MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE ABDELHAMID IBN BADIS - MOSTAGANEM

Faculté des Sciences Exactes et d'Informatique
Département de Mathématiques et informatique
Filière : Informatique

MEMOIRE DE FIN D'ETUDES

Pour l'Obtention du Diplôme de Master en Informatique

Option : **Ingénierie des Systèmes d'Information**

Présenté par :

BOUFERMA Bouchra

THEME :

SYSTEME DE CONTROLE D'ACCES POUR LES RESEAUX SOCIAUX

Soutenu le : Dimanche 02 octobre 2022.

Devant le jury composé de :

M.Zahmani	Université de Mostaganem	Président
Mme.Abid	Université de Mostaganem	Examinatrice
Mme.Hocine	Université de Mostaganem	Encadrante

Année universitaire 2021-2022

Résumé

Avec l'expansion d'Internet, le nombre d'utilisateurs des réseaux sociaux et les activités qui s'y déroulent ont encore augmenté. Les utilisateurs de ces réseaux partagent de nombreuses informations avec leurs amis, dont beaucoup sont confidentielles et privées. À cet égard, le maintien de la sécurité et de la confidentialité des utilisateurs est une exigence majeure dans les réseaux sociaux. Bien que les modèles de contrôle d'accès traditionnels puissent aider les utilisateurs à préserver leur vie privée en appliquant des niveaux d'accès initiaux, en permettant aux utilisateurs de mieux sécuriser leurs comptes ainsi que les données partagées. Certes qu'ils ne sont pas efficaces pour les réseaux sociaux, compte tenu de leur nature dynamique. C'est pourquoi, dans ce rapport de recherche bibliographique, on analysera les modèles récents de contrôle d'accès pour les réseaux sociaux, en particulier ceux qui sont basés sur le contenu et la vie privée des utilisateurs.

Mots-clés: Réseaux sociaux, Contrôle d'accès, confiance, classification du texte, vie privée.

Abstract

With the expansion of the Internet, the number of users of social networks and the activities that take place there have further increased. Users of these networks share a lot of information with their friends, much of which is confidential and private. In this regard, maintaining user security and privacy is a major requirement in social networks. Although traditional access control models can help users maintain their privacy by enforcing initial access levels, allowing users to better secure their accounts as well as shared data. Admittedly, they are not effective for social networks, given their dynamic nature. This is why, in this bibliographic research report, we will analyze recent models of access control for social networks, in particular those which are based on the content and privacy of users.

Keywords: Social networks, Access control, Trust, Text Classification, privacy.

ملخص

مع توسع الإنترنت ، ازداد عدد مستخدمي الشبكات الاجتماعية والأنشطة التي تتم هناك. يشارك مستخدمو هذه الشبكات الكثير من المعلومات مع أصدقائهم ، ومعظمها سري وخاص. في هذا الصدد ، يعد الحفاظ على أمان المستخدم وخصوصيته مطلبًا رئيسيًا في الشبكات الاجتماعية. على الرغم من أن نماذج التحكم في الوصول التقليدية يمكن أن تساعد المستخدمين في الحفاظ على خصوصيتهم من خلال فرض مستويات الوصول الأولية ، مما يسمح للمستخدمين بتأمين حساباتهم وكذلك البيانات المشتركة بشكل أفضل. من المسلم به أنها ليست فعالة للشبكات الاجتماعية ، نظرًا لطبيعتها الديناميكية. لهذا السبب ، في تقرير البحث الببليوغرافي هذا ، سنقوم بتحليل النماذج الحديثة للتحكم في الوصول للشبكات الاجتماعية ، ولا سيما تلك التي تستند إلى المحتوى وخصوصية المستخدمين .

الكلمات المفتاحية: الشبكات الاجتماعية ، التحكم في الوصول ، الثقة ، تصنيف النص ، حياة خاصة.

Dédicaces

A mes parents

Je vous dois ce que je suis aujourd'hui grâce à votre amour, à votre patience et vos innombrables sacrifices. Que ce modeste travail, soit pour vous une petite compensation et reconnaissance envers ce que vous avez fait d'incroyable pour moi.

Que Dieu, le tout puissant, vous préserve et vous procure santé et longue vie afin que je puisse à mon tour vous combler.

A mes très chers frères et ma sœur

Aucune dédicace ne serait exprimer assez profondément ce que je ressens envers vous, je vous dirais tout simplement, un grand merci, je vous aime.

A mes très chers amis

En témoignage de l'amitié sincère qui nous a liées et des bons moments passés ensemble.

Je vous dédie ce travail en vous souhaitant un avenir radieux et plein de bonnes promesses.

BOUFERMA Bouchra

Remerciements

En tout premier lieu, je remercie Allah le tout puissant, à la sagesse et au savoir infinis, « Gloire à toi ! Nous n'avons de savoir que ce que Tu nous as appris. Certes c'est Toi l'Omniscient, le sage, le tout miséricordieux le très miséricordieux » (Sourate al-Baqarah, verset 32).

Je tiens à remercier mon encadrant Mme HOUCINE Nadia pour le grand honneur qu'il m'a fait en me proposant le sujet de ce mémoire de fin d'étude. J'ai eu l'honneur et le privilège de travailler sous son assistance et de profiter de ses qualités humaines, professionnelles et de sa grande expérience, il m'a guidé tout au long de ce travail. L'élaboration avec amabilité et dynamisme le caractérisant. Que ce modeste travail puisse satisfaire mes examinateurs, pour qu'ils en témoignent ma gratitude et reconnaissance pour l'aide et les conseils qu'il ma prodigué, ainsi que pour la savoir qu'il ma inculqué.

Je remercie tous mes enseignants de l'université de Mostaganem.

Mes remerciements vont également aux membres de jury d'avoir accepté de juger mon travail.

Je remercie vivement toute ma famille, en particulier mes parents, pour m'avoir toujours soutenu au cours de mes études. Qu'ils trouvent ici le fruit de leur patience et du soutien permanent qu'ils m'ont prodigué pour affronter tous les moments difficiles.

Je tiens également à remercier M.DERRAR Abdelhamid pour son aide et ses conseils précieux.

Je tiens également à remercier M.MAMMAR Khireddine , Mlle.HERIZ Omaima et M.SETTOUF Omar pour ses aides et ses conseils précieux.

Merci pour tous ceux qui, m'ont aidé de près ou de loin à réaliser ce travail.

Liste des figures

Liste des tableaux

Liste des abréviations

Abréviation	Expression Complète
RS	Réseaux sociaux
AC	Contrôle d'accès
ReBAC	Le contrôle d'accès basé sur les relations entre utilisateurs
SNFTrust	modèle de confiance flou des réseaux sociaux
SI	Système d'informations
JSON	javaScript Object Notation
PHP	Hypertext Preprocessor
XML	langage de balisage extensible
API	interface de programmation d'application.
CMS	Système de gestion de contenu
DAC	le contrôle d'accès discrétionnaire
MAC	le contrôle d'accès obligatoire
RBAC	Le contrôle d'accès à base de rôles
ABAC	Le contrôle d'accès à base d'attributs
TBAC	Le contrôle d'accès à base des tâches
TMAC	Le contrôle d'accès basé sur l'équipe
AUTrust	le modèle de confiance des utilisateurs adjacents
TC	Le contrôle d'accès basé sur la confiance
MSN	les réseaux sociaux mobiles
OSNs	les réseaux sociaux en ligne
QoTN	Qualité du réseau de confiance
NLP/TALN/TAL	Traitement du Langage Naturel
RI	Recherche documentaire

EI	Extraction de l’nformation
TF-IDF	Term Frequency Inverse Document Frequency
TF	Term Frequency
IDF	Inverse Document Frequency
SVM	Support vector Machine
KNN	K-nearest neighbors
NB	Nive Bayes
UML	Unified Modeling Language
REST	representational state transfer
SOAP	Simple Object Access Protocol
WSDL	Web Services Description Language
TCP	Transmission Control Protocol
HTTP	Hypertext Transfer Protocol
WS	Web service
URI	Uniform Resource Identifier
OMG	Object Management Group
CBAC	Context based access control

Table des matières

INTRODUCTION GENERALE

1. Introduction

La croissance spectaculaire des réseaux sociaux (RS) dans le monde d'aujourd'hui est révélatrice de leur popularité parmi les personnes de tous âges et de toutes classes [1]. Dans les médias sociaux, les gens créent leurs propres espaces pour télécharger, partager leurs informations privées (photos, vidéos et audio) avec leur famille et leurs amis en utilisant diverses formes de réseaux sociaux, par exemple Facebook, Twitter, LinkedIn…etc. Les réseaux sociaux permettent aux utilisateurs d'étendre leurs relations et d'interagir avec des étrangers au-delà de la famille et des amis. Ils peuvent facilement partager leurs informations personnelles et avoir accès aux données personnelles de leurs amis et autres utilisateurs [2].

La protection de la vie privée des utilisateurs est l'objectif principal des services fournis par les plateformes de médias sociaux. Malheureusement, aujourd'hui, la divulgation de la vie privée à travers ces sites représente un danger indéniable pour les gens, d'autant plus que de nombreux utilisateurs sont présents sur ces sites et n'ont pas conscience des conséquences des informations qu'ils partagent [3]. Pour résoudre ce problème, un moyen courant consiste à utiliser des méthodes de contrôle d'accès(AC) afin de prendre en charge la confidentialité de l'utilisateur. Plusieurs techniques de contrôle d'accès sont mis en œuvre et qui seront traitées dans ce projet dans le but d'améliorer et de préserver la confidentialité des utilisateurs. Parmi les modèles de contrôle d'accès, on s'intéresse dans ce projet de Master au contrôle d'accès basé sur la confiance, ainsi que le contrôle d'accès basé sur les relations (ReBAC).

2. Problématique

Lorsqu'un utilisateur introduit une nouvelle donnée (statuts, photos, vidéos, liens) dans son espace personnel, il a le choix entre la rendre complètement publique ou la réserver à ses contacts, donc l'utilisateur est sensé renforcer les paramètres de confidentialité sur son profil afin de mieux contrôler ses données et limiter l'accès aux informations personnelles qu'il partage. Cependant, plusieurs études démontrent que de nombreux profils sont insuffisamment protégés, et cela, pour plusieurs raisons. En effet, les utilisateurs ont des difficultés à configurer correctement leurs paramètres de confidentialité ou à les modifier ; une multitude d'options de paramètres de confidentialité à gérer ; et la politique de gestion de confidentialité n'est pas stable.

Ainsi, on peut constater le manque de la considération des données de la vie privée réelle afin d'établir des politiques de contrôle d'accès robustes pour les systèmes collaboratifs centrés sur la communauté. En conséquence, de nombreux chercheurs ont proposés des modèles de contrôle d'accès basés sur les relations entre les utilisateurs du réseau social ainsi que le degré de confiance entre eux.

3. Objectif du projet de Master

Contrôle d'accès basé sur le contenu (messages, tweets, etc) et la vie privée des utilisateurs. L'objectif est de développer un outil qui permet d'aider les utilisateurs à paramétrer leurs fonctions de partage de ressources en se basant sur le contenu de leur message ainsi que leur profil (nombre d'amis, age, genre, …).

A l'inverse des techniques de contrôle d'accès qui permettent de gérer les ressources au niveau du serveur, la proposition de ce projet de Master se repose sur un contrôle d'accès basé sur la partie client en proposant de l'assistance à l'utilisateur. L'outil permet d'analyser le texte que l'utilisateur veut poster sur un réseau social (Facebook, twitter, linkedIn, etc) puis lui recommander différents paramètres de partage ou de reformulation du texte (conseil/feedback: partage avec amis/collègues, partage non conseillé puisque c'est sur le sujet de politique, santé, …). Cette recommandation est basée sur les concepts de vie privée et de confiance.

4. Organisation de ce rapport

Ce rapport de recherche bibliographique est composé d'une introduction générale suivie de deux chapitres :

- Chapitre I : présente le contexte général de ce projet qui est les réseaux sociaux et leurs caractéristiques, ainsi que la protection de la vie privée de l'utilisateur et l'accès à ses données dans les réseaux sociaux actuels.

- Chapitre II : on va présenter le résultat de l'analyse de l'état de l'art des modèles de contrôle d'accès existants pour les réseaux sociaux.

- Chapitre III : vise à présenter le processus de la catégorisation des textes et le prétraitement des textes, ainsi que les différents algorithmes d'apprentissage automatique supervisée. Nous avons également introduit les différents moyens d'évaluation d'un classificateur.

- Chapitre IV : en mettant l'accent sur l'algorithme utilisé dans notre travail : le naïve bayésienne. Ainsi que le schéma global de notre projet.

- Chapitre V : permettra d'évaluer les performances des différentes approches implémentées en présentant les résultats obtenus avec interprétation.

Et nous avons terminé par une conclusion qui nous voir ultérieurement.

Chapitre 1

RESEAUX SOCIAUX

1. Introduction

Depuis quelques années, les réseaux sociaux sont au cœur des nouvelles technologies de la communication. Ils ne servent plus seulement à favoriser l'interaction entre parents, amis et relations professionnelles, mais tendent à devenir aussi un canal de communication le plus utilisé pour s'adresser à l'opinion publique.

Les réseaux sociaux sont l'une des conséquences du Web 2.0 et intègrent différentes activités qui sont la technologie, l'interaction sociale et la création de contenus. Il vise à faciliter la créativité, promouvoir la collaboration et le partage entre les utilisateurs. Les médias sociaux sont considérés comme un ensemble d'applications en ligne (les réseaux sociaux, les blogues, sites de partage). Avec cette technologie, on assiste aujourd'hui au développement des objets connectés, l'apparition des nouvelles applications notamment les réseaux sociaux.

Les réseaux sociaux sont des dérivés des systèmes collaboratifs traditionnels qui permettent à la fois aux utilisateurs et aux organisations de créer des communautés pour promouvoir des intérêts communs et partager du contenu les uns avec les autres. A l'inverse d'un système collaboratif traditionnel qui est centré sur les tâches pour lesquels les utilisateurs se rassemblent en « groupes » pour les faires, un réseau social est centré sur la communauté, dans laquelle les « communautés » d'utilisateurs en ligne partagent des informations et des ressources d'intérêts communs.

Les systèmes centrés sur la communauté ont un certain nombre de caractéristiques uniques, en plus de celles des systèmes collaboratifs centrés sur les tâches. Premièrement, ces systèmes sont complexes et dynamiques, où les relations interpersonnelles régulent l'interaction entre les utilisateurs. Les utilisateurs créent une communauté en établissant des relations en ligne qui ressemblent à leurs relations interpersonnelles réelles. De plus, dans les systèmes centrés sur la communauté, les utilisateurs coopèrent pour créer, gérer et protéger les ressources au sein de la communauté. Enfin et surtout, les communautés peuvent rassembler des utilisateurs ayant des antécédents et des expertises différents [4] . Les systèmes de collaboration centrés sur la communauté ont émergé et ont gagné en popularité, le besoin de mécanismes appropriés pour protéger les ressources sensibles partagées dans ces systèmes devient une préoccupation majeure.

2. Les réseaux sociaux

Les réseaux sociaux (RS) sont des applications ayant comme objectif de relier des amis, des connaissances ou des associés. Les RS présentent des orientations plus ou moins personnelles ou professionnelles, c'est-à-dire que l'objectif des utilisateurs peut être de retrouver des amis et de partager du contenu avec eux (photos, messages, commentaires, applications ludiques...) ou de tisser un réseau professionnel (rencontrer des partenaires potentiels, trouver un nouvel emploi, trouver des collaborateurs, annoncer des événements ou des activités professionnelles...).

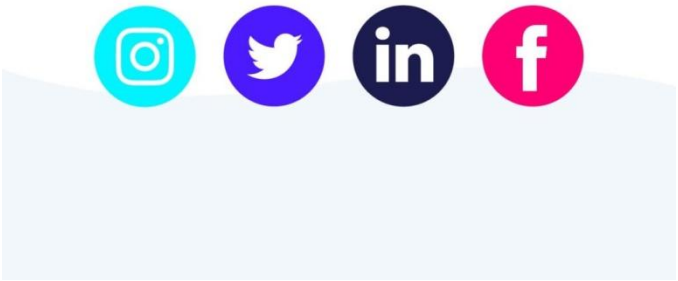

Figure 1: logos de réseaux sociaux

Un réseau social permet de :

- raccourcir les distances géographiques entre les personnes dans le monde entier ;
- sont utilisés pour le divertissement, l'éducation, la recherche d'emploi, etc. ;
- permettre aux utilisateurs de partager leurs données personnelles avec d'autres de manière relativement privée ;
- permettre aux utilisateurs de développer leurs relations sociales en liant leur profil à d'autres utilisateurs ayant des similitudes ;
- les utilisateurs, les entreprises ou les secteurs de l'éducation peuvent créer leurs propres pages et publier des informations pertinentes dans leur time line [2].

Partant de ce principe, les RS ont vu le jour. Ces derniers sont des services basés web qui permettent aux individus de construire un profil public ou semi-public dans un système fermé, de créer une liste d'utilisateurs avec lesquels ils partagent une connexion (lien entre deux utilisateurs), de visualiser et d'explorer leurs propres listes de contacts (ensemble de connexions) et celles des autres au sein du système. La nature et la nomenclature de ces connexions peuvent varier d'un site à un autre. Même si en parlant de RS, un mot nous vient automatiquement à l'esprit "Facebook", cependant il n'est pas le premier. Les géants actuels sont LinkedIn, Facebook, Twitter que nous allons parlés dans ce qui suit [3]. La figure 1 illustre quelques logos des réseaux sociaux les plus connus.

2.1 Exemples de réseaux sociaux actuels

Il existe de nombreux sites et applications de médias sociaux avec divers objectifs, services et types qui sont développés au fil des ans pour inclure différents types et catégories. Parmi les réseaux sociaux les plus connus sont Facebook, Twitter et LinkedIn :

- **Facebook** est un célèbre réseau social qui a été lancé en 2004 par Mark Zuckerberg. À la base, il n'était destiné qu'aux étudiants de l'Université d'Harvard. Depuis 2006, n'importe qui (âgé de minimum 13 ans) peut s'y inscrire afin d'y construire son réseau. Une fois inscrit sur le site, vous pouvez y retrouver des amis, des collègues, des membres de votre famille... Mais à quoi ça sert exactement ? Facebook permet de discuter avec vos proches, renouer avec des personnes, montrer son intérêt pour certaines choses, partager des photos, adhérer à des pages sur des sujets que vous appréciez. Vous pourrez ainsi rencontrer des personnes qui ont les mêmes passions ou opinions. Facebook c'est également une messagerie qui permet de chatter (discuter) en direct et d'envoyer des messages aux "amis" qui ne sont pas connectés. Le site est devenu incontournable au fil des années [5].

- **Twitter** est « un réseau d'information basé sur des messages de 140 caractères, appelés Tweets. C'est un moyen facile de découvrir les dernières actualités liées aux sujets qui vous intéressent. » Le slogan de Twitter est "Que faites-vous en ce moment ?". L'idée de départ était donc simplement de partager votre quotidien avec un certain nombre de personnes. Par la suite, Twitter est devenu également un outil de partage, car il est souvent très facile de publier sur Twitter une information que vous avez lue. Les messages que vous publiez sont vus par vos "abonnés", c'est-à-dire les personnes qui ont décidé de suivre votre fil d'actualité. Inversement, les "abonnements" sont les comptes que vous suivez. [5]

- LinkedIn est un réseau social à vocation professionnelle. Destiné aux entrepreneurs, salariés, recruteurs, personnes en recherche de stages ou d'emploi de tous les domaines possibles. Le principe de LinkedIn c'est qu'il permet de se faire connaître et se positionner en tant qu'expert. LinkedIn est également parfait pour élargir son réseau dans le but de trouver de nouveaux collaborateurs, de nouveaux clients ou un nouvel emploi (en mettant son CV sur LinkedIn par exemple). LinkedIn fonctionne sur le système de création de profil pour lequel il est recommandé d'indiquer un maximum d'informations. Vous pouvez diffuser du contenu (publications, articles, vidéo) directement depuis la plateforme, LinkedIn comprend un système de « like », commentaires et partages pour faciliter l'interaction entre les membres. Bien sûr, plus vous êtes, plus vous êtes visible .LinkedIn comprend également des groupes abordant de nombreuses thématiques et des hashtags qu'il est possible de suivre pour être rapidement au courant des derniers contenus diffusés selon un sujet précis [6].

D'autres réseaux sociaux sont aussi très utilisés. Le tableau 1 résume ces types de réseaux et fournit quelques exemples avec une description pour chaque type [2].

Catégories	Types	Exemple	Description
Réseaux sociaux	Professionnels	LinkedIn	Un réseau professionnel international qui permet la mise en relation entre professionnels
		Viadeo	Permet aussi de construire et gérer son réseau professionnel, mais beaucoup plus connu en France
	Généralistes	Facebook	il permet à l'internaute d'échanger avec sa communauté d'amis sur tout et n'importe quoi. Pour devenir ami avec quelqu'un il faut lui envoyer une demande et ce dernier doit l'accepter
		Twitter	Il permet de partager des messages avec d'autres internautes avec une limitation à 2800 caractères par message, la possibilité de suivre d'autres comptes
		MySpace	Site interactif qui offre à ses abonnés de multiples services combinant blogue, espace personnel, espace communautaire
	Partage Médias	YouTube	YouTube permet de déposer des vidéos, suivre des vidéos et faire des commentaires
		DailyMotion	Un site de partage de vidéo auquel les utilisateurs peuvent télécharger, regarder et partager des vidéos
		Instagram	Il permet aux utilisateurs de créer un compte et de pouvoir éditer, partager des photos, des vidéos et des messages avec son cercle d'amis ou famille
		Flickr	Flickr met à la disposition de son public un espace pour poster photos et vidéos. Il est cependant possible de géo localisé les endroits où les photos ont été prises

Tableau 1: Les types de réseaux sociaux

2.2 Caractéristiques des réseaux sociaux

- **Profils :** Les profils peuvent être considérés comme les briques de base du RS. Les profils contiennent généralement des informations démographiques basiques sur l'utilisateur tel que son nom, son sexe, sa ville natale et sa profession actuel... etc. Parallèlement à ces informations

personnelles considérées essentielles pour chaque profil, la plupart des RS encouragent également les utilisateurs à écrire une courte biographie sur eux-mêmes et de partager leurs goûts et leurs intérêts. Pourtant ces types d'informations ne sont pas obligatoires pour pouvoir s'inscrire sur les RS, de nombreux utilisateurs mettent beaucoup de détails facultatifs sur leurs profils [3].

- **Les amis** : La plupart des RS sont conçus et construits autour du concept d'«amis» ou «Friends». Sur un RS, un «ami» peut être un ami, un membre de la famille, une connaissance, un ami d'un ami, ou même quelqu'un que l'utilisateur n'a jamais rencontré auparavant, sauf en ligne. En 2013, le nombre moyen d'amis pour chaque utilisateur sur Facebook est de 175. Le RS permet à l'utilisateur de garder la trace des activités de ses amis. Par exemple, quand ils publient une nouvelle photo, mettent à jour leurs profils, changent leurs statuts ou lorsqu'ils achètent quelque chose de nouveau en ligne. Le RS a généralement une fonctionnalité de recherche qui peut aider l'utilisateur à trouver de nouveaux amis [3].

- **Fonctionnalités de réseautage** : En plus des relations d'amitié, certains RS proposent également des fonctionnalités de réseautage pour faciliter l'interaction entre les utilisateurs, tels que les groupes et la messagerie instantanée. Chaque RS a aussi des fonctionnalités particulières propres à lui tels que l'envoi des « pokes » sur Facebook [7].

- **Les groupes :** La plupart des RS s'appuient sur la notion de groupe pour aider les utilisateurs à trouver des personnes ayant des intérêts similaires ou à s'engager dans des discussions sur certains sujets. Parfois, les groupes sont appelés par d'autres noms, tel que «les réseaux sur LinkedIn [3].

- **Les évènements :** C'est une fonctionnalité de réseautage permettant aux «amis» de connaître les événements à venir dans leur communauté ainsi que d'organiser des rassemblements sociaux. Par exemple, sur MySpace, il est possible de publier un questionnaire ou de décorer la page de l'événement [7].

- **Les tags :** Un tag est un mot-clé ou terme assigné à un élément d'information. Par exemple, un tag peut être un bookmark en ligne, une photo numérique ou un fichier. Ce type de métadonnées décrit un objet et permet de le trouver par la recherche ou en navigant. Facebook et Friendster permettent aux utilisateurs d'associer un tag à une zone spécifique dans l'image. Par exemple, l'utilisateur peut taguer les personnes figurant dans une image d'une famille dans une place particulière par leurs noms et mettre un tag pour spécifier le nom de la place où la photo a été prise. Si le nom utilisé pour le tag est associé à un membre de Facebook ou à une page (région connue par exemple), le tag se transforme en un lien hypertexte vers le profil ou la page [3]

- **Flux d'actualité (News Feeds)** : Les flux d'actualité sont des outils utiles pour rester en contact avec les «amis». Par exemple, les mises à jour de profil, les messages sur le blog, les photos et vidéos publiées sont souvent diffusées sous forme de «news feeds» [7].

- **Les applications sociales** : Les RS comprennent un grand nombre d'applications sociales que les utilisateurs peuvent ajouter à leur profil. Ces applications peuvent être programmées à travers des

interfaces ouvertes aux développeurs tiers pour concevoir et mettre en œuvre des applications ou des jeux sur la plateforme. Les applications créées à l'aide de ces API[1] posent des problèmes pour la vie privée car elles demandent aux utilisateurs d'accéder à leurs informations personnelles (situation familiale...etc.), à leurs listes d'amis, aux informations personnelles de leurs amis... etc. Voire même de publier sur leur profil et avoir un accès permanent à ces derniers [3].

2.3 Gestion des droits et paramètres de sécurités des RS

L'intérêt croissant aux RS apporte de nombreuses questions sue l'impact qu'ont ces derniers sur la vie privée de leurs utilisateurs .En effet, dans un RS chaque utilisateur est son propre administrateur, c'est lui qui gère sa confidentialité, contrairement aux SI classiques .Plusieurs travaux portent sur la confidentialité et la vie privée des RS à cause de manque de AC .Qui a le droit de m'identifier ? Qui peut voir mes photos ou les commentées ? [3]

À cause des exigences toujours plus rigoureuses en matière de sécurité, les utilisateurs doivent s'authentifier et entrer leur paramètre de connexion (identifiant / mot de passe) propre à chacune de ces applications, et ce de façon courante. Dans le contexte de la gouvernance des systèmes et de l'application des mesures de sécurité d'informations , des dispositions ont été prises afin d'assurer la sécurité du réseau, par exemple par l'utilisation de mots de passe forts, et conservation d'historique de mot de passe. Avec tout cela, des problèmes liés à la réinitialisation des mots de passe et à "l'inconfortabilité" des utilisateurs se sont accentués. C'est dans ce cadre qu'ont été élaborées les solutions d'authentification unique pour les entreprises. Cette façon de s'authentifier permet aux utilisateurs de se loguer une seule fois, sachant que les processus de connexion aux applications dans la suite, sont gérés automatiquement [8].

Par exemple , Facebook a beaucoup évolué Sous la pression des utilisateurs et des autorités quand une information est accessible à tous ,alors tout le monde peut voir tout élément Lorsqu'il est public, , même les personnes qui ne font pas partie de vos amis sur Facebook ou en dehors en distinguant plusieurs types d'audiences , avec différentes catégories intermédiaires , en accédant aux réglages associés, pour chaque type d'information ou de publication .Certains éléments, comme votre nom et votre photo de profil, sont toujours publics sur Facebook ,vous pouvez modifier le paramètre de confidentialité Si vous avez partagé du contenu sur votre journal que vous ne voulez plus afficher, ou supprimer la publication etc. [9].

En théorie, il est donc possible de contrôler précisément l'accès à toutes les informations . En pratique, c'est plus délicat.les Publications visibles par certaines personnes et pas par d'autres, stories réutilisables ou non, accessible en mode public… par exemple: Lorsque vous publiez quelque chose sur votre journal, vous choisissez qui peut consulter votre profil, et qu'il peut voir vos publications selon s'il fait partie de l'audience pour des infos et des publications précises .Les options et les paramètres de confidentialité sont

[1] API : est un ensemble de définitions et de protocoles qui facilite la création et l'intégration de logiciels d'applications.

nombreux. Surtout, ils ne sont pas tous rangés au même endroit , la plupart figurent dans le menu paramètres de Facebook, et d'autres tout ne sont plus accessibles que via la page profil. Pour compliquer encore davantage les choses, les degrés de confidentialité varient d'un paramètre à l'autre : pour certains, le choix se limite à Visible par tout le monde ou Personne tandis que pour d'autres, on peut utiliser des listes d'amis pour affiner la vie privée. [8].

3. Framework existants pour la création des RS

Il existe de nombreux frameworks qui aident les développeurs à implémenter leur propre réseau social. Certains frameworks étant conçus pour des projets de grande envergure tandis que d'autres sont utiles pour accélérer le développement de petites tâches.

3.1 Elgg

C'est un Framework de développement rapide open source pour les applications Web socialement conscientes. C'est un excellent choix pour créer n'importe quelle application où les utilisateurs se connectent et partagent des informations [10]. Il a été utilisé pour créer toutes sortes d'applications sociales:

- réseaux ouverts (similaires à Facebook)
- d'actualité (comme la communauté Elgg)
- intranets privés/d'entreprise
- datation
- éducatif
- blog de l'entreprise

L'avantage de Elgg est qu'il contient des modules qui vous permettent d'étendre les fonctionnalités pour suivre, gérer, mettre à jour un réseau social. Les principaux modules suivants sont proposés :

- ***Activity :*** l'activité de votre réseau
- ***Profile :*** le profile de chacun des utilisateurs inscrits de votre réseau
- ***Notifications :*** la possibilité de notifier les utilisateurs inscrits de votre réseau
- ***Groups :*** les groupes à la manière de facebook
- ***Blog :*** le blog disponible pour chaque utilisateur inscrit
- ***Embed media :*** l'utilisateur peut enrichir de sa production digitale ses contributions au réseau
- ***Files : ellg*** vous permet de gérer une grande variété de fichiers qui peuvent être mis à disposition ou non sur le réseau.
- ***The wire :*** le fil d'information de votre réseau Enfin, d'un point de vue plus technique, vous disposez avec Elgg d'un accès facilité à une API en mesure de fournir des résultats en JSON, PHP sérialisé ou des données au format XML [7].

Le seul inconvénient d'Elgg est que la partie PHP doit s'exécuter en mode Apache, et non en mode CGI ce qui rend difficile le fonctionnement de Elgg sur un serveur mutualisé [7]

3.2 Hum Hub

Est une plate-forme de réseau social open source avec une grande variété de cas d'utilisation tels que l'intranet social, la communauté ou la plate-forme de collaboration. Hum Hub se compose d'une application de base, qui peut être étendue via des modules supplémentaires et ajustée à vos besoins par de nombreuses options de configuration [11].

Certains de ses principaux avantages sont [11]:

- utilisation gratuite.
- Auto-hébergé.
- Conservez vos données, sans avoir besoin de les partager avec des services externes.
- Installation et maintenance simples
- Ne nécessite qu'un environnement de serveur Web simple
- Hautement personnalisable
- Prise en charge de thèmes et des modules personnalisés
- De nombreuses options de configuration et de réglage fin
- Libre
- Développement et discussions transparents
- Soutien et contribution de la communauté
- Contact direct avec l'équipe de développement de base
- Principe de beaucoup d'yeux
- Traduit dans plus de 40 langues
- Interface utilisateur intuitive

Certes Hum Hub possède de nombreux avantages, mais on peut trouver certains inconvénients parmi eux on trouve [11] :

- Hum Hub n'est pas conçu pour prendre en charge une base d'utilisateurs aussi importante (big data).
- Hum Hub ne prend actuellement en charge qu'un concept de visibilité de contenu plutôt simple avec du contenu privé et public
- Il n'existe actuellement aucun concept disponible ou prévu pour les espaces imbriqués ou les sous-espaces
- Hum Hub ne peut actuellement pas être exploité comme un véritable système multi-client avec des bases d'utilisateurs entièrement séparées.
- Le mode invité est limité **,**il n'y a actuellement aucun moyen pour les invités de créer, commenter ou aimer tout type de contenu.
- Il n'y a actuellement aucune intégration de paiement disponible

3.3 Boonex Dolphin Pro

Est un CMS social intégré complet. Dolphin comprend un serveur multimédia, un convertisseur vidéo, des modules de partage de médias, des profils sociaux, un chat, un messager, une chronologie, des événements, des groupes, un magasin, des blogs, des forums et plus encore [12].

Les avantages de Boonex sont [13] :

- Conception personnalisable, textes, navigation, pages, blocs, actions et autorisations.
- Profils sociaux avancés, chronologies, likes, partages, votes, amis et commentaires.
- Mise en page de site réactive et adaptée aux mobiles, ainsi que des applications iOS et Android natives.
- Tout nouveau chat audio/vidéo WebRTC avec des salles privées et publiques, des groupes et plus encore.

Le seul inconvénient de l'utilisation de l'édition gratuite est qu'il existe des publicités intégrées pour Boonex Dolphin dans le produit qui ne peuvent pas être supprimées sans acheter une licence [14].

3.4 Ning

Est la plus grande plateforme SaaS au monde pour créer des sites de réseautage avec de grandes intégrations sociales.Les sites Ning sont évolutifs et offrent un hébergement rapide, des capacités analytiques poussées et des options de monétisation avancées [15]. Ils peuvent être aussi utilisés pour une variété de fins, y compris transactions commerciales, la promotion de bonnes causes et de discuter des intérêts spécialisés, ainsi que, bien sûr, faire de nouveaux amis [16].

De nombreux avantages de Ning qu'on peut trouver parmi eux [17] :

- Le service est gratuit :
- L'interface NING est claire, souple et on se l'approprie très facilement
- La gestion des droits sur NING reste sous la responsabilité de la communauté créée.
- La gestion de la plateforme est très simple et très bien conçue
- La conception des réseaux NING permet de créer des écosystèmes relationnels et informationnels.
- Le référencement de la plateforme devient intéressant.
- La personnalisation des plateformes est aujourd'hui un point fort.
- Les contacts avec l'équipe NING ont toujours été excellents et constructifs

Ning possède par contre quelques limites telles que :

Ning, c'est rassembler une communauté avec le moins d'investissement possible [18].

- Ning est vraiment un outil de marketing, c'est pourquoi il est si coûteux d'utiliser et est idéal pour les entreprises et les constructeurs de marque [18].

4. Conclusion

Les RS apportent des bénéfices claires à la société, non pas parce qu'ils marquent la fin des médias passifs, mais puisqu'ils apportant un contenu libre et interactif créé par l'utilisateur. On a présenté dans ce chapitre les principaux leaders mondiaux pour les réseaux sociaux, leurs caractéristiques, et leurs fonctionnements .Ainsi que des différents Frameworks et bases de données existantes pour les RS qu'on les utiliserait dans les chapitres suivants. Nous avons aussi voir les politiques de confidentialité qui ont un impact majeur sur la vie privée des utilisateurs qui nous poussons à découvrir des modèles de contrôle d'accès divers qu'on va les détaillera dans le chapitres suivants.

Chapitre 2

ANALYSE DE L'ETAT DE L'ART DES MODELES DE CONTROLE D'ACCES POUR LES RESEAUX SOCIAUX

1. Introduction

Le contrôle d'accès est un mécanisme de sécurité important qui est utilisé pour autoriser et limiter l'accès des utilisateurs aux informations en empêchant les utilisateurs de faire des accès illégaux en violation des politiques de sécurité [1]. Il existe principalement trois modèles traditionnels de contrôle d'accès : le contrôle d'accès mandataire (MAC), le contrôle d'accès discrétionnaire (DAC) et le contrôle d'accès par rôle (RBAC). Ces méthodes ont été confrontées à de nouveaux défis dans un environnement ouvert et distribué, et peuvent difficilement répondre aux besoins dynamiques car il existe de nombreux utilisateurs et ressources sur les réseaux sociaux, qui sont en constante évolution. Ainsi, l'augmentation linéaire du nombre d'utilisateurs et le contrôle local sont les principaux problèmes. Les utilisateurs des réseaux sociaux n'ont pas d'interactions préalables, ce qui rend plus difficile l'obtention d'informations fiables. Il est plus important d'établir un niveau acceptable de relation de confiance entre les utilisateurs participants. Par conséquent, le problème de la confiance entre les utilisateurs devient de plus en plus important dans les réseaux sociaux. Pour surmonter les lacunes des modèles de contrôle d'accès traditionnels et certains chercheurs ont introduit le calcul de degré la confiance entre les utilisateurs ainsi que la considération des relations entre les utilisateurs dans les modèles de contrôle d'accès. Ceci a ouvert une nouvelle voie vers la création de modèles de contrôle d'accès basés sur la confiance qu'on va les traiter dans ce chapitre [19].

2. Les modèles de contrôle d'accès classique

En général, la sécurité des systèmes dépend des moyens mis en œuvre pour garantir les objectifs ou propriétés de sécurité fondamentales que sont la confidentialité, l'intégrité, la disponibilité. Pour cela ,il ont proposés les modèles de contrôle d'accès dits classiques se sont rapidement dégagées. Les modèles de contrôle d'accès discrétionnaire (DAC) ont pour caractéristique principale que les utilisateurs (sujets) peuvent être propriétaires des données et par conséquent attribuer les permissions sur les objets qu'ils possèdent. Le modèle de contrôle d'accès mandataire(MAC) qu'est utilisé lorsque la politique de sécurité impose que les décisions de protection ne soient plus confiées au propriétaire des objets concernés [20].

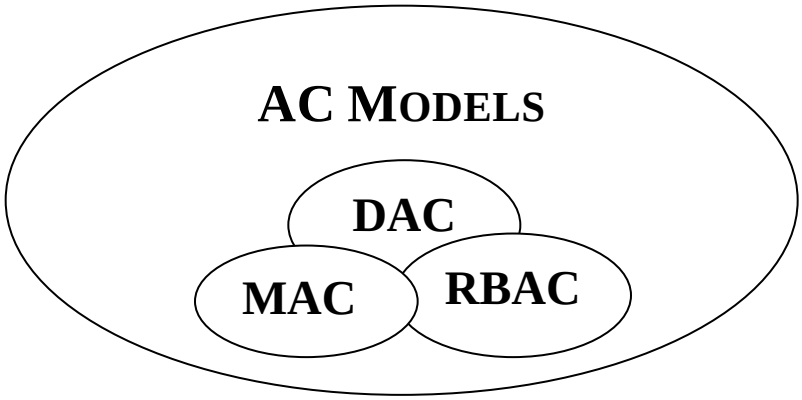

Figure 2 :Les modèles de contrôle d'accès traditionnels

Ces deux modèles reposent sur l'exploitation du triplet sujet (S), objet (O), Action (A)) pour représenter les politiques de contrôle d'accès.

Une politique de contrôle d'accès classique peut alors être élaborée à partir des entités fondamentales suivantes :

- Un ensemble de sujets (S) : Un sujet peut être un processus, un utilisateur, une application etc.

- Un ensemble d'objets (O) : Un objet est un conteneur d'informations, sur lequel un sujet peut effectuer des actions (exemples : fichiers, sockets de communication, périphériques matériels, etc.)

- Un ensemble d'actions (A) : représente l'action à traiter par le sujet sur l'objet. (exemples : lecture, écriture, exécution d'un fichier, etc.)

Leurs limites ont conduit à la proposition du modèle RBAC pour s'appuyer sur la notion de rôles dans les règles de politiques que l'on va les décrire dans les paragraphes suivants.

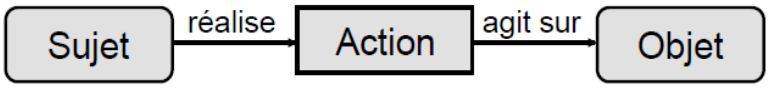

Figure 3:Les règles de AC

2.1 Le contrôle d'accès discrétionnaire(DAC)

C'est un modèle de contrôle d'accès qui restreint l'accès aux objets en fonction de l'identité des sujets et/ou des groupes auxquels ils appartiennent .Il est basé sur l'identité de l'utilisateur et règles de contrôle d'accès. Chaque demande d'un utilisateur d'accéder à un objet est évaluée avec les autorisations spécifiées dans la matrice de contrôle d'accès. S'il existe une autorisation déclarant que l'utilisateur peut accéder à l'objet dans le mode actuel, l'accès est donc autorisé, sinon c'est refusé [2]. Dans une matrice d'accès , les lignes représentent les utilisateurs, les colonnes représentent les objets et chaque entrée [s,o] indique les droits d'accès que l'utilisateur ''s'' a sur l'objet'' o'' [4].

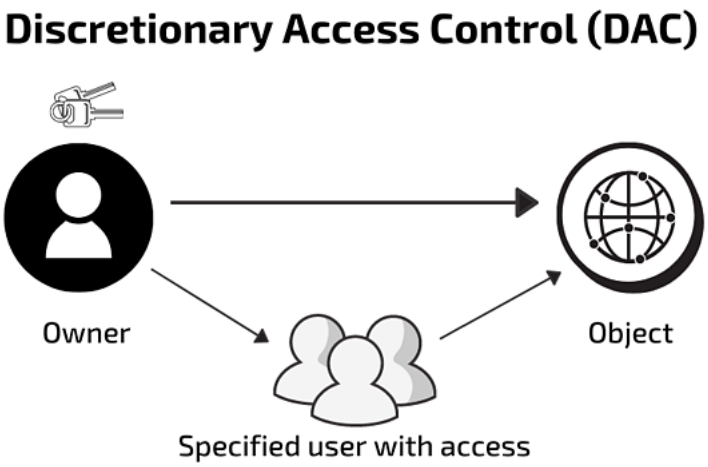

Figure 4: le contrôle d'accès discrétionnaire

Ce modèle est facile à implémenter et à entretenir, et offre une grande flexibilité, granulaire et convivial que les utilisateurs peuvent gérer leurs données et accéder rapidement aux données des autres utilisateurs [21]. Les politiques de contrôle d'accès discrétionnaires ont, malheureusement, l'inconvénient de ne pas fournir une assurance réelle sur le flux de l'information dans un système et il est souvent facile de contourner les restrictions d'accès [22]. Il est faible niveau de protection des données et obscur ; pas de gestion d'accès centralisés [21].En rajoutant d'autre limitations concernant la matrice, en particulier lorsque l'on considère les systèmes centrés sur la communauté et les environnements collaboratifs en général. Tout d'abord, la matrice d'accès est un modèle de contrôle d'accès basé sur l'identité, dans lequel les droits d'accès doivent être spécifiés pour chaque sujet et objet individuellement. Cela rend la spécification et la gestion des droits d'accès peu pratiques pour les systèmes ouverts et complexes comme les systèmes centrés sur la communauté. De plus, ce modèle ne tient pas compte des informations contextuelles et, par conséquent, n'est pas en mesure de saisir la nature dynamique de la collaboration et de la communauté [4]. Donc, il n'est pas adopté dans les sites de réseaux sociaux.

	Fichier1	**Fichier2.txt**	**Compilateur C**	**Sys_clk**	**Imprimante**
Utilisateur 1	ORW	R	X	R	W
Utilisateur 2	R	R	X	R	W
Administrateur	-	ORW	OX	ORW	O

Tableau 2:La matrice d'accès

2.2 Le contrôle d'accès mandataire (MAC)

Le modèle de contrôle d'accès mandataire (MAC) a été présenté pour le domaine militaire en 1970 pour inclure l'utilisation d'un noyau de sécurité par Bell et LaPadula (BLP) et BIBA.. MAC est basé sur l'idée de niveaux de sécurité qui sont associés à chaque sujet et objet c.à.d. basé sur la classification des sujets et des objets. Le principe consiste à attribuer une classe d'accès à chaque sujet et à chaque objet [22] .Dans ce modèle, les utilisateurs ne peuvent pas définir eux-mêmes les droits AC [2]. D'après la politique de contrôle d'accès, un fichier contenant des informations sensibles ne peut être accédé que par un utilisateur exécutant une application d'un niveau de sécurité [22].

MAC apporte de nombreux avantages. Tout d'abord, un haut niveau de protection des données que Les utilisateurs réguliers ne peuvent pas modifier les attributs de sécurité, même pour les données qu'ils ont créées. En outre, il est granulaire et Immunisé contre les attaques de cheval de Troie[2]. Enfin, il a une scalabilité aisée. Malheureusement, à causes de ces limitations qui sont la maintenabilité : La configuration manuelle des niveaux de sécurité et des autorisations nécessite une attention constante de la part des administrateurs .L'évolutivité : ne s'adapte pas automatiquement. Il est pas convivial (Les utilisateurs doivent demander l'accès à chaque nouvelle donnée) ; ils ne peuvent pas configurer les paramètres d'accès pour leurs propres données. il est très rigide et impose des restrictions sur l'accès des utilisateurs qui, conformément aux politiques de sécurité, ne permettent pas les modifications dynamiques. Qu'on ne peut pas l'utiliser dans les sites des réseaux sociaux [21].

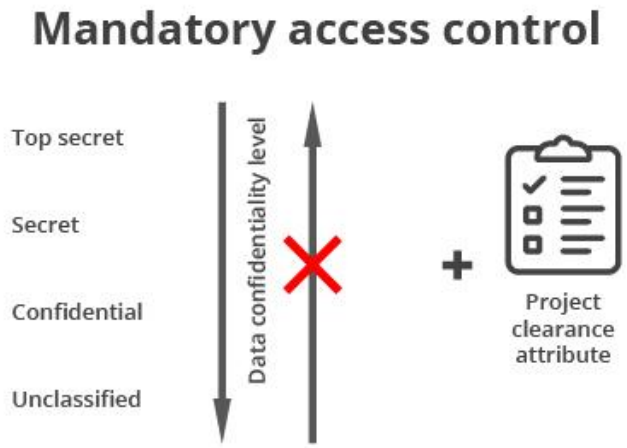

Figure 5: le contrôle d'accès mandataire

[2] Cheval de troie : ou trojan est un virus informatique, il est appelé ainsi car c'est un piège qui fonctionne de la même manière (il rentre discrètement puis fait des destructions).

2.3 Le contrôle d'accès à base de rôles (RBAC)

Ce modèle est utilisé pour répondre aux problèmes du grand nombre des objets, des sujets et des définitions des droits d'accès. Les utilisateurs sont affectés au rôle, les objets sont affectés aux groupes, et les rôles ont des droits définis et peuvent être organisés hiérarchiquement avec le soutien des droits de succession [19]. Dans RBAC, les autorisations ne sont pas attribuées directement aux utilisateurs ; au lieu de cela, les autorisations sont attribuées aux rôles et les utilisateurs héritent des autorisations attribuées au(x) rôle(s) dont ils disposent (directement ou via la hiérarchie des rôles) [4].Le modèle de contrôle d'accès à base de rôle permet de simplifier l'administration et augmenter la performance ainsi que faciliter la monté en charge ou scalabilité. RBAC peut nous servir pour réponde à la contrainte d'utilisabilité puisqu'il permet de minimiser et simplifier les tâches d'administration et de configuration concernant le contrôle d'accès [4].

Parmi ses limites est qu'il est n'est pas adapté à un contexte dynamique et distribué. RBAC attribue les rôles de manière statique à son utilisateur et il est impossible de modifier les droits d'accès d'un utilisateur sans modifier son rôle [21].

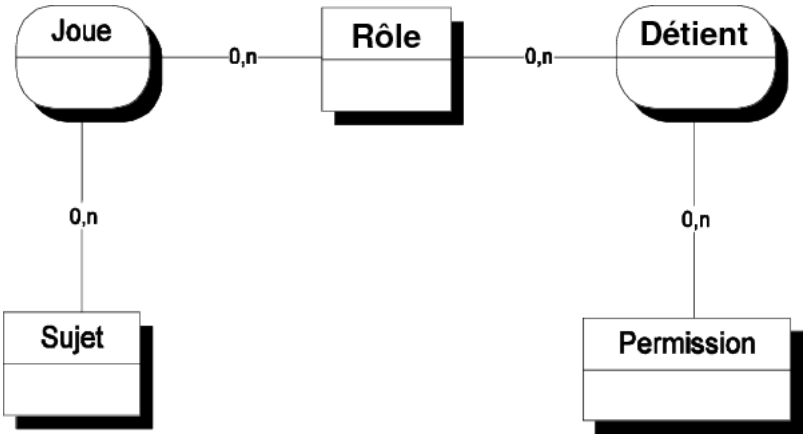

Figure 6: Le modèle RBAC

2.4 Le contrôle d'accès à base d'attributs (ABAC)

ABAC est un modèle qui a évolué à partir du RBAC .Un modèle de contrôle d'accès où la décision de contrôle d'accès d'un sujet à un objet, est déterminé par l'évaluation des attributs associés au sujet, à l'objet, aux opérations demandés ou dans certains cas des conditions d'environnement. Il exprime un ensemble de règles booléennes complexes pouvant évaluer de nombreux attributs différents. Il permet de regrouper ou jouer le rôle de plusieurs modèles de contrôle d'accès statique et dynamique. Cela peut être réalisé en associant des attributs statiques et dynamiques aux objets, sujets, ou opérations [19].

Le principal avantage d'ABAC est qu'il accorde l'accès en fonction non pas du rôle de l'utilisateur mais des attributs de chaque composant du système. De cette façon, vous pouvez décrire une règle métier de n'importe quelle complexité. Même si vous devez rendre certaines données accessibles uniquement pendant les heures de travail, cela peut être facilement fait avec une politique simple. De plus, les règles ABAC peuvent évaluer les attributs des sujets et des ressources qui doivent encore être inventoriés par le système d'autorisation [21].
En ce qui concerne les limitations ABAC, ce type de système est difficile à configurer en raison de la manière dont les politiques doivent être spécifiées et maintenues. Il est difficile d'effectuer un audit préalable et de déterminer les autorisations disponibles pour un utilisateur spécifique. Il pourrait être impossible de déterminer l'exposition au risque pour un poste donné [21]. Les attributs de l'environnement permettent à ce modèle d'être sensible au contexte, le rend ainsi adapté à une variété d'applications, y compris les systèmes centrés sur la communauté. De plus, les relations interpersonnelles peuvent également être codées en tant qu'attributs. Cependant, il a été observé à Crampton et Sellwood (2014) qu'il peut être difficile de saisir et de gérer la dynamique complexe des communautés (par exemple, les chaînes de relations interpersonnelles) dans ABAC [4].

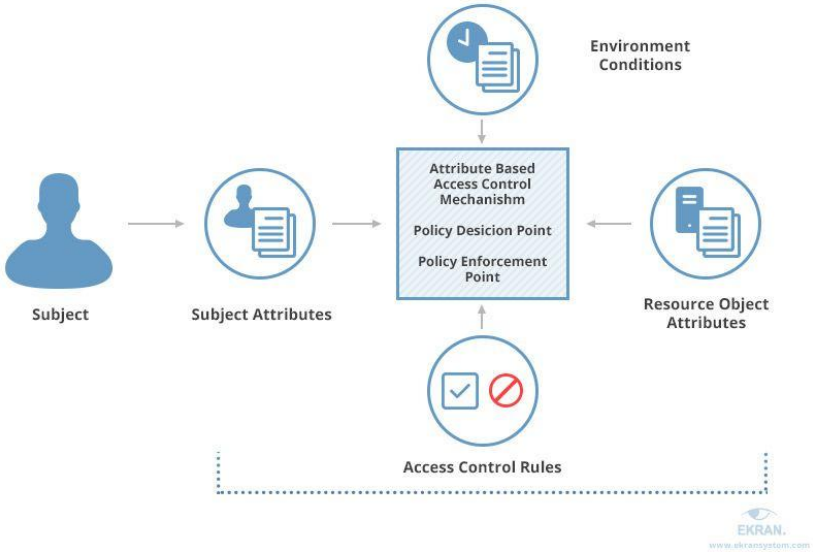

Figure 7: Le modèle ABAC

2.5 Le contrôle d'accès basé sur les tâches (TBAC)

TBAC a été conçu afin d'activer une permission par rapport aux tâches effectuées par l'utilisateur. L'idée essentielle de ce modèle consiste à ajouter la notion de tâche dans des règles d'autorisation. Cela permet de définir les permissions qu'un sujet peut activer selon la tâche qui est en cours. Chaque étape d'autorisation correspond à certaines activités ou tâches dans le contexte plus large. Le modèle TBAC fut le premier modèle à introduire le concept de tâche. TBAC sert à structurer et contrôler la réalisation d'actions composites, appelées tâches ou activités. Il ajoute une étape d'autorisation qui permet de définir les permissions qu'un sujet peut activer selon la tâche qui est en cours. Ainsi, TBAC offre une approche pour différencier l'affectation et l'activation des permissions par rapport à des tâches données aux utilisateurs [22].

TBAC propose d'organiser et de gérer les autorisations des utilisateurs par rapport aux tâches à exécuter comme il est illustré dans la figure 8. Bien que les modèles de contrôle d'accès susmentionnés fournissent certaines fonctionnalités de base pour faciliter la spécification et la gestion des politiques de contrôle d'accès pour les environnements collaboratifs, ils ne sont pas en mesure de gérer pleinement la complexité de la collaboration et des communautés [4].Il fournit un support très limité pour la modélisation et le raisonnement sur le contexte ; les informations contextuelles sont généralement limitées aux tâches et à la progression du flux. De plus, ils n'utilisent pas les relations interpersonnelles entre les utilisateurs dans la prise de décision d'accès [4].

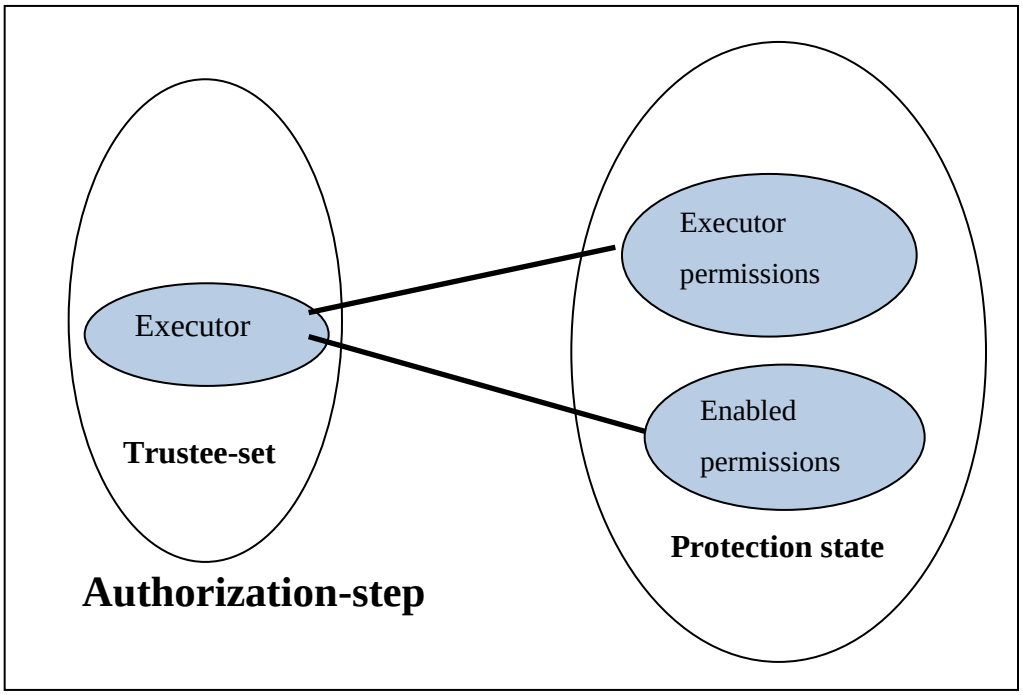

Figure 8 : L'autorisation dans le modèle TBAC

2.6 Le contrôle d'accès basé sur l'équipe (TMAC)

En 1997, Thomas a introduit la notion d'équipe dans le contrôle d'accès pour capter l'effet de la dynamique de groupe sur la protection des ressources sensibles. [23]

TMAC étend RBAC avec la notion d'équipe pour regrouper les utilisateurs à des fins de contrôle d'accès. Les équipes sont structurées en termes de rôles, offrant ainsi un moyen de définir le contexte de collaboration pour les activités à exécuter. Cependant, TMAC ne tient pas compte des relations interpersonnelles entre les membres de l'équipe, limitant son applicabilité aux environnements centrés sur la communauté. De la même manière que TMAC, Bullock et Benford (1999) présentent un mécanisme de contrôle d'accès qui vise à soutenir les personnes travaillant ensemble dans des équipes collaboratives. Il

suppose qu'il n'y a pas de communication entre les utilisateurs .Il repose sur des coalitions et des organisations, avec des rôles et des hiérarchies prédéfinis entre les utilisateurs. Cependant, dans les réseaux sociaux, les décisions d'accès sont souvent prises en fonction des relations ce qui rend ce modèle limité [4].

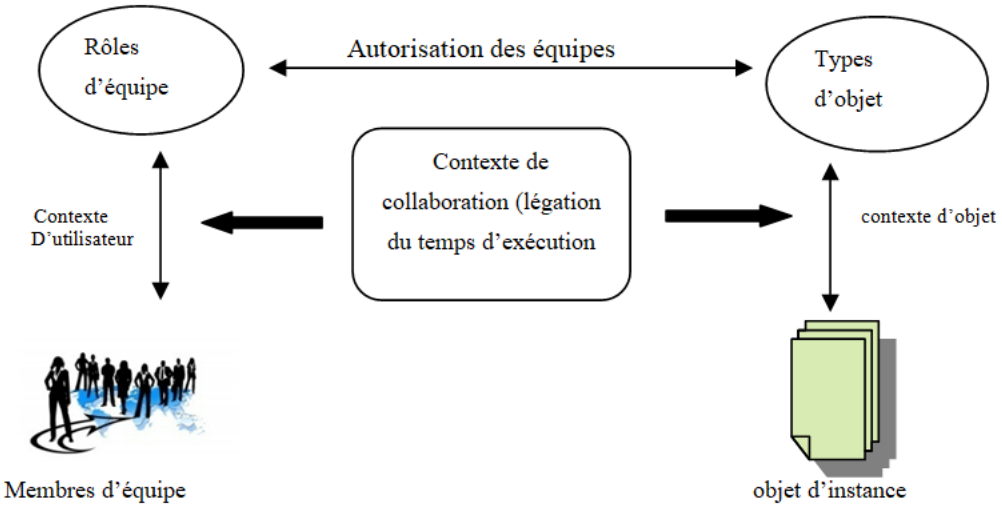

Figure 9: Illustration des concepts dans le TMAC

Les modèles de contrôle d'accès classiques définissent une relation directe entre les sujets et les objets. Ces modèles classiques sont développés pour résoudre des problèmes de sécurité traditionnels comme la confidentialité et l'intégrité. Mais, ils ont trouvé leurs limites : trop rigides, insuffisamment sûrs ou difficiles d'administration. A l'usage, une limite importante de ces modèles est apparue : la politique d'autorisation devient rapidement complexe à exprimer et administrer. Il est en effet nécessaire d'énumérer les autorisations pour chaque sujet, action ou objet. En particulier, lorsqu'un nouveau sujet ou objet est créé. Il est nécessaire de mettre à jour la politique d'autorisation pour définir les nouvelles permissions associées à ce sujet ou objet [20]. Ils ne supposent aucune connexion entre les utilisateurs ; ils sont basés sur des coalitions et des organisations bien définies, avec des rôles et des hiérarchies prédéfinis parmi les utilisateurs. Au contraire, dans les systèmes centrés sur la communauté, les décisions d'accès sont souvent prises selon des relations , qui peuvent évoluer dans le temps et ne sont pas régies par une unité administrative centrale .Bien que les modèles de contrôle d'accès susmentionnés fournissent certaines fonctionnalités de base pour faciliter la spécification et la gestion des politiques de contrôle d'accès pour les environnements collaboratifs, ils ne sont pas en mesure de gérer pleinement la complexité de la collaboration et des communautés [4].

Le tableau [4] résume la comparaison entre les différents modèles de contrôle d'accès classiques précédemment décrits.

Facteurs	DAC	MAC	RBAC	ABAC	TBAC
AC aux informations	Le propriétaire	règles fixes	Les rôles	Les attributs	Les des règles d'autorisation.
AC basé sur	l'identité d'utilisateur	classification des utilisateurs et des données	Classification des rôles	Evolution des attributs	Les tâches
Flexibilité	Haute	faible	haute	Très haute	Très haute
Complexité	Haute	faible	moyenne	moyenne	moyenne
Scalabilité	Oui	oui	Oui	non	oui
Granularité	Oui	oui	non	oui	oui

Tableau 3: Comparaison entre les différents modèles de contrôle d'accès

3. Modèles de contrôle d'accès pour les réseaux sociaux

3.1 Le contrôle d'accès basé sur les relations entre utilisateurs (ReBAC)

Pour surmonter les limitations des modèles de AC classiques, ces dernières années ont vu l'émergence d'une nouvelle tendance visant à aider les utilisateurs dans la spécification de politiques de contrôle d'accès fines pour les systèmes centrés sur la communauté. Comme indiqué pour la première fois par Carminati et al. (2006), que le seul recours aux relations interpersonnelles directes n'est pas assez flexible pour désigner les utilisateurs autorisés dans les systèmes centrés sur la communauté [24]. Sur la base de cette observation, Gates (2007) a introduit un nouveau paradigme de contrôle d'accès basé sur les relations interpersonnelles, appelé Relationship-Based Access Control (ReBAC), et plusieurs modèles ReBAC (U2U ,U2R, R2R) [25], [26], [27], [28] , [29], [30]ont été proposées.

Dans les Réseaux sociaux en ligne (OSNs), l'accès aux ressources est généralement contrôlé en fonction de la relation entre l'utilisateur accédant et l'utilisateur contrôlant la cible trouvée sur le sous-graphique social. Ce type de contrôle d'accès basé sur la relation [10] prend en compte l'existence d'une relation particulière ou d'une paire séquence particulière de relations entre les utilisateurs et exprime les politiques de contrôle d'accès en termes d'utilisateur à utilisateur(U2U) [31].
ReBAC est un modèle où les décisions d'accès sont basées sur les relations d'un sujet. Lorsque le sujet (souvent un utilisateur, mais peut-être aussi un appareil ou une application) souhaite accéder à une ressource, notre système autorisera ou refusera cet accès en fonction des relations spécifiques du sujet [32].

En particulier, les décisions d'accès sont prises sur la base de relations primitives (par exemple, ami) et composites (par exemple, ami d'un ami) entre le demandeur de ressource et le propriétaire de la ressource. Il convient de noter que les relations composites ressemblent à des chaînes de confiance et de délégation d'autorité, qui ont été largement étudiées dans le domaine de la gestion de la confiance.

Les exemples les plus connus de contrôle d'accès basé sur les relations sont probablement les réseaux sociaux. Dans Facebook, par exemple, vous pouvez autoriser l'accès à vos publications et photos à des amis d'amis. Mes amis peuvent voir mes messages. Les amis de mes amis peuvent voir mes messages. Mais les amis de ces amis ne peuvent pas parce qu'ils ont le mauvais chemin relationnel vers mon message (une étape de trop).Ainsi, dans ReBAC, nous n'autorisons pas l'accès parce que quelqu'un a un certain rôle (par exemple, un utilisateur du groupe (Ressources humaines). Nous autorisons l'accès car ils ont certaines relations avec d'autres entités de notre système. ReBAC est souvent expliqué dans la littérature académique en référence aux réseaux sociaux car, par définition, ils contiennent un réseau de relations [32].

Figure 10: Graphe sociale qui représente les relations internationales

Dans ce modèle, les graphes sociaux proposés par Bruns, al,Carminati ,Fong Crampton et Sellwood sont étendus en associant des arêtes à une étiquette indiquant le taper relation entre deux entités (par exemple, un ami, un collègue, une famille). Les modèles ReBAC expriment généralement les contraintes d'accès en termes de chemins dans le graphe social dont les nœuds désignent les entités au sein du système et les bords désignent les relations interpersonnelles entre ces entités [4], [28], [29], [26], [33].

Fong représente les contraintes d'accès sous forme de formules dans la logique modale fournissant ainsi une base mathématique solide pour le contrôle d'accès basé sur les relations et des capacités d'analyse formelle pour évaluer l'exactitude des politiques [33]. Cependant, il a été montré que la logique modale n'est pas complète sur le plan de la représentation et que plusieurs politiques basées sur les relations typiques des réseaux sociaux de style Facebook, abordant ce problème. [4].

Crampton et Sellwood proposent un modèle ReBAC générique basé sur les conditions de chemin. Les conditions de chemin, représentées sous forme de séquences de relations, sont utilisées pour mapper le demandeur de ressource à un (ensemble de) principal(s) d'autorisation [28].

Carminati et al, proposent d'utiliser la profondeur et force de relations. La profondeur des relations représente la longueur du chemin le plus court entre deux nœuds du graphe social, et elle est utilisée pour

contrôler le rayon du cercle social auquel l'accès doit être accordé, dans la lignée des travaux antérieurs dans les domaines de la gestion de la confiance et l'ingénierie des exigences de sécurité. De plus, Carminati et ses collègues annotent les relations interpersonnelles avec un niveau de confiance représentant la force des relations et spécifient les politiques de contrôle d'accès en termes de conditions d'accès [4], [29], [34].

3.2 Contrôle d'accès à base de confiance

3.2.1 La confiance

La confiance joue un rôle important dans les réseaux sociaux et est la base de toute relation car les réseaux sociaux créent un monde virtuel autour de nous, qui devient une partie de notre monde réel et de nos utilisateurs qui se font confiance et interagissent les uns avec les autres sont les principales sources de pouvoir et de croissance pour eux. Le manque de confiance dans les réseaux sociaux empêche les utilisateurs d'exprimer leurs opinions et de partager leurs idées [1].

La confiance a plusieurs caractéristiques, dans lesquelles spécifiques au contexte, dynamiques, propagatives, non transitives, subjectives, l'asymétrie et la sensibilité aux événements sont les plus importantes. Il peut être difficile d'exprimer la confiance par des théories mathématiques en raison de l'incertitude de la confiance. Dans relations humaines, la confiance s'exprime plus fréquemment de manière descriptive que forme numérique; par conséquent, la théorie floue peut être utilisée pour résoudre le problème de l'incertitude de la confiance [1].

La confiance est une notion subjective. Cependant, la confiance peut être quantifiée en analysant le comportement des utilisateurs et nous pouvons lui attribuer une valeur dans l'intervalle réel [0, 1], 0 signifie absolument pas digne de confiance, et 1 signifie totalement fiable. Chaque utilisateur possède une valeur de confiance particulière envers les autres à chaque instant ou au cours d'un certain intervalle de temps. La valeur de confiance d'un utilisateur change seulement comme conséquence d'une interaction avec les autres utilisateurs. Un utilisateur avec une valeur de confiance élevée sera assigné à un rôle principal, et effectue des privilèges supérieurs, tandis qu'un utilisateur avec une valeur de confiance faible peut s'être refusé l'accès au ressources du système [22].

La valeur de confiance d'un utilisateur peut être évaluée par sa valeur de la réputation. Si la valeur de la réputation locale d'un utilisateur est « t » et la valeur de la réputation globale est « T », alors la valeur de la confiance est souvent calculée comme suit :

$$\text{Trust} = \alpha * t + (1- \alpha) * T \quad 0 < \alpha < 1,$$

où α est le coefficient de poids défini par le système selon les applications, et « T » la réputation globale est l'ensemble des réputations des autres utilisateurs interagis avec cet utilisateur [22].

La fiabilité de chaque personne dans le réseau social est basée sur sa personnalité virtuelle parce que leurs activités se déroulent sur le Web. Ainsi, il existe une forte demande pour la gestion de la confiance, qui permet le calcul et analyse de la confiance entre les utilisateurs et les aide à prendre de meilleures décisions sur la base des informations de confiance. De plus, la gestion de la confiance réduit les

dommages causés à de nombreuses applications telles que le contrôle d'accès, authentification, présentation de services sécurisés, routage sécurisé, etc. [1].

3.2.2 Modèles de contrôle d'accès existants à base de confiance

Un système informatique de confiance présenté pour aborder le problème de la simulation de confiance. Ce système est capable de simuler la confiance entre deux individus directement connectés sur les réseaux sociaux. En intégrant des valeurs de confiance calculées à travers trois composants de calcul de confiance basés sur la similarité des profils, la fiabilité des informations et les opinions sociales. Le système renvoie finalement le score de confiance qui représente l'ampleur réelle d'un individu particulier à l'autre [1].

3.2.2.1 Modèle de confiance des utilisateurs adjacents(AUTrust)

Ce modèle a été proposé pour calculer la confiance entre les utilisateurs adjacents et pour construire un réseau de confiance sociale. Les facteurs affectant la confiance des utilisateurs peuvent être classés en trois dimensions de similarité, de familiarité entre les utilisateurs et de réputation sociale de l'utilisateur. La confiance sociale de chaque utilisateur est calculée par la moyenne de toutes les confiances de ses utilisateurs adjacents à lui. Un utilisateur des réseaux sociaux de confiance est une personne, dans laquelle ses voisins le considèrent comme un individu de confiance ; sinon, l'utilisateur n'est pas dans cette confiance des réseaux sociaux [35].

Pendant ce temps, AUTrust se concentre uniquement sur l'évaluation du niveau de confiance entre les voisins (utilisateurs adjacents), et non les non-adjacents [1].

3.2.2.2 Le modèle de confiance TC (TrustComp)

L'évaluation de la confiance qui a été présentée dans le modèle précédent est considérée comme un problème de classification, et une nouvelle approche utilisant une méthode d'apprentissage automatique a été présentée, appelée TC. Premièrement, le vecteur de caractéristiques de confiance est construit en fonction des facteurs liés à la confiance. Ensuite, en s'entraînant avec des échantillons de données collectés qui contiennent des vecteurs de caractéristiques de confiance et des évaluations de confiance, un classificateur de confiance a été établi. Enfin, le niveau de confiance de l'ensemble du réseau et la réputation sociale de tous les utilisateurs ont été évalués [36]. Il est difficile de calculer la réputation de l'utilisateur dans ce système lorsqu'un utilisateur a été commenté avec beaucoup de mauvaises critiques, là où sa réputation devrait tomber ; cette question, cependant, n'est pas reflétée dans cette approche. De plus, la mesure de la fiabilité du système TC est assez difficile. [1].

3.2.2.3 Le modèle de confiance K-FuzzyTrust

C'est un mécanisme d'inférence de confiance efficace basé sur la communauté floue. Ce mécanisme est appelé le Kappa-FuzzyTrust .Il est une nouvelle approche pour déduire des valeurs de confiance pour les réseaux sociaux mobiles (MSN) à grande échelle qui est basé sur l'utilisation d'un un algorithme de

détection de la structure communautaire dans les réseaux complexes sous degré flou κ et construire un graphe social implicite flou fortement connexe pour détecter les relations entre les utilisateurs et présentons une approche de calcul de confiance efficace, déterminer le niveau d'adhésion de ces relations, et pour améliorer la compréhension de la confiance entre les utilisateurs mobiles. Les propriétés MSN critiques sont analysées et construites dans un contexte de confiance mobile basé sur deux aspects : des attributs statiques et des modèles de comportement dynamiques, tels que le profil de l'utilisateur, le prestige, la familiarité avec les interactions, l'évaluation des interactions, les identités de lieu et de temps. Pour calculer les valeurs de confiance globales entre les nœuds connectés indirectement, la méthode considère l'agrégation et la propagation des valeurs de confiance pour les utilisateurs connectés indirectement et qui se chevauchent [1, 37].

Hao et al. a proposé un nouveau mécanisme d'inférence floue, à savoir MobiFuzzyTrust, pour déduire sémantiquement la confiance d'un utilisateur mobile à un autre qui peut ne pas être directement connecté dans le graphe de confiance de MSN. Dans cette approche, un contexte mobile, ainsi que le crédit des utilisateurs, l'emplacement, l'heure et le contexte social ont été créés, puis le modèle de confiance sensible au contexte mobile a été créé pour évaluer efficacement la valeur de confiance entre deux utilisateurs mobiles. Enfin, la technique linguistique floue a été utilisée pour exprimer la confiance entre deux utilisateurs mobiles et améliorer la compréhension humaine de la confiance. Il s'agit généralement de preuves de confiance simples agrégées pour calculer directement une valeur de confiance [37].

Liu et al, ont proposé une structure de réseau social axée sur la confiance contextuelle qui prend en compte les facteurs d'impact contextuels sociaux, notamment la confiance, le degré d'intimité sociale, le facteur d'impact communautaire, la similarité des préférences et la distance de localisation résidentielle [13]. Ces facteurs ont une influence significative à la fois sur les interactions sociales entre les participants et sur l'évaluation de la confiance. Ensuite, un nouveau concept QoTN (Quality of Trust Network) et un modèle d'extraction de réseau de confiance sensible au contexte social ont été proposés. Enfin, un algorithme d'extraction de réseau de confiance heuristique tenant compte du contexte social (H-SCAN-K) en étendant la méthode K-Best-First Search (KBFS) avec de nouvelles stratégies d'optimisation utiles a été suggéré [1].

3.2.2.4 Le modèle de confiance SNFTrust

A base de limitations des modèles précédents, cette nouvelle méthode SNFTrust de contrôle d'accès basée sur la confiance dans les réseaux sociaux est proposée en utilisant un système d'inférence floue. SNFTrust relie le contrôle d'accès à la confiance qui sont très pertinents les uns pour les autres. Elle est une méthode flexible car il permet de déterminer la confiance en fonction de n'importe quel ensemble d'attributs. De plus, il peut être étendu en termes de variables linguistiques de droit d'accès à plus d'éléments dans différentes situations. Faisant partie de SNFtrust, le module de contrôle d'accès est chargé de mapper le droit d'accès à l'autorisation d'accès, par conséquent, le nombre d'ensembles doit également être égal [1].

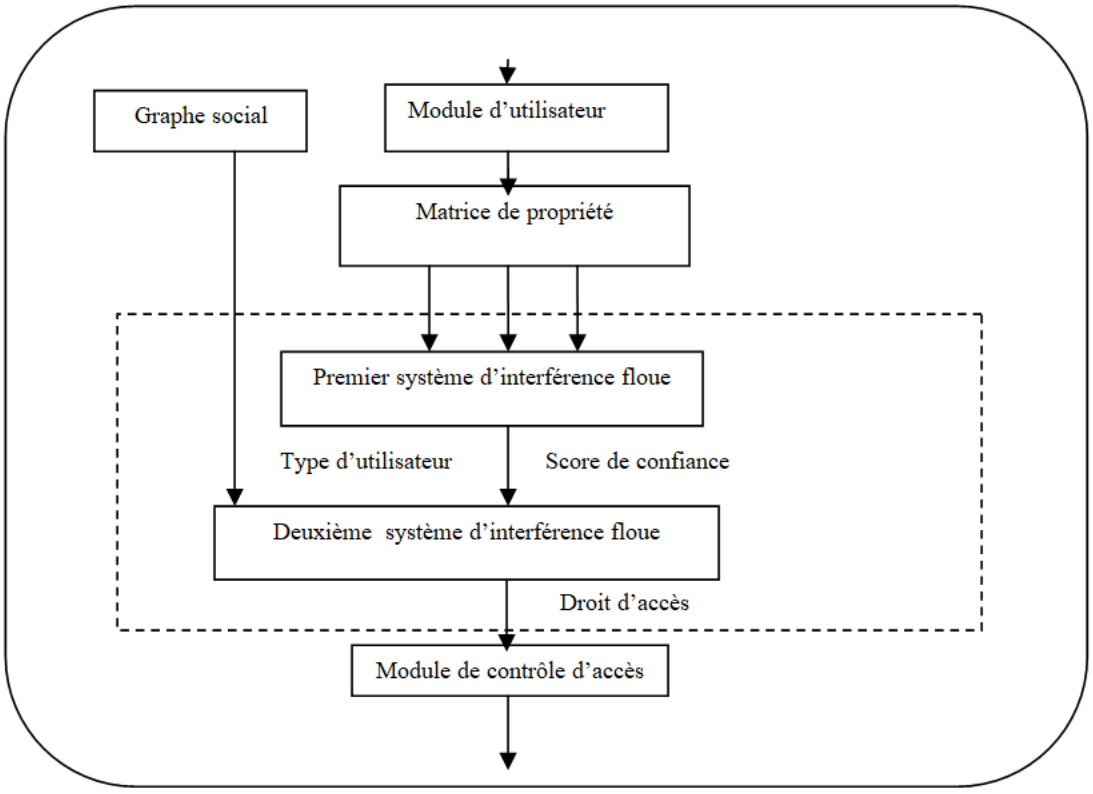

Figure 11: Architecture de la méthode SNFTrust

3.3 Modèle de contrôle d'accès basé sur le contenu (CBAC)

Comme pour le rôle, le contexte fournit un niveau d'indirection entre les utilisateurs et les autorisations.

Désormais, inspiré par la flexibilité de gestion du modèle RBAC, tout comme le rôle, le contexte peut fournir un niveau d'indirection entre les utilisateurs (Sujets) et les autorisations (Droits). Ainsi, au lieu de gérer les utilisateurs et leurs autorisations individuellement, un administrateur système définit pour chaque contexte l'ensemble des autorisations applicables. Lorsqu'un utilisateur opère dans un contexte spécifique, il acquiert immédiatement l'ensemble des autorisations actives pour le contexte associé. Lorsqu'il change de contexte d'exploitation, les autorisations précédentes sont automatiquement révoquées et les nouvelles autorisations acquises.

Bertin et al. ont souligné en 2011 que «les mécanismes d'application des politiques de contrôle d'accès basés sur les données contenus » sont nécessaires pour une protection complète des données.

La notion de contrôle d'accès basé sur le contenu a été utilisée dans la spécification de contrôle d'accès relationnel par Bertino e, al, Giuri et Iglio en1997, les bases de données multimédia par Tzelepi et al En 2001,et les bibliothèques numériques par Adam et al en 2002, etc.

En particulier, Kagal et al en 2003 ; Bhatti et coll en 2007; Reddivari et al en 2005 disent que la notion de contenu fait référence à des valeurs d'attributs ou à des concepts définitifs extraits d'objets de la bibliothèque numérique. Les privilèges d'accès sont spécifiés de manière statique en fonction des relations entre les informations d'identification de l'utilisateur et les attributs/concepts. De même, les modèles de contrôle d'accès basés sur des politiques lient les droits d'accès aux informations d'identification de l'utilisateur, cependant, la décision est toujours basée sur les valeurs définitives des attributs (par exemple, les utilisateurs avec title="physician" pourraient accéder aux dossiers des patients dans son service) [38].

En 2001 Tzelepi et al annoncent que RBAC est étendu à spécifier les politiques de contrôle d'accès sur le contenu de l'image (capturé en tant qu'attributs). Tran & Dang en 2007 imposent un contrôle d'accès aux bases de données vidéo basé sur des annotations textuelles sur les vidéos, tandis que Bertino et al en 2003b gère les vidéos en grappes (basées sur le contenu visuel), et prend en charge un contrôle d'accès plus flexible. Dans tous les cas, des règles explicites et statiques sont requises - les informations d'identification de l'utilisateur, le contenu vidéo et les politiques de contrôle d'accès sont tous explicitement définis a priori [38].

Plus récemment, Hart et al en 2007; Monte en 2010 ; et Hart en 2006 appliquent un contrôle d'accès dans le Web 2.0 basé sur les balises des messages, où les balises sont apprises à partir du contenu du message. Le contrôle d'accès est explicitement spécifié sur les balises, par exemple, il existe des règles explicites telles que : "[les membres de la famille] sont autorisé à accéder aux messages marqués avec [home] ». Pour gérer la dynamique des applications d'entreprise modernes, quelques propositions récentes tentent de déduire le provisionnement du contrôle d'accès à partir de décisions connues en utilisant l'apprentissage supervisé, lorsqu'une décision ne peut pas être directement prise à partir des politiques disponibles.

Cette approche est efficace lorsqu'un bon nombre d'échantillons d'apprentissage (décisions d'accès connues) sont disponibles, et que les échantillons d'apprentissage et de test suivent statistiquement la même distribution. D'autre part, le contrôle d'accès au niveau du concept a également été proposé pour le web sémantique. Enfin, les termes contexte et sémantique ont été utilisés dans diverses approches de contrôle d'accès. Le contexte fait principalement référence au contexte opérationnel de l'utilisateur, tandis que la sémantique est souvent utilisée pour indiquer la sémantique du schéma de données et des politiques

de contrôle d'accès, en particulier dans les applications d'intégration et de fédération de données par Fabien et al en 2012 [38].

4. Conclusion

Dans ce chapitre, nous avons fait une vu globale sur les différents modèles de contrôle d'accès traditionnels (MAC, DAC, RBAC, TBAC) et pour les réseaux sociaux basés sur la confiance .Nous avons traités leurs principes, avantages et leurs limites, qui ne sont pas adoptés pour les RS. Ensuite, on a abordé la notion de confiance qui est le fondement de la relation interpersonnelle et de l'interaction aussi bien dans les réseaux sociaux que dans le monde réel .Elle occupe un grand intervalle dans la le vie privée des utilisateurs.

Chapitre 3

Contrôle d’accès basé sur le contenu et la vie privée des utilisateurs

1. Introduction

Avec l'augmentation de l'utilisation d'Internet, la quantité de données textuelles disponibles a également continué d'augmenter rapidement. De plus, le développement d'ordinateurs plus puissants a rendu le traitement des données beaucoup plus facile. Le domaine des réseaux sociaux a un fort potentiel pour utiliser ces données disponibles sur Internet ; pourtant, d'un autre côté, une forte proportion des données disponibles sont non étiquetées et non traitées. Afin de les utiliser efficacement, de nouvelles méthodes et de nouvelles approches sont nécessaires. À cet égard, le domaine du traitement du langage naturel (TAL) aide les chercheurs à utiliser des données textuelles et à développer une compréhension de l'analyse de texte. En utilisant des approches d'apprentissage automatique, le potentiel d'exploration de texte peut s'étendre énormément, conduisant à des informations plus approfondies, à une meilleure compréhension des phénomènes sociaux et, par conséquent, à une meilleure base pour la prise de décision.

Le Traitement Automatique des Langues est une discipline à la croisée des chemins entre linguistique et informatique. L'objectif du TAL ou TALN (on parle alors de Langage Naturel) est de permettre à des dispositifs informatiques de reconnaître, comprendre et interpréter les langages « humains » de manière automatisée : langage parlé, langage écrit, langage des signes… [39]

Dans cette optique, un ensemble d'architectures, de démarches et d'outils a été regroupé en une forme homogène sous le terme de fouille de données. Donc c'est quoi la fouille de données ?

Dans ce chapitre nous allons exposer la classification automatique de texte, plus en détail la catégorisation de textes. Nous présentons quelques définitions sur la classification et les différents jeux de mots utilisés : fouilles de données, fouilles de textes classification, catégorisation ou clustering, puis nous décrivons le processus général de la catégorisation de textes avec toutes ces étapes.

2. Fouilles de données

Aujourd'hui, des milliards de données sont collectées chaque jour dans le monde. En effet, les faibles coûts des machines en termes de stockage et de puissance ont encouragé les sociétés à accumuler toujours plus d'informations. Les entreprises étaient jusqu'alors incapables de transformer leurs données en connaissance directement utilisable. [40]

La fouilles de données désigne le processus d'analyse de volumes massifs de données et des grandes bases de données (Big Data) sous différents angles afin d'identifier des relations entre les data et de les transformer en informations exploitables. [41]

Parmi les tâches réalisées en fouilles de données comme est illustré dans (la figure 12) :

- Descriptives : consiste à trouver les caractéristiques générales relatives aux données fouillées (Résumé/synthèse, Clustering, Règles d'association).

- Prédictives : Consiste à utiliser certaines variables pour prédire les valeurs futures inconnues de la même variable ou d'autres variables (Séries temporelles, Régression, Classification) [42].

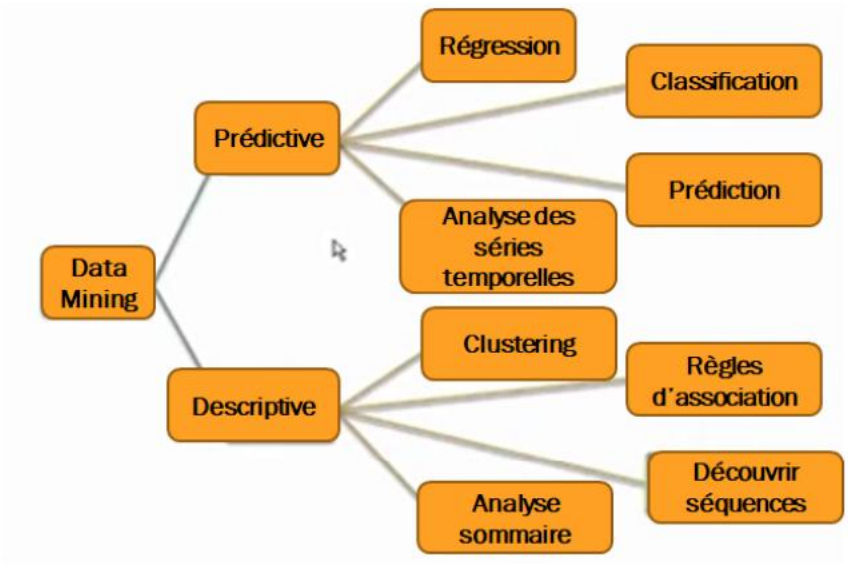

Figure 12: les tâches réalisées en fouilles de données

3. Fouilles d textes :

La fouilles de textes, également appelé fouille de textes ou extraction de à partir de textes, est un ensemble de méthodes, de techniques et d'outils pour exploiter les documents non structurés que sont les textes écrits, comme les fichiers bureautiques de type word(.docx), les emails, les documents.

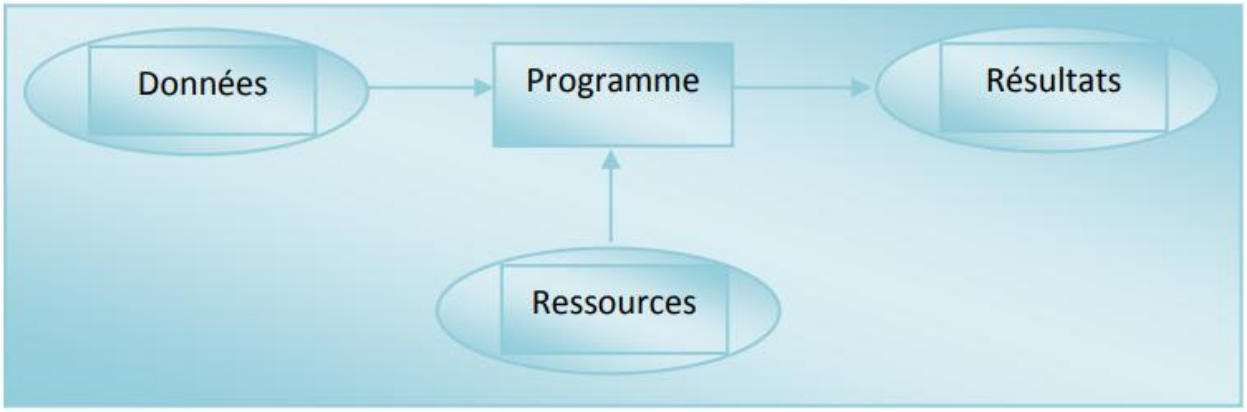

Figure 13: les techniques d'analyse en fouilles de textes

La fouille de texte s'appuie sur des techniques d'analyse linguistique (figure13). Elle est utilisé pour classer des documents, réaliser des résumés de synthèse automatique ou encore pour assister la veille stratégique ou technologique selon des pistes de recherches prédéfinies. [40].

$$\boldsymbol{Fouilles\ de\ textes = Linguistique + Fouilles\ de\ données}$$

3.1. Techniques de fouilles de textes :

La fouille de textes s'apparente à d'autres domaines avec qui elle est très complémentaire : le traitement automatique des langues (TAL) et la recherche documentaire (RI) et l'extraction de l'information (EI). [40]

Pour effectuer le processus de fouilles de textes sont [40] :

- L'acquisition : Source de données telle que : corpus textuels, bibliothèques électroniques, Web…etc.
- Le filtrage : Sélection des mots les plus pertinents (techniques de sélection d'attribut).
- Le nettoyage des données : Segmentation du texte, élimination des mots vides, lemmatisation.
- L'identification des mots pertinents : Analyse statistique (n-gram), analyse sémantique, analyse syntaxique ou structurelle (extraction d'attribut).
- L'extraction des connaissances : Application de l'un des algorithmes de la fouille de textes.

La fouille de textes consiste à utiliser le Machine Learning pour l'analyse de texte qui s'appuie sur des différent modèles qui sont classés en deux catégories [43] :

- Supervisé : consiste à apprendre à une fonction à faire correspondre une entrée à une sortie en se basant sur des exemples connus (des paires entrée-sortie).il est divisé en deux sous-classe qui sont la régression et la classification (qui modélise notre problème).

- Non supervisé : (ou ”clustering”) désigne des méthodes capables de regrouper entre elles dans des ”paquets” (”clusters”) des données, sans autre information que ces données elles-mêmes.

4. Classification automatique de texte

La classification de textes est une tâche générique qui consiste à assigner une ou plusieurs catégories, parmi une liste prédéfinie, ou non à un document.

Actuellement, la classification de textes est un domaine de recherche très actif et l'automatisation de cette opération est devenue un enjeu pour la communauté scientifique, les travaux évoluent considérablement depuis une vingtaine d'années et plusieurs modèles ont vu le jour comme le filtrage (classification supervisée bi-classe), le routage (classification supervisée multi-classe) ou le classement ordonnée (classement des textes par ordre de pertinence pour chaque catégorie).

4.1. Définition

La classification de textes(ou la catégorisation du textes) est une tâche générique qui consiste à regrouper de manière automatisée des documents qui se ressemblent suivant certains critères à savoir les critères observables tels que le type du document, l'année, la discipline, l'édition, etc. Ou le critère du contenu, et à assigner une ou plusieurs catégories, parmi une liste prédéfinie, ou non à un document (étiquettes, classes). [44]

La classification de textes est définie comme une opération qui identifie des classes d'équivalence entre des segments de textes en tenant compte de leur contenu informationnel (mots, n-gram, etc.).

Prenons l'exemple d'une application qui offre la possibilité à l'utilisateur de lire des articles sur l'actualité venant de sources différentes. Disons que l'on veut classer les articles par thème. On commence par définir les thèmes possibles : Sport, Politique, Économie, Santé, Technologie. Le rôle d'un classifieur ou outil de classification, dans notre cas, sera de « détecter », pour chaque article de l'application, s'il s'agit d'un article de sport, de politique, d'économie, de santé ou de technologie.

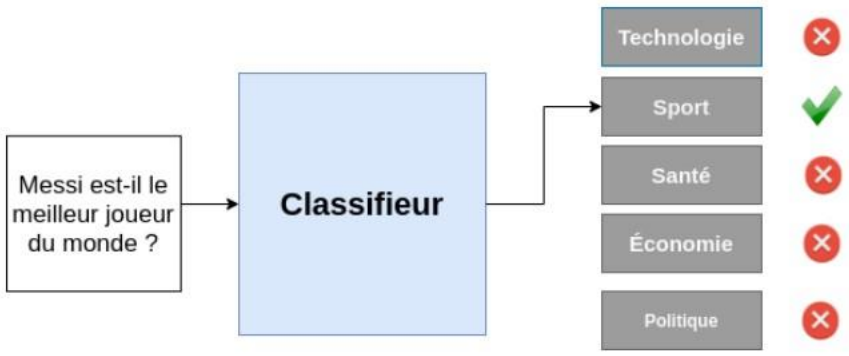

Figure 14:Schéma de classification de texte

4.2. Définition formelle

Formellement, la catégorisation de texte consiste à associer une valeur booléenne à chaque paire $(d_j, c_i) \in D \times C$, où D est l'ensemble des textes et c'est l'ensemble des catégories selon que dj∈ ci, ou non. Le but

de la catégorisation de texte est de construire une procédure (modèle, classifieur) $\Phi : D \times C \Rightarrow B$ qui associe une ou plusieurs étiquettes (catégories) à un document d_j avec la fonction F: $D \Rightarrow C$, la vraie fonction qui retourne pour chaque vecteur d_j une valeur c_i. [40]

4.3. Principe de classification du texte

Le processus de catégorisation intègre la construction d'un modèle de prédiction qui, en entrée, reçoit un texte et, en sortie, lui associe une ou plusieurs étiquettes. Pour identifier la catégorie ou la classe à laquelle un texte est associé, un ensemble d'étapes est habituellement suivie. Ces étapes concernent principalement la manière dont un texte est représenté, le choix de l'algorithme d'apprentissage à utiliser et comment évaluer les résultats obtenus pour garantir une bonne généralisation du modèle appris [45] ,comme il est résumé dans (la figure15).

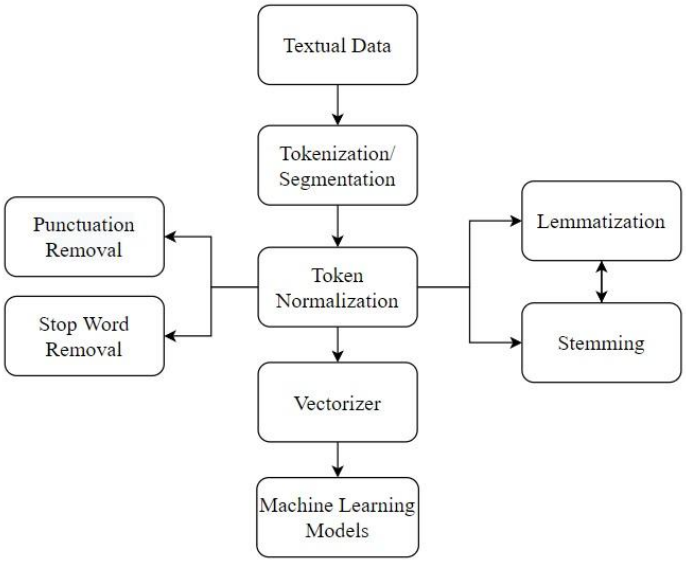

Figure 15: Principe de fonctionnement de la classification

Le processus reçoit en entrée un document textuel afin de lui trouver sa catégorie, pour cela plusieurs étapes doivent d'être suivies [46]. Ces étapes sont :

4.3.1. Données textuelles

Il s'agit du format de données d'entrée dans NLP. Il peut s'agir de données textuelles telles que des critiques de films, des commentaires, des chats, etc. Ces documents texte sont transmis à l'étape suivante - Tokenisation / Segmentation.

4.3.2. Tokénisation/segmentation

Cette étape consiste à décomposer un flux de texte en mots (segmentation de mots) et en phrases (segmentation de phrases). Ce processus de tokenisation/segmentation des données nous aide à gérer chacun des jetons séparément, facilitant ainsi la tâche de nettoyage de ces données. Par exemple:

Entrée : Il court très vite ! Elle court très lentement.

Sortie : ['Il court très vite !', 'Elle court très lentement.'].

4.3.3. Normalisation des jetons (tokens)

La normalisation dans ce contexte ou dans n'importe quel contexte signifie « rendre les choses normales ». N'aide pas vraiment, n'est-ce pas ! Permettez-moi de le dire de cette façon - C'est un processus de conversion de tout en une norme définie afin qu'aucun élément n'obtienne plus de préférence que les autres. Bien qu'il puisse y avoir des mots plus importants que d'autres, nous y reviendrons dans la partie vectorisation. Diverses méthodes utilisées pour normaliser les jetons :

- **Suppression de la ponctuation** : Cette étape implique la suppression de la ponctuation et d'autres sémantiques grammaticales qui ne sont pas nécessaires pour la PNL. Il réduit également l'ensemble du document de phrases à l'alphabet minuscule pour établir une norme.

- **Suppression des mots vides (Stop Word)** : C'est l'une des étapes les plus connues. Comme son nom l'indique, il s'agit de supprimer les "mots vides" des phrases comme :'la' ; 'le' ; 'un' ; 'il' etc. Les mots vides sont de tels mots qui ne fournissent pas d'informations précieuses dans la phrase. Des mots comme les articles, les conjonctions et les prépositions font partie des mots vides courants. La suppression de ces mots nous aide à filtrer les mots inutiles et laisse ceux qui contiennent le plus d'informations.

4.3.4. Stemming

Il s'agit d'une technique de normalisation symbolique qui réduit chaque mot d'une phrase à sa forme de base. Il supprime des choses comme les préfixes et les suffixes et nous laisse avec le mot racine. Cela prend moins de temps que la lemmatisation.

4.3.5. Lemmatisation

Il s'agit également d'une technique de normalisation de jeton similaire au stemming. Ceci est considéré comme un moyen plus sophistiqué et plus long de déterminer la racine d'un mot. Un avantage est qu'il est capable de capturer le contexte dans lequel le mot est utilisé en considérant des choses comme le temps et les mots autour du mot en considération. De plus, cela donne une racine qui est un mot significatif contrairement à la racine.

4.3.6. Vectoriseur

Maintenant que nous avons des phrases correctement formatées, nous devons les convertir en vecteurs. Nous utilisons des paramètres comme la fréquence d'un mot et des méthodes comme l'encodage pour les convertir en vecteurs. Il existe différentes méthodes de vectorisation comme sac des mots et TF-IDF qui sont largement utilisées.

- Sac des mots : Cette méthode consiste à représenter le document sous forme d'un vecteur de mots. Le processus qui permet de convertir le texte d'un document à un ensemble de termes est appelé l'analyse lexicale qui permet de reconnaître les espaces de séparation des mots, les ponctuations, les chiffres,…etc., pour qu'ils seront tous supprimés de la représentation.

- Le codage TFIDF :
  1. TF (Term Frequency) : La fréquence d'un terme est simplement le nombre d'occurrences de ce terme dans le document considéré.
  2. IDF (Inverse Document Frequency) : La fréquence inverse de document est une mesure de l'importance du terme dans l'ensemble du données.
  3. TF*IDF(Term Frequency Inverse Document Frequency): Le poids d'un terme T dans un document D est calculé comme suit :

$$TFIDF(Ti, Dj) = TF(Ti, Dj) * \log(N / DF(T))$$

Avec :

- TF (Ti, Dj) : la fréquence du terme dans le document.
- N : le nombre total de documents de la base documentaire.
- DF(Ti) : le nombre de documents contenant le terme.

  4. Les n-grammes : Cette méthode consiste à représenter le document par des n-grammes. Le n-gramme est une séquence de n caractères consécutifs. Elle consiste a découpé le texte en plusieurs séquence de n caractère en se déplaçant avec une fenêtre d'un caractère.

4.3.7. Modèles d'apprentissage automatique

Maintenant que nous avons converti des phrases en vecteurs, nous pouvons traiter chaque vecteur comme un point de données comme dans n'importe quel autre ensemble de données. Ces données sont ensuite fourni à un algorithme d'apprentissage automatique ou le modèle est formé sur ces données et évalués à l'aide de diverses mesures pour tester sa validité.

- Algorithme d'apprentissage automatique : Une fois les caractéristiques extraites, plusieurs algorithmes d'apprentissage peuvent être utilisés pour la classification automatique de textes. Pour citer les plus connus :

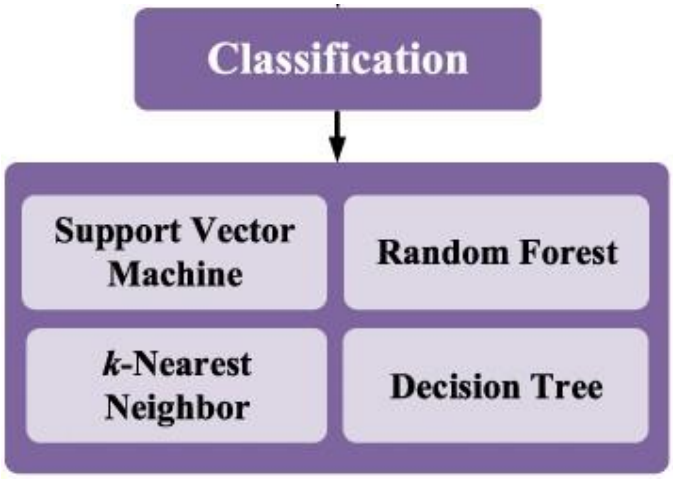

Figure 16: Différents algorithmes d'apprentissage automatique

4.3.7.1. Machines à vecteurs de support(SVM)

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais support-vector machine, SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination. Les SVM peuvent être utilisés pour résoudre des problèmes de discrimination, c'est-à-dire décider à quelle classe appartient un échantillon, ou de régression, c'est-à-dire prédire la valeur numérique d'une variable. La résolution de ces deux problèmes passe par la construction d'une fonction H qui a un vecteur d'entrée x fait correspondre une sortie y : $Y= h(x)$.

4.3.7.2. k-Nearest Neighbors k-NN

la méthode des k plus proches voisins k-NN est un algorithme standard de classification qui repose exclusivement sur le choix de la métrique de classification. Il est « non paramétrique » (seul k doit être fixé) et se base uniquement sur les données d'entraînement. KNN peut être utilisé pour les problèmes prédictifs de classification et de régression. Cependant, il est plus largement utilisé dans les problèmes de classification dans l'industrie. Pour évaluer une technique.

4.3.7.3. Arbre de décision

Les arbres de décision sont composés d'une structure hiérarchique en forme d'arbre. Un arbre de décision est un graphe orienté sans cycles, dont les nœuds portent une question, les arcs des réponses et les feuilles des conclusions ou des classes terminales. Un classificateur de texte basé sur la méthode d'arbre de décision est un arbre de nœuds internes qui sont marqués par des termes, les branches qui sortent des nœuds sont des tests sur les termes et les feuilles sont marquées par catégories.

4.3.7.4. Les forets aléatoires (Random Forest)

La forêt aléatoire (Pal, 2005) est basée sur des arbres de décision intégrés au bagging dans le processus de formation, ce qui introduit en outre la sélection aléatoire des caractéristiques. Après avoir utilisé une stratégie d'échantillonnage aléatoire pour créer environ 70 % des échantillons d'apprentissage à partir de l'ensemble de données d'origine et généré un arbre de décision pour chaque échantillon d'apprentissage séparément, les 30 % restants des échantillons d'apprentissage ont été utilisés comme données de validation pour des tests croisés internes afin d'évaluer la précision de la classification de la forêt aléatoire. [47]

4.3.7.5. Naive bayes

La classification naïve bayésienne s'apparente à une classification bayésienne probabiliste simple (dite naïve). Elle repose sur le théorème de Bayes, qui n'est autre qu'un modèle de probabilités. La méthode de classification naïve bayésienne est très utilisée dans le cadre du machine learning (ou apprentissage) supervisé. Un jeu de données (data set) va, par exemple, permettre à l'outil d'apprendre à classer des images de produits en fonction de leur caractéristique (forme, couleur...). [48]

4.3.8. Prédiction

L'étape de prédiction consiste à appliquer le modèle d'apprentissage construit lors de l'étape 7, sur les textes ou documents que l'on souhaite classer.

5. L'algorithme Naïve Bayes

La classification naïve bayésienne est un type de classification Bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses. Elle met en œuvre un classifieur bayésienne naïf, ou classifieur naïf de Bayes, appartenant à la famille des classifieurs Linéaires. Un terme plus approprié pour le modèle probabiliste sous-jacent pourrait être « modèle à Caractéristiques statistiquement indépendantes » [44].

Selon la nature de chaque modèle probabiliste, les classifieurs bayésiens naïfs peuvent être entraînés efficacement dans un contexte d'apprentissage supervisé.

Dans beaucoup d'applications pratiques, l'estimation des paramètres pour les modèles bayésiennes naïfs repose sur le maximum de vraisemblance. Autrement dit, il est possible de travailler avec le modèle bayésienne naïf sans se préoccuper de probabilité bayésienne ou utiliser les méthodes bayésiennes.

Malgré leur modèle de conception « naïf » et ses hypothèses de base extrêmement simplistes, les classifieurs bayésienne naïfs ont fait preuve d'une efficacité plus que suffisante dans beaucoup de situations réelles complexes.

5.1. Description du modèle Bayésienne

Il s'agit d'une technique de classification basée sur le théorème de Bayes avec une hypothèse d'indépendance entre les prédicteurs. En termes simples, un classificateur Naive Bayes suppose que la présence d'une caractéristique particulière dans une classe n'est pas liée à la présence de toute autre caractéristique.
Le théorème de Bayes fournit un moyen de calculer la probabilité a posteriori P(c|x) à partir de P(c), P(x) et P(x|c). Regardez l'équation ci-dessous :

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

- P (c|x) est la probabilité a posteriori de la classe (c, cible) compte tenu du prédicteur (x, attributs).
- P (c) est la probabilité a priori de classe .
- P (x|c) est la vraisemblance qui est la probabilité du prédicteur étant donné la classe .
- P (x) est la probabilité a priori du prédicteur .

Naive Bayes utilise une méthode similaire pour prédire la probabilité de différentes classes en fonction de divers attributs. Cet algorithme est principalement utilisé dans la classification de texte et avec des problèmes ayant plusieurs classes [49].

5.1.1. Avantages:

- Il est facile et rapide de prédire la classe de l'ensemble de données de test.
- Il fonctionne également bien dans la prédiction multi-classes
- Un classificateur naïve Bayes est plus performant que d'autres modèles comme la régression logistique et vous avez besoin de moins de données d'entraînement.
- Il fonctionne bien dans le cas de variables d'entrée catégorielles par rapport aux variables numériques.

5.1.2. Inconvénients:

- Si la variable catégorielle a une catégorie (dans l'ensemble de données de test), qui n'a pas été observée dans l'ensemble de données d'apprentissage, le modèle attribuera une probabilité de 0 (zéro) et ne pourra pas faire de prédiction.
- D'un autre côté, Bayes naïf est également connu comme un mauvais estimateur, de sorte que les sorties de probabilité de predict_proba ne doivent pas être prises trop au sérieux.
- Une autre limitation de naïve Bayes est l'hypothèse de prédicteurs indépendants. Dans la vraie vie, il est presque impossible d'obtenir un ensemble de prédicteurs complètement indépendants.

5.2. Applications des algorithmes naïfs de Bayes

- Prédiction en temps réel : Naïve Bayes est un classificateur avide d'apprentissage et il est certainement rapide. Ainsi, il pourrait être utilisé pour faire des prédictions en temps réel.
- Prédiction multi-classes : cet algorithme est également bien connu pour la fonction de prédiction multi-classes.
- Classification de texte/filtrage de spam/analyse des sentiments : les classificateurs Naïve Bayes principalement utilisés dans la classification de textes ont un taux de réussite plus élevé par rapport aux autres algorithmes. En conséquence, il est largement utilisé dans le filtrage anti-spam (identifier les spams) et l'analyse des sentiments (dans l'analyse des médias sociaux, pour identifier les sentiments positifs et négatifs des clients)
- Système de recommandation : Naïve Bayes Classifier crée un système de recommandation qui utilise des techniques d'apprentissage automatique et d'exploration de données pour filtrer les informations invisibles et prédire si un utilisateur aimerait ou non une ressource donnée.

5.3. Critères d’évaluation du modèle

La façon la plus simple d’évaluer un modèle est de le tester sur des données dont on connaît déjà les labels et de comparer les résultats du modèle aux vraies valeurs des labels.

Il existe un grand nombre de métriques qui permettent d'évaluer un modèle de classification. Celles-ci sont les plus utilisées pour mesurer la performance d'un classifieur.

5.3.1. La matrice de confusion

La matrice de confusion est un élément d'évaluation visuel. C'est une matrice carrée dont la taille est le nombre de classes. Et l'élément de la cellule i,j est le nombre d'observations (du jeu de données d'évaluation) qui appartient à la classe i mais a été classifié par le modèle comme appartenant à la classe j. La matrice de confusion contient assez d'informations pour calculer toute autre métrique. En plus, elle permet de juger de la performance du classifieur d'un simple coup d'œil. En effet, plus la matrice de confusion tend à être diagonale, plus la classification est précise [44].

		Reality	
Prediction	**Confusion matrix**	Negative : 0	Positive : 1
	Negative : 0	True Negative : TN	False Negative : FN
	Positive : 1	False Positive : FP	True Positive : TP

Tableau 4: Matrice de confusion

5.3.2. Accuracy

L'accuracy est une métrique qui permet de mesurer le pourcentage de réussite de notre classifieur sur le jeu de données d'évaluation. Sa formule est la suivante :

$$\textbf{Accuracy} = \frac{\textbf{Nombre de classifications correctes}}{\textbf{Nombre total d'observations évaluées}} * \textbf{100}$$

5.3.3. Précision

La précision est le pourcentage d'éléments correctement associés à un certain label par rapport à tous les éléments associés à ce label. Elle peut être vue comme la capacité du modèle à ne pas classifier dans une catégorie donnée un élément qui ne doit pas y être (qui n'est pas de la catégorie) [44]. Sa formule est la suivante :

$$\textbf{Précision}_{\textbf{i}} = \frac{\textbf{Nombre d'éléments correctement classifiés dans i}}{\textbf{Nombre total d'observations classifiées dans i}}$$

5.3.4. Rappel

Le rappel est le pourcentage de classifications correctes suivant un label donné. Soit la capacité du modèle à classifier correctement les éléments d'une catégorie donnée [44]. Sa formule est la suivante :

$$\textbf{Rappel i} = \frac{\textbf{Nombre d'éléments correctement classifiés dans i}}{\textbf{Nombre total d'observations dans i évaluées}}$$

5.3.5. F1-score

Le F1-score est une métrique qui prend en compte la précision et le rappel sur une catégorie donnée. Elle peut être vue comme une combinaison de ces deux métriques [44]. Elle vaut la moyenne harmonique du rappel et de la précision :

$$\mathbf{F1_i} = 2 * \frac{(\mathbf{pr\acute{e}cision_i} * \mathbf{rappel_i})}{(\mathbf{pr\acute{e}cision_i} + \mathbf{rappel_i})}$$

Le F1-score est compris entre 0 et 1. Le meilleur score est donc 1 et le pire 0.

6. Classification de contenu

Dans cette section, nous expliquons comment le texte peut être classé en fonction de son contenu. L'approche décrite est basée que sur les mots des textes. En utilisant un apprentissage automatique supervisé (classification), en appliquant l'algorithme du naïve bayes avec l'une de ses critères pour avoir un modèle fiable et performant qui nous donne des bonnes résultats pour la prédiction.

7. Conclusion

La classification supervisée de documents a fait beaucoup de progrès ces dernières années.

Nous avons présente les principales techniques de classification automatique supervisée, utilisées pour classer des unités textuelles en groupes homogènes.

Dans ce chapitre nous avons présente quelques techniques de la catégorisation automatique de texte. Nous basons dans notre travail sur la méthode Naïve Bayes et nous introduit les différents critères d'évaluation d'un classificateur qui seront appliqués dans le dernier chapitre (implémentation).

Chapitre 4

Conception du Projet

1. Introduction

Dans ce chapitre, nous allons d'abord définir les moyens utilisés pour élaborer un web service responsable de la création et gestion des politiques de contrôle d'accès, ainsi que le réseau social choisi pour montrer le fruit de notre travail.

Nous tacherons aussi d'expliquer comment se fait la classification (catégorisation) du texte en Python dans les réseaux sociaux à laide de l'un des framework représenter dans le premier chapitre.

On va aussi utiliser le langage UML pour modéliser les différents diagrammes afin de représenter les interactions entre les objets et schématiser le fonctionnement global de notre application.

2. Contrôle d'accès basé sur le contenu et la vie privée des utilisateurs au sein d'un RS

Dans le cadre de la gestion de la confidentialité et de l'intégrité des informations partagées des utilisateurs, la modélisation permet d'identifier plusieurs contextes d'attribution des droits d'accès aux différentes publications à travers les interactions entre les utilisateurs au sein d'un RS illustrées dans la (figure17).

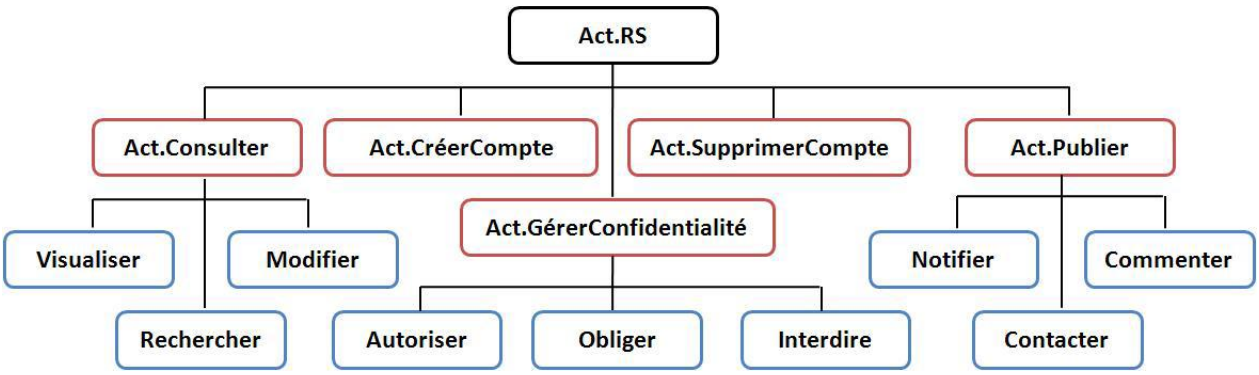

Figure 17:Actions effectuer au sein d'un réseau social

Ainsi comme on le voit dans le tableau suivant (tableau5), des activités et des actions réalisables au sein d'un réseau social.

Activités	Actions associées
Consulter	Visualiser
	Modifier
	Rechercher
Gérer Confidentialité	Autoriser
	Interdire
	Recommander
	Obliger
Publier	Notifier
	Ajouter
	Commenter
	Archiver
	Contacter
Créer Compte	Choisir login
	Choisir mot de passe
	Introduire informations
Supprimer Compte	

Tableau 5: Activités réalisable par les utilisateurs dans un réseau social

On peut définir un contexte professionnel, dans lequel un utilisateur RS peut personnaliser complètement son profil au niveau de l'accès aux informations personnelles (coordonnées, albums photos, groupes) en tant que profil professionnel. Et comme ça les contacts professionnels (exemple : collègues de travail) auront l'accès à ce profil uniquement, ainsi que les différentes correspondances et interactions avec ces utilisateurs seront spécifiques à ce contexte. Donc il fallait opter des bons paramètres (illustré dans le tableau 6) afin de protéger la vie privée.

Vues et Objets	Rôles					
	Réseaux	Amis d'amis	Amis	Famille	Etude	Propriétaire
Infos de compte						
Infos personnelles						
Infos professionnelles						
Photos et vidéos						
Identification						
Paramétrage						
Liste d'amis						
Mur						
Pseudo de messagerie						
Téléphone						
Adresse actuelle						
Site web						
@email						

	Déconseillé
	Recommandé
	Avec prudence

Tableau 6:Profil professionnel réduit pour un réseau social

3. Notre approche

L'idée du notre travaille est de répondre à la question : comment protéger la vie privée des utilisateurs dans les réseaux sociaux ?
Pour cela, Nous avons utilisés un outil de réseautage pour gérer la vie privée des utilisateurs en s'appuyant sur une architecture de service web REST(figure18).

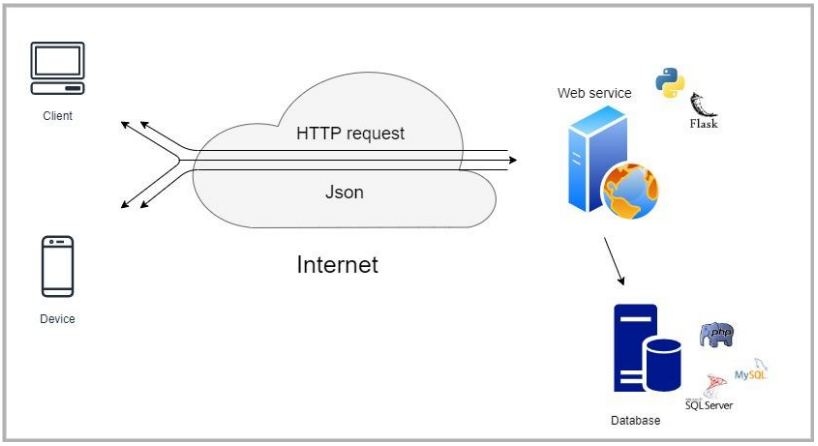

Figure 18: Architecture du service web REST

Cette approche permet de gérer des informations d'une publication en précisant son contenue.

Nous proposons dans ce mémoire une méthode qui effectue cette classification de manière automatique. Afin d'obtenir de telles connaissances, nous utilisons une base de données qui sera représenter et utilisé par la suite.

Pour gérer les paramètres de confidentialité et la vie privée, nous avons ajouté un traitement dont l'utilisateur peut spécifier la catégorie des personnes qui veut partager avec eux ses publications, en sélectionnant une parmi une liste des catégories qui contient : Publique, privé, amis et utilisateurs connectés dans ce réseau social illustré dans le graphe social (figure19).

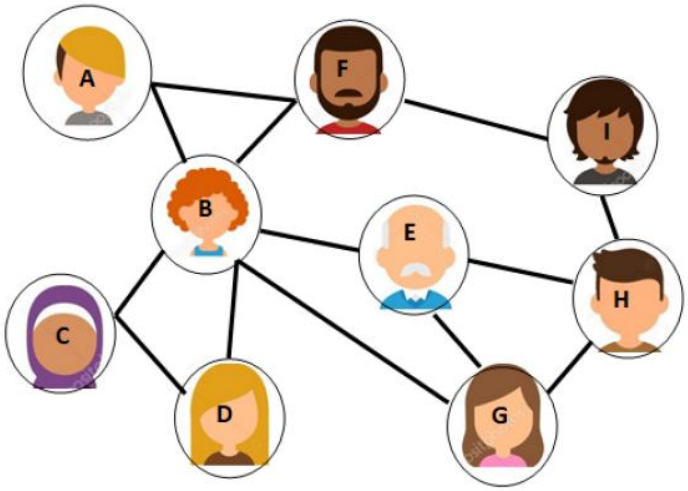

Figure 19: Graphe social définit les relations entre les utilisateurs

L'idée derrière notre approche est que la classification automatique du texte est appliquée sur le contenu de message publier par l'utilisateur en s'appuyant la méthode naïve bayésienne (spécifications des catégories) (figure 20), afin de garantir une protection légale sur les RS, en limitant l'accès à certaines publications ayant de contenu sensible (politique, religion, santé…etc.),et on autorise les autres.

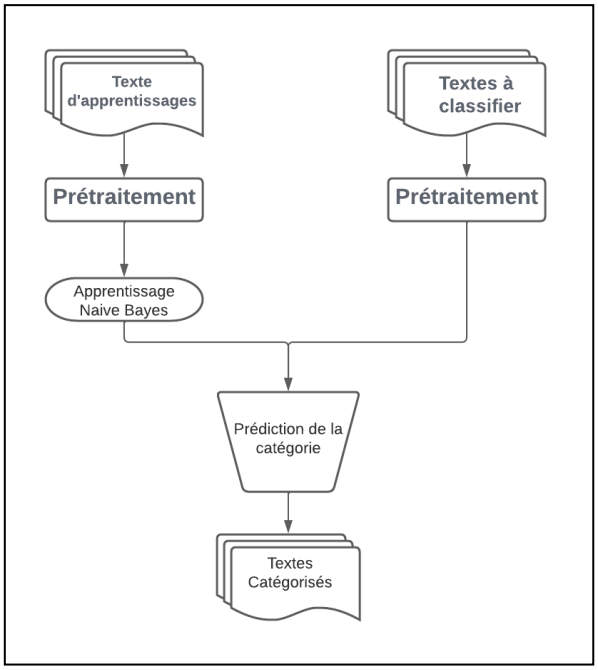

Figure 20: Les étapes de notre approche

4. Web service

Un service web est un protocole d'interface informatique de la famille des technologies web permettant la communication et l'échange de données entre applications et systèmes hétérogènes dans des environnements distribués. Il s'agit donc d'un ensemble de fonctionnalités exposées sur internet ou sur un intranet, par et pour des applications ou machines, sans intervention humaine, de manière synchrone ou asynchrone. Le protocole de communication est défini dans le cadre de la norme SOAP dans la signature du service exposé (WSDL). Actuellement, le protocole de transport est essentiellement TCP (via HTTP) [50].

Il existe plusieurs technologies derrière le terme services web :

- **les services web REST** : exposent entièrement ces fonctionnalités comme un ensemble de ressources identifiables par un URI et accessibles par la syntaxe et la sémantique du protocole HTTP. Les services web de type REST sont donc basés sur l'architecture du web et ses standards de base : HTTP et URI ;
- **les services web WS :** exposent ces mêmes fonctionnalités sous la forme de services exécutables à distance. Leurs spécifications reposent sur les standards SOAP et WSDL pour

transformer les problématiques d'intégration héritées du monde middleware en objectif d'interopérabilité.

4.1. Web service REST

C'est un style d'architecture logicielle définissant un ensemble de contraintes à utiliser pour créer des services web. Les services web conformes au style d'architecture REST, aussi appelés services web RESTful, établissent une interopérabilité entre les ordinateurs sur Internet. Les services web REST permettent aux systèmes effectuant des requêtes de manipuler des ressources web via leurs représentations textuelles à travers un ensemble d'opérations uniformes et prédéfinies sans état [49].

Le principe principal qu'il suit est le modèle client-serveur illustré dans la (figure21). Une application Web cliente communique avec les points de terminaison de l'API pour travailler sur certaines données, puis affiche les données côté client. L'API permet au client d'accéder aux données du backend de manière sécurisée.

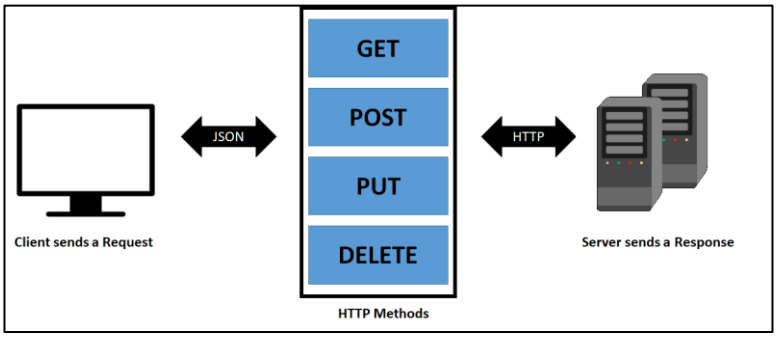

Figure 21:Architecture Client-serveur dans une application web

REST est un ensemble de règles qui vise à créer une application de Web Service selon les quatre règles de base ci-dessous :

1. Utiliser les méthodes HTTP explicitement.
2. Être apatride.
3. Exposer les URI de type structure de répertoires.
4. Transférer XML, JavaScript Object Notation (JSON), ou les deux.

REST donne une règle obligeant les programmeurs à spécifier leur but via la méthode HTTP. Les finalités comprennent normalement l'obtention, l'insertion, la mise à jour ou l'effacement des données. [49]

- Pour créer une ressource sur le serveur, vous devez utiliser la méthode POST. Cette méthode est opté pour notre application.
- Pour accéder à une ressource, utilisez GET.
- Pour modifier l'état d'une ressource ou pour la mettre à jour, utilisez PUT.
- Pour annuler ou supprimer une ressource, utilisez DELETE

5. UML

Figure 22:Logo UML

est un langage de modélisation graphique à base de pictogrammes conçu comme une méthode normalisée de visualisation dans les domaines du développement logiciel et en conception orientée objet.

L'UML est une synthèse de langages de modélisation objet antérieurs : Booch, OMT, OOSE. Principalement issu des travaux de Grady Booch, James Rumbaugh et Ivar Jacobson, UML est à présent un standard adopté par l'Object Management Group (OMG) [49].

UML est destiné à faciliter la conception des documents nécessaires au développement d'un logiciel orienté objet, comme standard de modélisation de l'architecture logicielle. Les différents éléments représentables sont :

- Activité d'un objet/logiciel
- Acteurs
- Processus
- Schéma de base de données
- Composants logiciels
- Réutilisation de composants.

5.1. Diagrammes UML

Les diagrammes sont dépendants hiérarchiquement et se complètent, de façon à permettre la modélisation d'un projet tout au long de son cycle de vie.

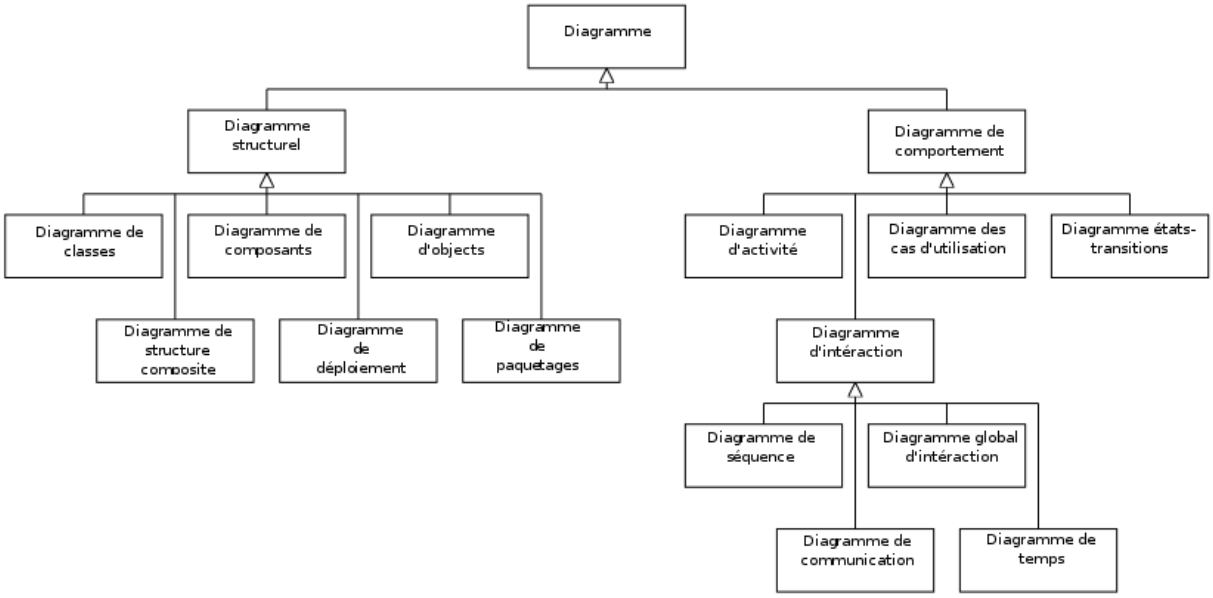

Figure 23: Hiérarchie de diagrammes UML

5.2. Diagrammes utilisés dans cette conception

Diagramme de cas d'utilisation (Use Case) : il permet d'identifier les possibilités d'interactions entre le système et les acteurs (intervenant extérieure au système), c'est-à-dire toutes les fonctionnalités que doit fournir le système. Le diagramme de cas d'utilisation est un diagramme comportemental [3].

Diagramme de séquence (sequence diagram) : représentation séquentielle du déroulement des traitements et des interactions entre les éléments du système et /ou de ses acteurs. Le diagramme de séquence est un diagramme d'interactions ou dynamique [3].

5.3. Diagramme de cas d'utilisation (Use Case)

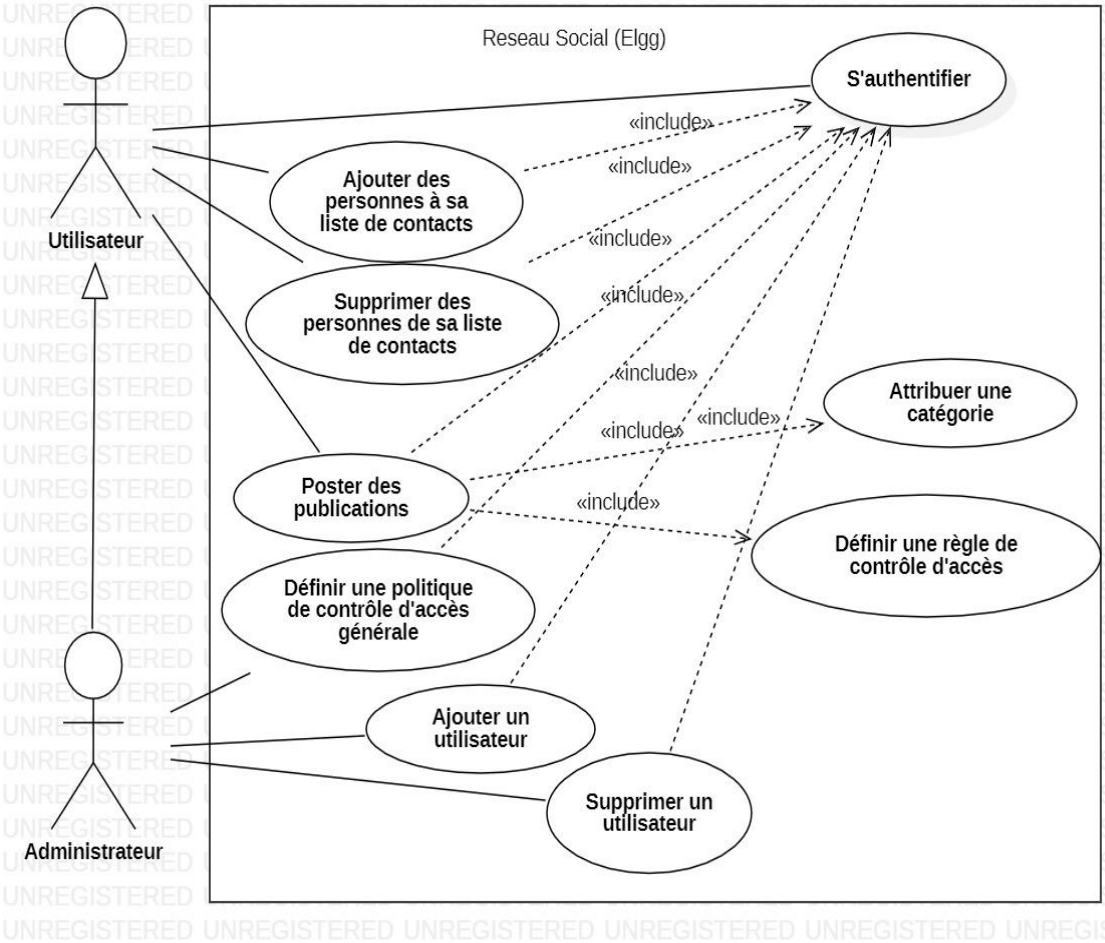

Figure 24:Diagramme de cas d'utilisation

5.4. Diagramme de séquence (sequence diagram)

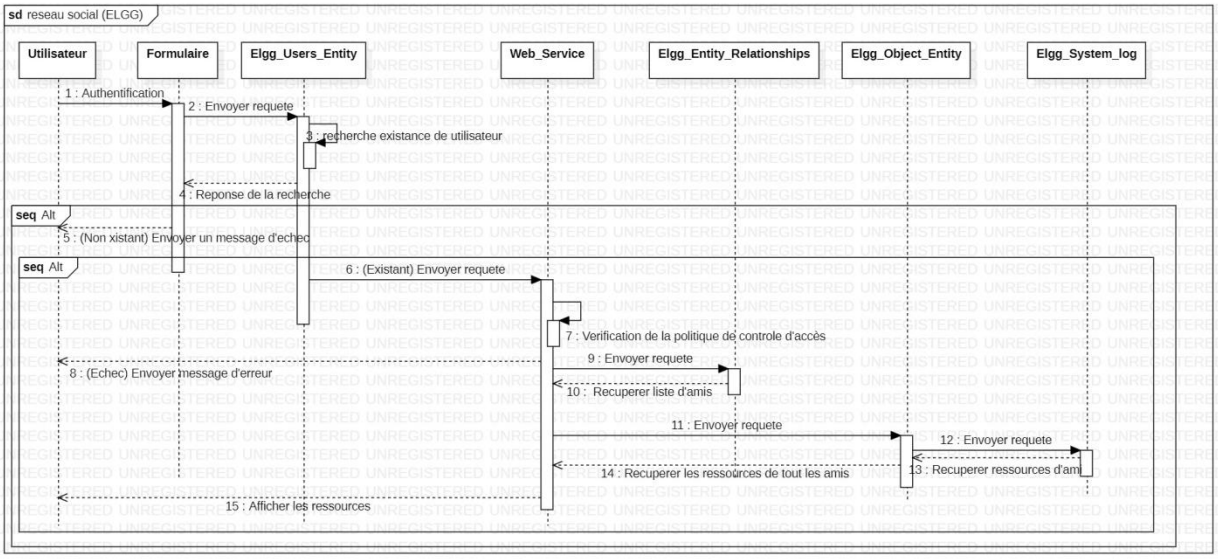

Figure 25: Diagramme de sequence"Fil d'actualités"

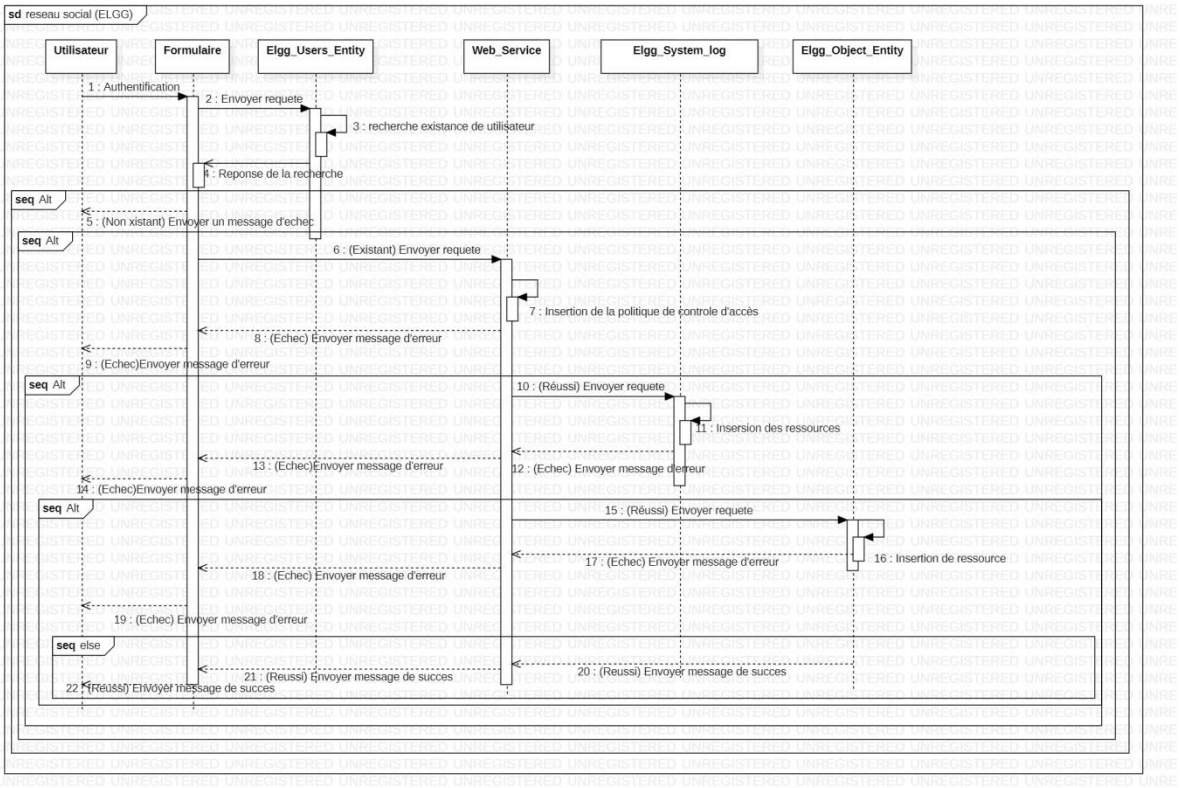

Figure 26: Diagramme de séquence "Publications"

6. Conclusion

Ce chapitre représente le cœur de notre projet, il nous a permis de concevoir et schématiser notre travail et représenter les différents éléments du langage avec les quels nous avons développé notre politiques de contrôle d'accès. Il nous a permis aussi de concevoir le fonctionnement global de notre projet afin que tout le monde puisse le comprendre. Dans le chapitre suivant on présentera l'implémentation et les tests de notre application.

Chapitre 5

Implémentation

1. Introduction

Dans ce chapitre on va présenter l'implémentation et les tests réalisés de notre web service sur un réseau social open source Elgg qui est représenter dans le premier chapitre, on va mettre en pratique nos classification des textes, nos politiques de contrôle d'accès basé sur le contenu du message et présenter notre interface pour puisse facilement gérer les paramètres de confidentialité et protéger la vie privée de chaque utilisateur.

Tout cela va être représenté grâce à des captures d'écran et des observations sur les résultats obtenus.

2. Outils de développement

Pour réaliser ce projet nous avons utilisé un ensemble d'outils, de logiciels et de langages que nous allons présenter :

2.1. WAMP

Figure 27: Logo WampServer

WampServer est une plate-forme de développement Web sous Windows pour des applications Web dynamiques à l'aide du serveur Apache2, du langage de scripts PHP et d'une base de données MySQL.
Il possède également PHPMyAdmin pour gérer plus facilement nos bases de données [3].

2.2. PHPMYADMIN

Figure 28:Logo phpMyAdmin

PhpMyAdmin(PMA) est une application web de gestion pour les systèmes de gestion de base de données MySQL réaliser en PHP et distribuée sous licence GNU GPL.

Il s'agit de l'une des plus célèbres interfaces pour gérer une base de données MySQL sur un serveur PHP [49].

2.3. PHP

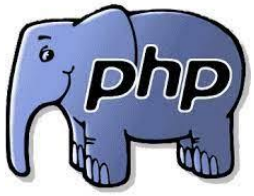

Figure 29: Logo PHP

PHP: Hypertext Preprocessor, plus connu sous son sigle PHP (acronyme récursif), est un langage de programmation libre principalement, utilisé pour produire des pages Web dynamiques via un serveur HTTP, mais pouvant également fonctionner comme n'importe quel langage interprété de façon locale [3].

2.4. PYTHON

Figure 30: Logo PYTHON

Python est un langage de programmation interprété, multiparadigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire et d'un système de gestion d'exceptions.
Il est utilisé dans l'apprentissage automatique, le développement Web, les applications de bureau et de nombreux autres domaines [50].

3. Entrainement de modèle de classification

Pour effectuer notre tache, nous sommes besoin de entrainer un modèle de classification du texte pour pouvoir gérer le contenu des applications

3.1. Jeu de données utilisé

L'ensemble de données avec lequel nous allons travailler fait suite à l'ensemble de données des articles originaux appartiennent à la BBC. Il contient 2225 lignes.
Ces données

- Se compose de 2225 documents du site Web d'actualités de la BBC correspondant à des articles dans cinq domaines d'actualité de 2004-2005.
- Étiquettes de classe : 5 (affaires, divertissement, politique, sport, technologie)

Catégorie	Nombre de titres
Affaires	510
Politique	417
technologie	401
Divertissement	386
Sports	511

Tableau 7: Les données de notre jeu de données

3.1. Création du modèle de classification

3.1.1. Prétraitement

Consiste à :

- Importer les bibliothèques nécessaires et charger la base de données,
- Visualiser les données et vérifier l'existence des champs vides,
- Convertir les majuscules en minuscules,

- Enlever les caractères non alphanumériques : les mots sont séparés par des espaces, des signes de ponctuations, des chiffres et les caractères spéciaux…
- Elimination des mots vides, les mots grammaticaux les mots de pays et les noms propres.
- Encodage du texte.
- Vectorisation des données par TF-IDF

```
import matplotlib.pyplot as plt
%matplotlib inline
import textblob

from keras.preprocessing import text, sequence
from keras import layers, models, optimizers
from sklearn import preprocessing, model_selection
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from textblob import TextBlob
from sklearn.decomposition import LatentDirichletAllocation
from sklearn.metrics import accuracy_score
from sklearn.naive_bayes import MultinomialNB
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
import warnings
warnings.simplefilter("ignore")
```

lecture de fichier csv

```
data = pd.read_csv("bbc-text.csv", encoding='utf-8')
```

```
data
```

Figure 31: Chargement des données textuelles

	category	text
0	tech	tv future in the hands of viewers with home th...
1	business	worldcom boss left books alone former worldc...
2	sport	tigers wary of farrell gamble leicester say ...
3	sport	yeading face newcastle in fa cup premiership s...
4	entertainment	ocean s twelve raids box office ocean s twelve...
...	...	...
2220	business	cars pull down us retail figures us retail sal...
2221	politics	kilroy unveils immigration policy ex-chatshow ...
2222	entertainment	rem announce new glasgow concert us band rem h...
2223	politics	how political squabbles snowball it s become c...
2224	sport	souness delight at euro progress boss graeme s...

2225 rows × 2 columns

Figure 32: Visualisation des données

```
train_x, valid_x, train_y, valid_y = model_selection.train_test_split(data['text'], data['category'])

encoder = preprocessing.LabelEncoder()
train_y = encoder.fit_transform(train_y)
valid_y = encoder.fit_transform(valid_y)
```

```
count_vect = CountVectorizer(analyzer='word', token_pattern=r'\w{1,}')
count_vect.fit(data['text'])

xtrain_count =  count_vect.transform(train_x)
xvalid_count =  count_vect.transform(valid_x)
```

Figure 33: Encodage des données textuelles

```
# word level tf-idf
tfidf_vect = TfidfVectorizer(analyzer='word', token_pattern=r'\w{1,}', max_features=5000)
tfidf_vect.fit(data['text'])
xtrain_tfidf =  tfidf_vect.transform(train_x)
xvalid_tfidf =  tfidf_vect.transform(valid_x)

# ngram level tf-idf
tfidf_vect_ngram = TfidfVectorizer(analyzer='word', token_pattern=r'\w{1,}', ngram_range=(2,3),
                                   max_features=5000)
tfidf_vect_ngram.fit(data['text'])
xtrain_tfidf_ngram =  tfidf_vect_ngram.transform(train_x)
xvalid_tfidf_ngram =  tfidf_vect_ngram.transform(valid_x)
```

Figure 34: Vectorisation des données par le module TF-IDF

3.1.2. Choix de l'algorithme et évaluation

Nous avons tester notre base de données avec des différents algorithmes d'apprentissage supervisé en calculant la prédiction par la suite, nous avons choisi la méthode naïve bayésienne car elle est parmi les meilleurs algorithme dans la classification.

Figure 35: test avec différents algorithmes d'apprentissage supervisé

3.1.3. Prédiction

Nous avons maintenant atteindre la phase de la prédiction. On va tester notre modèle par des différents textes de catégories variées

Figure 36: Prédiction de texte

Pour lier notre modèle avec notre framework nous avons suivi certaines étapes :

- Sérialisations des données en format JSON après la satisfaction du modèle de classification.
- à l'aide du Flask[3], nous avons crée une API web service, en passant en paramètres le texte pour r[4]écupérer le résultat de prédiction en JSON.
- Une fois le résultat est retourné, nous allons appelés l'API web service depuis notre framework en utilisant une commande CURL Post.

4. Test du application Web

Après la création de notre modele de classification bayésienne , nous avons serialisé nos données pour pouvoir les utilisés dans notre application web REST.

Maintenant passant à notre framwork Elgg , en connectant avec un compte administratif pour pouvoir gérer les paramètres de notre application web et de confidentialité.

[3] Flask : un micro-framework open-source de développement web en python. Il a pour objectif de garder un noyau simple mais extensible.
[4] CURL : est un utilitaire d ligne de commande qui permet aux utilisateurs de créer des requêtes réseau .Elle est utilisée pour tester les API, afficher les entêtes de réponse et effectuer des requêtes HTTP.

- S’enregistrer ou connecter à notre application web via un formulaire

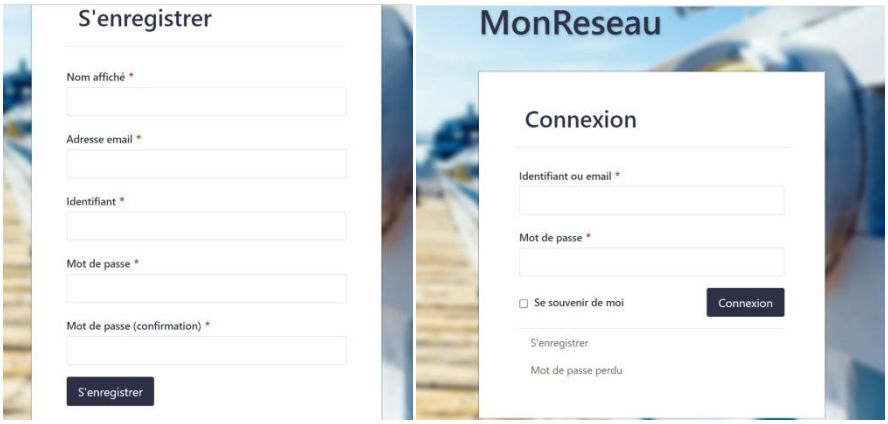

Figure 37: formulaire de connexion ou enregistrement

- La page d’accueil s’affiche dans le deuxième interface

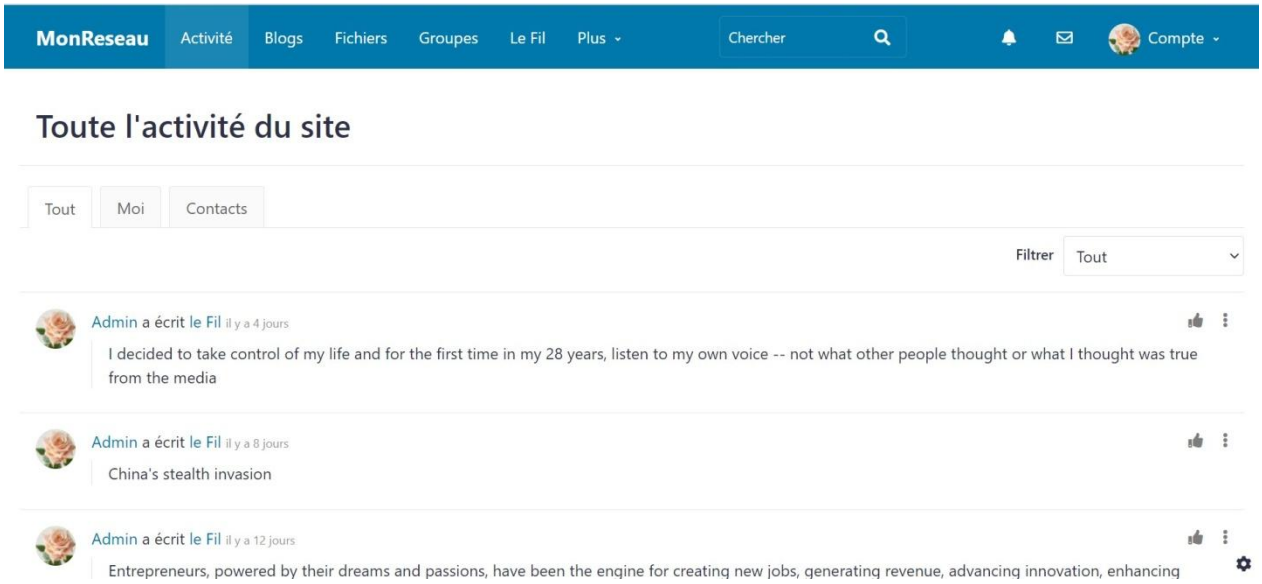

Figure 38: Page accueil

La partie qui nous intéresse dans notre application web est le «Fil » ou message d’actualités (figure39).

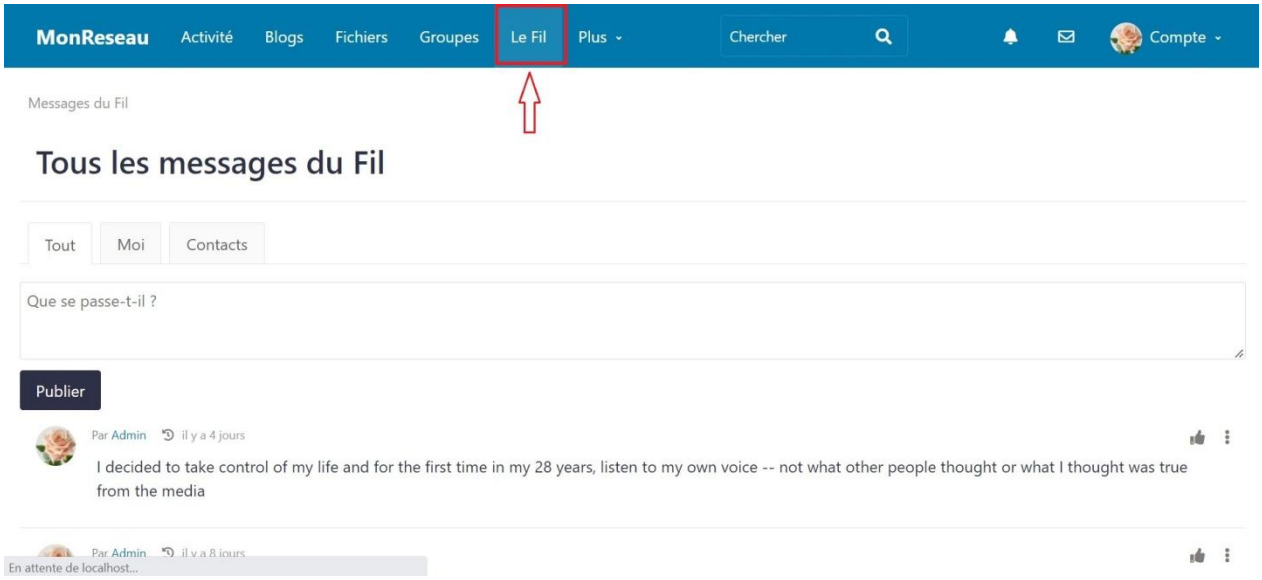

Figure 39: Page de Fil d'actualités

Nous avons ajouté une liste à choix multiples pour des différentes catégories, en permettant l'utilisateur à gérer la confidentialité de son post, afin de protéger la vie privée (figure40).

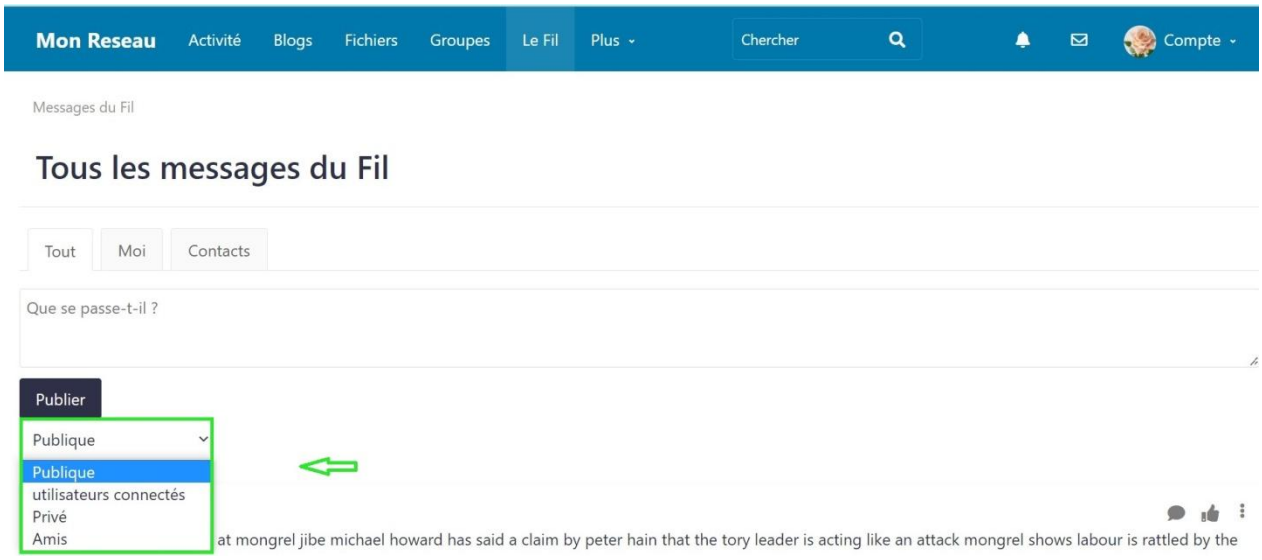

Figure 40: Liste des choix des relations entre les utilisateurs

On va essayer de poster un message (figure41), en sélectionnant une catégories des personnes qu'on veut partager avec eux notre message avant d'inclure les paramètres de confidentialité et limiter l'accès au message de contenu sensible.

Le contenue du message est le suivant :

> « *Entrepreneurs, powered by their dreams and passions, have been the engine for creating new jobs, generating revenue, advancing innovation, enhancing productivity, and improving business models and processes. Entrepreneurship is the cornerstone of the free enterprise system around the world. In fact more than 500 million adults around the globe are engaged in some form of entrepreneurial activity each year.* »

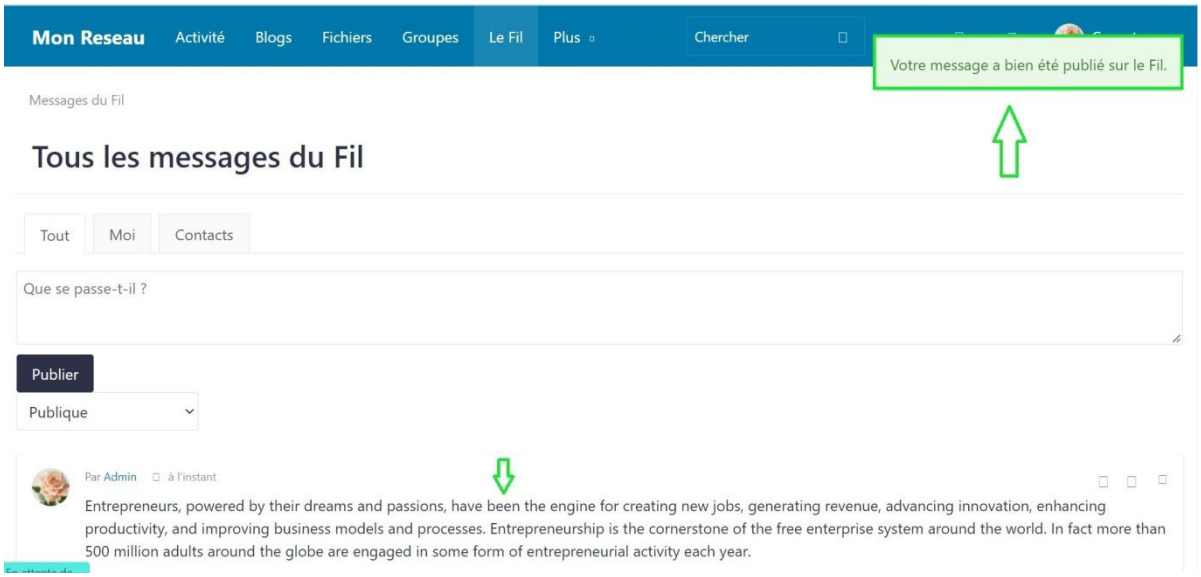

Figure 41: Poster une publication

Parmi les catégories qu'on a sur notre jeu de données, nous avons supposé que les textes dont ses contenus est du catégorie « Politique » sont sensible à partager et risque du violation de la vie privée. Le résultat est illustré dans la (figure 42).

Nous avons essayé de poster le texte suivant :

« *The former president failed to derail the criminal investigation into his hoarding of sensitive documents and is stuck paying for a costly process that threatens to undermine his public claims.*»

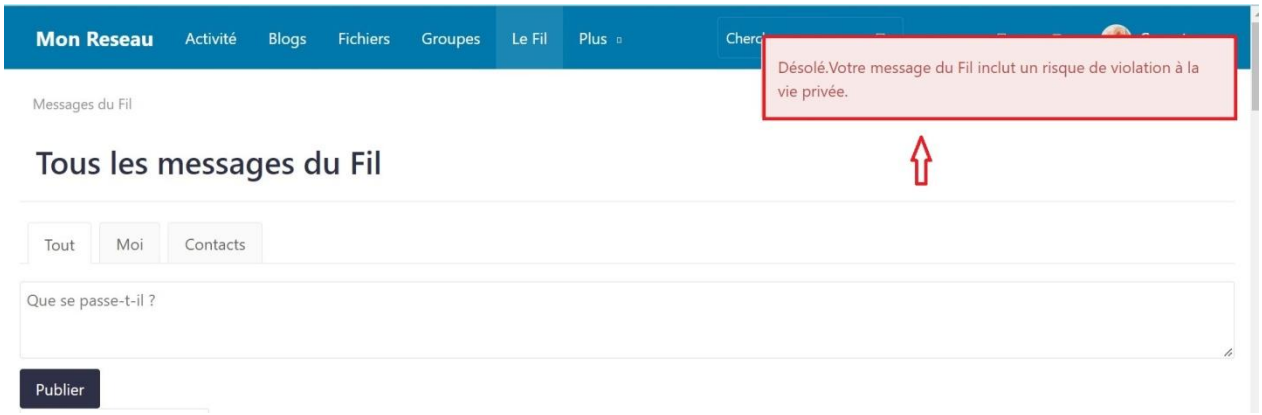

Figure 42: Message inclut un risque de violation

5. Conclusion

Ce chapitre nous a permis de présenter les différents tests et captures d'écran de nos politiques de contrôle d'accès et montrer les améliorations apportées par apport au réseau social open source de base, On a pu présenter les différentes interfaces et bases de données qui nous ont permis de réaliser ce projet ainsi que les différents outils et langages utilisés.

Conclusion générale

Dans ce mémoire, nous nous sommes intéressés à concevoir un système de contrôle d'accès pour les réseaux sociaux en basant sur le contenu du message publier par les utilisateurs à l'aide de la catégorisation des documents avec la méthode des naïve bayes .

Rappelons que le but de notre application est de protéger la vie privée des utilisateurs contre toutes divulgation au sein d'un réseau social .Ainsi que la catégorisation est utilisé pour apprendre à une machine de classer un texte dans la bonne catégorie en se basant sur son contenu.

Nous avons opté pour une démarche expérimentale qui est la catégorisation textuelle qui consiste à classer des documents selon leurs contenus avec le classifieur NB avec texte normal d'un jeu de données, en utilisant un web service pour effectuer un contrôle d'accès et garantir la gestion de confidentialité des données et la vie privée.

Après l'analyse des résultats obtenus, on a pu constater que les taux de classification sont acceptables ; mais ça n'empêche pas l'existence de quelques problèmes parmi les : l'absence de règles absolues permettant de déterminer avec exactitude.
On a rencontré aussi quelques problèmes dans les phases de prétraitement et de représentation des textes en particulier le développement d'un algorithme efficace pour la lemmatisation et le stemming.
Même si notre travail augmente le niveau de sécurité, l'utilisateur reste la première barrière contre la violation de sa vie privée.

BIBLIOGRAPHIE

[1] S. V. &. R. Ravanmehr, A novel trust-based access control for social networks, Département de génie informatique, Branche centrale de Téhéran, Université islamique Azad, Téhéran, Iran: Springer Science+Business Media, LLC, part of Springer Nature 2019, 22 January 2019.

[2] M. Adda, M. Atieh, H. Ibrahim et N. Kashmar, ACCESS CONTROL IN CYBERSECURITY AND SOCIAL MEDIA,Cybersécurité et médias sociaux (pp.69-105), Université Laval, February 2021.

[3] B. W. &. S. Mehdi, Implémentation D'un Modèle De Contrôle D'accès Dynamique Pour Un Réseau Social, Blida: universite saad dahleb, 2015.

[4] P. FEDERICA, A. SQUICCIARINI et N. ZANNONE, Survey on Access Control for Community-Centered, ACM Computing Surveys, Vol. 51, No. 1, Article 6., January 2018.

[19] B. C. Ismail, Mise en place d'un réseau de télétravailleurs sécurisé, MOSTAGANEM: UNIVERSITE ABDELHAMID IBN BADIS , 2019-2020.

[20] M. CHEAITO, Un cadre de spécification et de déploiment de politiques d'autorisation, Toulouse: l'Université Toulouse III - Paul Sabatier, 2012.

[22] S. JEMILI, ANALYSE DE RISQUE DANS, OUTAOUAIS: Laboratoire de Recherche en Sécurité Informatique_UNIVERSITÉ DU QUÉBEC , 2013.

[23] R. K. Thomas., Contrôle d'accès basé sur l'équipe (TMAC), D. d. 2. a. A. s. l. c. d. b. s. l. r. (RBAC'97), Éd., ACM, New York,, 1997., pp. 13-19.

[24] E. F. e. A. P. Barbara Carminati, Contrôle d'accès basé sur des règles pour les réseaux sociaux, Springer-Verlag, Berlin, 1734-1744, 2006.

[25] J. C. S. S. H. T. e. A. W. Evangelos Aktoudianakis, «Modèles de politique,» chez *DansActes de la conférence internationale annuelle sur la confidentialité, la sécurité et la confiance. ,* 2013.

[26] P. W. F. I. S. e. M. H. Glenn Bruns, «Contrôle d'accès basé sur les relations,» chez *: son*

expression, ACM, New York, 2012.

[27] M. M. A. e. Z. Z. Philip WL Fong, *Un modèle de préservation de la vie privée pour un réseau social de type Facebook*, Springer-Verlag, Berlin,: DansActes du 14e Symposium européen sur la recherche en sécurité informatique, Notes de cours en informatique, , 2009.

[28] F. P. e. S. S. Anna Squicciarini, *PriMa : une approche globale de la protection de la vie privée,* Anne., 2014.

[29] E. F. e. A. P. Barbara Carminati, «Application du contrôle d'accès dans les réseaux sociaux basés sur le Web.,» *ACM ,Trans, Inf, sécurisé 13,1 (article6),* p. 38, 2009.

[30] J. d. H. e. N. Z. Stan Damen, «CollAC: Collaborative access control,» chez *International Symposium on Collaborative Technologies and Systems*, Minneapolis, MN, USA, 2014.

[31] Y. Cheng, P. Jaehong et R. Sandhu, Un modèle de contrôle d'accès basé sur la relation d'utilisateur à utilisateur pour les réseaux sociaux en ligne, San Antonio: Institut pour Cyber Security, Université de Texas , juillet 2012.

[33] B. C. e. E. Ferrar, «Contrôle d'accès collaboratif dans les réseaux sociaux en ligne,» chez *7e Conférence internationale sur l'informatique collaborative : mise en réseau, applications et partage de travail*, Orlando, États-Unis, 2011.

[34] F. M. J. M. e. N. Z. Paolo Giorgini, Ingénierie des exigences pour l'homme de confiance: modèle, méthodologie et raisonnement, Trente, Italie: Département des technologies de l'information et de la communication, Université de Trente, 2006.

[35] G. J. F. C. S. L. X. H. X. Yin, «Autrust : une mesure de confiance pratique pour les utilisateurs adjacents dans,» chez *Deuxième conférence internationale sur le cloud et l'informatique verte (CGC), pp.360-367*, 2012.

[36] K. P. L. Zhao, «Un cadre d'évaluation de la confiance basé sur l'apprentissage automatique pour les réseaux sociaux en ligne.,» chez *13e conférence internationale de l'IEEE sur la confiance, la sécurité et la confidentialité dans l'informatique et les* , 2014.

[37] S. W. G. J. W. Chen, *κ-FuzzyTrust : calcul de confiance efficace pour les réseaux sociaux mobiles à grande échelle,* Inf. Sci.318, 123-143 , 2015.

[38] W. Zeng, Content-Based Access Control, Graduate Faculty of the University of Kansas: the

Department of Electrical Engineering and Computer Science and the, 2015.

[40] OualiChouayb, «Classification automatique de textes,» FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE, UNIVERSITE DE M'SILA, 2013/2014.

[42] M. NEMICHE, «Data Mining,» Faculté des Sciences d'Agadir, 2014/2015.

[45] A. Mohammed, *CATEGORISATION AUTOMATIQUE,* GUELMA: FACULTE DES SCIENCES ET DE L'INGÉNIEUR, 2009.

[47] W. P. X. W. YifanZhaoa, «Classification of Zambian grasslands using random forest feature importance selection during the optimal phenological period,» p. 5/135, Février 2022.

[48] B. S. a. L. A. Korab Rrmoku, «Application of Trust in Recommender Systems—Utilizing naive bayes classifier,» pp. 5-6 sur 14, 12 janvier 2022.

[50] H. Zhang, «The Optimality of Naive Bayes,» University of New Brunswick, Fredericton, New Brunswick, Canada, 2004.

[51] R. e. N.-M. A. Caruana, «Une comparaison empirique des algorithmes d'apprentissage supervisé.,» chez *Actes de la 23e conférence internationale sur l'apprentissage automatique.*, Pittsburgh, 25-29 juin 2006.

Webographie

[5] «pmtic,» https://www.pmtic.net/.

[6] «blog waalaxy,» https://blog.waalaxy.com/quel-est-le-principe-de-linkedin/.

[7]«open source guide,» https://open-source-guide.com/Solutions/Applications/Reseaux-sociaux-d-entreprise-rse/Elgg.

[8]«CCM,»https://www.commentcamarche.net/applis-sites/reseaux-sociaux/1159-regler-les-parametres-de-confidentialite-sur-facebook/.

[9] «facebook,» https://www.facebook.com/about/basics/manage-your-privacy/profile.

[10] «elgg,» http://learn.elgg.org/en/stable/.

[11]«HumHub documentation,» https://docs.humhub.org/docs/about/humhub/#what-is-humhub.

[12] «sourceforge,» https://sourceforge.net/projects/boonex-dolphin/.

[13] «boonex,» https://www.boonex.com/features.

[14] «cmscritic,» https://www.cmscritic.com/boonex-dolphin-review.

[15] «Ning,» https://www.ning.com/fr.

[16]«saloninnovationsinc,» https://www.saloninnovationsinc.com/2gyEQJg1/.

[17]«paperblog,» https://www.paperblog.fr/1665907/creer-son-propre-reseau-social-ning-13/.

[18] «4meahc,» https://fr.4meahc.com/what-is-ning-is-it-worth-using-37626#menu-5.

[21] «ekransystem,» https://www.ekransystem.com/en/blog/mac-vs-dac.

[32]«scaledaccess,»https://www.scaledaccess.com/whitepapers/the-developers-guide-to-relationship-based-access-control.

[39]«Erdil,» https://www.erdil.fr/blog/traitement-automatique-langues/. [Accès le 18 08 2022].

[41] «Talend,» https://www.talend.com/fr/. [Accès le 12 08 2022].

[43] «moncoachdata,» https://moncoachdata.com/. [Accès le 08 2022].

[44]«ledatascientist,»https://ledatascientist.com/introduction-a-la-categorisation-de-textes/. [Accès le 11 08 2022].

[46]«Machine Learning Concepts,» https://ml-concepts.com/2022/03/14/processing-textual-data-an-introduction-to-natural-language-processing/. [Accès le 02 aout 2022].

[49] «Wikipedia,» https://fr.wikipedia.org. [Accès le 08 2022].