées (données relatives au document HTML. Le navigateur n'affichera pas ces métadonnées, mais il les utilise afin d'améliorer l'ergonomie de la page).

Les métadonnées définissent généralement le titre du document, le codage des caractères, les styles (un style est un ensemble de caractéristiques de mise en forme), les scripts (Un script est chargé d'exécuter une fonction bien précise lorsqu'un utilisateur réalise une action ou lorsqu'une page web est en cours d'affichage) et bien d'autres méta-informations.

Les balises suivantes décrivent les métadonnées : <titre>, <style>, <méta>, <lien>, <script> et <base>.

<body> Cet élément définit le corps du document, il contient tout le contenu d'un document HTML, tel que du texte, des hyperliens, des images, des tableaux, des listes, etc.

L'élément <body> est unique dans un document HTML, et il est toujours placé après l'élément <head>.

Le navigateur affiche tout son contenu, ce qui permet aux utilisateurs de voir tout ce qui est placé dans cet élément.

<title> Cet élément est requis dans tous les documents HTML, c'est le seul élément HTML qu'est obligatoire, il est placé dans la tête du document, et il définit le titre de ce dernier.

Cet élément est l'une des métadonnées exploitées par le navigateur, il l'utilise pour le titre de la fenêtre ou l'onglet, afin de fournir un titre pour la page lorsqu'elle est ajoutée aux favoris, et ainsi que dans les résultats des moteurs de recherche.

1.5.7 Un document HTML formel ou valide

Dans l'exemple suivant un document HTML formel

- !DOCTYPE html >
- <html>
- <head>

— <title>First HTML</title>
— </head>
— <body>
— <!–Content for the user here–>
— </body>
— </html>

Maintenant si on supprime du code source de l'exemple précédent toutes les balises qui sont optionnelles, on aura alors un document HTML valide

— <!DOCTYPE html>
— <title>First HTML</title>

1.5.8 Confirmer la validation d'un document HTML

Les recommandations de bonnes pratiques consistent à réviser et valider son code afin de garantir son bon fonctionnement plus tard.

Il existe pour cela des validateurs du document HTML en ligne, qui analysant le code, en recherchant des balises HTML manquantes ou non équilibrées, des caractères parasites, des identifiants en double, des attributs manquants…etc

Si le validateur trouve une erreur, il l'indicera en affichant des notifications [**?**].

1.5.9 Le DOM HTML

L'analogie avec l'arborescence constitue un bon moyen de décrire les multiples couches d'un document HTML. C'est-à-dire qu'on représente la structure d'un document html à l'aide d'un arbre. On parle d'arbre DOM (Document Object Model) du document.

Cette forme de DOM permet aux langages de programmation comme par exemple le JavaScript d'accéder au document et de manipuler le contenu et les styles. Cet arbre DOM se construit comme suit :

— On parcourt le document séquentiellement et chaque élément est un nouveau nœud.
— Si l'élément <élé2> est imbriqué dans l'élément <élé1>, le nœud élé2 est fils du nœud élé1.

— Tous les nœuds des éléments qui se trouvant au même niveau (c-à-dire tous les éléments qui sont imbriqué au même élément) sont frères, et Ils se représentent sur l'arbre « de gauche ‘a droite » selon leur ordre d'apparition.

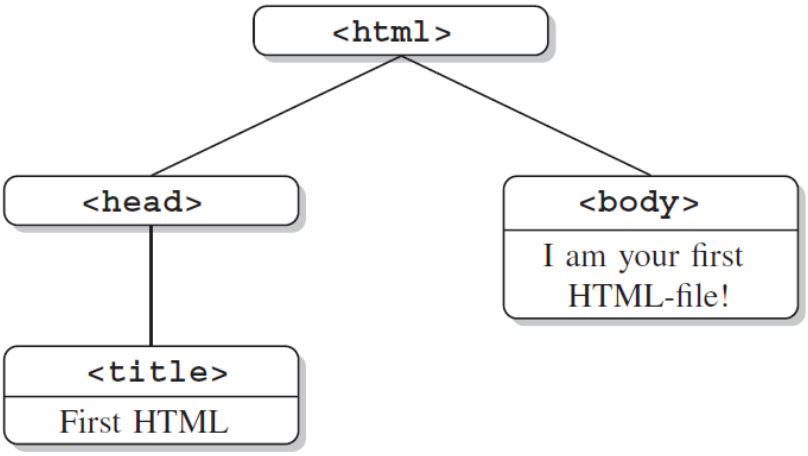


FIGURE 1.4 – L'arbre DOM correspondant au document html dans la figure 1.2 [1].

1.6 Conclusion

Dans ce chapitre, on a présenté les différents aspects du web qui est un réseau maillé dont l'entité de base est la page web. Les pages Web sont un moyen facile de naviguer sur Internet, Elles permettent d'effectuer une infinité d'actions. Le web est en évolution continue. Le navigateur un logiciel qui permet de parcourir un site web, le service web se base principalement sur le model client-serveur, dans le chapitre suivant, On va présenter l'activité de modération.

CHAPITRE 2

MODERATION DU CONTENU

2.1 Introduction

Depuis la naissance du Web, mais surtout depuis la montée des réseaux sociaux, la modération de contenu est devenue l'une des fonctions les plus importantes et les plus confidentielles de l'industrie de supervision.

Une armée de modérateurs autour du monde à Manille et à Bangalore, et même aussi aux États-Unis, au Canada et en Europe cible la violence, le terrorisme, la pédophilie à travers l'écran, discours de haine et des montagnes d'autres contenus illicites, illégaux ou inappropriés. Les estimations varient beaucoup, mais la plupart indiquent que plus de 100 000 personnes modèrent le contenu à l'échelle mondiale.

L'intelligence artificielle offre la promesse et la menace de pouvoir prendre en charge une grande partie du travail, Sur certaines plateformes sociales, l'intelligence artificielle est plus capable que les gens à détecter des images indésirables. D'autres travaillent sur une intelligence artificielle capable de comprendre ce qui se passe dans une vidéo en direct. D'autres travaillent encore pour identifier le type de harcèlement et d'intimidation, cela nécessite une compréhension contextuelle plutôt que de simplement identifier un mot ou une image inappropriés. Il ne fait aucun doute que la façon de modérer le contenu numérique change et développe chaque jour.

Contrairement aux prévisions selon lesquelles l'intelligence artificielle supplantera toute modération humaine, ACCENTURE [7] estime que l'avenir est construit sur une relation synergique entre les modérateurs humains et l'intelligence artificielle [6].

2.2 C'est quoi un modérateur sur internet ?

Les personnes qui sont chargés de modération sur internet sont des internautes appelés modérateurs ou modos dont le rôle est de surveiller la teneur des messages échangés et distribués entre les internautes ou laissés un commentaire sur un site web disposant d'un espace de commentaire (les réseaux sociaux, les forums, les réactions aux articles publiés, les sites de jeu vidéo ,etc…), Les modérateurs possèdent des outils supplémentaires lorsqu'ils parcourent le forum pour mener à bien leurs tâches, la politique de modération est explicite dans la charte d'utilisation (voir la paragraphe du titre 2.5). Les modérateurs préviennent les utilisateurs, contrôlent leur langage, suppriment les messages qui n'ont pas leur place sur le forum, soit

étant donné qu'ils contreviennent à la loi, soit qu'ils désobéissent les règles explicites ou même les règles implicites du forum [?] .

2.3 Quelle est la différence entre un administrateur et un modérateur dans un groupe ?

Deux rôles sont disponibles pour les individus qui gèrent des groupes : administrateur et modérateur. Ces deux rôles sont présentés dans le tableau suivant, ainsi que les habilitations correspondantes à chaque rôle [9] .

	Administrateur	Modérateur
Attribuer le rôle d'administrateur ou de modérateur à un autre membre	√	
Gérer les paramètres du groupe (par exemple, changer le nom, la photo de couverture ou les paramètres de confidentialité du groupe)	√	
Approuver ou refuser les demandes d'adhésion	√	
Approuver ou refuser les publications dans le groupe	√	√
Supprimer les publications et les commentaires sur les publications	√	√
Supprimer et exclure des personnes du groupe	√	√
Épingler ou détacher une publication	√	√

2.4 Les tâches du modérateur

Avec l'émergence des forums et des blogs et l'explosion des réseaux sociaux le rôle du modérateur est devenu fondamental, En effet, le modérateur est là pour garder le bon maintien des échanges sur le site. Il a une double fonctionnalité, d'une part contrôler les échanges du côté verbal (langage, disputes, insultes) et d'autre part garantir le bon déroulement de son forum.

En générale, tous les sites internet qui proposent une interaction entre ces membres, on trouve une équipe de modération chargé de prendre garde au respect des règles et à la conservation de la bienséance. Pour atteindre ce but, cette équipe a besoin d'avoir à sa disposition des outils de contrôle dont seul les administrateurs et les modérateurs peuvent y avoir accès.

Le modérateur du site doit d'être le plus possible présent, et il doit lire pas juste les sujets et les commentaires qui l'intéressent mais l'intégralité de ce qui est publié sur le site web qu'il surveille. Il doit vérifier que toutes les données (messages, images, contenu d'une séquence vidéo, etc.) publié sur un site internet, un forum ou un blog, sont conformes au règlement et respecte la charte du site. Il fait attention que tous les échanges entre les utilisateurs soient poli et aimable et il se réserve du droit de modifier ou même supprimer tous contenu inapproprié. Le modérateur est le garant de l'image du site internet qu'il surveille, il rend possible à la communauté des abonnées d'un site d'échanger d'une manière sereine sur le sujet qu'ils choisissent. Le modérateur peut pareillement intervenir sur les media sociaux, dans ce cas il sera gestionnaire d'une page ou d'un groupe et il doit filtrer les publications des participants.

Le modérateur est chargé aussi de :

- Créer le plan de modération.
- Ecrire et afficher le règlement intérieur du site conformément aux conditions d'utilisation.
- Surveiller le comportement des utilisateurs sur le site, y compris la détention (ou le blocage ou l'interdiction des utilisateurs) si et quand cela est nécessaire.
- Editer ou supprimer en fonction du type de modération sélectionné (pré, post ou réactif) et conforme aux règles éditoriales, aux conditions d'utilisation et aux règles de charte.
- S'efforcer de maintenir l'espace interactif sécurisé pour tous les utilisateurs.
- Surveiller le langage et le ton des contributions pour que cet espace reste attractif pour les utilisateurs potentiels. Ils doivent veiller à ce que la langue des messages publiés et des conversations ne devienne pas abusive, agressive, intimidante, ou contienne des

attaques personnelles inappropriées.

- Agir en tant qu'hôtes Web qui publient des messages : de bienvenue et de salutations, et de clôture remarques.
- Publier des "avis de modérateur" dans des articles et des discussions rappelant aux utilisateurs les conditions d'utilisation, les règles de charte du site, ou des problèmes spécifiques de modération.
- Apporter des réponses aux membres en besoin de réponses.
- Etre neutre dans tous les cas (car son poste lui exige de poser ses propres opinions à côté), sociable, et surtout présent.
- Maintenir de très bonnes relations avec tous les membres d'une part, et d'autre part il faut avoir une certaine sévérité et fermeté avec eux, de manière à prévenir toute violation sans avoir besoin de passer à l'acte.
- Les modérateurs ne sont pas censés de communiquer régulièrement les raisons de la modération, de la suppression ou de la modification à des membres du forum. Cependant, ils devraient envisager de conserver des enregistrements pour les occasions où les décisions de modération sont remises en question par les utilisateurs ou font l'objet d'une plainte formelle. Il peut être approprié de conserver des notes de dossier sur des incidents significatifs ou lorsqu'un membre de l'audience a continué à ignorer les règles publiées de la charte, les conditions d'utilisation, etc. Ces notes peuvent inclure des captures d'écran.

Notons que le modérateur n'est pas présent pour un objectif répressif ou censuré, alors qu'il est présent à titre préventif : il fournit avant tout une assistance, améliore la qualité des échanges et punit comme dernier recours [10] [11] .

2.4.1 Le Plan de modération

Les modérateurs doivent développer un plan de modération, Cela indiquera :

- Qui est responsable éditorial du contenu publié.
- Les ressources, y compris les détails de la modération à différents moments de la journée, de la semaine et de l'année.
- Durée de vie prévue du contenu généré par l'utilisateur.
- Les modérateurs et les intervenants.
- Approches de modération spécifiques adoptées pour des périodes différentes.

— Déterminer les règles de la charte spécifiques aux besoins du site.

Si l'approche de modération suit une modération post-réactive ou réactive, le plan peut indiquer les retards maximums probables dans l'examen des contributions ou des alertes des membres [10] .

2.4.2 Les processus de modération

— Un utilisateur publie une nouvelle participation sur le site (message, article, image, vidéo...).
— Une fois la nouvelle participation de l'utilisateur et validée, le système va alerter le modérateur (par un email par exemple) qu'une nouvelle publication est disponible.
— Le modérateur va décider via l'interface de modération (une interface spéciale au modérateur) si la nouvelle participation postée par l'utilisateur sera acceptée où rejetée.
— Selon le site web, parfois l'utilisateur est prévenu de la décision du rejet du modérateur par un message déterminant le motif [10] .

2.5 La charte de modération

Pour que le modérateur fasse son travail, il trouvera la politique à suivre et les grands principes de modération bien expliqués et bien indiqués dans la charte d'utilisation du site.

Cette charte détermine la ligne directrice de tous ce qui est permis, interdit ou enduré sur le site. Le modérateur la considère comme son document de référence, elle lui garantit sa légitimité dans le site surtout concernant ces ingérences qui peut être considérée comme une « violation de la liberté d'expression ».

La charte contient particulièrement tous les termes, tous les propos, et toutes les expressions à interdire : insultants, racistes, dégradants ou pouvant d'une façon ou d'une autre porter une préjudiciable à la sécurité des autres. Sur cette base, le modérateur a complètement le droit de refuser de poster un commentaire à caractère violent ou d'informer le membre que son commentaire ne sera pas posté.

Puis la charte indique le rôle exact du modérateur sur le site dont il est responsable d'assurer que les règles de la charte seront appliquées à sa communauté modérée, tout en gardant le maximum d'espace d'expression possible.

L'administrateur est généralement régi par une charte, publiée dans des forums, des sites et des blogs, afin que les choses soient entièrement transparentes et claires pour toutes les parties. Sa mission est de respecter les utilisateurs indifféremment, de les aider en toute indulgence, et de gérer les alertes émis par les utilisateurs du site à propos d'une publication posté par un membre.

Parmi les informations pertinentes de la charte c'est de déterminer le type et la méthode de modération utilisée sur le site web.

Dans certains cas non spécifiés dans la charte, les modérateurs se consultent pour déterminer si un message doit être supprimé [12] .

2.6 Les types de modérations

Trois types de modération sont utilisés :

2.6.1 La modération à priori

Dans cette approche, toutes les publications sont modérées avant de paraître en ligne. La modération à priori permet au modérateur de gérer des sites où le risque lié à la publication d'un contenu inapproprié est élevé. Cela peut inclure des sites d'informations et d'actualité, des discussions sur des sujets sensibles ou controversés, des sites destinés aux enfants ou des sites invitant à soumettre des photos, des vidéos ou des fichiers audio.

Les règles de la charte sur certains sites modérés à priori peuvent indiquer que toutes les contributions et tous les messages soumis pendant la nuit ou le week-end ne peuvent pas apparaître en ligne avant le prochain jour ouvrable.

Le choix de la modération à priori n'est pas disponible sur la plupart des sites web, y compris Facebook.

2.6.2 La modération a posteriori

A l'inverse de l'approche précédente, dans cette approche toutes les publications apparaissent immédiatement en ligne, ce qui permet aux utilisateurs d'avoir une conversation fluide sans être gênés par les retards qui surviennent lorsque chaque publication doit être approuvée.

Cela signifie que toutes les contributions et tous les messages soumis au site sont modérés après leur apparition en ligne.

La modération a posteriori peut être le choix éditorial le plus approprié dans les cas suivants :

- Quand il y a peu de risque que le sujet peut générer contenu offensant ou abusif.
- Le risque que les utilisateurs publient du contenu inapproprié est considéré faible.
- Quand il y a des ressources adéquates pour la modération sont disponibles pour garantir un examen rapide et opportun de toutes les soumissions.

Le rythme de révision reflétera la nature du contenu, le public en ligne et le risque que du contenu inapproprié reste en ligne pendant une longue période.

Dans certaines circonstances, il peut être approprié de revenir à la modération a posteriori la nuit ou le week-end [10] .

2.6.3 La modération réactive

Dans cette approche, toutes les contributions apparaissent immédiatement en ligne. Les publications ne sont modérées qu'après le modérateur reçoit une alerte d'un utilisateur. Cette forme de modération est souvent utilisée conjointement avec la modération a posteriori.

La modération réactive peut être appropriée lorsqu'une communauté particulière a démontré un haut degré de responsabilité et une capacité continue à se gérer. Il est plus susceptible de convenir à des sujets qui ne tendent pas à attirer des réponses polarisées ou extrêmes.

La décision concernant le type de modération approprié est basée sur des facteurs tels que le public cible, l'historique du comportement et de la gestion du public, la nature des sujets, la disponibilité des ressources de modération, la durée pendant laquelle un modérateur répond à une alerte d'un utilisateur.

Chaque fois que les alertes sont parvenus, les responsables de la rédaction doivent s'assurer qu'il existe un système bien géré d'accès et de gestion des alertes dès leur envoi par courrier électronique. Le modérateur doit surveiller le compte de messagerie régulièrement et réagir rapidement aux alertes. Une bonne pratique consiste à éviter les retards prolongés. De tels retards pourraient avoir des conséquences juridiques.

Les responsables éditoriaux et les modérateurs devraient élaborer un plan de modération permettant de surveiller et de gérer les alertes reçues en temps voulu.

Notant que ce type de modération est appelé aussi la Modération distribuée.

La modération de l'utilisateur

Cette approche consiste à permet à chaque utilisateur de modérer les participations de tous les autres utilisateurs. c'est une méthode fonctionne très bien dans la modération des sites web ayant une forte population active (par exemple le site d'actualités en langue anglaise "Slashdot").

L'idée c'est d'attribuer à chaque utilisateur modérateur un nombre limité de points dit "points de modération", chacun de ces points peut être utilisé afin de modérer un commentaire vers le haut ou vers le bas d'un seul point à la fois.

Les commentaires accumulent ainsi un score, qui est en plus lié à la plage de -1 à 5 points. Lors de la consultation du site, un seuil peut être choisi à partir de la même échelle et seuls les commentaires atteignant ou dépassant ce seuil seront affichés.

2.6.4 La Modération automatique

Cette approche de modération diffère des approches précédentes, car dans ce type il n'implique aucune intervention humaine. Il consiste à déployer divers outils techniques (principalement des filtres) pour traiter le contenu généré par l'utilisateur et appliquer des paramètres prédéfinis.

Les commentaires seront afficher ou non selon des règles de rejet ou d'approbation, dans ce qui suit quelques unes :

L'un de ses outils les plus couramment utilisés est le filtrage de mots, dans lequel une liste de mots interdits est prédéfinis et une fois l'outil trouve un mot interdit dans commentaire, il le supprime ou le remplace par une alternative spécifique, ou même il bloque ou il rejette le commentaire complètement.

Un outil similaire est la liste de blocage IP qui supprime les liens externes inappropriées, ou supprime le contenu provenant d'adresses IP interdites.

Il existe d'autres filtres plus sophistiqués pour la modération automatisée, cette modération globale implique un coût initial, mais ne comprend presque aucun coût opérationnel [13] .

2.7 Conclusion

Les technologies modernes de l'information et de la communication ont considérablement modifié le contenu du web. Les réseaux sociaux sont l'un des changements les plus profonds qui aient eu lieu. Chaque utilisateur a maintenant la possibilité de devenir un producteur de contenu. Ce nouveau type de production de contenu présente de nombreux défauts et problèmes qui suscitent de nombreuses préoccupations (diffamation, discours de haine, propriété intellectuelle...etc).

La solution à ces problèmes est la gestion correcte et efficace du contenu généré par l'utilisateur. Cela peut être réalisé en enregistrant les utilisateurs qui contribuent au contenu. Ce processus d'enregistrement est bien connu aux utilisateurs car il est utilisé depuis de nombreuses années dans de nombreux services Internet (services de messagerie, réseaux sociaux, ...etc.).

Mais la modération peut prendre beaucoup de temps et nécessite beaucoup de ressources humaines. Il serait donc très utile de fournir des outils informatiques afin de garantir l'efficacité de cette tache de modération, toute en réduisant les couts couts matériels et temporelles.

Dans ce contexte, nous parlerons dans le prochain chapitre du web mining notamment le web content mining, ses taches et ses outils, surtout le scraping.

En suit on parlera de la classification de textes et ses concepts de bases, et quelques outils de l'apprentissage automatique utilisés pour ce là.

CHAPITRE 3

WEB MINING ET CLASSIFICATION DE TEXTES

3.1 Introduction

Dans de nombreux domaines, des décisions décisives sont nécessaires, parfois dans un contexte difficile et dans un temps limité.

Par exemple, si un médecin est face à une situation urgente, il doit prendre une décision rapide pour traiter le cas, en utilisant ses connaissances et ses expériences pour prendre sa décision. Mais il est incapable de se souvenir de tous les cas qu'il a traités et de tous les dossiers qu'il a étudiés depuis des années.

Dans ce cas, l'informatique peut donner une aide précieuse à travers ses outils, parce qu'ils peuvent prendre en considération un grand nombre de cas déjà traités et proposer à un nouveau cas une décision basée sur l'agrégation de tous les cas précédents.

À la base de ces précieux outils, nous trouvons, l'apprentissage automatique, la fouille de données (data mining), et la fouille du web (web mining) [14].

3.2 L'apprentissage automatique

L'apprentissage automatique fait référence au développement, à l'analyse et à l'implémentation de méthodes permettant à l'appareil (au sens large) d'évoluer grâce à un processus d'apprentissage, réalisant ainsi des tâches difficiles, voire impossibles, à accomplir avec des moyens algorithmiques plus conventionnelles.

Afin de permettre au logiciel de générer des solutions de manière indépendante, l'action préalable des utilisateurs est nécessaire.

Par exemple, les algorithmes et les données nécessaires doivent être introduits à l'avance dans les systèmes, et les règles d'analyse respectives pour la reconnaissance des modèles à partir des données doivent être définies. Une fois ces deux étapes terminées, le système peut effectuer les tâches suivantes par l'apprentissage automatique :

- Rechercher, extraire des données pertinentes.
- Faire des prédictions basées sur les données d'analyse.
- Calculer des probabilités pour des résultats spécifiques
- S'adapter à certains développements de manière autonome.
- Optimiser les processus en fonction de modèles reconnus.

— Résoudre des problèmes de classification.

«L'apprentissage dénote des changements dans un système qui ... lui permet de faire la même tâche plus efficacement la prochaine fois» Herbert Simon

«L'apprentissage automatique est le domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être programmé explicitement .» Arthur Samuel [15].

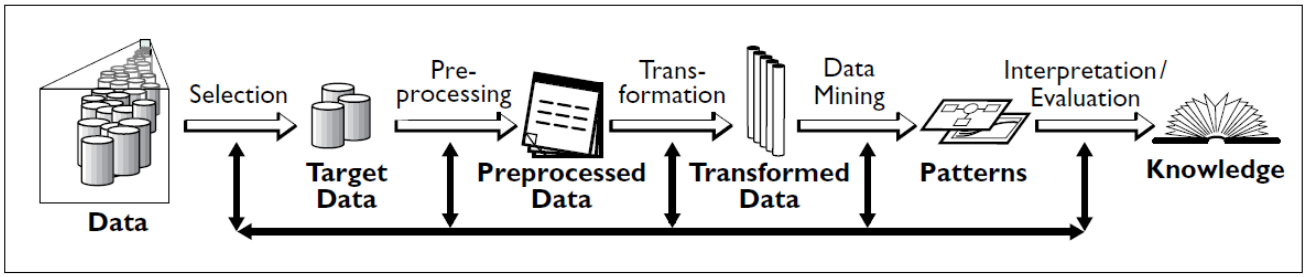


FIGURE 3.1 – l'apprentissage automatique et le data mining au centre de processus d'extraction de connaissances à partir des données [16] .

3.2.1 Algorithme d'apprentissage

Algorithme d'apprentissage est un algorithme prenant en entrer un ensemble de donné D (cette ensemble s'appelle : "l'ensemble d'entrainement" ou "l'ensemble d'apprentissage") et retourne une fonction F (modèle, patron).

On dit que le modèle obtenue F a été entrainer sur l'ensemble D.

L'ensemble D, contient sous forme de vecteur l'information nécessaire pour résoudre un problème (par exemple classification).

L'algorithme d'apprentissage permet l'adaptation automatique à la réalisation d'une tâche où chaque algorithme est connue (en quelque sorte destiné) pour fonctionner sur certain type de problème et pour réaliser certain tache et non pas sûr d'autre.

Le choix d'un algorithme d'apprentissage dépend généralement de l'objectif et du domaine de l'application. Par exemple les applications conçu pour le domaine médicale nécessite des algorithmes d'apprentissage les plus précis que possible qui différent des algorithmes d'apprentissage utilisés pour concevoir un moteur de recherche dans un site web d'une librairie, et dépond aussi de la tâche à résoudre (classification, estimation de valeurs, ...etc).

Notant qu'il existe toute une gamme d'algorithmes, pour chaque tâche spécifique.

3.2.2 Méthode d'apprentissage automatique

L'apprentissage supervisé et l'apprentissage non supervisé sont deux des méthodes d'apprentissage automatique les plus répandues, mais il existe également d'autres méthodes d'apprentissage automatique. Voici un aperçu des types les plus connus.

L'apprentissage supervisé

Dans l'apprentissage supervisé, l'ordinateur est fourni avec des exemples d'entrées qui sont étiquetés avec les sorties souhaitées. Le but de cette méthode est que l'algorithme puisse «apprendre» en comparant sa sortie réelle avec les sorties «enseignées» pour trouver des erreurs et modifier le modèle en conséquence. L'apprentissage supervisé utilise donc des modèles pour prédire les valeurs d'étiquettes sur des données non étiquetées supplémentaires.

Par exemple, avec un apprentissage supervisé, un algorithme peut être alimenté avec des images de requins étiquetés Poisson des images d'océans étiquetés comme Océan. En étant formé sur ces données, l'algorithme d'apprentissage supervisé devrait être capable d'identifier plus tard des images de requin non marquées comme Poisson des images océaniques non étiquetées Océan. Un cas d'utilisation de l'apprentissage supervisé consiste à utiliser des données historiques pour prédire des événements futurs statistiquement probables. Il peut utiliser les informations historiques sur les marchés boursiers pour anticiper les fluctuations à venir ou être utilisé pour filtrer les courriers indésirables [17].

Une brève définition : L'apprentissage supervisé c'et à partir de l'observation de données (Xi ; Yi). Tel que :

$$Yi = F(Xi) + e\ i$$

Où F est la fonction cible (inconnue), estimer F afin de faire des prédictions de F(X).

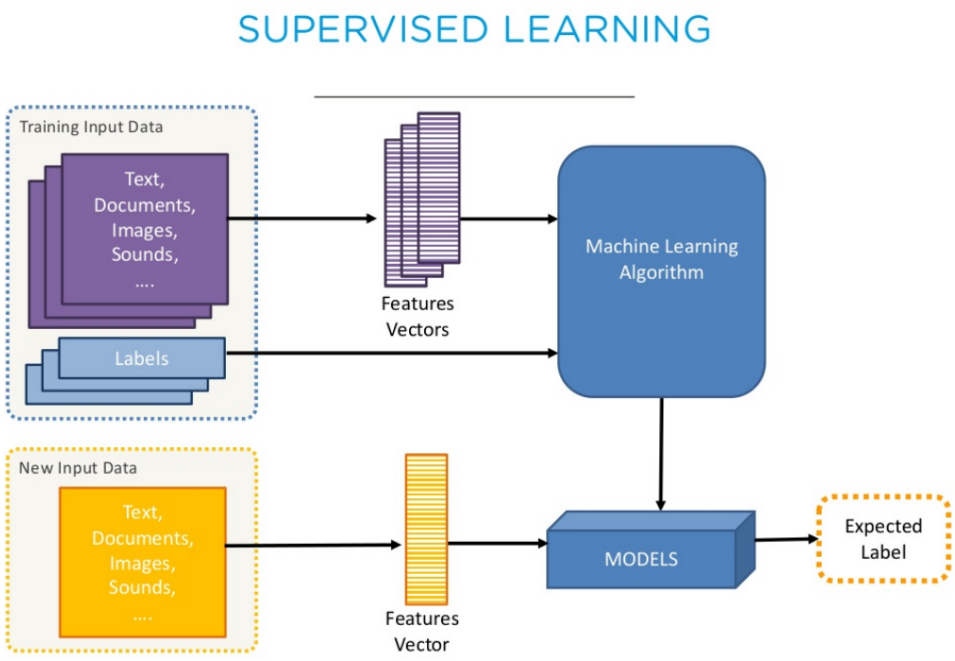


FIGURE 3.2 – L'apprentissage supervise est ses étapes [18] .

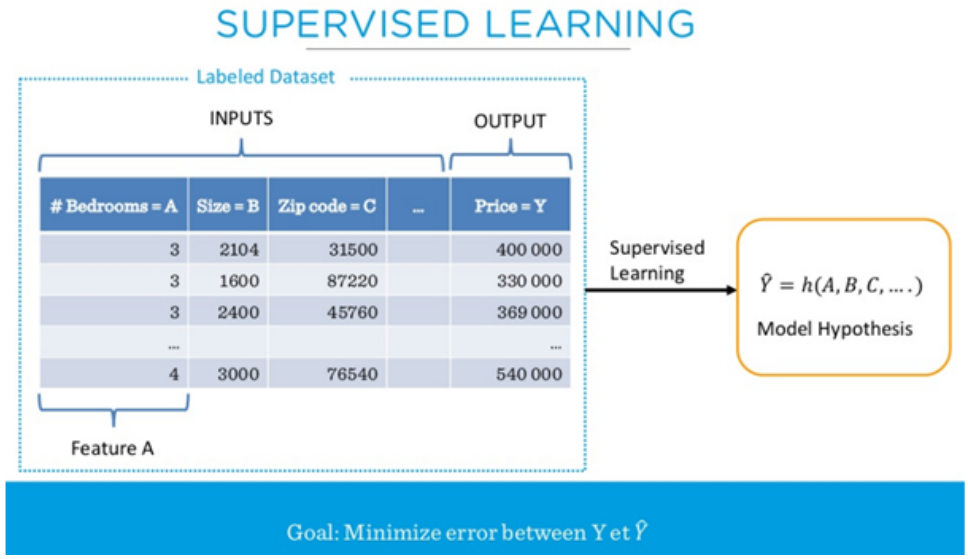


FIGURE 3.3 – Illustration d'une base de données labellise (étiquetées) [18] .

Les problèmes résolus par l'apprentissage supervisé On distingue en général deux grands types de problèmes auxquels l'apprentissage supervisé est appliqué : la classification supervisée (catégorisation), et la régression.

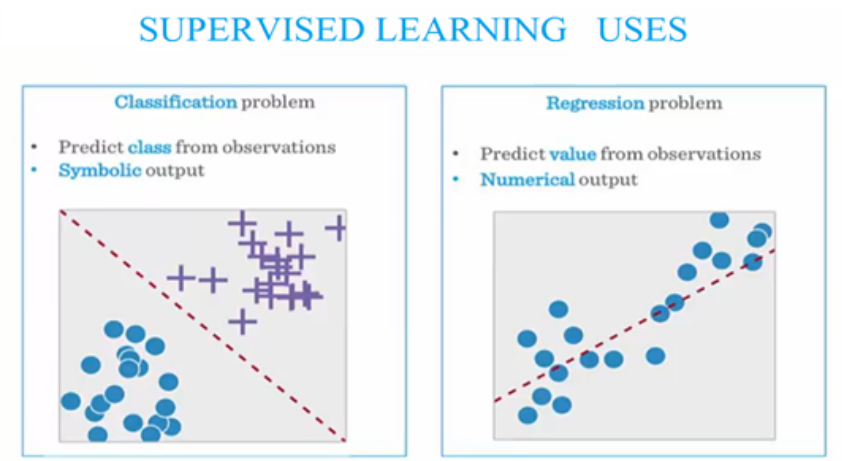


FIGURE 3.4 – L'apprentissage supervise s'applique sur la classification et la régression [18] .

Exemples :

Catégorisation de texte (classification supervisé), elle permet de prédire si un texte est membre d'un groupe ou d'une classe prédéfinie .C'est la classification de chaque texte $x \in X$ parmi un ensemble de classe préexistantes (connu à l' avance)

Le principe général c'est de chercher un modèle de prédiction estimé par un apprentissage supervisé.

Ce typer de classification peut être utilisé dans plusieurs domaines par exemple Filtrage anti-spam : identifier les e-mails comme spam ou non-spam Diagnostic médical : diagnostiquer un patient comme un malade ou non malade d'une maladie Prévision météo : prédire, par exemple : si il va pleuvoir demain ou non [17] .

L'algorithme de la sélection négative

Les systèmes immunitaires artificiels (AIS) sont un type d'algorithme intelligent inspiré des principes et processus du système immunitaire humain. Parmi ces algorithmes l'algorithme de la sélection négative, qu'est un algorithme d'apprentissage supervisé introduit par Forrest et al [19] .

L'idée sur laquelle se base cet algorithme est que seules les cellules T qui ne s'attaquent pas aux cellules du soi sont autorisées à quitter le thymus et auront pour tâche de reconnaitre les cellules du non soi.

Cette notion est très intéressante, surtout pour les applications de surveillance des systèmes et la détection d'utilisations anormales ou inhabituelles.

L'algorithme de la sélection négative qui reflète ce principe est un processus de détection d'anomalies composé de trois phases principales :

- La définition du soi.
- La génération des détecteurs.
- Le contrôle d'occurrence des anomalies.

L'algorithme se déroule comme suit :

- On considère un ensemble donnée de modèles de soi (P).
- Générer un ensemble de détecteurs (M) qui n'identifient aucun élément appartenant a l'ensemble (P).
- La génération des détecteurs (M) se passe par les étapes itératives décrites comme suit :

Calculer l'affinité entre chaque cellule C et tout l'ensemble de soi P.

Si l'affinité entre un élément C et au moins un élément P est supérieur ou égal a un seuil d'affinité prédéfini, alors cet élément C sera supprimé (il est considéré comme un élément de soi) .

Sinon il sera considéré comme un détecteur de non soi et sera ajouté à l'ensemble de détecteur M

Après avoir obtenu l'ensemble de détecteur, la prochaine étape sera de détecter la présence du modèle de non soi [20] .

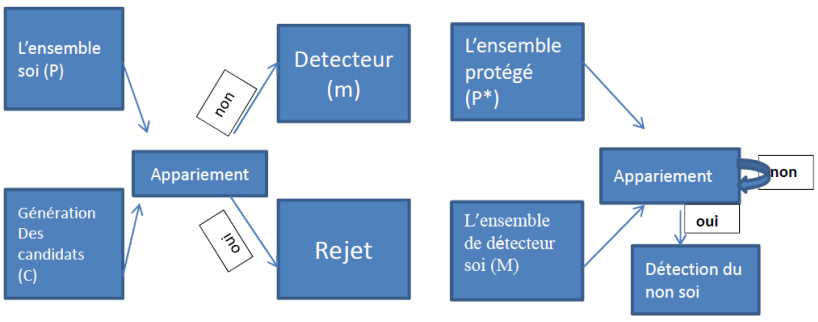


FIGURE 3.5 – La structure générale de l'algorithme de la sélection négative [20] .

L'apprentissage non-supervisé

En revanche, l'apprentissage automatique non supervisés est utilisés lorsque les données utilisées pour entraîner ne sont ni classées ni étiquetées. L'apprentissage non supervisé étudie comment les systèmes peuvent inférer une fonction permettant de décrire une structure cachée à partir de données non étiquetées. Le système ne trouve pas forcement le bon résultat, mais il explore les données et peut tirer des déductions à partir de jeux de données pour décrire les structures masquées à partir de données non étiquetées.

Autrement dit, c'est l'algorithme l'apprentissage qui doit découvrir par lui-même les similitudes et divergences entre les données et vise à caractériser la distribution des données, et les relations entre les variables.

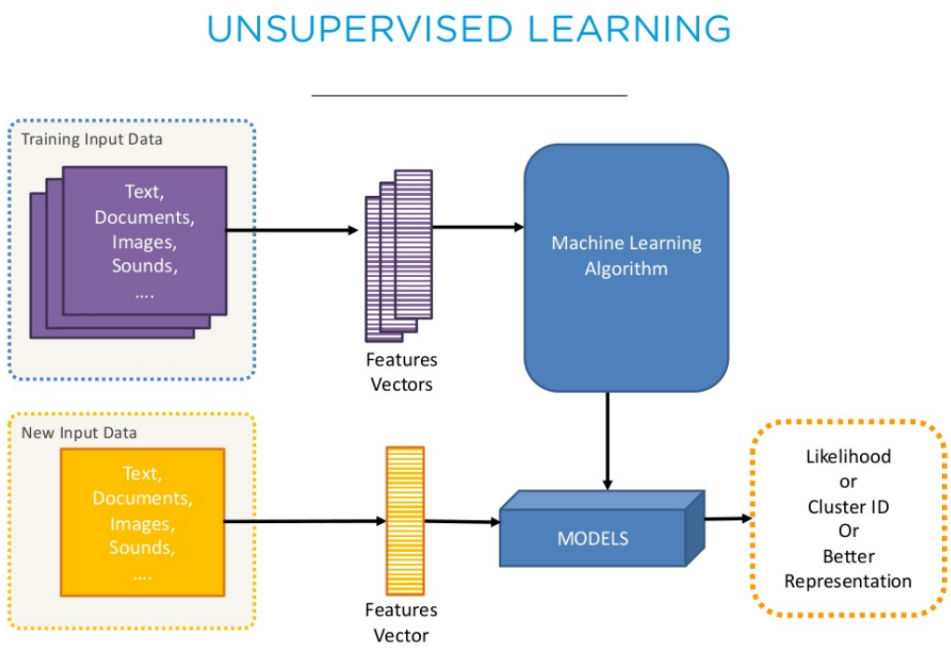


FIGURE 3.6 – l'apprentissage non-supervise est ses étapes [18] .

la figure ci-dessous présente un algorithme d'apprentissage automatique qui reçoit trois entrées X1,X2,X3 et fournit en deux sorties Y1 et Y2 (X1 et X2 sont étiquette par Y1, et X3 par Y2).

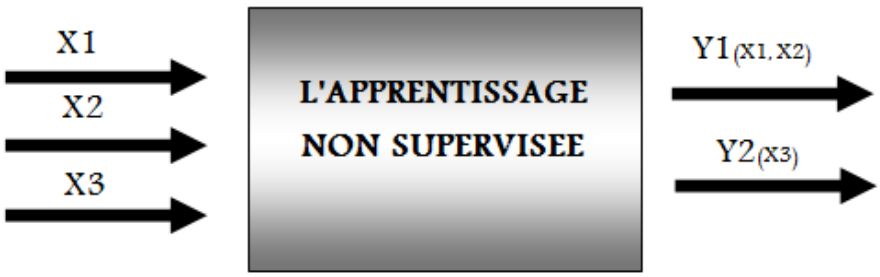


FIGURE 3.7 – le fonctionnement d'apprentissage non supervise

Les Problèmes résolus par l'apprentissage non supervisé L'apprentissage non superviser peut être adapté a des problèmes comme :

- Le Clustering. (Rassembler les données qui peuvent se ressembler).
- L"association, après avoir Extraire des caractéristique d'association, ça permet de découvrir des règles de regroupement. Comme le fait les moteurs de recommandation.
- Réduction de dimensionnalité.
- La détection d'anomalies.

Exemples

- Dans la segmentation client, on a un panel de client, on veut le diviser en des groupes, Mais on ne sait pas sur quels critères. Alors l'algorithme non supervisé est très fort pour faire ça.
- Pour faire de la reconnaissance d'image. Quand on prend un Smartphone pour faire une photo, un petit carré qui détecte les visages, là typiquement on est dans du non supervisé.
- Dans les systèmes de recommandation [17].

La classification non-supervisé (clustering) La classification non-supervisé (clustering en anglais) où pas de classes prédéfinies, c'est tous simplement la recherche d'une typologie, ou d'une segmentation, c'est-à-dire d'une partition, ou une répartition des individus en classes homogènes, ou en catégories.

Cela est fait grâce à l'optimisation des critères visant à regrouper les individus dans des classes, à l'intérieure de chacune de ces classe le plus homogène possible (les éléments similaires au sein d'un même ensemble), mais entre les classes au contraire, les plus distinctes possible (les éléments dissimilaires quand ils appartiennent à des ensembles différents).

Qu'est ce qu'un bon regroupement ? Une bonne méthode de regroupement doit garantir :

- Une grande similarité intra-groupe.
- Une faible similarité inter-groupe.

Donc la qualité d'un regroupement, dépend de la mesure de similarité utilisée par la méthode et de son implémentation.

Comment mesurer la qualité d'un cluster Par un métrique pour la similarité. La similarité est exprimée par le biais d'une mesure de distance.

Les définitions de distance sont très différentes que les variables soient des intervalles (continus), catégories, booléennes ou ordinales.

En pratique, on utilise souvent une pondération des variables.

Plusieurs choix sont laissés à l'initiative de l'utilisateur :

- Une mesure d'éloignement (dissemblance, dissimilarité ou distance) entre individus.
- Le critère d'homogénéité des classes à optimiser : il est, dans le cas de variables quantitatives, généralement défini à partir de la trace d'une matrice de variances-covariances.
- La méthode : classification ascendante hiérarchique, ré-allocation dynamique et DBSCAN sont les plus utilisées, seules ou combinées.
- Le nombre de classes : c'est un point délicat. Enfin, différents outils recherchent une interprétation, ou des caractérisations, des classes obtenues [17].

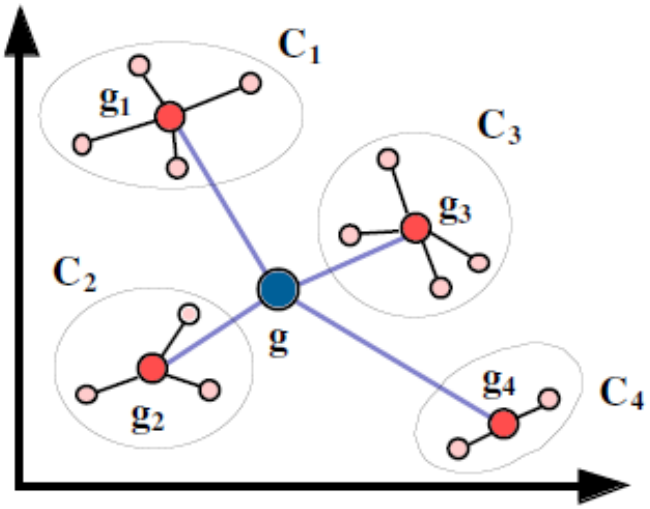


FIGURE 3.8 – Une bonne partition par minimiser l'inertie intra-cluster et maximiser l'inertie inter-cluster [18] .

L'apprentissage semi supervisé

L'apprentissage automatique semi-supervisé se situent quelque part entre l'apprentissage supervisé et l'apprentissage non supervisé, dans la mesure où il utilise à la fois des données étiquetées et non étiquetées - généralement une petite quantité de données étiquetées et une grande quantité de données non étiquetées.

De nombreux chercheurs en apprentissage automatique ont constaté que les données non étiquetées, lorsqu'elles sont utilisées avec une petite quantité de données étiquetées, peuvent améliorer considérablement la précision de l'apprentissage.

L'acquisition de données marquées pour un problème d'apprentissage supervisé nécessite souvent un agent humain qualifié (un superviseur) qu'il sera partiellement disponible pour étiqueter les données d'entrainement. Donc le coût associé au processus d'étiquetage peut rendre impossible la réalisation d'un ensemble de données d'entrainement entièrement étiqueté lorsque les jeux de données deviennent très grands, alors que l'acquisition de données non étiquetées est relativement peu coûteuse. Dans cette situations, l'apprentissage semi-supervisé peut être d'une grande utilité pratique. L'apprentissage semi-supervisé présente également un intérêt théorique pour l'apprentissage automatique et sert de modèle à l'apprentissage humain [17].

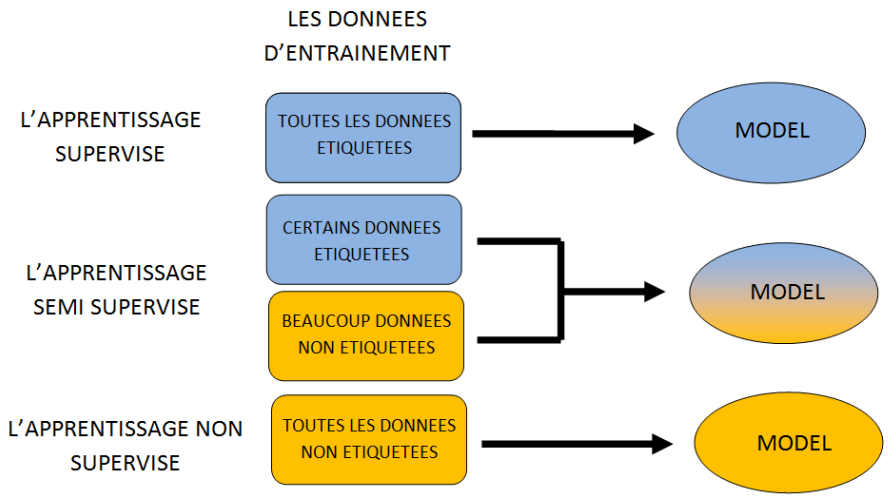


FIGURE 3.9 – comparaison entre l'apprentissage : supervise, semi-supervise, et non-supervise

3.3 Fouille de données (data mining)

Le concept de fouille de données (DATA MINING en anglais) est parfois confus avec un autre concepts plus générale qui est l'extraction de connaissances à partir de données ECD (KDD pour Knowledge Discovery in Databases en anglais), ils sont des fois même considérés comme synonymes. Alors que la fouille de données ce ni qu'une étape dans le processus d'extraction de connaissances à partir de données. Mais elle est considéré formellement l'étape centrale de ce processus [14].

3.4 Apprentissage automatique, fouille de données, quelle déférence ?

L'apprentissage automatique est le champ d'étude intéressée par le développement des algorithmes informatiques pour transformer les données en action intelligente, ce champ est né dans un environnement où d'une part, les données sont devenues plus disponibles, et d'autre part les méthodes statistiques et la puissance de calcul ont rapidement évolués.

Les données sont désormais deux fois plus nombreuses que par le passé, Ce qui nécessitait des puissances de calcul supplémentaire. Qui à son tour a stimulé le développement de méthodes statistiques pour analyser ces grosses données.

C'est le rôle de la fouille de données (data mining),qui est préoccupé par la génération de nouvelles connaissances (concepts) à partir des grandes bases de données. C'est une autre science jumelle étroitement liées à l'apprentissage automatique.

Malgré le grand chevauchement entre la fouille de données et l'apprentissage automatique, il existe toujours un désaccord sur la manière dont chacun d'entre eux intervient.

Un point de distinction potentiel est que :

L'apprentissage automatique se concentre sur l'enseignement aux ordinateurs comment être capables d'utiliser les données pour résoudre un problème.

Mais la fouille de données se concentre sur l'enseignement des ordinateurs afin de déterminer les modèles (Patrons, formes ou concepts) à travers des données, que l'homme peut utiliser par la suite ces modèles pour résoudre un problème.

Pratiquement toute fouille de données implique l'utilisation de l'apprentissage automatique, mais pas tout apprentissage automatique implique l'utilisation da la fouille de données [23] .

3.5 Le Web mining

Le web mining, est 'une application des techniques du Data Mining aux données du web. Les principales applications de ce type d'exploration de données consistent à collecter, classer, organiser et fournir les meilleures informations possibles disponibles sur le Web à la demande de l'utilisateur [24].

3.5.1 Les objectifs du web mining

Le web mining poursuit deux principaux objectifs :

- L'amélioration et la valorisation des sites Web, par l'analyse et la compréhension du comportement des internautes sur le web, ce qu'il permet de valoriser le contenu des sites en améliorant l'organisation et les performances de ces sites web.
- Personnaliser le contenu proposé aux internautes en tenant compte de leurs préférences en appliquant les techniques de Data Mining aux données collectées sur le Web, ce qu'il permet d'extraire des informations intéressantes relatives à l'utilisation du site par les internautes.

3.5.2 Les axes de développement du web mining

Les axes de développement actuels du web mining sont les trois axes suivant :

- **Le Web Content Mining (WCM) :** consiste en une analyse textuelle avancée (traitement linguistiques, classification des pages, segmentation thématique...etc.).
- **Le Web Usage Mining (WUM) :** s'intéresse à l'analyse des comportements de navigation sur les sites web notamment l'analyse du clickstream(c'est l'ensemble des clics exécutés sur un site).
- **Le Web Structure Mining (WSM) :** consiste à analyser l'architecture des sites web, et les liens entre eux [24].

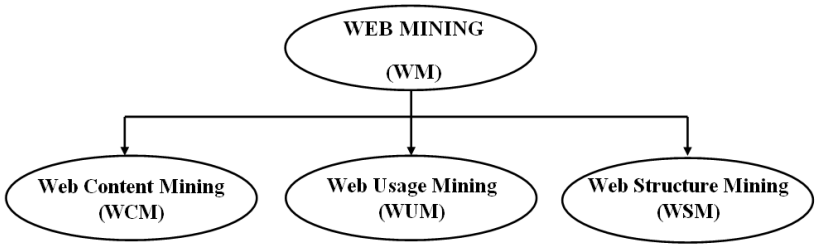


FIGURE 3.10 – les axes de développement du web mining.

3.5.3 Le processus du web mining

Il existe trois étapes dans le processus du web mining

- **Le prétraitement (preprocessing) :** cet étape consiste à abstraire, et transformer de données nécessaires pour la découverte de modèles.
- **La découverte de modèles de navigation (pattern discovery) :** cet étape nécessite l'utilisation de plusieurs méthodes et algorithmes provenant de branches différentes, telles que les statistiques, l'apprentissage automatique, et le data mining.
- **L'analyse des modèles (pattern analysis) :** c'est la dernière étape processus du Web Mining, elle consiste à analyser les résultats et extraire des connaissances.

Ce processus illustré dans (Figure 14), s'adapte à chacun des axes du Web mining.

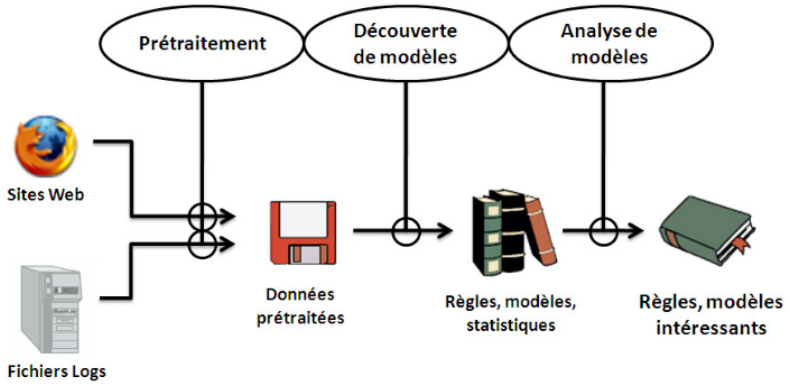


FIGURE 3.11 – Processus du web mining [24].

3.5.4 Le Web Content Mining (WCM)

On va basé dans notre étude sur l'axe Web Content Mining (WCM), qui est pour objectif d'extraire des connaissances à partir du contenu des pages web. ce contenu du web est représentés sous différentes formes : images, vidéo, audio,... une bonne partie de l'information accessible est sous forme textuelle (bibliothèques électroniques, pages HTML, forums de discussion...).

Comme le Web Content Mining (WCM) est un axe du Web Mining (WM), alors il suit les mêmes étapes du processus du web mining, qui sont :

- La premier étape est le prétraitement des données (nettoyage, structuration,...), en fait cette étape varie selon le type des données (textes, images,....).
- Puis la deuxième étape est l'application des techniques de data mining pour découvrir des modèles, afin d'extraire des connaissances, de même le choix de la méthode de data mining varie selon l'objectif de l'analyse.
- Finalement la dernière étape qu'est l'analyse et la validation.

On a choisit de basé sur l'axe Web Content Mining (WCM) dans notre étude, car l'objectif de l'analyse textuelle est d'identifier le contenu thématique (les sujets ou les thèmes) des commentaires, ce qui permet les classifier par la suite.

Les types de données traitées par le Web Content Mining (WCM)

Sur le web, il existe plusieurs types de données :

- **Les données structurées :** ces données sont disposées de façon à être traitées automatiquement et efficacement par un logiciel et pas forcément par l'homme. telles que les tables et les bases des données.
- **Les données semi-structurées :** dans la page web, une partie de contenu est destinée à l'homme, comme le texte (informations non structurées), alors qu'une autre partie est destinée à la machine, telles que les balises HTML (informations structurées).
- **Les données non structurées :** c'est celles de tous les documents sur support numérique qui ne peuvent être utilisés que par l'homme. Exemple : documents textes et multimédias [24].

3.6 Le Web Scraping

Le web scraping, également appelé récolte du Web ou même l'extraction du webest une technique permettant d'extraire des données provenant des sites web (ces données sont principalement au format HTML), et de les enregistrer dans un système de fichiers ou une base de données pour les récupérer ou les analyser ultérieurement afin de les utiliser dans un autre contexte.

Ceci est accompli soit manuellement par l'utilisateur mais ça sera un travail dur qui nécessite beaucoup d'efforts, et surtout beaucoup de temps, qui peut s'étaler sur plusieurs jours voire semaines selon la taille des sites web ciblés. Soit automatiquement par un robot qui effectuera la même tâche dans une fraction de temps.

Pour s'adapter à une variété de scénarios, les techniques actuelles de web scraping ont été personnalisées, passant des procédures simples assistées par l'homme à l'utilisation de systèmes entièrement automatisés capables de convertir des sites Web entiers en un ensemble de données bien organisé.

Les outils de web scraping à la pointe de la technologie sont non seulement capables d'analyser les langages de balisage tels que HTML, mais également de s'intégrer à l'analyse visuelle (Les méthodes d'analyse visuelle permettent aux décideurs de combiner leur flexibilité humaine, leur créativité et leurs connaissances avec les énormes capacités de stockage, de traitement, et d'analyse des données des ordinateurs actuels, ce qui leur permet de prendre des décisions éclairées dans des situations complexes) et au traitement du langage naturel pour simuler la façon dont les utilisateurs humains parcourir le contenu Web.

Le processus de web scraping peut être divisé en deux étapes séquentielles ; acquérir ressources web et extraire ensuite les informations souhaitéesà partir des données acquises.

Plus précisément un programme de web scraping commence par composer une requête HTTP pour acquérir les ressources d'un site Web ciblé. Une fois la requête reçue et traitée avec succès par le site Web ciblé, la ressource demandée sera extraite du site Web, puis renvoyée au programme collecte de données web.

La ressource peut être sous plusieurs formats, tels que des pages Web construites en HTML, des sources de données au format XML ou JSON, ou des données multimédia telles que des images et des fichiers audios ou vidéos.

Une fois les données Web téléchargées, le processus d'extraction continue à analyser, reformater et organiser les données de façon structurée.

Dans un programme de web scraping, il existe deux modules essentiels :

- Un module pour la composition des requêtes HTTP, tel que "Urllib2". Ce module définit un ensemble de fonctions permettant de traiter les requêtes HTTP, telles que l'authentification, les redirections, les cookies, ...etc.
- Un deuxième module pour l'analyse et l'extraction d'informations à partir du code HTML brut, tel que "Beautiful Soupe" et "Pyquery". Ce module est un toolkit pour décomposer et extraire les données désirées du fichier HTML, il permet de naviguer, rechercher et modifier un arbre DOM [?].

3.6.1 Types d'extraction de données via le scraping

Les modalités d'extraction des données via le scraping sont :

- **Extraction manuelle** , c'est un simple copier/coller. Un humain navigue sur internet et extrait des informations pertinentes aux intérêts depuis les pages qu'il visite.
- **Extraction semi-automatique** , utiliser un logiciel ou une application web pour aspirer et nettoyer les informations d'une ou plusieurs pages web.
- **Extraction automatique** , dans ce cas l'extraction se fait automatiquement via un navigateur web qui visite des pages et suit les différents liens qui se trouvent dans ces pages web afin de générer automatiquement un corpus de pages.

3.6.2 Les cas d'utilisation du web scraping

Le web scraping peut être utilisé pour une grande variété de scénarios, tels que la collecte de contacts comme l'email, le numéro de téléphone, l'adresse ; la surveillance et la comparaison des changements de prix, la collecte des critiques de produits, la collecte de listes de biens immobiliers, la météo, ...

Surveillance des données et détection des modifications d'un site web et intégration des données web. Par exemple :

- à petite échelle, le prix d'une action peut être régulièrement collecté afin de visualiser l'évolution du prix au fil du temps.

Dans les flux des médias sociaux peuvent être collectés ensemble pour enquêter sur les opinions publiques et identifier les leaders d'opinion.

— à grande échelle, les métadonnées de presque tous les sites web sont constamment collectées pour constituer les résultats des moteurs de recherche sur Internet, tels que "Google" et "Bing" [?].

3.7 La classification de textes

La classification de textes, définie comme le processus permettant d'associer une catégorie (ou classe) à un texte libre, en fonction des informations (termes, mots,...) qu'il contient, est un élément important des systèmes de gestion de l'information.

Mais, associer une classe à un texte libre est une procédure coûteuse et longue, par conséquent, l'automatisation de cette procédure est devenue un enjeu pour la communauté scientifique.

On peut dire aussi que la classification de texte consiste à chercher une liaison entre un ensemble de textes et un ensemble de classes (étiquettes, catégories). Cette liaison, que l'on appelle également modèle de prédiction, est estimée par un apprentissage automatique.

Pour ce faire, il est nécessaire de disposer d'un ensemble de textes préalablement classés, dit ensemble d'apprentissage, à partir du quel on estime les paramètres du modèle de prédiction le plus performant possible, c'est-à-dire le modèle qui produit le moins d'erreur en prédiction. Par exemple, on prend la tâche de classifier l'ensemble de documents comme bon ou mauvais. Dans ce cas, les catégories (ou étiquettes) « bon » et « mauvais » représentent les classes.

Formellement, la classification de texte consiste à associer une valeur booléenne à chaque paire $(d_j, c_i) \in \mathrm{D} \times \mathrm{C}$, où D est l'ensemble des textes et C est l'ensemble des classes.

La valeur V (Vrai) est alors associée au couple (d_j, c_i) si le texte dj appartient à la classe ci tandis que la valeur F (Faux) lui sera associée dans le cas contraire.

Le but de la classification de texte est de construire une procédure (modèle, classifieur) qui associe une ou plusieurs étiquette (classe) à un document d_j. Une fonction qui retourne pour chaque vecteur d_j une valeur c_i [26].

$$\mathrm{F} : \mathrm{D} \rightarrow \mathrm{C}$$

3.8 La représentation du texte

La machine considère le texte comme des données non structurées. Par conséquent, il est nécessaire de représenter le texte du document sous une forme structurée et formelle.

La technique la plus courante pour représenter le texte est proposé au milieu des années70, c'est le modèle de représentation vectorielle du texte qui reste à ce jour très utilisé et souvent désigné sous le nom de modèle de Salton ou de représentation de Salton.

Ce modèle, appelé Vector Space Model(ou term vector model), est un modèle algébrique utilisé pour représenter un document textuel par un vecteur d'identifiants, comme par exemple des termes. Ce modèle a été utilisé en recherche d'information, en indexation de documents ou encore pour faire du filtrage d'information. Dans ce modèle, un document est représenté par un vecteur dont chacune des dimensions est un terme.

Si ce terme apparaît dans le document, alors la valeur associée à cette dimension est non-nulle. À la base, les n termes sont les n différents mots apparaissant dans les textes de l'ensemble d'entraînement (l'ensemble d'apprentissage). Cette représentation des textes exclut toute analyse grammaticale et toute notion de distance entre les mots. C'est pour quoi elle est appelée "sac de mots" ou "bag of words".

Ce processus est également appelé Tokenisation, car le document est divisé en jetons qui sont des mots individuels [24].

3.8.1 Extraction des termes (Tokenization)

C'est la tâche qui consiste à découper le texte d'un commentaire en pièces (jetons) appelés par abus de langage mots ou termes tout en éliminant éventuellement quelques caractères tels que la ponctuation. Un jeton (token en anglais) est une instance d'une séquence de caractères dans un commentaire particulier qui sont regroupés comme une unité sémantique utile pour le traitement.

Exemple :

Texte du commentaire : « Ce blog est intéressant, il donne beaucoup d'astuces de jardinage »
Résultat d'extraction des tokens :

Jeton	Ce	blog	est	intéressant	il	donne	beaucoup	d'astuces	de	jardinage

L'extraction des jetons c'est l étape initiale [27].

3.8.2 La réduction des vecteurs

Afin de garantir performance et efficacité lors de la modération de commentaires sur le web, une bon représentation de ces commentaires-là est plus que nécessaire pour détecter l'existence ou non d'une information au sein des commentaires examinés.

Il est possible de représenter un commentaire par tout l'ensemble de ses mots quant on un petit ensemble de commentaire, ce qui donnera une vue complète et logique de ce commentaire, à mesure où cette représentation ne soit pas coûteuse.

Par contre, avec une large collection on doit réduire l'ensemble de mots représentatifs pour un commentaire. Cette tâche est accomplie à travers :

- L'élimination des mots vides qui ne sont pas porteur de sens relatif au commentaire tels que les conjonctions et les pronoms, ...etc.
- L'utilisation de la racinisation qui réduit des mots distincts à leur racine grammaticale commune.
- L'identification des groupes de noms ce qui élimine les verbes, adverbes et adjectifs.
- Des compressions supplémentaires sont applicables [28].

Elimination des mots vides (Stop Words)

Parfois, quelques mots extrêmement courants s'avèrent d'une importance symbolique voire nulle dans la détection de commentaires suspects vis-à-vis de la modération souhaité par le propriétaire du blog, de ce fait ils sont entièrement exclus du vocabulaire. Ces mots sont appelés mots vides (stop words). La stratégie générale de détermination de la liste des mots vides est de trier les termes par ordre de leurs nombre d'occurrences dans la collection de commentaires par la suite prendre les termes les plus fréquents filtrés souvent manuellement, ayant peu ou aucun trait avec les thématiques faisant objet de modération. Les mots de cette liste seront alors éliminés durant la représentation.

Notant qu'Il existe toutefois des système de modération qui n'utilisent pas de liste des mots vides [27].

Exemple de mots vides pour différentes langues.

La langue	Liste des mots vides
Arabe	ذلك، مثل، حيث، في، عن، إلي، الذي، ....إلخ
Anglais	which, this, , at , thus, with about, …etc
Français	mais, de, toutefois, désormais, compris, donc, … etc.

FIGURE 3.12 – Exemple de mots vides pour différentes langues

La normalisation

Dans ce qui suit quelques formes couramment utilisés pour la normalisation et leurs applications. Dans de nombreux cas, ils semblent utiles, mais dans certains cas, ils peuvent avoir des effets indésirables. En fait, dans la normalisation, de nombreux détails peuvent être abordés, mais souvent, si le traitement est appliqué d'une façon manière cohérente à la modération des commentaires, alors dans ce cas les détails les plus fins peuvent ne pas avoir d'impact global considérable sur les performances de la modération.

Unification de la casse : cette technique est couramment utilisée pour unifier la casse, en réduire tout le mot en minuscule pour faire correspondre entre "Commentaire" et "commentaires" par exemple, et cette méthde est très utile pour les comparaisons.

D'autre part beaucoup de noms propres sont dérivés de noms communs et ne sont donc distinguées que par le caractère majuscule au début. Dans la langue anglais une autre technique est appliqué à la place de la technique sus-décrites, elle ne convertir en minuscule que les mots en début de phrases ou dans les titres car en général, ce sont des mots ordinaires.

La normalisation est parfois largement sujette aux spécificités d'une langue comme pour ‘’colour” (dictée anglaise) et ‘’color” (dictée américaine) ou les pronoms le, la et les en langue française.

Prise en charge des accents et diacritiques : Diacritiques sur les caractères en anglais ont une incidence marginale, ce qui nous permettra de faire correspondre les deux mots : cliche et cliché, ou naïfs et naïve par exemples.

Cela peut être fait en normalisant les jetons pour supprimer les signes diacritiques. Dans de nombreuses autres langues, les diacritiques font régulièrement partie du système d'écriture

afin de distinguer les différentes prononciations. Parfois, les accents distinguent des mots différents.

Prenant par exemple la langue espagnol, on a les mots Peña et Pena qui sont complètement distinguent où : "Peña" signifie une falaise, alors que "Pena" signifie chagrin [27].

Racinisation et lemmatisation

Un mot se présente dans les textes sous formes différentes et ce pour des raisons linguistiques bien évidentes par exemple : travail, travailler, travaillent. et peut aussi être sujet aux dérivations tel que : concret, concrétisation et concrètement. Dès lors il serait judicieux de détecter les différentes formes d'un mot dans un texte surtout pour la modération en introduisant en entrée qu'une seule forme de ce mot.

La racinisation (stemming) : Enlèvement de quelques caractères à la fin des mots afin de correspondre à une forme commune, incluant souvent l'élimination des suffixes et des préfixes.

Lemmatisation : Utilisation d'analyses morphologiques et du vocabulaire des mots pour retourner la base du mot ou son entrée de dictionnaire.

Prenons par exemple en anglais, le mot préparation "preparation" le stemming peut retourner "prepar" alors que la lemmatisation retournera "prepare" infinitif du verbe [27].

Exemple d'un algorithme de racinisation :

Algorithme de Porter :

L'algorithme de racinisation le plus courant pour l'anglais est l'algorithme de porter, celui-ci consiste en 5 phase de réduction du mot appliquées séquentiellement et pour chacune d'elles il y a des conventions d'application des règles de réduction [**?**].

Exemple de règles :

Etape 1 :

Règle	Résultat après réduction	Exemple	Résultat
SSES	SS	Caresses	caress
IES	I	Ponies	poni
SS	SS	Caress	caress
S		Cats	cat

Table 3.1 – Etape 1 de Algorithme de Porter

Etape 2 :

Si m (nombre de syllabes) > 2

Règle	Résultat après réduction	Exemple	Résultat
eed	ee	Agreed	Agree
y	i	Happy	Happi
Ant		Irritant	Irrit
ement		Replacement	Replac
ment		Investment	invest

Table 3.2 – Etape 2 de Algorithme de Porter

Etape 3 :

Si m 0

Règle	Résultat après réduction	Exemple	Résultat
Ational	Ate	Relational	relate
Tional	Tion	National	Nation

Table 3.3 – Etape 3 de Algorithme de Porter

Etape 4 :

Si m < 1

Règle	Résultat après réduction	Exemple	Résultat
ance		Allowance	Allow
Able		Ajustable	Ajust
Ible		Defensible	Defens
Ion		Adoption	adopt

Table 3.4 – Etape 4 de Algorithme de Porter

Etape 5 :

Règle	Résultat après réduction	Exemple	Résultat
e		Probate	probat

Table 3.5 – Etape 5 de Algorithme de Porter

Fréquence des termes (TF)

" TF " fait référence à la fréquence de l'existence d'un mot dans un texte donné, à savoir le nombre d'occurrences d'un mot particulier dans un document. C'est un calcul de fréquence très simple, mais qui s'avère efficace et pratique. on l'utilise souvent en association avec d'autres fréquences. Plus la fréquence du mot augment, plus le poids de la fonction TF augmente (relation directe).

Nous dénombrons plusieurs manières de calcul de la TF :

TF absolue : c'est le nombre de fois qu'un terme apparait dans un texte donné.

$$\text{TF} = \text{NT}$$

Où " NT " est le nombre de fois où le terme est apparu dans le texte . **TF relative :** C'est le rapport entre le nombre de fois qu'un terme est apparu dans le texte sur le nombre de tous les termes du texte. Cette dernière est utilisée généralement pour limiter l'impacte de la longueur des textes. En effet, un terme qui apparait 5 fois dans un texte de 100 termes n'a pas le même degré de discrimination que celui qui apparait le même nombre de fois (5 fois) dans un texte de 30 termes.

$$\text{TF} = \frac{NT}{ST}$$

Où " NT " est le nombre de fois que le terme est apparu dans le document. Et" ST "est le nombre de tous les termes du document.

TF booléenne : il s'agit juste de la présence ou de l'absence du terme dans un texte.

$$TF = 1 \text{ ou } 0$$

Le principal inconvénient de la fréquence des termes " TF " est le fait qu'il est possible qu'un terme apparait avec une fréquence assez grande dans tous les documents du corpus. Par conséquent le terme en question perd toute sa notion de discrimination relative au degré de présence.

Ce cas est très probable en pratique, alors pour le rectifier on réduire l'effet du terme en question, par une autre notion nommée IDF (Inverse documents frequancies) [26].

Fréquence documents inverses (IDF)

Cette notion mesure en quelque sorte le degré de rareté d'un terme, non pas dans un document, mais dans tous les documents d'un corpus elle est définie par l'équation suivante :

$$IDF = \log \frac{N}{DF}$$

Où la variable " N " désigne le nombre total des documents dans le corpus. Et la fréquence de document " DF " est le nombre de documents de la collection dans lesquels le terme est apparu.

Si le terme est très présent dans tout le corpus alors le rapport sera égal à 1 et IDF = 0 donc le terme est neutralisé. Si par contre il apparait dans un seul document la valeur est maximale, car l'équation sera :

$$IDF = \log N$$

Cette pondération à elle seule ne peut pas définir le degré de discrimination d'un terme dans un document puisque elle est relative au corpus. Mais l'association de IDF avec TF donne des résultats intéressants [26].

TFIDF (Term Frequency - Inverse Document Frequency)

On a vu que la fréquence d'un terme dans un document joue un rôle important dans le calcul de son degré de discrimination. En revanche, la représentation d'un texte, dans le but de le

classifier, ne dépend pas seulement de son contenu, mais elle est liée étroitement au corpus auquel le texte appartient, la rareté de ce terme au sein des autres documents du corpus s'avère aussi importante que sa fréquence dans le document en question.

Cette combinaison de ces deux principes (abondance particulière et rareté générale) a engendré la pondération dite TFIDF, qu'est calculée comme suit [26] :

$$\text{TFIDF} = \text{TF} \times \log \frac{N}{DF}$$

CHAPITRE 4

MODELISATION ET IMPLEMENTATION

4.1 Approche proposée

Dans le but de détecter les commentaires suspects et les présenter au modérateur d'un blog afin de décider s'il veut les supprimer ou pas, des étapes successives sont suivie pour comparer le texte de ces commentaires avec une base de données des termes suspects et décider de leur degré de suspicion.

Dans un premier lieu on procède grâce à une fonction de web scraping à la collecte des commentaires d'un blog, celle-ci scannera les pages web et analysera le contenu de ces pages et retourne les commentaires.

Ensuite, un traitement est appliqué à ces commentaires afin de lister les termes y existant, Ce traitement comporte d'abord la tokénisation du texte du commentaire puis la suppression des mots vides et enfin le stemming de ces mot afin de tomber sur les mêmes mots clés présent dans la base des termes suspects. Ceux-là étant déjà racinisés.

Après le traitement textuel des commentaires, on calcul les poids (calculé en utilisant le schéma de pondération TF-IDF) des différents termes dans les commentaires. On calcul aussi le poids des termes dans la liste des termes communs avec un commentaire donné.

Pour chaque commentaire on calcul la similarité entre celui-ci et la liste des mots suspects qui s'y trouvent. Si cette similarité est supérieure à un paramètre : "Seuil" que l'utilisateur aurait fixé le programme considère que le commentaire en question est suspect et le retourne au modérateur afin de le supprimer ou l'approuver.

On a choisie pour le calcul de la similarité la formule suivante [28] :

$$SIM(C_j, L) = \frac{\sum_{i=1}^{n} w_{ij} \; x \; w_{il}}{\sqrt{\sum_{i=1}^{n} {w_{ij}}^2} \; x \; \sqrt{\sum_{i=1}^{n} {w_{il}}^2}}$$

Où :

- C_j : Commentaire j.
- L : Liste des termes communs entre le commentaire et la liste des termes suspects.
- W_{ij} : Poids du terme i dans le commentaire j
- W_{il} : poids du terme i dans la liste L.
- n : nombre de termes.

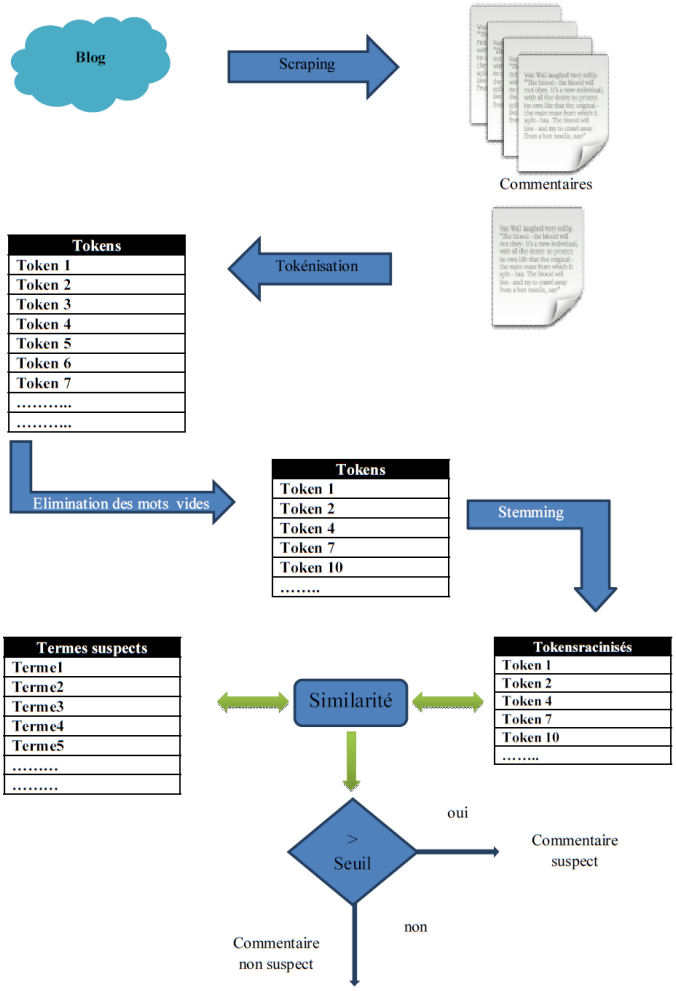
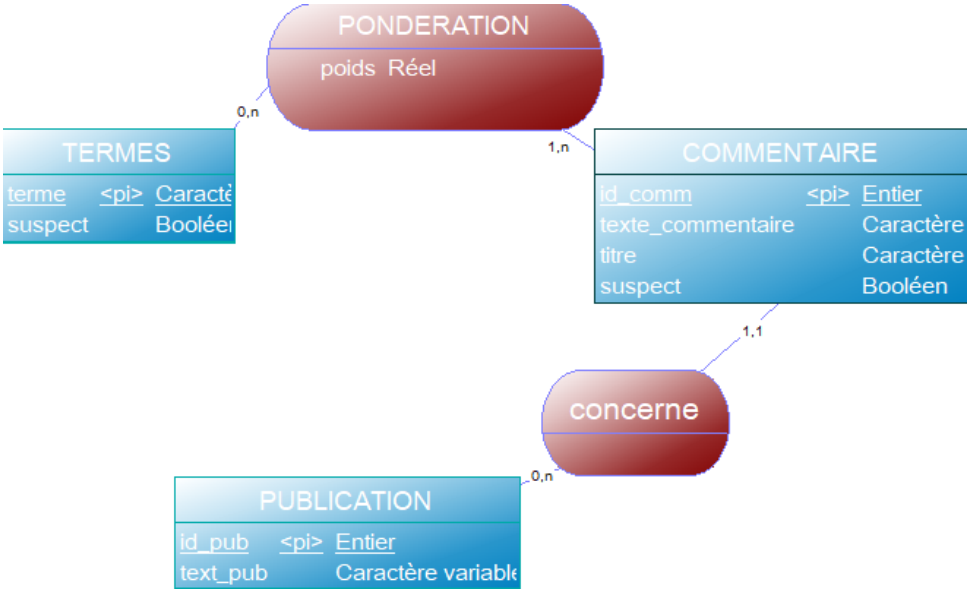


FIGURE 4.1 – Schéma explicatif de l'approche

Pseudo-algorithme
Suspect: booléen; S:réel; Début Scraping(blog); Pour i :=1 à nb_commentaires faire Debut Tokéniser_comment(i); Stoplist_comment(i); Stemming_comment(i); L:= extraire_liste_suspects_comment(i); S:= calcul_sim(commentaire[i],L); Fin pour Si S > Seuil alors Suspect := vrai; sinon suspect = faux; Fin

4.2 Modèle entité association relatif



4.3 Réalisation du prototype

4.3.1 Base de données

Une base de données sous le système de gestion de base de données Mysql est utilisée pour stocker les commentaires et effectuer les calculs nécessaires.

Cette base de données possède la structure suivante :

Table	Champ	Type de champ	Représentation
Publication	Id_pub	Entier	Identifiant de la publication
	Text_pub	Texte	Texte de la publication
	Id_comm	Entier	Identifiant du commentaire
Commentaire	Id_comm	Entier	Identifiant du commentaire
	Texte_comm	Varchar	texte du commentaire
	Titre_comm	Varchar	texte du commentaire
	Suspect	Booléen	Par défaut non suspect, il indique après examen si le commentaire est suspect ou pas
Termes	Terme	Varchar	Texte racinisé du terme et aussi identifiant
	Suspect	Booléen	Indique si le terme est dans la liste des termes suspects
Pondération	Id_comm	Entier	Identifiant du commentaire
	Terme	Varchar	Identifiant du terme
	Poids_terme	Réel	Poids du terme dans le commentaire

4.3.2 Outils de développement

L'application est réalisé avec les langages : HTML 5, javascript, Css ; Le serveur d'application est apache.

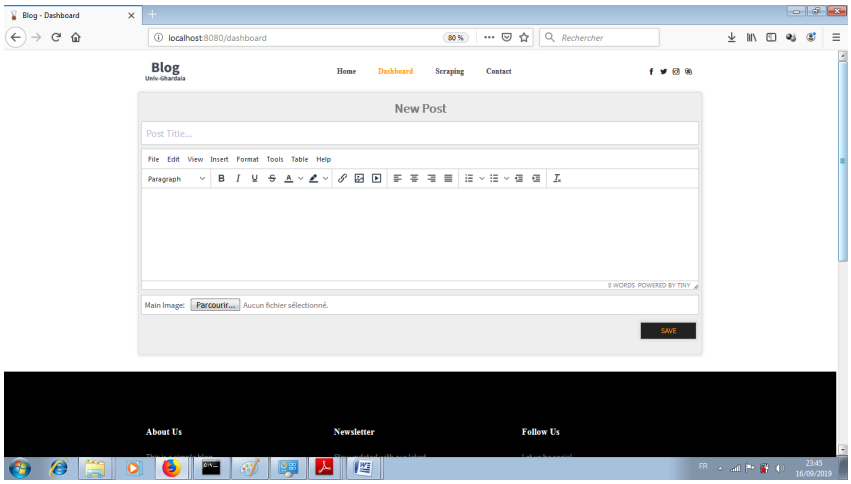
4.4 Présentation de l'application réalisée

L'application réalisée est un blog universitaire contenant plusieurs thèmes d'études. Elle permet aux étudiants de commenter les différentes publications.

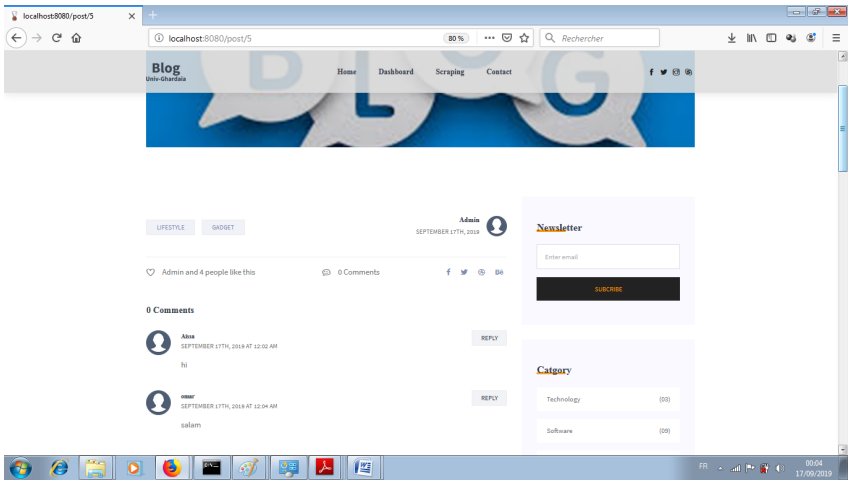
La modération utilisée est de type a priori, ainsi elle intervient avant la publication d'un commentaire relatif à l'une des publications.

Cette application contient deux volets :

— Administration : permet de gérer les publications.



— Utilisateur : permet aux étudiants de consulter et commenter les différentes publications.



CONCLUSION

Dans les récentes années, le contenu du web a connu d'importants changements et des outils modernes comme les systèmes de gestion de contenu ou les réseaux sociaux sont apparus rendant la manipulation de ce contenu accessible à de plus larges catégorie d'utilisateurs. Cette révolution informatique a soulevé de nouveaux défis tels que la propagation des discours de haines, et la facilité de diffusion de campagnes de diffamation, et aussi la diffusion de contenu à caractère violent.

Le présent mémoire présente la problématique de modération de ce contenu et en particulier le contenu textuel dans les blogs et les sites web. En exploitant quelques techniques de recherche d'information afin de traiter l'information présente sur ces blogs on a conçu une approche pour détecter les commentaires susceptibles de contenir des propos indésirable ou inapproprié. Cette approche comprend 2 phases :

- Représentation des commentaires : extraction du texte des commentaires et relevé des termes.
- Calcul de la similarité entre les commentaires et les termes de la liste des termes suspects.

Ainsi un blog été conçu afin de concrétiser cette approche.

La présente étude ne constitue qu'une introduction à la modération de contenu dans le web. D'autres aspects de cette modération méritent d'être étudiés à savoir, la modération automatique des documents multimédias tels que les images et les vidéos.

BIBLIOGRAPHIE

[1] S. Iacus, "Automated data collection with r-a practical guide to web scraping and text mining," *Journal of Statistical Software*, vol. 68, no. b03, 2015.

[2] C. Lablanche, F. Seine, S. Gastaud, and H. Chang, "Les web services," *Université de Nice-Sophia Antipolis, Rapport de TE, 52p*, 2005.

[3] R. Berrabah and S. Brahimi, *Administration des services de partage et de transfert des fichiers (samba, nfs et ftp) par l'outil webmin.* PhD thesis.

[4] M. Abdel-Salam, A. Ahmed, and M. Mahrous, "Transient analysis of gridconnected wind-driven pmsg, dfig and scig at fixed and variable speeds," *Innovative systems design and Engineering*, vol. 2, no. 3, pp. 135–152, 2011.

[5] S. Ambrosini, M. Jeannin, and S. Joao-Vidal, "Cosadoca, consortium de sauvetage du patrimoine documentaire en cas de catastrophe : un site web pour la sauvegarde du patrimoine documentaire," tech. rep., 2005.

[6] K. Gollatz, F. Beer, and C. Katzenbach, "The turn to artificial intelligence in governing communication online," 2018.

[7] D. T. S. L. GROUPE, "Du secteur financier," *Pour un encadrement intégré et simplifié du secteur financier au Québec.*

[8] C. Maizonniaux, "Lire des textes littéraires hybrides puis écrire son texte en l2 : quelle place pour l'image et pour les «langues en réserve» de l'apprenant ?," *Lidil. Revue de linguistique et de didactique des langues*, no. 57, 2018.

[9] T. Stenger and A. Coutant, "Community management et community managers : Cheval de troie marketing pour le web social ?," *Web social, communautés virtuelles et consommation*, vol. 140, 2011.

[10] A. Bruns and M. Bahnisch, "Social media : tools for user-generated content : social drivers behind growing consumer participation in user-led content generation [volume 1 : state of the art]," 2009.

[11] N. Smyrnaios and E. Marty, "Profession «nettoyeur du net»," *Réseaux*, no. 5, pp. 57–90, 2017.

[12] S. Wojcik, "Les modérateurs des forums de discussion municipaux. des intermédiaires démocratiques ?," *Questions de communication*, no. 12, pp. 335–354, 2007.

[13] A. Veglis, "Moderation techniques for social media content," in *International Conference on Social Computing and Social Media*, pp. 137–148, Springer, 2014.

[14] B. Brahimi, "Extraction de connaissances à partir de données incomplètes et imprécises," *Université de M'Sila*, 2011.

[15] A. Butter, G. Kasieczka, T. Plehn, and M. Russell, "Deep-learned top tagging with a lorentz layer," *arXiv preprint arXiv :1707.08966*, vol. 43, pp. 2–3, 2018.

[16] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The kdd process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[17] M. TAFFAR, "Initiation al'apprentissage,"

[18] P. Domingos, *The master algorithm : How the quest for the ultimate learning machine will remake our world.* Basic Books, 2015.

[19] E. Bendiab and M. K. Kholladi, "The negative selection algorithm : a supervised learning approach for skin detection and classification," *International Journal of Computer Science and Network Security*, vol. 10, pp. 86–92, 2010.

[20] M. H. BENDIABDELLAH, "Système immunitaire artificiel pour la reconnaissance du diabète," 2011.

[21] G. Allain, *Prévision et analyse du trafic routier par des méthodes statistiques.* PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, 2008.

[22] X. Tian, G. Gasso, and S. Canu, "A multiple kernel framework for inductive semi-supervised svm learning," *Neurocomputing*, vol. 90, pp. 46–58, 2012.

[23] D. Zouache, *Fouille de données basée algorithmes bio-inspirés.* PhD thesis, 2018.

[24] M. Charrad, *Une approche générique pour l'analyse croisant contenu et usage des sites Web par des méthodes de bipartitionnement.* PhD thesis, Paris, CNAM, 2010.

[25] A. Cho, J. Graves, and M. A. Reidy, "Mitogen-activated protein kinases mediate matrix metalloproteinase-9 expression in vascular smooth muscle cells," *Arteriosclerosis, thrombosis, and vascular biology*, vol. 20, no. 12, pp. 2527–2532, 2000.

[26] R. Jalam, "Apprentissage automatique et catégorisation de textes multilingues," *PhD Tesis, Université Lumiere Lyon*, vol. 2, 2003.

[27] M. Sanderson, "Christopher d. manning, prabhakar raghavan, hinrich schütze, introduction to information retrieval, cambridge university press. 2008. isbn-13 978-0-521-86571-5, xxi+ 482 pages.," *Natural Language Engineering*, vol. 16, no. 1, pp. 100–103, 2010.

[28] R. Baeza-Yates and B. Ribeiro-Neto, "Modern information retrieval addison-wesley longman," *Reading MA*, 1999.

[29] J. Pérez, G. Garrido, A. Rodrigo, L. Araujo, and A. Peñas, "Information retrieval baselines for the respubliqa task," in *Working Notes for the CLEF 2009 Workshop, Corfu, Greece*, 2009.