République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur
et de la recherche scientifique
Université Larbi Tébessi - Tébessa
Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie
Département : Mathématiques et Informatique

Mémoire de fin d'étude
Pour l'obtention du diplôme de ***MASTER***
Domaine : ***Mathématiques et Informatique***
Filière : ***Informatique***
Option : ***Systèmes d'information***

Thème
Présenté Par :

Analyse et controle de diffusion de l'information en ligne sur les réseaux sociaux basé sur big data

Hama soltani

Devant le jury :

Mr. Bendjenna Hakim	*Prof*	*Univérsité Larbi Tébessa*	*Président*
Mr. Betouil ali abdellatif	*MCB*	*Univérsité Larbi Tébessa*	*Examinateur*
Mr.makhlouf Derdour	*MCA*	*Univérsité Larbi Tébessa*	*Encadreur*

Date de soutenance : 23/06/2019

ملخص

يلعب تحليل نشر المعلومات على الشبكات الاجتماعية دورًا كبيرًا في العديد من المجالات مثل الصحة،والتسويق، والخدمات المصرفية، والكشف عن الاحتيال، وتحليل السوق، وهذه التحليلات التي تم إجراؤها باستخدام تقنيات التعلم الآلي،واستخراج البيانات، ومعالجة اللغة الطبيعية(NLP) .يساعد الكم الهائل من المعلومات التي يتم تداولها على الشبكات الاجتماعية الأشخاص على حل مشكلاتهم وإيجاد إجابات لأسئلتهم. الوصول إلى هذه المعلومات ليس مهمة بسيطة حتى مع طرق البحث التقليدية.

مساحة البحث وزمن الاستجابة هما عاملان رئيسيان، مع البيانات التي توفرها الشبكات الاجتماعية مع خصائصها (التباين، عدم التجانس) أصبح التحكم في هاته العوامل أمر صعب

نقترح هنا عملية تعتمد على تقنيات التعلم الآلي وتقنيات معالجة البيانات الضخمة(BIG DATA) .من بين هذه التقنيات، التقنية التي أثبتت فعاليتها مع البيانات الكبيرة وهي: التعلم العميق.

الكلمات المفتاحية: *تحليل نشر المعلومات، الشبكات الاجتماعية، البيانات الضخمة، التعلم الالي، التعلم العميق.*

Abstract

Analysis of the dissemination of information on social networks play a big role in several areas such as health, marketing, banking, fraud detection, market analysis, these analyzes made using the techniques of machine learning, data mining, and natural language processing (NLP). The huge amount of information circulating on social networks helps people to solve their problems and finds answers to their questions. Access to this information is not a simple job even with traditional search methods.

Research space and response time are key factors in data that provide social networks with their characteristics (contrast, heterogeneity), which makes it difficult to control these factors. In this thesis proposes a process based on machine learning techniques and massive data processing techniques (BIG DATA). Among these techniques, the technique is proven effective with large data is: deep learning.

***Key Word**: analyse diffusion of information, social network, Big Data, machine learning, deep learning*

Résume

L'analyse et contrôle de la diffusion d'information sur les réseaux sociaux jouent un grand rôle dans plusieurs domaines comme la sante, le marketing, les banques, la détection de fraude, l'analyse des marches, Ces analyses sont effectuées à l'aide des techniques d'apprentissage automatique, d'exploration de données, et de traitement de langage naturelle (NLP). La quantité énorme d'information qui circule sur les réseaux sociaux aident les gens à résoudre leurs problèmes et à trouver des réponses à leurs questionnes. L'accès à ces informations n'est pas un travail simple surtout avec les méthodes de recherche traditionnel.

L'espace de recherche et le temps de réponse sont les deux facteurs principaux influençant les données fournis par les réseaux sociaux avec qui souffrent des caractéristiques de variabilité et d'hétérogénéité el de leur traitement.

Dans ce travail nous proposons un processus basé sur les techniques d'apprentissage automatique et les techniques de traitement des données massives (BIG DATA). Parmi ces techniques, la technique qui est avérée efficace avec les données volumineuses est l'apprentissage approfondi.

***Mots clés** : Analyse de la diffusion de l'information, Les réseaux sociaux, données massives (Big Data), Apprentissage automatique, Apprentissage approfondi*

Remerciements

Avant tout, je remercie Dieu le tout puissant de m'avoir donné le courage et la patience de terminer ce travail.

Je tiens à exprimer mes sincères remerciements et ma reconnaissance à mon encadreur de mémoire Monsieur le Docteur **Makhlouf Derdour** qui m'ont encadré et soutenu tout au long de ce travail de mémoire. Leurs grandes qualités humaines, leurs conseils scientifiques, et leurs critiques constructives ont rendu ce travail particulièrement enrichissant. Je tiens à les remercier pour m'avoir communiqué leurs passions pour la recherche scientifique.

Mes remerciements s'adressent aux **membres du jury** pour l'honneur qu'ils m'ont accordé en acceptant de juger notre travail.

Mes sincères remerciements s'adressent à **mes parents**, **mes frères**, **mes sœurs** ainsi qu'à toute la famille pour leur soutien moral, leur encouragement inconditionnel et leurs aisés financiers. Sans oublier de remercier tous les **enseignants** et **enseignantes** qui, pendant mon cursus universitaire, ont veillé pour ma formation et ma réussite.

Tous les mots restent faibles pour exprimer ma profonde reconnaissanc à tous ceux qui m'ont aidé de près ou de loin pour la réalisation de ce travail,en particulier tous **mes ami(e)s** pour leur soutien moral et leur présence à mes côtés.

Dédicaces

Je rends grâce au **mon Dieu** de m'avoir donné la force, la volonté et la sagesse afin de parvenir à cette conclusion de mon cycle.

Dans cet espace

je souhaiterai dédier ce travail a l'esprit de mon **neveu Ishak**

je dédier ce travail à mes très chers parents

En premier lieu mes dédicaces vont droit à ma **chère mère**. Tes encouragements et tes prières ont été d'un grand soutien pour moi je te remercie infiniment.

Je remercie également mon **cher père** pour sa présence dans ma vie, de son soutien et tous ses sacrifices et ses précieux conseils, j'espère avoir réussi à te j'espère avoir réussi à te

rendre fière chose que je tâcherai de continuer à faire.

Je le dédie aussi à mes adorables **soeurs** et leurs **maris**, mes **freres** et leurs **epouse** pour leur patienceet,et j'oublier pas **mes neveux**.

A les personnes qui m'a toujours soutenu et a été présent à mes côtés **Seraya** et **Habiba**.

Enfin je le dédie à tous mes amis(mes freres)

Hossam, **Azzeddine**, **Lamin**,**Belgacem**,**Kamel**, **Yacin** et mon pote **Imad_Aimen**

je dédie ce travaile à mon amie **Wissem**

et à toute personne qui ma aider et encourager de prêt ou de loin toute au long de mes étude.

Liste de figure

Sommaire

INTRODUCTION GENERAL

Introduction Géneral

Ces dernières années, les médias sociaux sont devenus une partie importante de notre vie quotidienne. L'utilisation de différents médias sociaux modifie la façon dont nous communiquons, collaborons et rassemblons des informations et réalisons ainsi le monde qui nous entoure. Les réseaux sociaux sont désormais la principale source d'information, car chacun partage ses opinions, ses sentiments, ses problèmes et ses intérêts en partageant ses publications, tweets, vidéos, mots et photos de sa propre vie.

L'analyse de la diffusion de l'information sur réseaux sociaux jouent un grand rôle dans plusieurs domaines comme la sante, le marketing, les banques, la détection de fraude, l'analyse des marches. Les techniques utilisées pour mener ses analyses : l'apprentissage automatique, l'exploration de données, et le traitement de langage naturelle (NLP).

La quantité énorme d'information qui circule sur les réseaux sociaux aident les gens à résoudre leurs problèmes et trouve des réponses à leurs questionnes. Cependant pour trouve une information pertinente dans cette quantité énorme des informations est tache lourde. L'information dans les réseaux sociaux est inutile si nous ne pouvons pas extraire des informations utiles à bon moment.

Lors que la quantité fournit par les réseaux sociaux est énorme, les méthodes de stockage et de traitement traditionnelle ne peuvent pas traiter et adopter ces données. C'est la justification de l'utilisation du Big-Data et ses technologies et qui est devenu la meilleure solution pour gérer les problèmes : des grandes masses de données et d'hétérogénéité des données et de variabilité produit par les réseaux sociaux avec un cout accessible (le temps, l'outils et l'espace de recherche).

Le Big Data qui circule dans les réseaux sociaux ne peuvent être utile que si nous voyions au-delà de toutes ces données. En faisant cela, beaucoup d'opportunités s'offriront à nous. Il faut donc savoir s'y retrouver dans toute cette quantité de données toujours plus importantes. L'enjeu est de transformer une donnée en connaissance, d'adapter nos actions en fonction de cela.

Notre objectif est de minimiser l'espace de recherche à travers la suppression des données inutile et inutilisable, ainsi l'amélioration de la qualité des résultats, sans oublier le facteur principal qui est le temps (les résultats en temps réel ou un temps acceptable).

INTRODUCTION GENERAL

Nous avons proposé un processus pour exploiter les données diffusées sur les réseaux sociaux en utilisant les techniques d'apprentissage automatique et le traitement de langage naturelle.
Ce processus repose sur deux filtres : le premier est le filtre des données massives (BIG DATA) afin d'extraire les données liées au domaine de la requête en utilisant les méthodes de l'apprentissage automatique. Ensuite, un deuxième filtre des données selon le contexte de la requête en utilisant le profil de l'utilisateur pour la personnalisation des données.
Nous discutons dans la première chapitre c'est quoi les reseaux sociaux et leur caracteristique,et parler sur les big data et leurs outils et techniques,et en en fin de ce chapitre sur l'exploration de données et leurs methodes comme des techiniques pour analyse de ces les données massives. Et en deuxième chapitre nous discutons un peut sur comment modeliser la diffusion de l'information et quelle sont les modeles de la diffusion de l'information,et nous rappelons les domaines interssant qui utilise l'analyse et controle de la diffusion de l'information comme l'analyse des sentiments et detection des evenement. Dans le troisième chapitre on va détaillé notre contribution et les methode et les techniques utilisée pour obtenir nos objectives,nous discutons sur la phase de pretraitement et l'apprentissage automatique et approfondie et les methodes our represente le profile d'utilisateur pour personaliser les données obtenu. Le quatrieme chapitre c'est une vue pour represente les techniques et outils utiliser pour simuler l'etape de filtrage par domaine dans le cas la source du données est Twitter est en terminer notre travail avec un conlusion generale sur tous qui tourne dans notre mémoire.

Chapitre I :

Big Data & Analyse de données

1 Introduction

Une vérité maintenant, le monde vers un petit village en raison de l'influence apparente des réseaux sociaux (Facebook, Twitter, Instagram, YouTube, LinkedIn, Pinterest, medium, Quora, Google+, Viadeo, Slideshare, Reddit, ...). Il relie des personnes de différentes régions du monde, d'âges et de nationalités et leur permet de partager leurs opinions, leurs expériences, leurs sentiments, leurs loisirs, des images et des vidéos. Cela a ouvert la porte des offres et nouveaux travaux pour promouvoir, exploiter, analyser, apprendre et améliorer les organisations publiques et privées de tous les domaines sur la base des données fournies dans les médias sociaux.

Les médias sociaux fournissent une énorme quantité de données continues en temps réel. La structure de ces données n'est pas organisée et apparaît sous différentes formes, telles que : texte, voix, images et vidéos. En outre, les médias sociaux fournissent une quantité considérable de données continues en temps réel qui rendent les méthodes statistiques traditionnelles inappropriées pour analyser ces données volumineuses. **Ces données sont inutiles si elles ne sont pas converties en informations utiles.**

pour l'exploitation des données massives qui sont extrait de les réseaux sociaux à la lumière de ses défis en raison de leurs caractéristiques telles que: volume, vélocité, variété, véracité et valeur on peut appliquer des Diverses techniques d'exploration de données pour résoudre les différents problèmes des tréteaux sociaux tels que la détection d'influence, la détection de communauté ou de graphes, la recherche d'experts, la prévision de liens, les systèmes de recommandation, la prédiction de la confiance et la méfiance des individus, le comportement et l'analyse d'humeur, l'exploration d'opinions.

2 Les Réseaux Sociaux

Apparaissions des réseaux sociaux en l'année 1997 avec la naissance de première réseau sociaux "SixDegrée" permet de crée de profile, partager les avis, navigation entre les profiles. Cette réseau sociaux a été fermer en 2001 malgré le nombre intéressant des utilisateurs par manque d'argent. De 1997 à 2001, beaucoup de plateformes par communautés se sont créées comme par exemple AsianAvenue, BlackPlanet, MiGente.... Dans l'année 2002 nouvelle réseau apparu appelée Freindester, après cette réseau social plusieurs plateforme a été apparu comme "MySpace" en 2003, et d'autre réseau social particulier sous le nom LinkedIn c'est un réseau social Professional.

Une nouvelle démarche commence à Harvard en 2004 avec la naissance Facebook (a été reste jusqu'en 2006 un réseau social fermé), après ça, plusieurs réseau sociaux a été apparu YouTube en 2005 (pour le partage de vidéo. Il a été acheté par Google en 2006), Twitter en 2006 (microblogging. Tweets de 140 caractères, inspirés de la popularité des SMS),et en 2010 lancement de Pinterest (pour le partage et l'enregistrement d'images dans des tableaux avec des fonctionnalités sociales d'interactions et de suivi) ,et Instagram (pour le partage de contenu image et vidéo et a été acheté par Facebook en 2012).et Google en 2011 lance leur propre réseau social sous le nom Google+(comme concurrent réel de Facebook) , dans la même année Snapchat a été née(permet de Messagerie d'images et courtes vidéos). Après 2011 des centaines des réseaux sociaux (les blogs, les forums, ...).

2.1 Définition d'un réseau social :

Un réseau social est défini comme : « une structure définie par des relations entre des individus », Un réseau social représente un système d'entités en interaction. On le modélisera comme un graphe G = (S, A) où S est un ensemble d'entités (les sommets ou nœud du graphe) et A est l'ensemble des arcs (ou connexions) représentant les interactions entre ces sommets [1].

Par exemple Facebook est le réseau le plus célèbre maintenant, on peut être représenter de cette façons un nœud est un membre et si deux relies (connecte) les membres sont des amis par exemple ou bien le premier membre a été fait un partager/commentaire d'une publication de l'autre, les entités peuvent être de tous sorte : des page web, utilisateurs (membres, visiteurs) d'un site, des comptes..., et les liens peuvent représente les interactions(relations)variées entre les entités(liens d'amitié sur Facebook, j'aime sur Instagram ,Flower sur Twitter, hyperliens sur le web).

Figure 1:Les relations internationales sur le réseau Facebook*[1]*

2.2 Les caractéristiques des réseaux sociaux

Les réseaux sociaux avoir plusieur caracteristiques comme :

- **Taille de réseau**

Les réseaux sociaux sont souvent très grands : nombre massive de nœuds et de liens, et en plus, Il y a en général de nombreux attributs sur les nœuds (nom_utlisateur, ...), voire sur les liens (nombre des messages entre les deux nœuds par exemple) ;

- **L'Effet petit-monde**

La principale caractéristique d'un réseau social est « l'effet petit monde », Le réseau forme un petit monde : la longueur moyenne du plus court chemin entre deux nœuds est petite (la longueur du plus court chemin entre deux nœuds est la distance, ou degré de séparation, entre ces nœuds). Stanley Milgram réalisa une expérience en 1967 pour démontrer cette règle du petit monde qu'on appelle aussi règle des six degrés de séparation. Cette expérience suggère que deux nœuds, choisies au hasard parmi les nœuds de réseaux social(graph) sont reliées en moyenne par une chaîne de six relations [2].

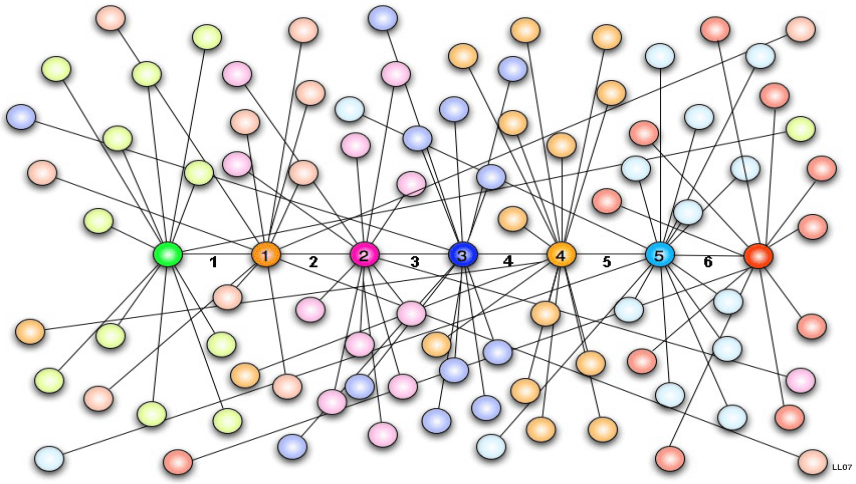


Figure 2: Les six degrés de séparation [a]

[a] : https://fr.wikipedia.org/wiki/étude_du_petit_monde

- **L 'homophilie**

Les nœuds qui portent des attributs similaires ont tendance à être reliés : ce phénomène est nommé l'homophilie [1].

- **Les communautés**

Un réseau social peut souvent être décomposé en sous-groupes, ou bien communautés, une communauté est ensemble des nœuds ou il y'a beaucoup de communication entre eux et peu de communication avec les autre nœuds (un ensemble de nœuds fortement liés entre eux et faiblement liés avec les autres nœuds) [1].

2.3 Analyse de réseaux sociaux :

- L'analyse des réseaux sociaux est basée sur les acquis de la théorie des graphes pour formaliser le réseau social en tant qu'ensemble de nœuds et de liens, chaque nœud représentant un acteur et reliant une relation entre deux acteurs. Une valeur peut être assignée à un lien puis en représenter la force. Il peut être utilisé pour représenter l'importance d'une relation, que ce soit en comptant simplement le nombre d'occurrences de cette relation ou en prenant en compte d'autres processus de pondération [1].
- L'analyse des réseaux sociaux comporte des outils mathématiques dérivés directement de la théorie des graphes mais également d'outils et de techniques qui lui sont spécifiques. Nous allons donc trouver dans l'analyse des réseaux sociaux les terminologies de la théorie des graphes telles que le degré, la force ou le poids d'un lien.

- L'analyse des médias sociaux est un domaine naissant qui a émergé après l'avènement du Web 2.0 au début des années 2000. La principale caractéristique de l'analyse des médias sociaux modernes est son caractère centré sur les données. La recherche sur l'analyse des médias sociaux couvre plusieurs disciplines (psychologie, sociologie, anthropologie, informatique, mathématiques, physique et économie). Le marketing a été la principale application de l'analyse des médias sociaux au cours des dernières années. Cela peut être attribué à l'adoption généralisée et croissante des médias sociaux par les consommateurs du monde entier [3].
- Le contenu généré par l'utilisateur (sentiments, images, vidéos ...), ainsi que les relations entre les entités du réseau constituent les deux sources d'informations des réseaux sociaux. L'analyse des médias sociaux peut être classée en deux groupes : analyse basée sur le contenu et analyse basée sur la structure.
- Diverses techniques ont récemment émergé pour extraire l'information de la structure des réseaux sociaux.
    - La détection de communauté, également appelée découverte de communauté, extrait des communautés implicites dans un réseau. Une communauté est un sous-réseau d'utilisateurs qui interagissent davantage entre eux qu'avec le reste du réseau [1]. La détection de communauté aide à résumer d'énormes réseaux, ce qui facilite ensuite la couverture du comportement existant et la prédiction des propriétés émergentes du réseau, une technique d'exploration de données utilisée pour partitionner un jeu de données en sous-ensembles disjoints en fonction de la similarité des points de données. La détection communautaire a trouvé plusieurs domaines d'application, notamment le marketing et le Web [3].
    - L'analyse d'influence sociale fait référence à des techniques de modélisation et d'évaluation de l'influence d'acteurs. Naturellement, le comportement d'un acteur dans un réseau asocial est affecté par d'autres. Il est donc souhaitable d'évaluer l'influence des participants, de quantifier la force des connexions et de mettre au jour les schémas de diffusion d'influence dans un réseau. Les techniques d'analyse d'influence sociale peuvent être utilisées dans le marketing viral pour améliorer efficacement la notoriété et l'adoption de la marque [3].

3 Big data

Big data (concept, definition): Ç'est un terme décrire une collection d'un ensemble de données très vaste et complexe qu'il devient très difficile de traiter à l'aide des outils de systèmes de base de données ou une application de traitement de donnés traditionnel.

Big data peut-être définie à l'aide de cinq V (5V) : Volume, Vélocité, Variété, Valeur, Véracité [4] [5]

Volume : la caractéristique la plus célèbre, avec le développement des technologies chaque seconde une masse de données est générer (Réseaux sociaux (Facebook, Twitter, YouTube, LinkedIn …), internet des objets, éducation, santé, …).

Vélocité : les données et résultats souvent disponible en temps réel, Les données peuvent être analysées, traitées, stockées et gérées rapidement [5].

Variété : hétérogénéité des ressources, différent format de données (des données structurelle, semi-structurelle, non-structurelle) (documents, vidéos, audio, email).

Valeur : la transformation de big data à une autre valeur donne une force pour déduire l'importance de big data pendant l'analyse et l'exploitation (perspective commercial) [4].

Véracité : mesurer la fiabilité et la confiance des données, parce qu'il y'a des données incomplètes des données valide se sont des données correct et claire, exacte pour aide à prendre des bonne décisions [4].

3.1 Comparaison entre les données traditionnelles et Big data

Le big data donne un autre sens pour le stockage et analyse des données, maintenant nous explique quelque différence big data et les données traditionnelles :

Structure de données : dans la majorité des cas la donnée traditionnelle utilise une base de données centraliser, et architecture centraliser elle est très couteuse et complexe et peut-être inefficace pour massive de données. Mais le big data utilise une architecture distribuée, elle très efficace pour traite une grande quantité de données dans temps peut être idéal par rapport la précèdent.

Type de données : seulement traitent les données structurées et ignorer les autres types mais avec le big data tous les types sont inclus (structurée, semi-structurée, non-structurée).

Cout : pour gérer une masse de données dans système de base de données traditionnel on a besoin souvent de matériel et logiciel très couteuse. Mais avec le big data on peut utiliser matériel standard et logiciel open source pour analyse les données (Hadoop, ...).

Facteur de comparaison	Traditionnel	Big data
Architecture de données	Centralise	Distribue
Type de données	Structurées	Structurées ,Semi-Structurées Non-Structurées
Volume	Gb->Tb	Tb->Pb->Eb->Zb
Cout	Élevée	Moyenne
Précision &confidentialité	Faible	Élevée

Tableau 1::comparaison entre big data et données traditionnelles

3.2 Les problèmes de big data

Le big data est venu poser plusieurs problèmes, parmi ces problèmes :

Stockage des données énormes et en croissance continue

Dans les années les quatre-vingt-dix on a parlé sur giga-octets mais maintenant on parle sur zettaoctets (1 ZB = 10^{21}octets). Dans chaque seconde, il y'a des méga-octets de données été généré à partir des source différente (réseau sociaux, capteurs, …). Ce phénomène pose une questionne

" Comment peut stocker ces données massive (dans le cas elle est toujours en croissance continue) "

Traitement des données ayant une structure complexe

Parmi les différences entre le big data et les données traditionnel la forme de données. Dans le big data il y'a des données structure (table, schéma fixe, format organise, …), semi structure (XML, fichier JSON, ...), non-structure (vidéo, audio, ...). Ça pose une question :

" Comment traite les données ayant une complexe "

La performance (traitement des données plus rapide)

Avec des données massives les applications de l'analyse des données ne répond pas à les besoin (temps de réponse, tolérance de faux, ...) comme avec les données traditionnelles.

Ce manque pose les questionnes

"comment peut répond à ces besoin "

" Comment pouvons-nous diminuer le temps de réponse "

4 Hadoop comme solution

Figure 3:hadoop logo

Hadoop (Définition) :Hadoop Framework java libre permettant de créer des applications distribuées, scalables et très tolérant aux fautes. Hadoop est composé de deux composants : HDFS, système de fichiers distribué et MapReduce qui permet d'effectuer des calculs parallèles.

4.1 HDFS (définition) :

HDFS (Hadoop Distributed File System) est un système de fichier distribué permettant de stocker et de récupérer des fichiers en un temps record [6].

4.2 HDFS (Comment fonctionne) :

Chaque fichier hdfs découpe en bloc de taille fixe

Les blocs de même fichier ne sont pas forcément dans la même machine (nœud)

Ils sont copiés dans diffèrent machine (par défaut 3 copies)

HDFS ayant deux de node : Master-node, Worker-node(slaves)

HDFS est constitué de machines jouant différents rôles entre eux :

- **Namenode** Cette machine contient tous les noms et blocs des fichiers, métadonnées (master node).
- **Secondary namenode** une sorte de namenode de secours (master node)
- **Datanodes**. Elles stockent les blocs du contenu des fichiers (slaves)

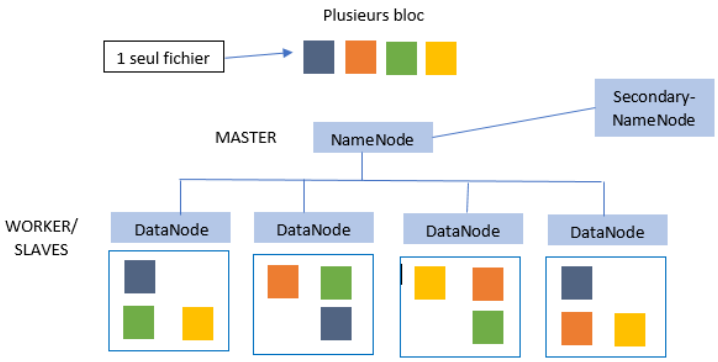


Figure 4 :HDFS fonctionnement (inspirer le[11][6])

4.3 MapReduce :

MapReduce est un Framework de programmation qui nous permet d'effectuer des traitements distribués et parallèles sur de grands ensembles de données dans un environnement distribué [7]

4.4 Fonctionnement de MapReduce :

MapReduce effectuer essentiellement deux taches MAP et REDUCE :

MAP : la tache de mapping est lue et traitée les entrée (les entrées à partir HDFS) et les sorties est une séquence de pair (clé, valeur).

REDUCE : peut dire c'est une tache de combinaison .l'entrée de cette taches sont les sortie de la tache mapping ,et la sortie est un seul valeur (combinent tous les valeur associe au clé dépend le code écrite) [7]

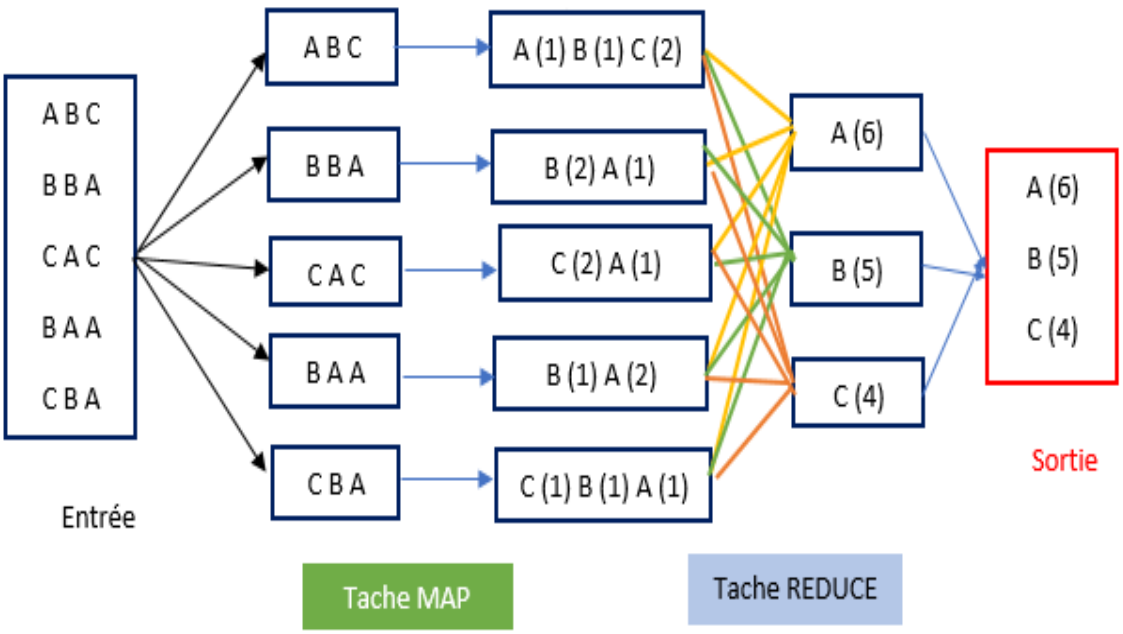


Figure 5: fonctionnement de Map Reduce

5 Apache SPARK

5.1 Analyse en temps réel

Dans chaque minute une grande masse de données est génère par les réseaux sociaux (Facebook Twitter, Instagram, YouTube, …).

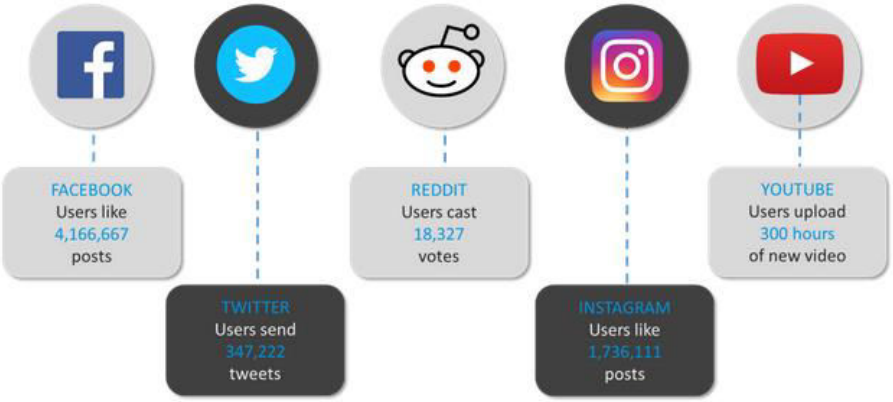


Figure 6:Quantité de données générées chaque minute*[8]*

Ces données nécessitent de traiter dans le minimum de temps, c'est pour ça on a besoin d'un Framework de traitement en temps réel. Apache Spark a été développé pour répondre à ces besoins. Maintenant le traitement en temps réel de big data est enracinée dans tous les aspects de nos vies (détection des fraudes dans les secteurs bancaires, systèmes de prévision sur le marché boursier, etc. Parmi les domaines nécessitent un traitement en temps réel la sante, les banques, télécommunication gouvernement, la bourse.

Sante : utilise l'analyse en temps réel pour vérifier en permanence l'état médical des patients avoir un état critiques.

Banque : Il devient très important d'assurer des transactions tolérantes aux pannes sur l'ensemble du système. La détection des fraudes est rendue possible grâce aux analyses en temps réel dans le secteur bancaire.

….

5.2 Apache Spark

Apache Spark est une Framework open source de calcule distribuer pour traitement en temps réel .il s'agit l'un des projets les plus réussies d'APACHE FOUNDATION SOFTWARE. Aujourd'hui, Spark est adopte par les grandes organisations comme Amazon, eBay, Yahoo. Spark offre la possibilité d'accéder aux données de différentes sources. Apache Spark fournit une suite complète d'outils complémentaires comprenant une bibliothèque d'apprentissage machine complète (MLlib), un moteur de traitement graphique (GraphX) et un traitement de flux [9].

5.3 Spark (quand Hadoop est déjà là)

Hadoop est basé sur le concept de traitement par lots qui consiste à traiter des blocs de données déjà stocke sur une période donnée. Jusqu'à 2014 Hadoop dépasse toutes les attentes avec le Framework MapReduce. Mais après Spark dépassé Hadoop et sera capable de traiter les données en temps réel et était environ 10 fois plus rapide que Hadoop MapReduce dans les traitements par lots [10].

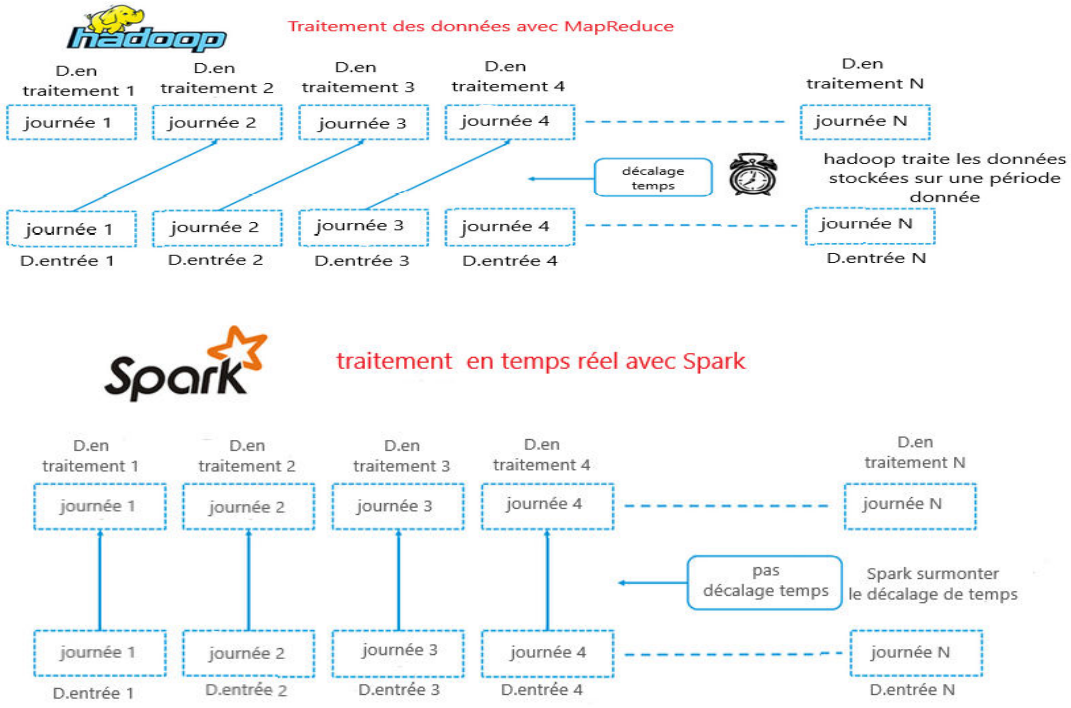


Figure 7:la différence entre Hadoop & Spark(inspirer de [8])

5.4Caractéristiques d'Apache Spark :

L'apache Spark avoir plusieurs caractéristique comme :

Polyglotte : Spark fournit des APIs de haut niveau en JAVA, SCALA, PYTHON, R. le code des Spark peut être écrit dans l'un de ces quatre langages.

La Vitesse : Spark est plus rapide que Hadoop MapReduce pour le traitement de données à grand échelle. Cette vitesse grâce à un partitionnement contrôle,Il gère les données à l'aide de partitions permettant de paralléliser le traitement des données distribuées avec un trafic réseau minimal.

Intégration Hadoop : Apache Spark offre une compatibilité parfaite avec Hadoop. C'est une aubaine pour tous les ingénieurs Big Data qui ont commencé leur carrière chez Hadoop. Spark peut remplacer les fonctions MapReduce de Hadoop.

Apprentissage automatique (Machine Learning) : MLlib de Spark est le composant d'apprentissage automatique très utile pour le traitement de big data Il remplace utilisation plusieurs outils, un pour le traitement et l'autre pour l'apprentissage automatique. Spark fournit aux les analystes et les ingénieurs un moteur puissant et unifié, rapide et facile à utiliser.

Spark a d'autre caractéristique comme : traitement en temps réel, multiple format (supporte les données diffèrent source).

5.5 Les Composants Spark

Les composants Spark sont ce qui donne Apache Spark rapidité et fiabilité.il comprend les composants suivants :

1. Spark Core
2. Spark Streaming
3. Spark SQL
4. GraphX
5. MLlib (Machine Learning)

Spark Core est le moteur de base du traitement de données parallèle et distribué à grande échelle. Le noyau est le moteur d'exécution distribué et les API Java, Scala et Python offrent une plate-forme pour le développement d'applications ETL distribuées [11].

Spark Streaming est le composant de Spark utilisé pour traiter les données de diffusion en temps réel. Il permet le traitement des flux de données en direct à haut débit et à tolérance de pannes. L'unité de flux fondamentale est DStream, qui est essentiellement une série de RDD (Resilient Distributed Datasets) (ensembles de données distribuées résilientes) pour traiter les données en temps réel [11].

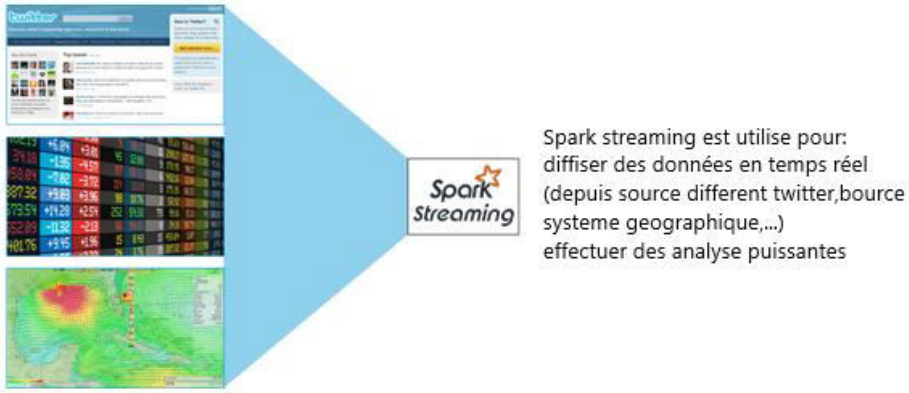


Figure 8: Spark streaming*[12]*

Spark SQL est un nouveau module de Spark qui intègre le traitement relationnel à l'API de programmation fonctionnelle de Spark. Il prend en charge l'interrogation des données via SQL ou via le langage Hive Query. Il fournit un support pour diverses sources de données et permet de tisser des requêtes SQL avec des transformations de code, ce qui donne un outil très puissant [11].

Les quatre bibliothèques de Spark SQL :

Data Source API, DataFrame API, Interpreter & Optimizer, SQL Service.

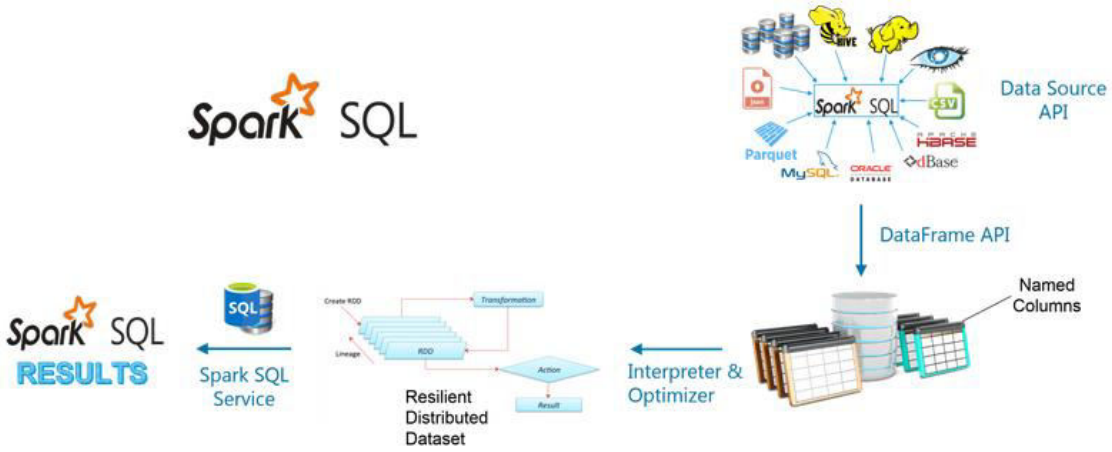


Figure 9:Un processus Spark SQL utilisant les quatre bibliothèques en séquence*[12]*

GraphX est un nouveau composant de Spark pour les graphes et le calcul parallèle de graphes. À un niveau élevé, GraphX étend le RDD Spark en introduisant une nouvelle abstraction Graph : un multi-graphe dirigé avec des propriétés attachées à chaque sommet et arête [11].

MLlib est l'abréviation de Machine Learning Library. Spark MLlib est utilisé pour effectuer un apprentissage automatique dans Apache Spark [11].C'est une bibliothèque d'apprentissage machine fournie par Apache Spark pour rendre / l'apprentissage machine évolutif et facile. Il fournit divers outils [6], par exemple :

ML Algorithmes : algorithmes d'apprentissage courants, tels que la classification, la régression, la classification et le filtrage collaboratif.

Featurization (Fonctions): extraction, transformation, réduction de la dimensionnalité et sélection.

Pipelines: outils pour la construction, l'évaluation et le réglage de ML Pipelines.

Persistance : sauvegarde et chargement des algorithmes, des modèles et des pipelines.

Utilitaires: algèbre linéaire, statistiques, traitement des données, etc.

6. Data mining

Data mining **:** Il existe diverses définitions de l'exploration de données (data mining en anglais) dans la littérature. On peut appeler cela la tâche d'obtenir une information "précieuse" parmi de grandes quantités de données [13]. L'exploration de données, en d'autres termes la découverte de connaissances dans des bases de données (KDD : Knowledge Discovery from Data en anglais), consiste à extraire des informations potentiellement utiles et compréhensibles qui n'ont jamais été découvertes auparavant dans de grandes quantités de données. Les techniques d'analyse de données telles que les systèmes de gestion de base de données traditionnelles, les statistiques, l'intelligence artificielle, l'apprentissage automatique, les processus parallèles et distribués sont également appelés exploration de données [14].
L'exploration de données est un outil puissant qui facilitera la recherche de modèles cachés et de diverses relations entre les données. Le traitement des données découvre des faits cachés dans des bases de données volumineuses. L'objectif général de la technique d'exploration de données est d'extraire des informations d'un vaste ensemble de données et de les transformer en une structure compréhensible pour une utilisation accrue [15].

6. Processus de data mining

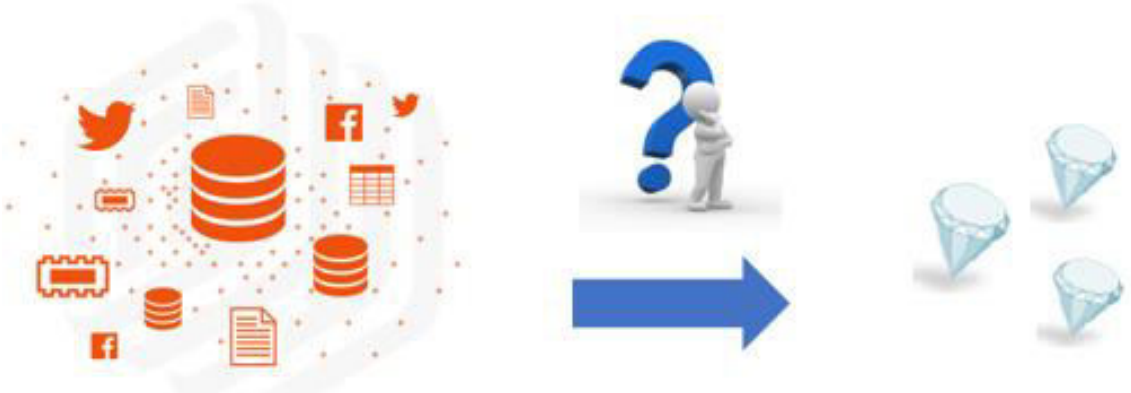


Figure 10: (data mining)la recherche des connaissance(inspirer de[13])

L'exploration de données (data mining) ne consiste pas uniquement à rechercher des modèles dans une masse de données. Il ne s'agit que d'une étape dans tout un processus suivi par des scientifiques visant à extraire des connaissances des données. Data mining un processus composé de cinq étapes (taches) sous le standard CRISP-DM (CRoss-Industry Standard Process for Data Mining).

Les phases de ce processus sont interactives et itératives, vous devrez peut-être revenir aux étapes précédentes pour ajouter ou corriger des données, les phases sont illustres dans la figure dessous.

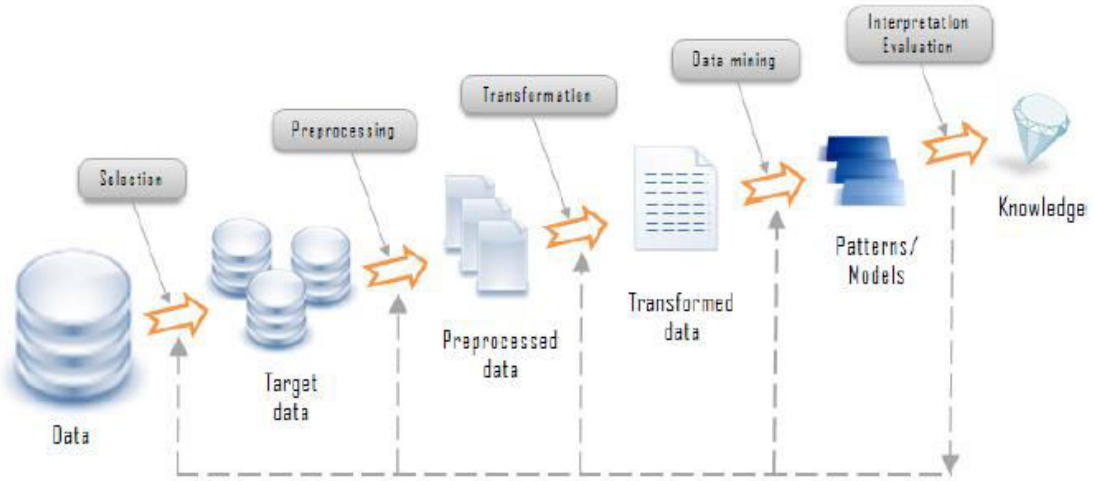


Figure 11:processus de KDD*[16]*

La sélection : en sélectionnant un ensemble de données, ou en se concentrant sur un sous-ensemble de variables ou d'échantillons de données, sur lesquels la découverte doit être effectuée.

Prétraitement : Les données collectées et sélectionne doivent être "préparées". Tout d'abord, elles doivent être nettoyées car elles peuvent contenir plusieurs types d'anomalies : les données peuvent être omises à cause d'erreurs de frappe ou d'erreurs dues au système lui-même. Le prétraitement implique également une réduction des données, ce qui réduit le nombre d'attributs pour accélérer les calculs et représenter les données dans un format optimal pour l'exploration.

Transformation : les données sont transformées et consolidées dans des formes appropriées à l'opération en effectuant une opération de synthèse ou d'agrégation.

Data mining un processus essentiel où des méthodes intelligentes sont appliquées pour extraire des modèles de données. Dans cette étape, on doit choisir la bonne technique pour extraire les connaissances (exploration) des données. Des techniques telles que les réseaux de neurones, les arbres de décision, les réseaux bayésiens, le clustering, ... sont utilisées. Généralement, l'implémentation se base sur plusieurs de ces techniques, puis on choisit le bon résultat.

Evaluation : identifier les modèles vraiment intéressants représentant la connaissance sur la base de mesures d'intéressement.

6.2 Les techniques de data mining

L'exploration de données utilise des outils et des techniques sophistiqués d'analyse de données pour le but de rechercher des modèles et des relations valides et obscurs dans de grands ensembles de données.

L'association parmi les techniques d'exploration de données les plus connues. En association, Aussi appelé technique des relations car c'est un modèle basé à la découverte d'une relation entre les éléments d'une même opération [17]. La technique d'association est utilisée dans l'analyse du panier de marché (réel ou virtuel) pour reconnaître un ensemble de produits que les clients achètent souvent ensemble. Les vendeurs utilisent la technique des associations pour identifier les habitudes d'achat des clients. En se basant sur les données historiques des ventes, les vendeurs pourraient découvrir que les clients achètent toujours du lait quand ils achètent du pain. Ils peuvent donc placer le pain et le lait l'un à côté de l'autre pour gagner du temps et augmenter les ventes.

L'analyse de modèles séquentiels est l'une de ces techniques d'exploration de données qui essaie de trouver ou de reconnaître des modèles similaires, des événements réguliers ou des tendances dans les données de transaction sur une période donnée. Dans les ventes, avec les données de transaction historiques, les entreprises peuvent identifier un ensemble d'articles que les clients achètent ensemble à différents moments de l'année. Les entreprises peuvent ensuite utiliser ces informations pour suggérer aux clients de l'acquérir avec de meilleures offres en fonction de leur fréquence d'achat passée [17].

Clustering est une technique d'exploration de données qui permet de constituer une grappe d'objets significative ou utile [17]. La technique de clustering définit les classes et place les objets qui leur sont liés dans une classe, tandis que les objets de classification sont placés dans des classes prédéfinies. La mise en cluster consiste à placer les objets ayant des propriétés similaires dans un groupe et les objets ayant des propriétés différentes dans un autre groupe [18]. Il est couramment utilisé dans les études de marché, la reconnaissance de formes, le traitement d'images, etc.

La classification est une technique classique d'exploration de données à la lumière de l'apprentissage automatique [17]. La classification permet d'organiser chaque élément d'un ensemble de données en un ensemble prédéfini de classes ou de groupes. Ceci est utilisé pour analyser un ensemble de données donné et en prend chaque instance. Il attribue cette instance à une classe particulière. Il est utilisé pour extraire des modèles. La méthode de classification

utilise des techniques mathématiques telles que les arbres de décision, la programmation linéaire, le réseau de neurones et les statistiques. Dans la classification, nous construisons un logiciel capable de comprendre comment classer les éléments de données en groupes. Par exemple, segmenter le marché en découvrant des groupes de clients distincts à partir de bases de données d'achat.

La prédiction est l'une des techniques d'exploration de données les plus utiles car elle vous permet de projeter les types de données que vous verrez ultérieurement. Des fois, il suffit de reconnaître et de comprendre les tendances historiques pour fournir une prévision assez précise de ce qui se passera dans le futur. La prédiction elle-même est calculée à partir des données disponibles et modélisée conformément à la dynamique existante. Par exemple : en utilise la technique prédiction pour faire des Prévisions météorologiques [17].

La régression est une technique d'exploration de données utilisée pour prédire une plage de valeurs numériques (également appelées valeurs continues), en fonction d'un ensemble de données particulier. Par exemple, la régression peut être utilisée pour prédire le coût d'un produit ou d'un service, en fonction d'autres variables. La régression est utilisée dans plusieurs secteurs pour la planification des activités et du marketing, les prévisions financières, la modélisation environnementale et l'analyse des tendances.il y'a deux types de régression linéaire et non linéaire [17].

6.4 Les domaines d'applications de l'exploration de données (data mining)

Parmi les domaines d'application de DM est :

La santé : l'exploitation de données offre un potentiel considérable pour améliorer les systèmes de santé. Il utilise des données et des analyses pour identifier les meilleures pratiques permettant d'améliorer les soins et de réduire les coûts. Les chercheurs utilisent des méthodes d'exploration de données telles que des bases de données multidimensionnelles, l'apprentissage automatique, l'informatique douce, la visualisation de données et les statistiques [19].

Education : Un nouveau domaine émergent, appelé (Educational Data Mining EDM), concerne le développement de méthodes permettant de découvrir des connaissances à partir de données provenant d'environnements éducatifs. Les objectifs de l'EDM sont les suivants : prédire le comportement futur des élèves en matière d'apprentissage, étudier les effets du

soutien pédagogique et faire progresser les connaissances scientifiques sur l'apprentissage [20].

Détection des fraudes : Des milliards de dollars ont été perdus à cause des fraudes. Les méthodes traditionnelles de détection des fraudes prennent du temps et sont complexes. L'exploration de données aide à fournir des modèles significatifs et à transformer les données en informations. Toute information valable et utile est un savoir [21].

CRM:(Customer Relationship Management) la gestion de la relation client consiste à acquérir et à fidéliser des clients, à les fidéliser et à mettre en œuvre des stratégies orientées client. Pour entretenir de bonnes relations avec un client, une entreprise a besoin de collecter des données et de les analyser [22].

Marketing

L'exploration de données permet aux entreprises de comprendre les modèles cachés dans les données historiques des transactions d'achat, facilitant ainsi la planification et le lancement de nouvelles campagnes marketing de manière rapide et rentable [23].

Analyse du panier de marché : L'analyse du panier de marché est une technique de modélisation basée sur une théorie selon laquelle, si vous achetez un certain groupe d'articles, vous êtes plus susceptible d'acheter un autre groupe d'articles. Cette technique peut permettre au détaillant de comprendre le comportement d'achat d'un acheteur.

Les entreprises de vente au détail (détaillants) utilisent l'exploration de données pour identifier les habitudes d'achat du comportement du client.

Figure 12: les domaines d'applications de data mining

7. Conclusion

Dans ce chapitre on discutant

- Le réseau social est défini comme : « une structure définie par des relations entre des individus », Un réseau social représente un système d'entités en interaction.
- Les resaux sociaux fournit un grand masse de données continu
- Les big data et leur caracteristique et les methodes d'analyse les big data
- Les apaches Hadoop & Spark comme des outils pour traiter et utiliser les big data
- Dataminig et leurs techniques pour traiter et l'explorer et l'analyser et le contrôle de données
- Les domaines d'applications dataminig (la sante , education ,banque,…..)

Chapitre II :

Methodes & Approches d'analyse & contrôle de données

1. Introduction

L'analyse des informations dans réseaux sociaux est devenue un outil important dans tous aspects de la vie parce que les gens partagent leur vie avec tous en utilisant à partir des réseaux sociaux (partage les photos, les opinions, les sentiments, les événements Qui se produisent dans leurs vies, ...)

l'analyse de la information qui circule dans les réseaux sociaux ouvert la porte de plusieurs axe de recherche tel que l'analyse des sentiments (analyse des opinions),la détection des événements, la détection de communautés ,parmi les travaux récent la détection des événement de trafic [32],gérer les sondage (sur le vote ,nouvel loi, un événement d'actualité, ...),et l'analyse des réseaux sociaux changer complètement la méthodologie de marketing (la prédiction de la produit de future , marketing ciblé [cible des gens qui intéressent à des produits spécifique]).

Les médias sociaux contiennent de nombreuses informations géographiques et constituent l'une des sources de données les plus importantes pour limiter des risques. Comparés aux méthodes traditionnelles de collecte d'informations géographiques liées aux catastrophes, les médias sociaux présentent les caractéristiques suivantes : fourniture d'informations en temps réel et faible coût. En raison du développement des technologies d'extraction de données volumineuses, il est maintenant plus facile d'extraire des informations géographiques utiles liées aux catastrophes des données volumineuses des médias sociaux. Les informations émotionnelles publiques contenues dans les médias sociaux pourraient nous aider à comprendre les catastrophes de manière plus détaillée que ne le permettent les méthodes traditionnelles [33].

2.La diffusion de l'information

La diffusion de l'information c'est la circulation d'information d'un individu ou d'une communauté à un autre dans le réseau. Nombreuses recherches ont été consacrées à l'analyse de la diffusion de l'information. Les modèles et les méthodes de la diffusion de l'information aide nous de comprendre " comment l'information diffuse et quelle est l'information diffuse plus rapidement " en générale la compréhension du phénomène de diffusion de l'information.
Le model de diffusion a une valeur de référence pour diverse applications (contrôle de rumeur [21], analyse de comportement, mesure l'opinion publique). Les modèles de diffusion peuvent être classé en deux catégories des modèles explicatifs et des modèles prédictifs [22].

2.1Modèles explicatifs

L'information est diffusée par d'interactions entre différents membres de la société. Ces individus (membres) peuvent être considérés comme des nœuds dans les réseaux sociaux. Le nœud du réseau social est abstraction de l'utilisateur dans une "vraie" communauté. Les interactions ou relations sont représentées par des arêtes entre deux nœuds dans des réseaux sociaux. Par conséquent, un véritable groupe social peut être cartographié par un vaste réseau social et une information peut être diffusée par ces nœuds au sein de celui-ci [22].

2.1.1 Les modèles de base épidémiques

Le processus de diffusion de l'information peut être considéré de la même manière qu'un processus de propagation épidémique. Dans la transmission d'épidémies, il y a des utilisateurs infectés par des agents pathogènes et des utilisateurs susceptibles aux agents pathogènes. Le virus peut se transmettre des utilisateurs infectés aux utilisateurs vulnérables, et les informations peuvent être diffusées des communicateurs aux destinataires de la même manière.
Les modèles de base des épidémies sont le modèle SI (Suceptible Infected), le modèle SIS (Suceptible Infected Suceptible), le modèle SIR (Suceptible Infected Removed)) et le modèle SIRS (Suceptible Infected Removed Suceptible).

i. A. Le modèle SI

Le modèle SI pour les réseaux complexe propose par Pastrosatoras en 2001. Le modèle suppose que le nombre total de personnes est égal à N. N est divisé en deux catégories : S (susceptible) et I (infecté). Au temps t, s (t) représente la proportion susceptible de la population totale, i (t)

représente la proportion infectée [Ns(t)+Ni(t)=N] et λ représente le taux de contact quotidien. Le modèle SI peut être décrit par les équations (1) et (2).

$$\frac{di}{dt} = \lambda i(i-1) \qquad (1)$$

$$i(0) = i_0 \qquad (2)$$

i. B. Le modèle SIS

Parmi les inconvénients de modèle SI ne permet pas aux utilisateurs infectés d'être guéris après avoir été infectés. Newman et Gross [23]propose le modèle SIS pour résoudre les problèmes de modèle SI, le model SIS garde tous paramètres de model SI (S(t), I(t), N, λ) avec la même définition de model précèdent, Newman et Gross propose le modèle SIS pour résoudre les problèmes de modèle SI , le model SIS garde tous paramètres de model SI (S(t),I(t),N, λ)avec la même définition de model précèdent, en ajoutant μ paramètre . μ Représente la proportion des utilisateurs infectés qui ont été guéris dans la population totale.

Le modèle SIS peut être décrit par les équations (3) et (4).

$$\frac{di}{dt} = \lambda i(i-1) - \mu i \quad (3)$$

$$i(0) = i_0 \qquad (4)$$

Le modèle SIS est une généralisation du modèle SI (si u = 0 on obtient le modèle SI).

i.C. Le model SIR

Lorsqu'un individu est guéri peut-être deviendra un immunitaire. Dans les modèles précédents (SI, SIS) ce cas n'est pas compté. [23] [24]Kermack et Mc Kendrick ont établi un modèle SIR, ajoutant une autre catégorie R ("Removed'', "Recovered''' : pour les personnes immunisées après la cure), et préservant les deux catégories des modèles précédents (S : susceptible, I : infecté), donc au temps t : s(t)+i(t)+r(t)=1. Dans le model SIR on suppose que s (0) = s0, i (0) = i0, r (0) = 0.

Le modèle SIR peut être décrit par les équations (5) et (6), (7).

$$\frac{ds}{dt} = -\lambda is = -\lambda i(i-1) \quad (5)$$

$$\frac{di}{dt} = \lambda i(i-1) - \mu i \qquad (6)$$

$$\frac{dr}{dt} = \mu i \qquad (7)$$

/μ : Représente la proportion des utilisateurs immunisée dans la population totale.

i. D. Le model SIRS

Le model SIRS ajouter l'idée la probabilité d'un individu (utilisateur) peut-être revenu à l'état susceptible [25] .Le modèle SIRS est une généralisation de modèle SIR ou la probabilité d'un individu guérie revenu à l'état susceptible est zéro. Le modèle garde les même catégories (S, I, R) avec les mêmes définitions. Le modèle SIRS peut être décrit par les équations (8) et (9), (10).

$$\frac{ds}{dt} = -\lambda is - \alpha i \quad (8)$$

$$\frac{di}{dt} = \lambda i(i-1) - \mu i \quad (9)$$

$$\frac{dr}{dt} = \mu i - \alpha i \quad (10)$$

/α : Représente la proportion des utilisateurs revenu à l'état Susceptible dans la population totale.

ii. Comparaison entre les modèles

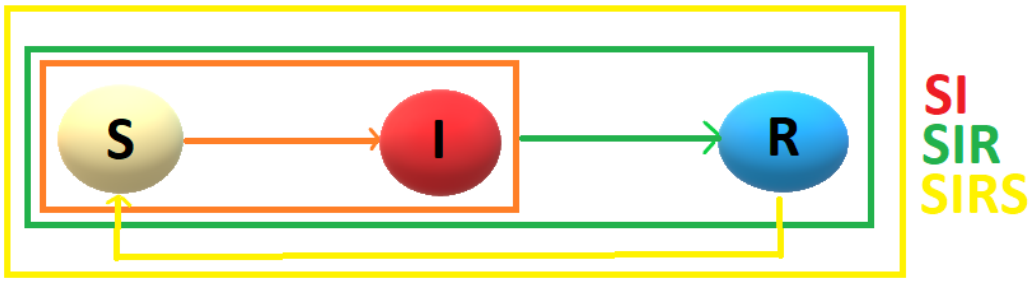


Figure 1:compraison entre les modèles épidémiques

En peut associe les catégories des modèles(SIR) avec les statuts des utilisateurs de réseaux

S(susceptible) : L'individu peut devenir un diffuseur de l'information.

I(infecté) : un communicateur.

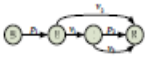
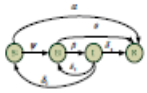
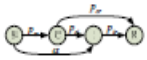
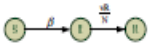
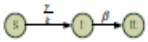
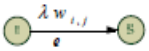
R(immunité) : L'individu ne peut pas devenir un diffuseur de l'information.

iii. Les modèles épidermiques dans les réseaux sociaux

Le modèle SEIR a établi en ajoutant le catégorie(nœud) Exposed (E) basé sur le model SIR [26],le modèle SEIR décrire avec précision le processus de propagation des informations, en analysant l'impact de la fréquence de connexion des utilisateurs et du nombre d'amis sur la diffusion des informations [22].

Le modèle S-SEIR c'est le résultat de modification de modèle SEIR parce que ont constaté que la diffusion de l'information n'était pas seulement liée au comportement de l'utilisateur, mais aussi à la valeur de l'information elle-même. Le modèle S - SEIR peut bien simuler le processus de diffusion d'informations sur un réseau social [22].

Le modèle SCIR base sur le model SIR pour les micro-blogs en ajoutant un statut Contacted (C). Cela suppose que tous les fans ont le statut Contacté lorsqu'un utilisateur publie un message. Ensuite, le statut des fans changera selon une certaine probabilité, devenant soit des utilisateurs de la transmission, soit des utilisateurs immunisés après un certain temps. Ce modèle peut représenter la régularité de la diffusion en ligne de sujets [27]

Scalable Model	Method	Consider the User's Different Behaviors	Expression of the Diffusion Process	Dynamic Infected Rate and Recovery Rate	Performance Metrics	Applications
SEIR	add Exposed node	-		-	distribution of nodes density	detect the affect factors: login frequency and number of friends
S-SEIR	information value is considered	√	 δ = user behavior	-	distribution of S, E, I, and R	simulate the diffusion process
SCIR	add Contacted node	-		-	distribution of I and R	represent the regularity of online topic spreading
irSIR	add Infection Recovery dynamics	-	 v = an infectious recovery rate	√	degree of fitting with real data	describe OSN abandonment
FSIR	consider the behavior of the neighbors	√	 k = node degree	√	degree of fitting with real data	detect the affect factors: information numbers and friends numbers
ESIS	consider the information weight with emotion	-	 λ = the probability of I to S; $w_{i,j}$ = the strength of edge e from i to j	√	degree of fitting with real data	detect the affect factors: propagation probability and transmission intensity

OSN: online social network. S: suceptible. E: exposed. I: infected. R: removed.

Figure 2:les modèles épidémiques dans les réseaux sociaux [22]

2.1.2. Les modèles d'influence

Sur la base de l'influence, les modèles de la diffusion de l'information sont classifiés en trois class : influence individuelle, influence de communauté [28].

i. Influence individuel :

Influence individuel : les réseaux sociaux est représenté come des nœuds relie avec des liens (arête), l'influence individuel c'est l'opération de de trouve leader de l'opinion c'est à dire la recherche des nœuds avoir un grand rôle dans la diffusion de l'information. Les travaux sur l'influence des leaders d'opinion comprennent des méthodes basées sur la structure du réseau, les attributs de l'utilisateur.

Plusieurs méthodes proposé pour influence individuel.
Une méthode basée sur la structure des réseaux pour modéliser et mesure l'influence individuel des leader d'opinion dans les micro-blog. Le processus de la diffusion est décrit par graphe direct dynamique. [22]

Une méthode basée sur le comportement et les interactions d'un utilisateur à l'autre pour de recherche de leaders d'opinion. [22]

Un modèle efficace pour maximiser la diffusion d'informations et minimiser le temps de contagion à partir la rechercher les nœuds d'influence, Ce modèle prend en compte les interactions entre les nœuds, la structure topologique du réseau. Dans ce model il classe les nœuds en fonction des poids entre différents nœuds. L'influence entre les utilisateurs est déterminée par les interactions temporelles de l'utilisateur et de ses voisins. Les nœuds d'influence des K supérieurs seront sélectionnés par les voisins de code, les voisins de voisins et les connexions topologiques. [29]

Researcher	Network Structure	User Interactions	User Attributes		Method	Quantitative Criterion	Applications
			User behaviors	Other features			
Chenxu	√	-	-	-	social network analysis	out-degree	identify opinion leaders and prediction
Bo	-	√	√	centrality	competency	activists, centrality and intermediary	identify opinion leaders and influence maximization
Jiaxin	√	-	√	access time	social network analysis	capability of diffusion	influence predicting
Xianhui	√	√	√	topic and weight	page-rank	coverage and coreratio	mining topic opinion leader
Ullah	√	√	√	neighbors-of-neighbors	social network analysis	activists	identify influential nodes

Figure 3:les méthodes de l'influence individuelle *[22]*

ii. Influence de la communauté

Une communauté est un ensemble (groupe) de personnes ayant des propriétés communes. Dans les réseaux sociaux les individus construit La détection communautaire est la base de la recherche sur l'influence communautaire, la recherche sur la détection de communautés inclut des méthodes basées sur des liens, des contenus ou des attributs et des sentiments. Cependant, les méthodes basées sur un seul critère (les méthodes basées sur les liens n'est pas précise dans les réseaux dynamiques, les méthodes basées sur le contenu, un contenu falsifie induiront en erreur dans le processus de détection) ne convient pas à l'étude des réseaux sociaux dynamiques. Des différentes communautés à cause de différents intérêts de ces individus .la communauté est une sous-ensemble de réseaux où les utilisateurs de ce groupe sont connectés et ayons des attributs similaires, malgré que la structure de réseau peut être change avec le temps les communautés restent stables. Le défi est comment détecter les communautés qui ont une grande influence au sein d'un réseau social. Parmi les solutions trouve l'utilisation des liens, le contenu, attributs.

Model	Links	Attributes or Contents	Sentiment	Method
PCL-DC	√	√	-	probability
SA-Cluster-Inc	√	prolific and topic	-	cluster
CODICIL	√	stemmed words, title and context, tags	-	cluster
sentiment-topic based	√	user, text	√	probability
SVO	√	interests	√	cluster

Figure 4: algorithmes de l'influence de communauté *[22]*

C'est pour ça, la plupart des méthodes propose utilisent une combinaison de ces deux méthodes (les méthodes de cluster, liens de manière itérative), la première étape consiste à construire la structure de réseau en fonction du contenu ou des attributs. Deuxièmement, la structure initiale sera mise à jour par les liens de manière itérative. L'objectif principal est d'améliorer la précision de la détection de la communauté et de réduire la consommation de temps. En termes de précision, divers attributs sont pris en compte.

2.2 Les modelés prédictifs

Être capable de prédire comment les informations se répandront dans le réseau à l'avenir sera utile (Tout vendeur voudra connaître l'article le plus courant dans les années à venir, ...). Les modèles prédictifs sont utilisés pour prédire le futur processus de diffusion de l'information dans les réseaux sociaux en fonction de certains facteurs. Ces modèles sont également souvent utilisés pour maximiser l'influence. Ce sont le modèle IC (independant cascade), le modèle LT (Linear Threshold).

2.2.1 Le model indépendance cascade(ICM)

Dans le model IC le nœud active est le nœud accepté l'information, le nœud peut changer leur état inactive vers un état active. Lorsque le nœud u devient actif pour la première fois à l'étape t, il dispose d'une seule chance d'activer chaque enfant v actuellement inactif v et réussit avec

la probabilité p, v. Ici, pu, v est une constante indépendante de l'historique du processus, et le nœud v est appelé un enfant du nœud u et le nœud u est appelé un parent du nœud v s'il existe un lien dirigé (u, v) de u vers v. Si u réussit, alors v deviendra actif à l'étape t + 1. Si plusieurs parents de v deviennent actifs à l'étape t, leurs tentatives d'activation sont séquencées dans un ordre arbitraire, mais effectuées à l'étape t. Qu'il réussisse ou non, il ne pourra plus tenter d'activer v dans les sous-centres. Le processus se termine si aucune autre activation n'est possible [30].

Un exemple de ce modèle est présenté à la Figure 17 . Les nœuds actifs sont indiqués en trait pointillé jaune. Au moment initial, deux nœuds C et D sont activés. Au pas de temps suivant, les nœuds C et D ont la possibilité d'activer leurs trois voisins (A, G et H) et (B, E et F) respectivement. Selon la figure 17.b, seuls trois nœuds A, H et E ont été activés avec succès et les nœuds actifs initiaux deviennent gris (ce qui signifie qu'il reste actif mais ne permet pas d'activer les autres). Au prochain pas de temps, deux nœuds G et F deviennent actifs et les nœuds actifs précédents A, E et H deviennent gris. À heure = 2, deux nœuds F et G deviennent actifs. Les voisins du nœud G étant actifs, il n'a aucune chance d'activer des nœuds. Le nœud F a une option pour activer le nœud I. Toutefois, il échoue, comme indiqué dans l'exemple de la figure 17.d. Comme il n'y a plus de nouveau nœud actif, le processus de diffusion s'est arrêté.

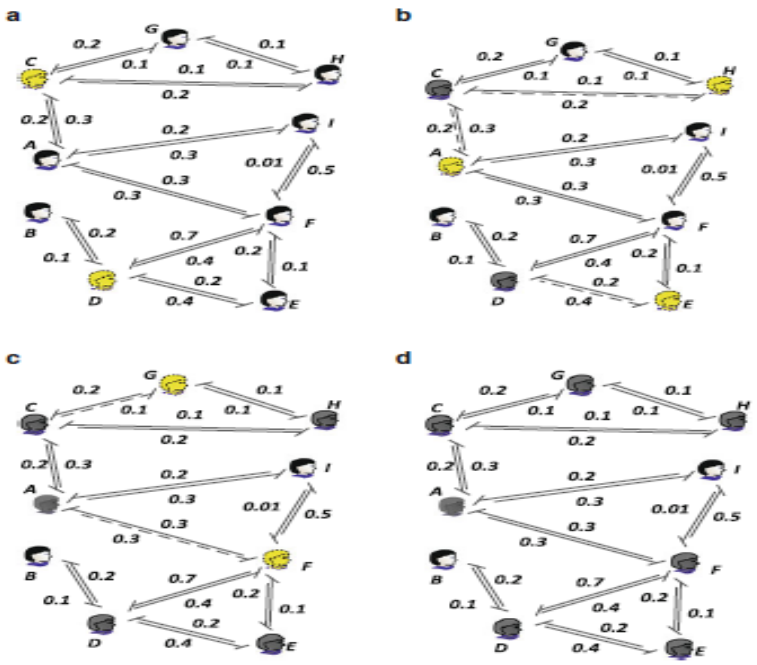


Figure 5:model IC(Independent cascade) *[31]*

2.2.2 Modèle de seuil linéaire (Linear Threshold Model)

Le modèle à seuil linéaire étend le modèle de basculement à sa variante naturelle pondérée, où chaque bord dirigé (u, v) E a un poids non négatif b (u, v). Pour tout nœud v ∈ V, la somme

$$\sum_{u \in \eta^{in}(v)} b(u,v) \leq 1$$

totale des poids des fronts entrants est inférieure ou égale à un, c'est-à-dire [31]:

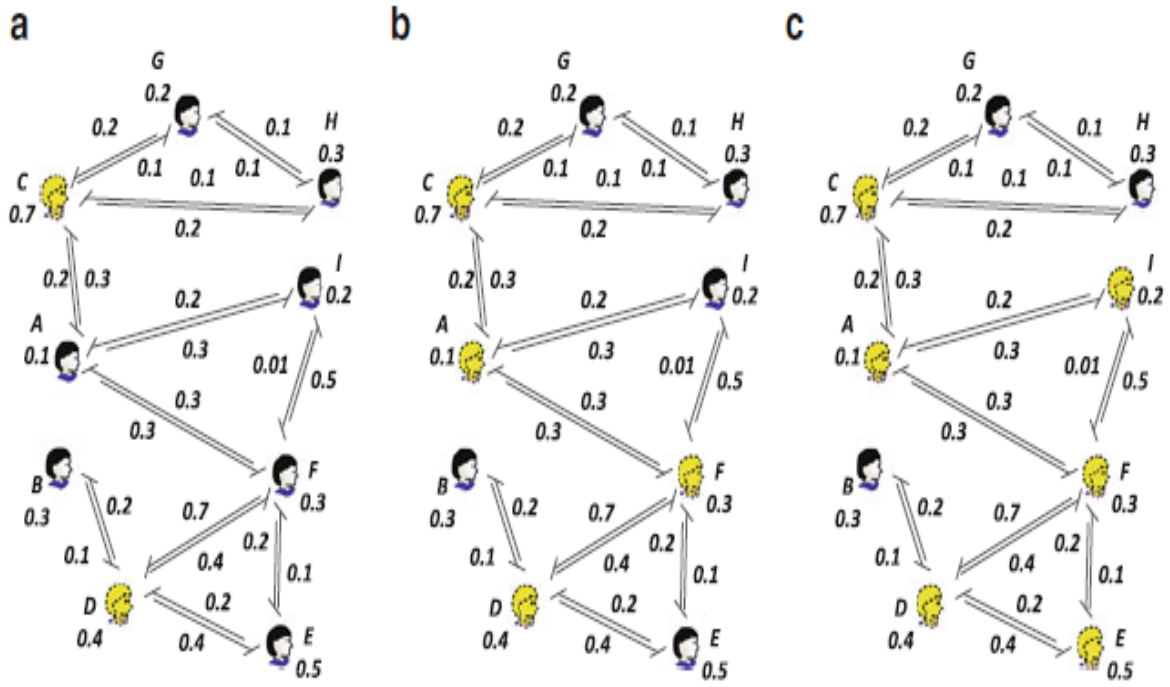


Figure 6:exemple model LT *[31]*

Un exemple de ce modèle est illustré à la Figure 18. Un seuil aléatoire est attribué à chaque nœud dans [0,1] et deux nœuds en pointillés jaunes C et D sont activés initialement. Nœud C est incapable d'activer deux nœuds G et H car son poids d'influence n'est pas assez important, mais il est capable d'activer le nœud A (0.2≥0.1). Le nœud D active également le nœud F (0.4≥0.3), mais pas B et E. À l'étape suivante (Figure 6.b), il y a quatre nœuds actifs et ils permettent d'activer le nœud I (0.3 + 0.5 ≥ 0.2) et E (0,4 + 0,2 ≥ 0,5). Au prochain pas, aucun nouveau nœud actif existe ; alors, le processus de diffusion se termine.

2.3 Discussion

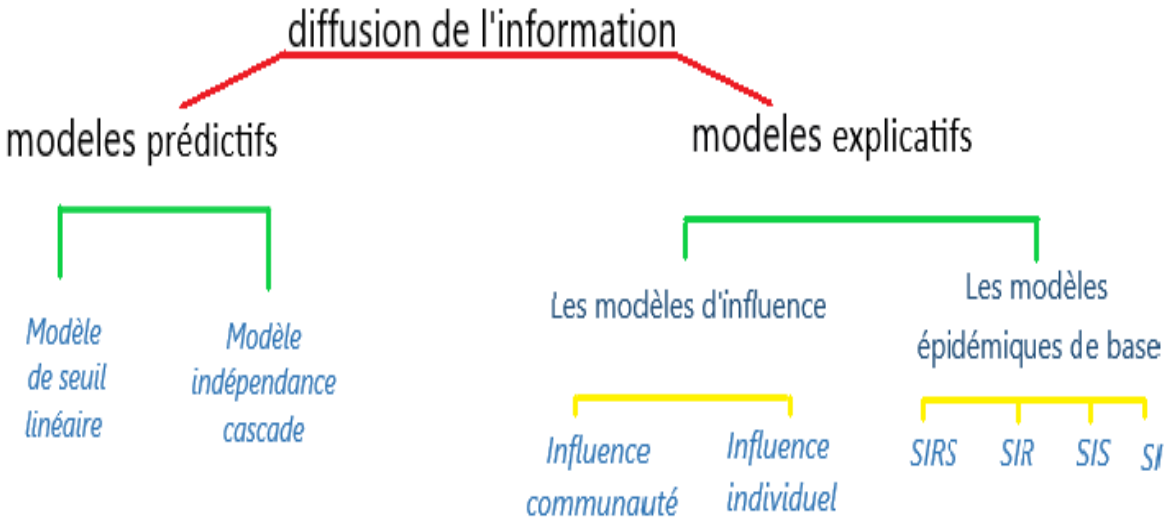


Figure 7:les modèles de diffusion de l'information

Les réseaux sociaux en ligne clonée la structure société humaine. Avec une opération de diffusion de l'information, il peut extraire plusieurs informations cachées peuvent déployée ces informations pour la prédiction de vente, marché, contrôle des rumeurs, analyse des opinions et les sentiments sur un sujet quel qu'on. Et pour accès a ces informations, les chercheurs basés sur les modèles de diffusion, ces model regroupes en deux groupes explicatifs (les modèles épidémiques, les modèles de l'influence) et prédictifs (IC, LT). Les modèles prédictifs ne sont pas indépendants a les modèles explicatif et l'inverse. Les modèles IC, LT ne sont pas utilisés seulement pour la prévision, mais également pour la propagation et la maximisation de l'influence.

3. Analyse des sentiments

Les gens s'expriment des sentiments et l'analyses de ces sentiments aide nous de comprendre leurs attitudes et leurs réactions. L'analyse des sentiments est utilisée d'une façons très large et puissante pour capture des opinions et émotions publiques sur un sujet quel qu'on. Et maintenant avec les réseaux sociaux comme un grand libre espace pour donne les opinions et les vues qui représente une grande ressource de données continu, l'application de l'analyse des sentiments sur les réseaux sociaux est très utiles et adoptée par les entreprises pour augmenter leur bénéfice [34].

Selon Bing Liu dans [35] :

« L'analyse des sentiments, également appelée extraction d'opinion, est le domaine d'étude qui analyse les opinions, les sentiments, les évaluations, les appréciations, les attitudes et les émotions des personnes vis-à-vis d'entités telles que des produits, services, organisations, individus, problèmes, événements et sujets. "

La recherche sur l'analyse des sentiments s'est principalement concentrée sur deux choses : déterminer si une entité textuelle donnée est subjective ou objective et identifier la polarité de textes subjectifs. La plupart des études d'analyse des sentiments utilisent des approches d'apprentissage automatique [36].
Dans le domaine de l'analyse des sentiments, les textes appartiennent à des classes positives ou négatives. Il peut également y avoir des classes à valeurs multiples ou binaires comme positives, négatives et neutres (pertinentes ou non pertinentes). La complexité fondamentale de la classification des textes dans l'analyse des sentiments par rapport à celle d'autres catalogues thématiques est due à la non-utilisation des mots-clés [36].

Les différentes approches pouvant être appliquées à l'analyse des sentiments [36] [37] :

- Approches lexicale
- Approches basée sur l'apprentissage automatique
- Approches hybride / combinée

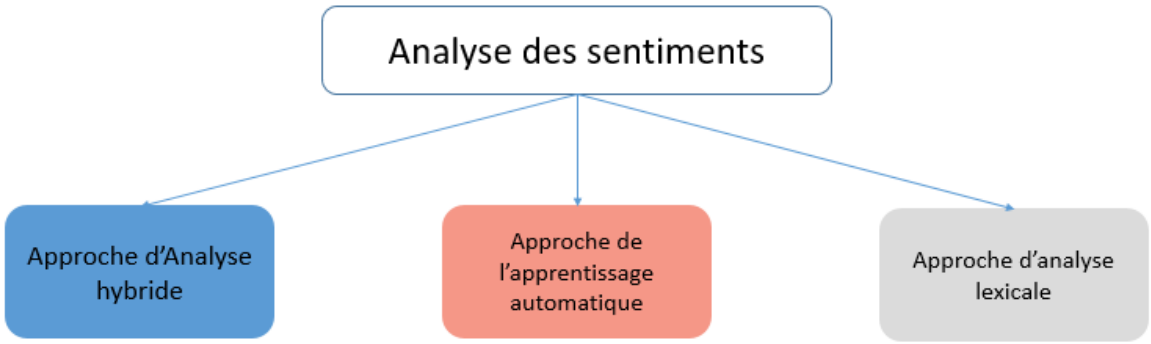


Figure 9: les approche de l'analyse sémantique

3.1 Approche de l'analyse lexicale

La méthode basée sur Lexico utilise un dictionnaire de sentiments avec des mots d'opinion et les associe aux données pour déterminer la polarité. Ils attribuent des scores de sentiment aux mots d'opinion décrivant la valeur positive, négative et objective des mots contenus dans le dictionnaire. Les approches basées sur le lexique s'appuient principalement sur un lexique des sentiments, c'est-à-dire un ensemble de termes, d'expressions et même d'idiomes connus et précompilés, développés pour les types de communication traditionnels, tels que le Opinion Finder Lexicon [37].

Cette technique est régie par l'utilisation d'un dictionnaire composé de lexiques pré-étiquetés. Le texte saisi est converti en jetons par Tokenizer. Chaque nouveau jeton rencontré est ensuite mis en correspondance avec le lexique du dictionnaire. Si la correspondance est positive, le score est ajouté au pool total de scores pour le texte saisi. Par exemple, si « dramatique » est une correspondance positive dans le dictionnaire, le score total du texte est incrémenté. Sinon, le score est décrémenté ou le mot est marqué comme négatif. Bien que cette technique semble être de nature amateur, ses variantes se sont prouvée dignes.

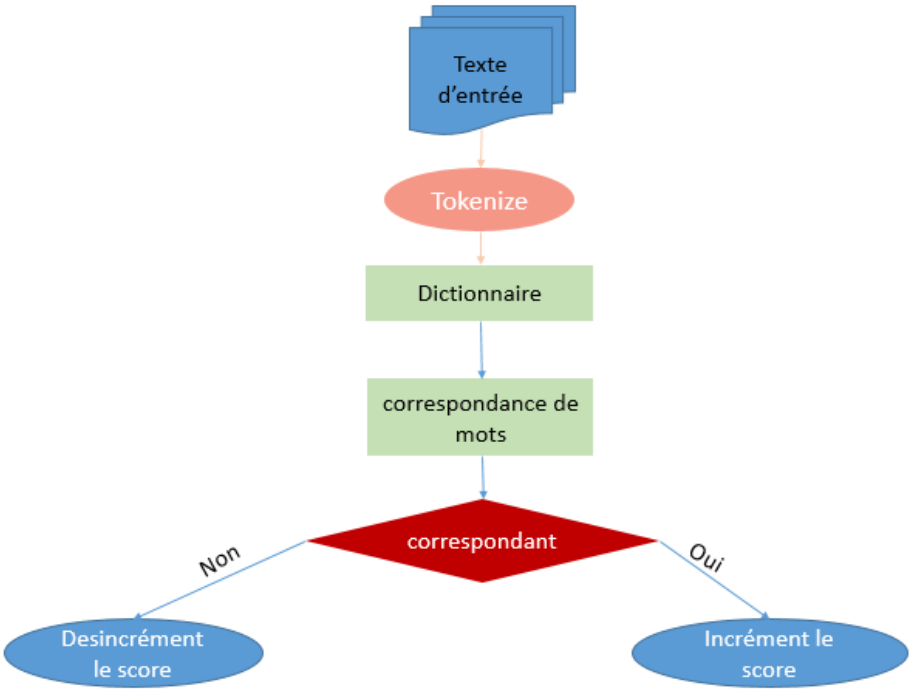


Figure 10:processus de travail dans approche lexicale(inspire de *[36]*)

Il existe deux sous-classifications pour cette approche :

3.1.1 À base de dictionnaire :

Il est basé sur l'utilisation de termes (graines) qui sont généralement collectés et annotés manuellement. Cet ensemble se développe en recherchant les synonymes et antonymes d'un dictionnaire. WordNet est un exemple de ce dictionnaire. Il est utilisé pour développer un thésaurus appelé SentiWordNet [37].

Inconvénient : Ne peut pas traiter les orientations spécifiques au domaine et au contexte.

Dans l'approche de dictionnaire un ensemble des mots a une orientation soit positive/négative qui est collecte manuellement, comme premier phase, après en ajoutant chaque synonyme de chaque mot dans le premier ensemble, le processus est terminé c'est aucun nouveau mot ajouté. Plusieurs travaux sur ce domaine concentré sur nettoyer le dictionnaire (plus précisément les mots colleté) parmi les méthodes utilise est les méthodes de similarité, et d'autre utilisée les méthodes de probabilité pour minimiser l'erreur. Plus de ça d'autre ont également exploité de nombreux modèles d'affixes générant des antonymes tels que Y et DisY (par possible-impossible, pertinent - non-pertinent) pour augmenter la couverture. Avec l'utilisation de

Wordnet (Kamps et al,2004) proposé une méthode basée sur la distance entre les mots, cette distance aide nous de décide l'orientation de chaque mot soit positive soit négative, cette distance est calculé en utilisant deux terme de référence (un terme positive et l'autre négative), autre méthode propose mettre en compte les mots n'a pas une orientation exacte (neutre) ,cette méthode utilise les graphe sémantique pondre dirigé ,les nœuds voisins sont des synonymes ou antonymes [38].

D'autres auteurs ont mis au point une méthode qui utilisait trois dictionnaires différents (traditionnellement, un seul est utilisé) pour obtenir des synonymes et des antonymes basés sur des mots clés. Ensuite, le lexique développé a été utilisé pour la classification des tweets.la technique proposée permettait de classer les tweets, contrairement à la méthode traditionnelle basée sur un dictionnaire. Néanmoins, l'approche suggérée présente plusieurs inconvénients. Le problème principal est une collection de synonymes et les antonymes nécessitent beaucoup de temps. En outre, les dictionnaires contiennent généralement des mots formels, mais les données (tweets) sont remplies de lexiques informels [39].

Généralement, l'inconvénient majeur de l'approche basée sur les dictionnaires est l'incapacité de détecter des mots de sentiment avec des orientations de polarité propres à un domaine ou à un contexte.

3.1.2 À base de corpus :

L'approche basée sur un corpus a pour objectif de fournir des dictionnaires liés à un domaine spécifique. Ces dictionnaires sont générés à partir d'un ensemble de termes d'opinion initiaux qui se développent grâce à la recherche de mots apparentés au moyen de techniques statistiques ou sémantiques.

- Méthodes basées sur des statistiques : Analyse sémantique latente (LSA).
- Les méthodes basées sur la sémantique telles que l'utilisation de synonymes et d'Antonymes ou les relations issues de thésaurus comme WordNet peuvent également représenter une solution intéressante [37].

L'approche basée sur le corpus aide à résoudre le problème de la recherche de mots d'opinion avec des orientations spécifiques au contexte. Ses méthodes dépendent de schémas syntaxiques ou de schémas qui apparaissent conjointement avec une liste de semences (les mots initiaux) de mots d'opinion pour trouver d'autres mots d'opinion dans un grand corpus.

Une de ces méthodes proposées. Elle a commencé par dresser une liste d'adjectifs d'opinion relatifs aux semences et les ont utilisés avec un ensemble de contraintes linguistiques pour

identifier d'autres mots d'opinion adjectifs et leurs orientations. Les contraintes concernent les connecteurs tels que ET, OU, MAIS ... .la conjonction ET, par exemple, indique que les adjectifs associés ont généralement la même orientation. Cette idée s'appelle la cohérence des sentiments, ce qui n'est pas toujours cohérent dans la pratique. Il existe également des expressions adversatives telles que mais qui sont toutefois indiquées lorsque l'opinion change. Afin de déterminer si deux adjectifs conjoints ont des orientations identiques ou différentes, l'apprentissage est appliqué à un grand corpus. Ensuite, les liens entre les adjectifs forment un graphique et la classification est effectuée sur le graphique pour produire deux ensembles de mots : positif et négatif [40].

La recherche des mots de sentiments d'un domaine spécifique est utile, mais pas assez en pratique, parce que dans le même domaine le mot peuvent avoir plus d'orientations contextuelles (signification) par exemple long, court, grand, petit ...etc. dans le champ de la téléphone, le mot' long ' exprime clairement les points de vue opposés dans les phrases suivantes: "la durée de vie de la batterie est longue" (positif) et "le mettre à jour prend long temps" (négatif).Il ne suffit donc pas de trouver des mots de sentiment basés sur le domaine et ses orientations. Pour résoudre ce problème Les auteurs Ils ont proposé d'utiliser la paire (aspect, mot_de_sentiment) (aspect, sentiment_Word) comme contexte d'opinion, par exemple (« la vie de la batterie », « longue »). Leur méthode détermine ainsi les mots de sentiment et leurs orientations en fonction des aspects qu'ils modifient [38].

3.1.3 Discussion

- L'utilisation d'une approche basée sur le dictionnaire présente l'avantage de pouvoir trouver facilement et rapidement un grand nombre de mots de sentiments avec leurs orientations. Bien que la liste résultante puisse comporter de nombreuses erreurs, une vérification manuelle peut être effectuée pour la nettoyer, ce qui prend du temps, mais ce n'est qu'un effort ponctuel. Le principal inconvénient est que les orientations de sentiment des mots collectés de cette manière sont générales ou indépendantes du domaine et du contexte.

- L'utilisation de la seule approche basée sur un corpus n'est pas aussi efficace que l'approche basée sur un dictionnaire car il est difficile de préparer un large corpus couvrant tous les mots, mais cette approche présente un avantage majeur qui permet de rechercher leurs orientations en utilisant un corpus de domaine.

3.2 Approche de l'apprentissage automatique :

La méthode d'apprentissage automatique s'appuie sur les célèbres algorithmes d'apprentissage automatique pour résoudre l'analyse des sentiments en tant que problème de classification de texte classique qui utilise des caractéristiques syntaxiques (et /ou) linguistiques. Il y'a deux types d'apprentissage automatique : supervisée et semi & non supervisées [40].

Apprentissage automatique comportent deux phases : phase d'entrainement et phase de prediction.la génération de modèle de classification à partir des données étiquetées manuellement (pour extraire les caractéristiques)(généralement) c'est la première phase et la deuxième phase est prédiction (ex : la prédiction des sentiments des données non étiquetées via le modèle génère, c'est la phase de test de modèle) [34].

3.2.1. Apprentissage automatique supervisée :

Les méthodes d'apprentissage automatique supervisées nécessitent la présence de données d'apprentissage étiquetées qui sont utilisées pour le processus d'apprentissage. En tant qu'ensemble de données d'entraînement, les données étiquetées doivent être utilisés. Le modèle de sac de mots est utilisé pour représenter une donnée sous forme de vecteur de caractéristiques = (1, $w2$, …, wi, …, wN), où N est un ensemble de tous les termes uniques dans le jeu de données d'apprentissage et Wi correspondent au i-ème terme. Pour convertir les données d'apprentissage en vecteur de caractéristiques (avec taille de N /N : le nombre de mots unique) n'importe modèle utilise peut-être utiliser, parmi ces modèles : Le modèle binaire (1 si le mot existe dans le document si non 0), le modèle TF (terme frequency, le nombre d'occurrences d'un terme dans le document.), le modèle TF/IDF (terme frequency/ inverse document frequency, donne l'importance de chaque mot). Après conversion de données de l'entrainement vers un vecteur de caractéristiques, il peut être utilisé par le classificateur pour entrainer et estimer des étiquettes, parmi ces classificateurs : Naïve Bayes, Support Vector Machine (SVM), Neural Networks (NN)... [34] [37] [39] [40] .

A. Naïve Bayes

Le classifieur Naïve Bayes est le classificateur le plus simple et le plus couramment utilisé. Le modèle fonctionne avec l'extraction de caractéristiques BOWs (Bag Of Words, sac de mots), qui ignore la position du mot dans le document [40]. Le modèle est basé sur le théorème de Bayes, supposant que les caractéristiques sont indépendantes. Le classifieur Naïve Bayes définit la probabilité que le document appartienne à une classe particulière [41].

B. Support vector machine (SVM)

Les machines à vecteurs de support sont une machine à apprendre pour les problèmes de classification à deux groupes. Il est utilisé pour classer les textes en tant que positifs ou négatifs. SVM fonctionne bien pour la classification de texte en raison de ses avantages tels que son potentiel de gestion de grandes fonctionnalités. Un autre avantage est que SVM est robuste quand il y a peu d'exemples et aussi parce que la plupart des problèmes sont linéairement séparables [42].

C. Le classifieur d'arbre de décision

Le classifieur d'arbre de décision fournit une décomposition hiérarchique de l'espace de données d'apprentissage .la division des données est effectuée de manière récursive jusqu'à les feuilles de l'arbre contient les fins pour la classification, la division se faite avec des conditions [43] [40].

D. Les réseaux de neurones

Les principes des réseau neuronal artificiel similaire les principes du réseau neuronal biologique. Dans les réseaux neurones les problèmes résoudre d'une manière similaires d'humaine [39]. Le RN est un ensemble de neurones interconnectés. En général, RN a plusieurs couches. Le réseau de neurones est capable d'apprendre en ajustant les poids des neurones [40].

E. Discussion :

La majorité des travaux sur l'analyse des sentiments basée sur la l'apprentissage automatique supervisée, les méthodes et les techniques de cette approche tel que naïve de bayse, SVM, RN (NN) sont des méthodes d'actualité et donne des résultats efficace et utile.

Dans [41]utilise les techniques de naïve de baise et la technique SVM pour l'analyse des sentiments sur Twitter, parmi les problèmes trouve est le problème épineux : les données erronée et les données en argot, pour résoudre ce utilise un vecteur de caractéristiques.

3.2.2 Apprentissage automatique non supervisée :

L'approche d'apprentissage non supervisé utilise des ensembles de données non étiquetés afin de découvrir la structure et de rechercher les modèles similaires à partir des données d'entrée. La méthode non supervisée est habituellement utilisée lorsqu'il le manque de données non étiquetées, mais la collecte de données non étiquetées est plus facile [39].

L'apprentissage non supervisé ou le « clustering » construire des groupes (clusters) de données similaires à partir d'un ensemble hétérogène de données, il y'a plusieurs algorithmes d'apprentissage automatique non supervisée, comme : K-moyennes (KMeans), Fuzzy KMeans, Regroupement hiérarchique [44].

3.3 Approche hybride :

L'approche hybride comprend la méthode d'apprentissage automatique et la méthode basée sur le lexique contenant des règles linguistiques écrites manuellement. Différents classificateurs de sentiments basés sur des méthodes d'apprentissage basées sur le lexique ou machine sont utilisés en cascade de sorte qu'en cas d'échec d'un classificateur, le suivant commence à classer, et ainsi de suite, jusqu'à ce que le document restant soit catégorisé [45].

In [46] proposé un algorithme appelé ASSAY (qui signifie analyse), pour trouver la polarité au niveau du document. Dans l'algorithmes en utilisant des algorithmes des naïves de baise et SVM de l'approche de l'apprentissage automatique pour classifier les avis et les opinions de chaque domaine, puis déterminent la polarité au niveau du document à l'aide de l'algorithme de HARN, qui relève de la méthode du lexique.

In [45] présente une approche hybride de la méthode de classification des sentiments pour les textes coréens. Il est basé sur un système en cascade par lequel la classification basée sur le lexique effectue d'abord la détection des sentiments avec l'analyse locale des constituants des sentiments, et un algorithme d'apprentissage automatique supervisé trie les textes hors du lexique.

3.4 Discussion

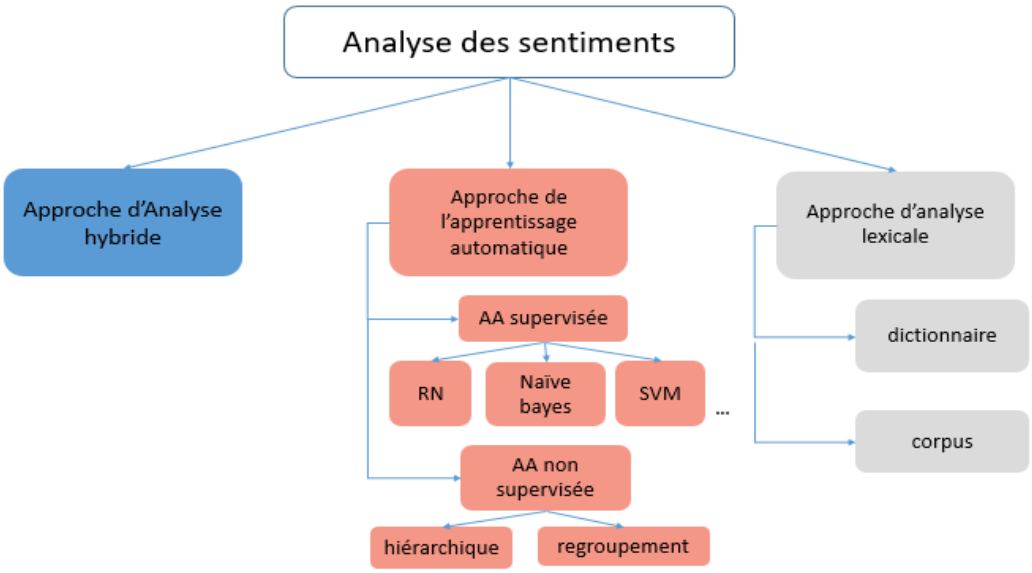


Figure 11 :les approches de analyse des sentiments

L'analyse des sentiments a une grande variété d'applications telles que la classification des critiques, la distinction des synonymes et des antonymes, qui sont utilisées pour une recherche intelligente sur le Web, résumant les critiques, surveillant les opinions au moyen de discussions en ligne et analysant les réponses à une enquête, Il existe trois approches pour l'analyse des sentiments: l'apprentissage lexical et l'apprentissage automatique, la troisième étant combinée aux avantages des deux approches précédentes , d'autres travaux prouvent les performances de la méthode d'apprentissage automatique dans la résolution de problèmes d'analyse de sentiments

4.Détection des évènements :

La détection d'événements a pour le but de détecter des occurrences du monde réel qui se déroulent dans l'espace et dans le temps. En tant que service de réseautage social en ligne et de microblog en croissance rapide, les réseaux sociaux fournissent un contenu sans précédent, généré par les utilisateurs, qui peut être transformé en connaissances exploitables et contextualisées. Plus important encore, les messages et les publications postés - dépassant actuellement plus de Des dizaines de millions de post (commentaire, ...) par jour - pourraient révéler des informations sur les événements du monde réel à mesure qu'ils se déroulent. Toutefois, la détection d'événements à partir de données de réseaux sociaux doit révéler de manière efficace et précise des informations pertinentes sur des événements d'intérêt général ou spécifique, qui sont enfouis dans une grande quantité d'informations banales (par exemple, messages sans signification, pollués et rumeurs).

Dans [47] représentent de nombreuses techniques proposées pour la détection d'événements à partir de données de Twitter. Ces techniques sont classées en fonction du type d'événement (spécifié ou non spécifié), de la tâche de détection (détection d'événements rétrospective ou nouvelle) et de la méthode de détection (supervisée ou non supervisée). La détection d'événements dans les flux Twitter est un domaine de recherche dynamique qui fait appel à des techniques de différents domaines tels que l'apprentissage automatique, le traitement du langage naturel, l'exploration de données, l'extraction de l'informations et l'extraction de texte.

3.1 Détection d'événements selon type d'événement

La détection d'événements à partir de données de Twitter en fonction du type d'événement .
est classées à dectection des evénement (spécifié ou non spécifié)

3.1.1 Détection d'événements spécifié

La détection d'événements spécifiés inclut les événements sociaux connus ou planifiés. Ces événements peuvent être partiellement ou entièrement spécifiés avec le contenu associé ou les informations de métadonnées telles que le lieu, l'heure, le lieu et les artistes interprètes. En d'exploiter le contenu textuel de Twitter ou les informations de métadonnées, ou les deux, en

utilisant un large éventail de techniques d'apprentissage automatique, d'exploration de données et d'analyse de texte pour détecter ces événements.

en [48] ont exploité les tweets pour détecter des types d'événements spécifiques tels que des tremblements de terre et des typhons. Ils ont formulé la détection d'événements comme un problème de classification et utilise un technique de l'apprentissage automatique, formé un SVM à un ensemble de données Twitter étiquetées manuellement comprenant des événements positifs (tremblements de terre et typhons) et des événements négatifs (autres événements ou événements non événementiels, événement contient des bruit).

3.1.2 Détection d'événements non spécifié

Les réseaux sociaux comme twitter sont particulièrement utiles pour la détection d'événements inconnus. Les incompréhensions d'intérêt sont généralement liées aux événements émergents, aux dernières nouvelles et à des sujets généraux qui attirent l'attention d'un grand nombre d'utilisateurs de Twitter. Comme aucune information sur les événements n'est disponible, les événements inconnus sont généralement détectés en exploitant les modèles ou le signal des flux Twitter. Les nouveaux événements d'intérêt général présentent une multitude de fonctionnalités dans les flux Twitter générant, par exemple, une utilisation accrue et soudaine de mots clés spécifiques. Les fonctionnalités en rafale qui apparaissent fréquemment dans les tweets peuvent ensuite être regroupées en tendances [47].

[49]Ont proposé TwitterStand, un système de traitement de nouvelles destiné à Twitter, destiné à capturer les tweets liés aux dernières nouvelles qui prennent en compte à la fois la similarité textuelle et la proximité temporelle. Ils ont utilisé un classificateur naïf de Bayes pour séparer les nouvelles des informations non pertinentes.

3.3. Détection d'événements selon la méthode détection

La détection d'événements dans les flux Twitter s'appuie sur des techniques de différents domaines, généralement l'apprentissage automatique et l'exploration de données, le traitement du langage naturel, l'extraction d'informations, l'extraction de texte et la récupération d'informations. Les techniques de détection d'événements sont classées en apprentissage supervisé et non supervisé.

3.3.1. Approches de détection non supervisées.

La plupart des techniques de détection d'événements non spécifiées dans les flux Twitter reposent sur des approches de regroupement (clustring). Les approches de clustring conviennent naturellement aux NED (new event detection) non spécifiés de Twitter, car elles ne nécessitent pas de données étiquetées pour l'entrainement [47].

3.3.2. Approches de détection supervisées

La approches NED pour des événements non spécifiés nécessite une classification non supervisée de nouveaux tweets, les techniques NED qui se concentrent sur la détection d'un type d'événement spécifique reposent principalement sur des approches d'apprentissage supervisées. Bien qu'étiqueter manuellement un grand nombre de messages Twitter demande beaucoup de temps et d'intensité, il est plus faisable pour des événements spécifiés que pour des événements spécifiés à l'avance. Lorsque certaines descriptions d'événements sont connues, des techniques de filtrage pourraient être utilisées pour réduire le nombre de messages non pertinents et faciliter la tâche d'un expert humain pour annoter un ensemble de données d'une taille « raisonnable ». En outre, le filtrage en fonction de descriptions d'événements spécifiques, tels que les mots clés, l'emplacement ou l'heure, permettrait également de réduire le nombre de messages Twitter devant être traités pendant le fonctionnement du système et de permettre à l'algorithme de détection de se concentrer sur un ensemble restreint de tweets [47].

4.Conclusion

- La diffusion de l'information c'est la circulation d'information d'un individu ou d'une communauté à un autre dans le réseau. Nombreuses recherches ont été consacrées à l'analyse de la diffusion de l'information.
- Les modèles de diffusion peuvent être classé en deux catégories des modèles explicatifs et des modèles prédictifs.
- L'analyse des sentiments a une grande variété d'applications telles que la classification des critiques, la distinction des synonymes et des antonymes, qui sont utilisées pour une recherche intelligente sur le Web, résumant les critiques, surveillant les opinions au moyen de discussions en ligne et analysant les réponses à une enquête, Il existe trois approches pour l'analyse des sentiments: l'apprentissage lexical et l'apprentissage automatique, la troisième étant combinée aux avantages des deux approches précédentes , d'autres travaux prouvent les performances de la méthode d'apprentissage automatique dans la résolution de problèmes d'analyse de sentiments
- La détection d'événements a pour le but de détecter des occurrences du monde réel qui se déroulent dans l'espace et dans le temps.
- nombreuses techniques proposées pour la détection d'événements. Ces techniques sont classées en fonction du type d'événement (spécifié ou non spécifié), de la tâche de détection (détection d'événements rétrospective ou nouvelle) et de la méthode de détection (supervisée ou non supervisée).

Chapitre III :

Un processus d'analyse & optimisation des données pour la diffusion de l'informations sur les Reseaux Sociaux

1. Introduction

Les réseaux sociaux sont désormais la principale source d'information pour nous maintenant, car chacun partage ses opinions, ses sentiments, ses problèmes et ses intérêts en partageant ses publications, tweets, vidéos, mots et photos de sa propre vie.

Les données partager sont des données massives avoir beaucoup de propriétés hétérogène variabilité ... la recherche dans cette plage de données un travail très lourd dans plusieurs aspects l'espace de recherche, le temps à cause de la nature des données diffuse.

L'objectif de notre travail est chercher une méthode aide nous minimiser l'espace de recherche et mettre en échelle et nous essayons de faire ça dans temps acceptable (dans temps réel) a causé vivacité de notre source de données (les réseaux sociaux).

C'est pour ça nous proposons un processus pour réduire l'espace de recherche le plutôt possible. Ce processus repose à deux filtre, le premier filtre pour le but de l'extraction des données selon un domaine dans la quantité énorme des données fournit. On applique un deuxième filtre sur le résultat de la premiers pour l'extraction des selon le contexte.

2. Notre contribution

La recherche des informations dans grand espace une tâche très difficile et lourd, surtout avec les méthodes traditionnelles et avec une source de données vive comme les réseaux sociaux (Facebook, Twitter, ...). Notre processus a le but faciliter ces recherches et basé sur la réduire et diminution de l'espace de recherche, la chose qui rendre le temps de réponse de requête un temps accessible.

Ce processus constitue sur plusieurs phases sont la détection des domaine, filtrage par domaine, filtrage par contexte, et la dernière phase c'est d'optimisation c'est la phase ou requête exécute sur données filtre, ce processus est présent dans la figure dessous.

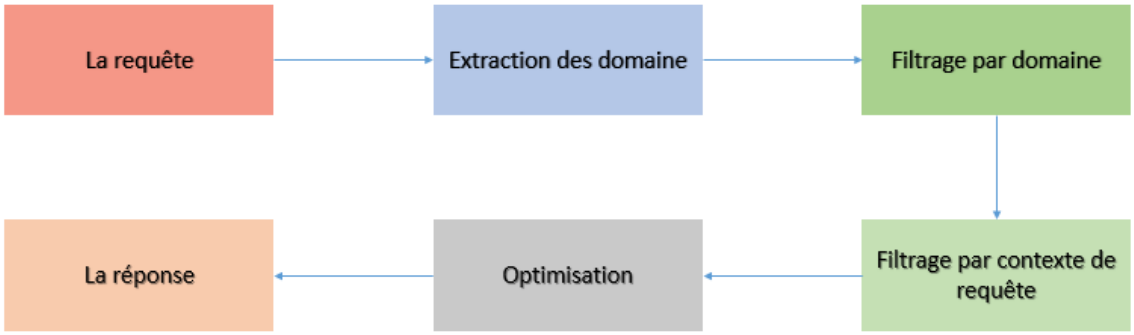


Figure 1: processus d'analyse & optimisation des données pour la diffusion de l'information sur les RSs

2.1 Extraction des domaines

Le but de cette phase est extraire le domaine qui la requête est entouré. Pour extraire le domaine, nous appliquons quelque méthodes et fonction pour traiter la requête et connu le domaine

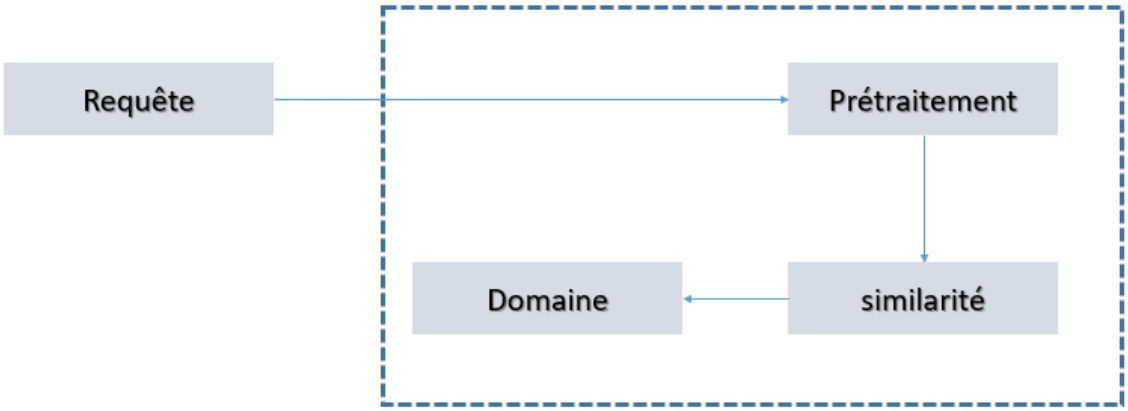


Figure 2: étape de l'extraction du domaine

➢ 2.1.1 Prétraitement :

Prétraitement (preprocessing) est une étape très importante qui consiste en plusieurs techniques visant à traiter les données pour les structurer et faciliter leurs utilisations. Elles convertissent les données textuelles originales dans une structure d'exploration de données prêtes.

But de prétraitement : Le prétraitement des données est le processus de nettoyage et de préparation du texte pour des opérations après.

Les textes en ligne contiennent notamment beaucoup de bruit et des parties inutile telles que des balises HTML. De plus, au niveau des mots, de nombreux mots dans le texte n'ont pas d'impact sur l'orientation générale de celui-ci. Le prétraitement des données : réduire le bruit dans le texte aider à améliorer les performances et accélérer le processus de traitement après, aidant ainsi à l'analyse en temps réel.

Dans la phase de prétraitement contient plusieurs fonctions sont applique, un pour le nettoyage, et autre pour éliminer les mots d'arrête (Stop Words) après une opération de segmentation, …ex.

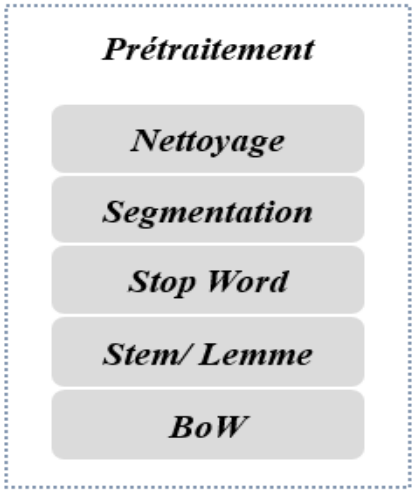


Figure 3: les prétraitements

- **Nettoyage :** c'est une opération pour éliminer les données inutile et supplémentaire dans les données d'entrée, parmi ces données, et corrige d'autre mots comme les mots contient des lettres répétées (les erreurs de frappe), tout ça après un passage au minuscule.

- **Segmentation (Tokenisation) :** Il s'agit de décomposer une phrase, et donc un document, en segment (Token). Un segment (token) est un élément correspondant à un mot.

  Exemple :

  Le traitement automatique du langage naturel vise à créer des outils de traitement de la langue naturelle pour diverses applications.

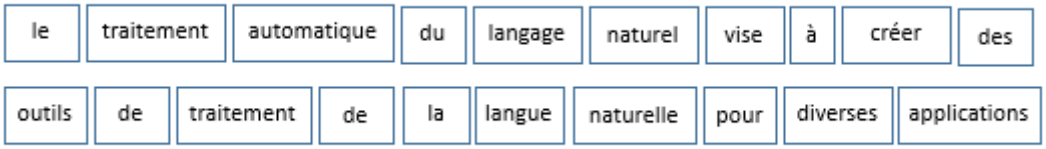


Figure 4: La segmentation

- **Enlever les mots vides (StopWords) :** on supprime les mots appartenant aux StopWords. Il s'agit de listes de mots définies au préalable soit par l'utilisateur soit dans des librairies existantes. Ces listes se composent de mots qui n'apportent aucune information, qui sont en général très courants, par exemple : je, nous, ...ex. La suppression de ces StopWords permet de ne pas polluer les représentations des documents afin qu'elle ne contienne que les mots représentatifs et significatifs.

Figure 5:retire les StopWords

Stem/ Lemme : cette étape contient deux parties : La stemmatisation, La Lemmatisation. Les termes restant notamment écrit au pluriel, au singulier ou avec différents accords et les verbes peuvent être conjugués à différents temps et personnes. C'est pour ça on applique l'un de deux phase (stem ou lemme). Stemmatisation : réduit les mots à leur radical ou racine. Le résultat n'est pas forcément un mot existant.

Lemmatisation : qui prend en considération le contexte dans lequel le mot est écrit, a pour but de trouver la forme canonique du mot, le lemme. Le lemme correspond à l'infinitif des verbes et à la forme au masculin singulier des noms, adjectifs et articles.

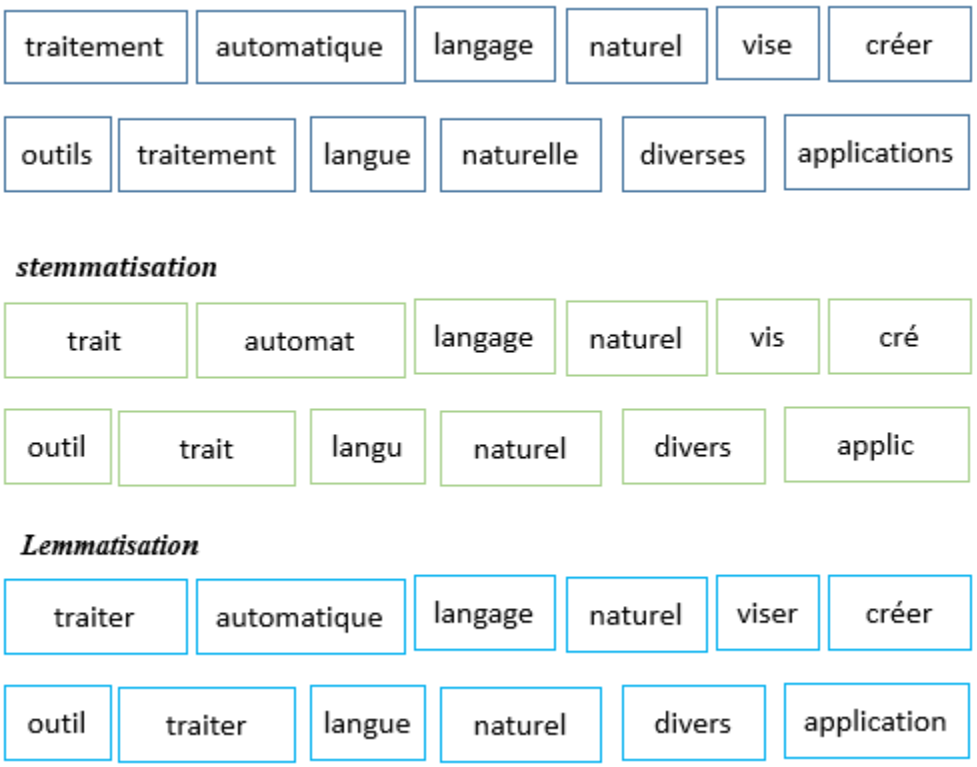


Figure 6 : Stemmatisation / Lemmatisation

- **Sac de mots (Bag of Words : BoW)** : nous voulons représenter un document par un vecteur.

  À cette fin, chaque composant du vecteur de document est associé à un mot du dictionnaire de corpus. Un composant contient donc une valeur pour chacun des mots existant dans tous les textes que nous traitons. Cette valeur peut être, par exemple, le nombre d'occurrences du mot dans le document. Si un mot n'est pas présent, il recevra la valeur 0. C'est une approche appelée sac de mot.

2.1.2 Similarité

Dans cette étape on calculer la pourcentage similarité de la requête aves les autres domaines et décide le domaine de requête à partir ces pourcentages (le pourcentage maximum : c'est le domaine de requête).

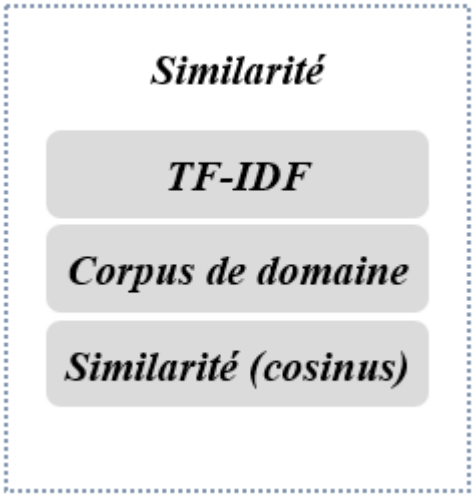


Figure 7: Similarité

- **TF-IDF : (terme frequency / inverse terme frequency)**

Cet acronyme anglais correspond à un poids calculé et affecté pour chaque mot de chaque document du corpus.

Tf : La fréquence d'apparition d'un mot dans un document.

df : Le nombre de document dans lequel le mot apparaît une fois ou plus

Idf : $\text{idf}_i = \log\left(\frac{D}{d_i}\right)$ /

D : le nombre total des documents (domaines) ;

d_i : le nombre des documents contient le terme ;

$$\text{tf–Idf} = \text{tf} * \text{Idf}$$

La combinaison de ces deux indicateurs donne le TF-IDF. Ce score présente l'importance d'un mot dans un document et prend en compte sa rareté dans l'ensemble du corpus. Les termes les moins présents dans le corpus ont donc un poids plus important car ils sont plus discriminants. On peut donc utiliser ce score comme valeur des vecteurs représentant nos documents (domaines).

- **Corpus des domaines :** sont des corpus représente les domaines chaque vecteur représente un domaine spécifique (chaque vecteur est passé à l'étape de prétraitement).

- **Calcule la Similarité :** il y'a plusieurs méthodes pour calculée la similarité comme la différence euclidienne, la distance, et la méthode célèbre et la similarité avec cosinus.
  La similarité cosinus est fréquemment utilisée [Baeza-Yates and Ribeiro-Neto, 1999] en tant que mesure de ressemblance entre deux documents d1 et d2. Il s'agit de calculer le cosinus de l'angle entre les représentations vectorielles des documents à comparer [50].

$$\mathbf{sim_{Cosinus}}\ (\mathbf{d1}, \mathbf{d2}) = \frac{\vec{\mathbf{d}}_1 . \vec{\mathbf{d}}_2}{||\vec{\mathbf{d}}_1|| . ||\vec{\mathbf{d}}_2||}$$

  Dans cette étape on va calcule la similarité de vecteur de requête avec tous les vecteurs du domaine existe, et le vecteur a la meilleure similarité est le vecteur de domaine correspondant à la requête.

2.2 Filtrage selon le domaine

Dans cette phase on filtre les données d'un domaine spécifique à partir une quantité énorme d'information.

Ce filtrage constitue à partir plusieurs étapes comme l'acquisition et la collection des données à partir de différente réseaux sociaux et une étape de prétraitement pour transformer ces données à une forme exploitable à la fin de cette phase on applique les technique d'apprentissage automatique et apprentissage approfondi pour classifier les données acquise/collectée.

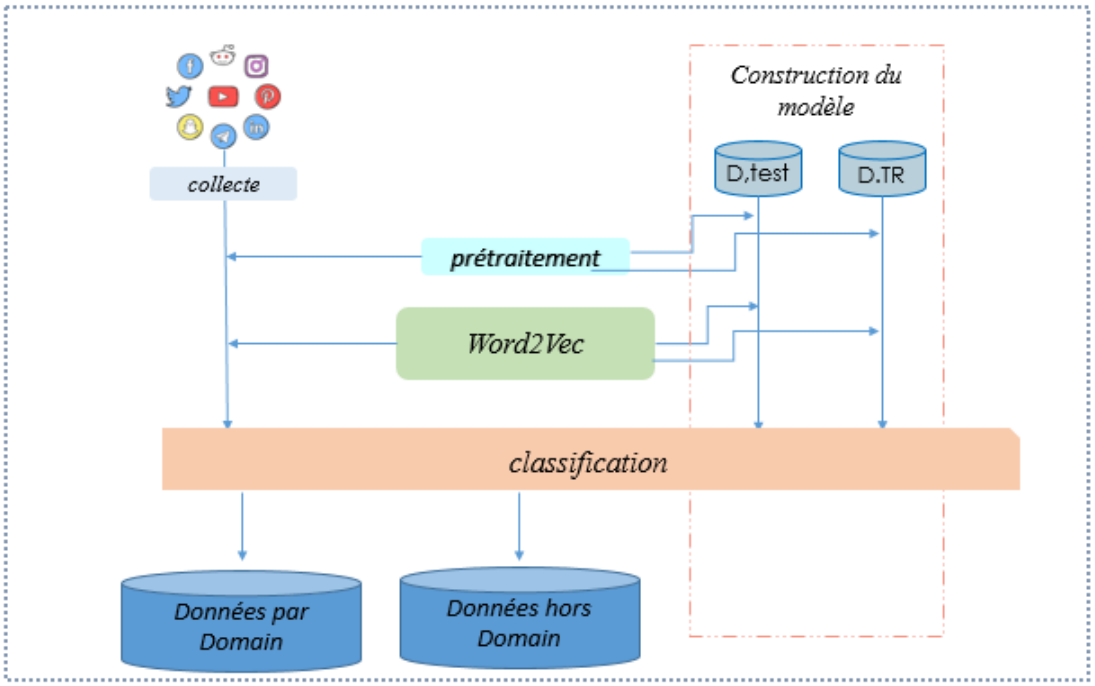


Figure 8:filtre selon le domaine

➢ 2.2.1 Collecte :

Les données sont des pièces nécessaires pour des informations utiles. La collection de données est nécessaire pour des individus tels que des chercheurs. La collecte de données est une étape importante pour toute recherche ou expérience. La collecte de données peut être définie comme le processus de collecte et de traitement de l'information pour évaluer les résultats et les utiliser pour les recherches [51]. Les réseaux sociaux sont l'une des meilleures et grande sources de données maintenant.

Il existe différents sites de réseaux sociaux sur lesquels nous pouvons collecter des données telles que Twitter, Facebook, …ex. En général, pour obtenir les données, vous pouvez procéder de différentes manières, par exemple en contactant l'administrateur du site pour obtenir le jeu de données, en téléchargeant le jeu de données créé à des fins académiques ou autres.

Il existe des APIs (Application Programmation interface) associés à chaque site social pour aider le collecteur de données à demander les services de ceux des sites. Les procédures communes pour ces sites installent bibliothèques correspondantes, en obtenant l'autorisation et puis de décider de la plate-forme sur laquelle le collecteur peut utiliser pour écrire le code [51].

- **Facebook** est un site de réseau social qui fournit certains services tels que l'état de publication, la publication d'images et la possibilité de se faire des amis. C'est une excellente source de données volumineuses. Il existe différentes techniques pour collecter les données. Par exemple, il offre Graph API Explorer, un excellent outil pour générer des données à l'aide de l'API Graph. Selon [52] la page développeur Facebook...

- **Twitter** est un service de micoblogging qui permet aux utilisateurs de poster des messages. Twitter4j est une bibliothèque java pour les API Twitter. Lors de la création d'une application à l'aide de l'API Twitter, le développeur obtient OAuth qui consiste en une clé de consommateur, un secret de consommateur, un jeton(Token) d'accès et un secret de jeton(Token) d'accès. Celles-ci sont censées être utilisées pour autoriser le développeur lors de la collecte de données sur Twitter. Le développeur doit avoir un compte sur Twitter pour pouvoir utiliser la bibliothèque [51]. Il y a plus de bibliothèques pour toutes les plateformes tels que tmhOAuth (PHP), tweepy (python)... [53]

2.2.2 Prétraitement

En raison de la nature variable et imprévisible du langage utilisé dans les données des réseaux sociaux, il est probable que des techniques de prétraitement pourraient être utilisées pour normaliser certains jetons(token) de données. Il est fort probable que la plupart des données contiennent une forme ou une autre de fautes de grammaire ou d'orthographe, un acronyme, des expressions familières et des argots.

Le prétraitement est un passage de données ambiguës à des données propres, pour obtenir ces données en applique quelque étapes comme la suppression les liens URL, les espaces vide successives, les symboles, ponctuations, les mots vides (stop words) … ex cette phase similaire à l'étapes de prétraitement

2.2.3 Word2Vec

Dans cette étape en va représenter les données en des vecteurs, Ces vecteurs capturent des informations cachées sur une langue, telles que les analogies de mots ou la sémantique. Pour meilleur performance et enrichir le côté sémantique, parmi les techniques utilisées le Word Embedding.

Word Embedding

Word Embedding : est une approches permettant de représenter des mots et des documents à l'aide d'une représentation vectorielle, et d'autre part, de prendre en compte la notion de contexte, facilitant l'analyse sémantique et syntaxique [54].

Les méthodes bien connues de production de modèles Word Embedding sont Word2Vec et Globe. Ces méthodes ont beaucoup attiré l'attention et ce sont les plus efficaces pour apprendre les représentations vectorielles des mots [55].

- **Word2Vec** : Mikolov et l'équipe de chercheurs ont développé la technique en 2013 chez Google. Leur approche a été publiée pour la première fois dans l'article intitulé «'Efficient Estimation of Word Representations in Vector Space» [56]. Ils ont affiné leurs modèles pour améliorer la qualité de la représentation et la vitesse de calcul. Ce travail a été publié dans l'article intitulé « Distributed Representations of Words and Phrases and their Compositionality » [57]. Word2Vec est un moyen efficace de représenter des mots en vecteurs. Il existe deux types de Word2Vec, Skip-gram et sac de mots continu (CBOW : Continuous Bag of Words).
    - **Skip-gram :** prédire le contexte d'un mot cible particulier, les contextes sont des voisins directs de la cible [55].
    - **CBOW :** est très similaire au Skip-gram, sauf qu'il change les entrées et les sorties. L'idée est que, compte tenu du contexte, nous voulons savoir quel mot est susceptible d'apparaitre [55].

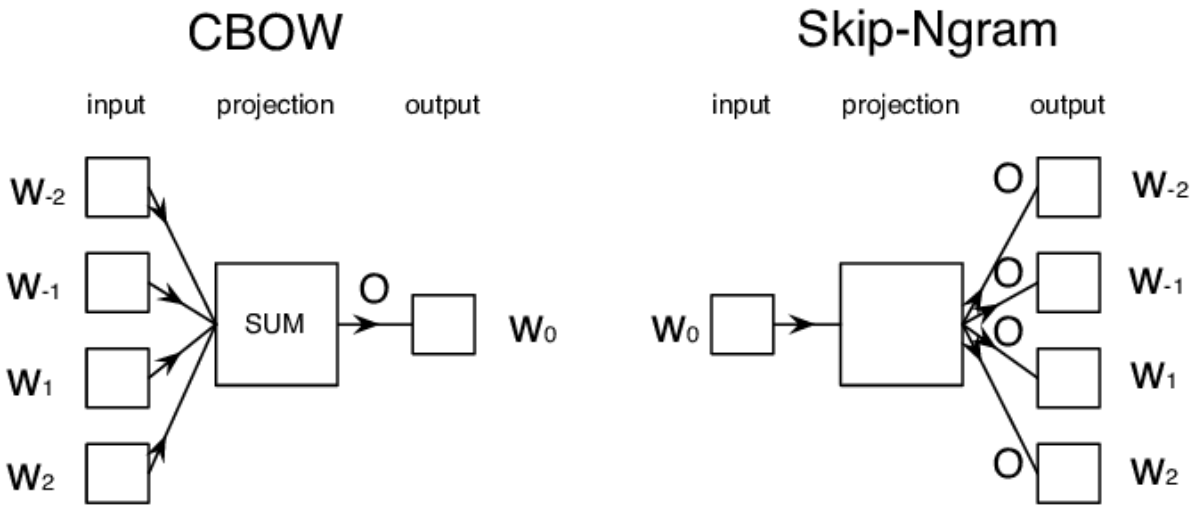


Figure 9:Skip-gram &CBOW [55].

- **Glove (Global Vectors) :** est l'une des méthodes les plus connues d'apprentissage des représentations de mots proposée par Pennington et al. À Stanford [58]. Pour générer des mots incorporés en agrégeant une matrice globale de cooccurrence mot-mot à partir d'un corpus [59].

- Il y'a d'autre méthode comme FastText (FastText est une extension de Word2Vec proposée par Facebook en 2016. Au lieu d'insérer des mots individuels dans le réseau de neurones, FastText divise les mots en plusieurs n-grammes (sous-mots)) [60].

2.2.3 Classification :

Quand il s'agit d'analyser et de classer d'énormes quantités de données texte, la tâche est trop lourde pour être effectuée manuellement. C'est également fastidieux, chronophage et donc coûteux, et trier manuellement de grandes quantités de données risque davantage de générer des erreurs et des incohérences.

Les modèles de l'analyse des sujets (Topic analysis models.) et la classification vous permettent de parcourir de grands ensembles de données et d'identifier les sujets de manière très simple, rapide et évolutive.

Les techniques d'apprentissage automatique permettent d'organiser et de comprendre de grandes collections de données textuelles en attribuant des étiquettes ou des catégories en fonction du sujet ou du thème de chaque texte. En d'autres termes, ils permettent de trouver des modèles et de déverrouiller des structures sémantiques derrière chacun des textes individuels.

Il existe de nombreuses techniques et des algorithmes que vous pouvez utiliser pour analyser automatiquement les sujets d'un ensemble de de données (documents, tweets...). Il y'a des algorithmes non supervisés (ce qui signifie que vous leur alimentez les textes et les paramètres d'apprentissage et qu'ils font le reste), et les algorithmes supervisés (Les machines reçoivent des exemples de données étiquetées en fonction de leurs sujets afin d'apprendre à baliser le texte individuellement, par sujet).

i. Les techniques d'apprentissage automatique /approfondi

Parmi les algorithmes d'apprentissage automatique pouvant être utilisés pour la classification.

Naïve Bayes est une famille d'algorithmes simples qui donnent généralement d'excellents résultats pour de petites quantités de données et des ressources de calcul limitées. Le membre le plus populaire de la famille est probablement le Multinomial Naïve Bayes (MNB).

Support Vector Machines (SVM) L'idée de base pour SVM est : une fois que tous les textes sont vectorisés (donc ils sont points dans l'espace mathématique), pour trouver la meilleure ligne (dans un espace de dimension grande appelé hyperplan) séparant ces vecteurs dans les sujets souhaités. Puis, quand un nouveau texte arrive, vectorisée et regardez de quel côté de la ligne cela se termine : c'est le sujet de sortie.

L'apprentissage en profondeur (Deep Learning) est en fait un terme fourre-tout pour une famille d'algorithmes vaguement inspirés par le fonctionnement des neurones humains. Bien que les idées sous-jacentes aux réseaux neuronaux soient assez anciennes (elles remontent aux années 50), ces algorithmes ont connu une forte reprise ces dernières années grâce à la baisse des coûts informatiques, à l'augmentation de la puissance de calcul et à la grande disponibilité des données.

L'apprentissage en profondeur offre d'excellents résultats en échange de certaines exigences informatiques draconiennes. Ce n'est pas inhabituel pour que les modèles d'apprentissage en profondeur s'entraînent pendant des jours, des semaines.

Pour la classification des sujets, les deux principales architectures d'apprentissage en profondeur utilisées sont les réseaux de neurones convolutionnels (CNN) et les réseaux de neurones récurrents (RNN).

Les algorithmes d'apprentissage en profondeur nécessitent beaucoup plus de données d'apprentissage que les algorithmes traditionnels d'apprentissage automatique, dans le royaume de millions d'exemples étiquettes. Cependant, il est important de noter que les algorithmes traditionnels d'apprentissage automatique tels que SVM et NB atteignent un certain seuil, après quoi l'ajout de plus de données d'entraînement ne permet plus d'améliorer la précision. En revanche, les classificateurs d'apprentissage en profondeur continuent de s'améliorer avec l'augmentation que les données qu'ils disposent

.

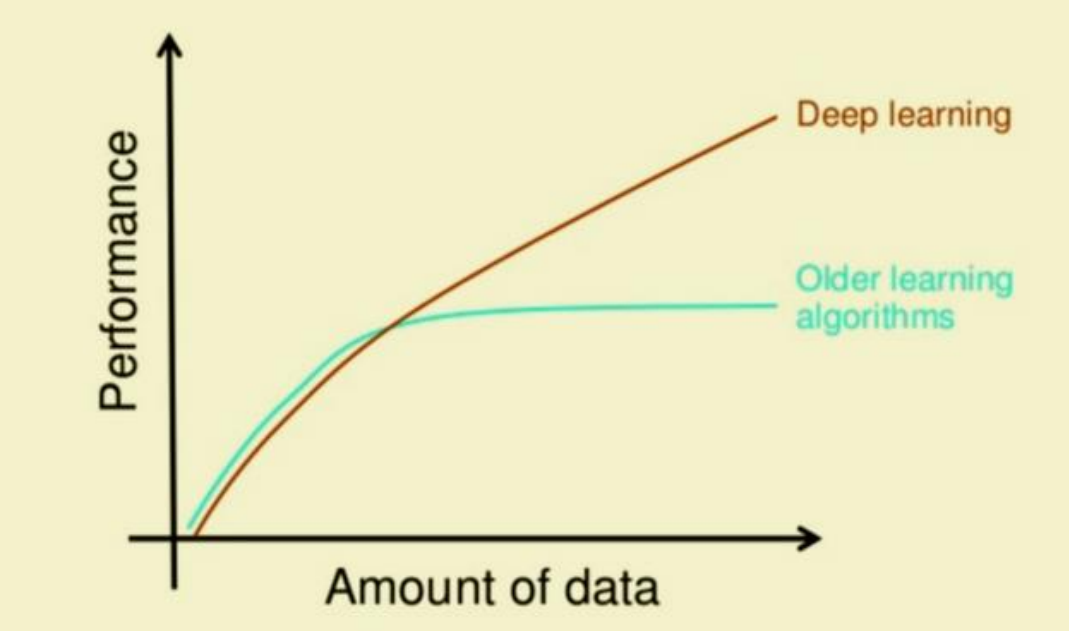


Figure 10:apprentissage approfondi & apprentissage automatique [*]

[*]https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063

L'apprentissage en profondeur fournit des modèles informatiques efficaces utilisant des combinaisons d'éléments de traitement non linéaires organisés en couches. Cette organisation d'éléments simples permet au réseau total de généraliser (c'est-à-dire de prédire correctement sur de nouvelles données).

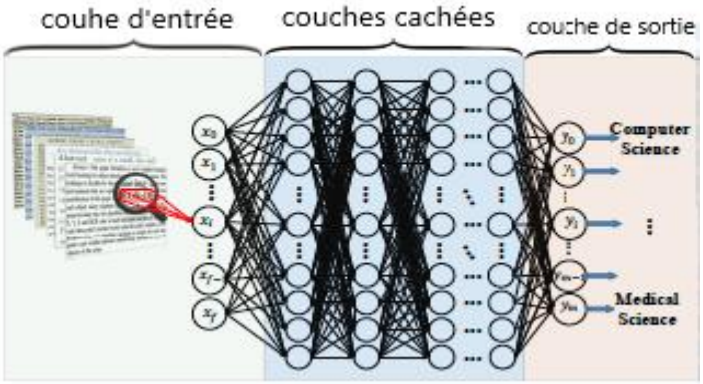


Figure 11: Architecture de l'apprentissage approfondi[**]

[**]https://medium.com/text-classification-algorithms/text-classification-algorithms-a-survey

Les trois architectures d'apprentissage approfondi que nous avons utilisées : réseaux de neurones profonds (DNN), réseaux de neurones récurrents (RNN) et réseaux de neurones convolutionnels (CNN).

ii. Réseaux de neurones profonds (Deep Neural Network, DNN)

Dans l'architecture DNN, chaque couche reçoit uniquement les entrées de la couche précédente et est transmise à la couche suivante. Les couches sont entièrement connectées. La couche d'entrée comprend les entités de texte et la couche de sortie comporte un nœud pour chaque étiquette de classification ou un seul nœud s'il s'agit d'une classification binaire. Cette architecture est le DNN de base [61].

Réseaux de neurones récurrents (Recurrent Neural Network, RNN)

Est un type de réseau de neurones dans lequel les sorties de l'étape précédente sont alimentées en entrée de l'étape en cours. Dans les réseaux de neurones traditionnels, toutes les entrées et toutes les sorties sont indépendantes les unes des autres, mais dans des cas comme lorsqu'il est nécessaire de prédire le mot suivant d'une phrase, les mots précédents sont nécessaires et il est donc nécessaire de se rappeler les mots précédents. C'est ainsi qu'est né RNN, qui a résolu ce problème à l'aide d'une couche cachée. La caractéristique principale et la plus importante de RNN est l'état caché, qui garde en mémoire certaines informations sur une séquence [62].

Réseaux de neurones convolutionnels (Convolutional Neural Networks, CNN)

Construits à l'origine pour le traitement d'images, les CNN ont également été utilisés efficacement pour la classification de texte. La couche convolutive de base d'un réseau CNN se connecte à un petit sous-ensemble d'entrées. De la même manière, la couche convolutionnelle suivante ne se connecte qu'à un sous-ensemble de la couche précédente. De cette manière, ces couches de convolution, appelées cartes de caractéristiques, peuvent être empilées pour fournir plusieurs filtres sur l'entrée. Pour réduire la complexité des calculs, les CNN utilisent le regroupement pour réduire la taille de la sortie d'une pile de couches à la suivante dans le réseau. Différentes techniques de mise en commun sont utilisées pour réduire les sorties tout en préservant les caractéristiques importantes [61].

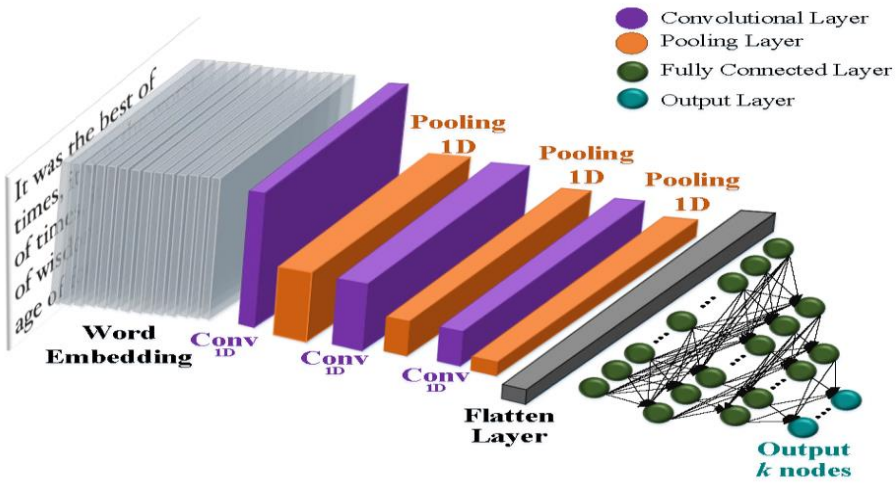


Figure 12:Convolutional Neural Networks (CNN)[***]

*[***]*https://medium.com/text-classification-algorithms/text-classification-algorithms-a-survey

2.3 Filtre selon le contexte

Après d'avoir les données selon le domaine, ensuite une phase similaire pour extraire des données selon le contexte.

Dans la tâche de classification, un contexte formé à partir le profil de l'utilisateur pour la classification. Dans ce phase l'utilisation d'apprentissage approfondi et les technique d'apprentissage automatique elle plus performent par rapport à les méthodes traditionnelle.

2.3.1 Profil utilisateur

Le profil d'utilisateur est modèle d'un ensemble d'informations essentielles décrivant l'utilisateur [63] [64].

2.3.2 Contenu d'un profil utilisateur

Parmi les informations qui se trouve dans le profil utilisateur

Intérêts de l'utilisateur : Les intérêts de l'utilisateur peuvent être classés en deux types : les intérêts à court terme et à long terme Le profil est à long terme si l'utilisateur est toujours intéressé par le sujet (par exemple, l'utilisateur est intéressé par l'informatique). Il est à court terme si l'utilisateur est intéressé par le sujet durant une

période limitée (par exemple, l'utilisateur est intéressé parle football pendant une coupe du monde) [64].

Contexte. Il existe plusieurs types de contexte selon le domaine d'application. Les différents types incluent les contextes environnementaux, les contextes personnels, les contextes sociaux et les contextes spatio-temporels. Le contexte environnemental capture les entités situées à proximité d'un utilisateur telles que les objets, la température, les personnes, la lumière, etc. Le contexte personnel comprend le contexte physiologique (comme le poids, la couleur des cheveux, etc.) et le contexte mental (comme l'humeur, le niveau d'angoisse ou de stress, etc.). Le contexte social peut contenir des informations telles que les amis, les voisins, les collègues, etc., des informations qui décrivent les aspects sociaux de l'utilisateur. Enfin, le contexte spatio-temporel est une combinaison des attributs suivants : temps, emplacement, ou direction [64].

2.3.4 Les Modèles de représentation d'un profil

i. ***Représentation sous forme de vecteurs*** : Il s'agit de l'une des représentations de profil la plus utilisée. Le vecteur profil correspond à un ensemble de caractéristiques avec pour chacune son poids (la valeur de la coordonnée pour une dimension). La façon de calculer le poids varie selon l'application [63] [64].

ii. ***Représentation sous forme d'ontologies.*** Il permet d'avoir une représentation plus sémantique en associant des liens entre les termes ou les items du profil de l'utilisateur [63] [64]

Il y'a d'autre représentation comme la représentation par graphe et représentation hiérarchique

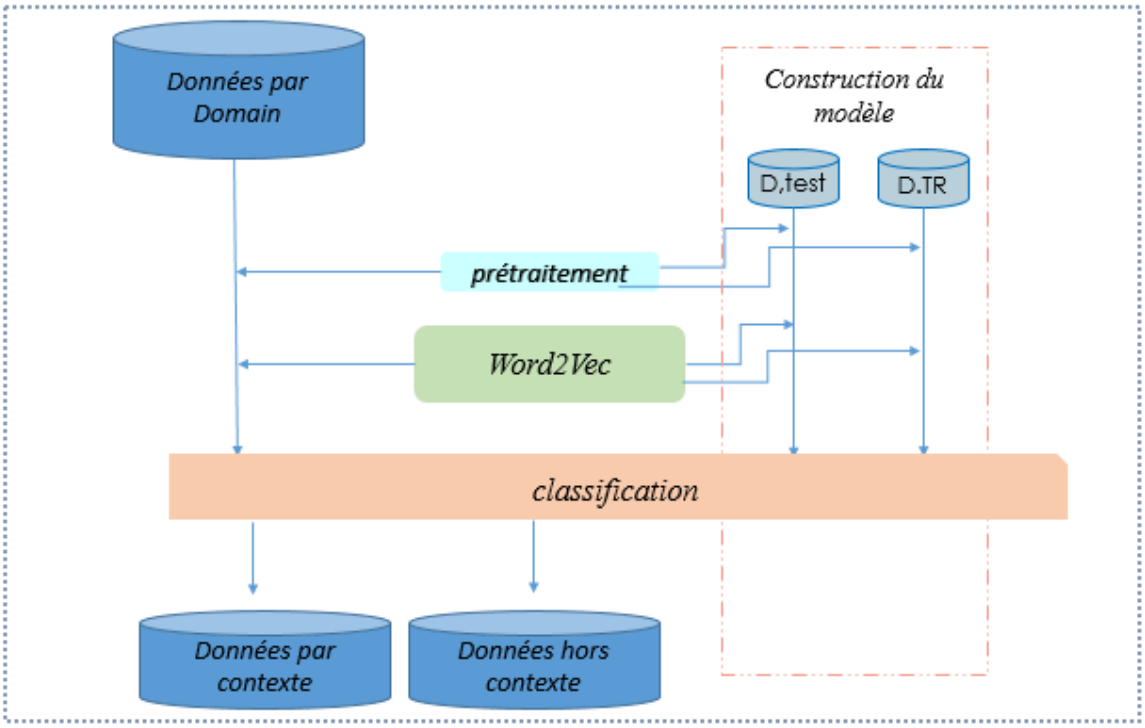


Figure 13:filtre selon le contexte

2.4 L'Optimisation

Dans cette phase on va exécute la requête sur la base de données générée après les deux filtres. Cette base de données sur la mesure de l'utilisateur et la requête étant donnée. Exécution de cette requête sur base de données générée plus rapide, plus fiable et la durée d'exécution moins Parce que la quantité d'informations n'est pas aussi grande qu'au début.

3. Conclusion

Les réseaux sociaux comme source d'information importante dans notre époque, c'est le premier fournisseur de données pour les BIG DATA. L'analyse et contrôle de la diffusion de l'information sur les réseaux sociaux est un grand défi pour les chercheurs, toujours chercher comment utilisée les bigdata fournit par les réseaux sociaux sans oublier les facteurs principaux de défi, le temps et l'espace de stockage.

Notre objective est réduire l'espace de recherche (implique l'espace de stockage) et traiter les informations dans temps réel (temps acceptable).

Notre proposition filtre les big data (le fournisseur de ces données est le réseaux sociaux) selon un domaine et un autre filtrage selon le contexte. Les deux filtrages donnent une données utile et pertinent à notre requête et Personnalisé à l'utilisateur.

Figure 14:filtrage selon Domaine &Contexte

Ce filtrage à l'aide des techniques de traitement de langage naturelle pour rendre les données dans forme exploitable et les technique d'apprentissage automatique et l'apprentissage approfondi et en utilisant le profil d'utilisateur pour personnaliser les données obtenues. Ce processus répondre au besoin de l'utilisateur avec respect de les facteurs principale le temps et le stockage. Ce processus aide nous a la réduction de l'espace de stockage et le temps de réponse et exécution. Le schéma détaillé du processus dans la figure dessous.

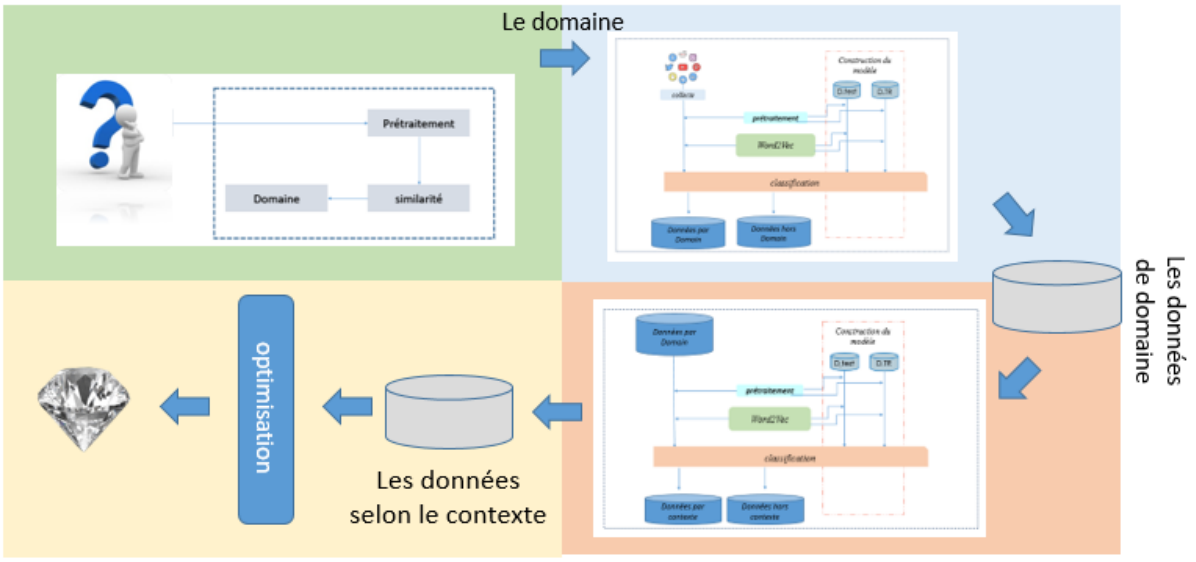


Figure 15:processus d'analyse & optimisation de données pour la diffusion de l'l'information sur les RSs détaille

Chapitre IV

Un outil pour l'optimisation des grandes source de données

(filtre basé domaine & contexte)

1 Introduction

Après nous être assuré de notre idée finale du processus proposée dans le dernier chapitre, ne pouvons d'implémente tout le processus à cause du manque des ressources matérielles et logicielles, et la nécessité d'un grosse dataset pour entrainer les modèles utilise (les modèles d'apprentissage approfondi)dans toutes le processus. Un essai d'implémentation a été réalisé pour les deux phases premiers de notre processus,en utilisant les model de l'apprentissage automatique. En utilisant le réseaux sociaux Twitter pour extraire les données et en applique un parmi trois modèles propose pour classifier les tweets et extraire les tweet d'un domaine donnez. Avant de détailler nos implémentations, nous allons d'abord parler des outils d'implémentation fournis, nous présenterons l'environnement d'implémentation, puis nous développerons une implémentation de notre méthode et nous terminerons notre chapitre par une conclusion.

2 Les outils d'implémentation

Dans notre implémentation utilise le twitter comme source de données et en utilisant le langage de programmation python à l'aide de Framework Anaconda(Spyder) pour programmer notre processus.

2.1 Le langage Python

Python est un langage de programmation puissant et facile à apprendre. Il dispose de structures de données de haut niveau et permet une approche simple mais efficace de la programmation orientée objet. Parce que sa syntaxe est élégante, que son typage est dynamique et qu'il est interprété, Python est un langage idéal pour l'écriture de scripts et le développement rapide d'applications dans de nombreux domaines et sur la plupart des plateformes.

L'interpréteur Python et sa vaste bibliothèque standard sont disponibles librement, sous forme de sources, pour toutes les plateformes majeures depuis le site Internet https://www.python.org/ et peuvent être librement redistribués[65].

Figure 1: logo python[65]

Parmi les bibliothèques utilise dans notre implémentation

Package Nltk : Nltk (Natural Language Toolkit) est pour la création de programmes Python pour travailler avec des données de langage humain.

Package re : (Regular expressions) Ce module fournit des opérations correspondant aux expressions régulières.

Package Sklearn : est un module en Python pour l'apprentissage automatique.

Tweepy : Une bibliothèque Python facile à utiliser pour accéder à l'API Twitter

Tkinter : tkinter est la méthode la plus couramment utilisée pour développer une interface utilisateur graphique (GUI). Python avec tkinter fournit le moyen le plus rapide et le plus simple de créer les applications à interface graphique. La création d'une interface graphique à l'aide de tkinter est une tâche facile.

2.2 Anaconda

Anaconda Distribution Open Source est le moyen le plus simple pour programmer avec Python / R et l'apprentissage automatique sur Linux, Windows et Mac OS X. Avec plus de 11 millions d'utilisateurs dans le monde entier, il s'agit du standard de l'industrie pour le développement, les tests, et l'entrainement sur une seule machine[66].

Figure 2 : anaconda logo

2.3 Spyder

Spyder est un environnement scientifique puissant écrit en Python, pour Python, et conçu par et pour les scientifiques, les ingénieurs et les analystes de données. Il offre une combinaison unique des fonctionnalités avancées d'édition, d'analyse, de débogage et de profilage d'un outil de développement complet avec des capacités d'exploration de données, d'exécution interactive, d'inspection approfondie et de visualisation[67].

Figure 3:Spyder logo

3 Les modèles d'apprentissage automatique

Notre démarche de la classification des tweets s'inscrit dans l'approche d'apprentissage automatique supervise. Nous avons utilisé l'un des algorithme d'apprentissage (Naïve Bayes, Naïve Bayes Multinomial, linear SVC) qui sera utilisé dans l'étape de prédiction. Concernant le côté implémentation, nous avons utilisé l'implémentation de ces algorithmes d'après le package Sklearn, l'appel du classfieur pour l'apprentissage se fait par le biais du code algorithme.

```
classifier = nltk.NaiveBayesClassifier.train(train_set)

LinearSVC_classifier = SklearnClassifier(LinearSVC())
LinearSVC_classifier.train(train_set)

MNB_classifier = SklearnClassifier(MultinomialNB())
MNB_classifier.train(train_set)

Train_set : les données d'entrainement
```

Pour entrainer les modèles précèdent en utilisant un jeu de données divise en catégorie (chaque catégorie représente un domaine).

affaire	divertissement	santé	politique	technologie	sport
139 ko	96 ko	63 ko	30 ko	96 ko	90 ko

La comparaison entre les trois modelés

Les modelés	Naïve de bayes	linearSVC	NBMultinomial
accuracy	91	97	94

4 Le réseau social : Twitter

Twitter est un réseau social de microblogage géré par l'entreprise Twitter Inc. Il permet à un utilisateur d'envoyer gratuitement de brefs messages, appelés tweets, sur internet, par messagerie instantanée ou par SMS. Ces messages sont limités à 280 caractères (140 caractères auparavant).

Figure 4: Twitter logo

Twitter a été créé le 21 mars 2006 par Jack Dorsey, Evan Williams, Biz Stone et Noah Glass, et lancé en juillet de la même année. Le service est rapidement devenu populaire, jusqu'à réunir plus de 500 millions d'utilisateurs dans le monde fin février 2012. Le 5 mars 2017, Twitter compte 313 millions d'utilisateurs actifs par mois avec 500 millions de tweets envoyés par jour et est disponible en plus de 40 langues.

Le siège social de Twitter Inc. se situe aux États-Unis à San Francisco. L'entreprise dispose de plus de 35 bureaux supplémentaires à travers le monde8 et de serveurs informatiques à New York[68].

Figure43:profil d'utilisateur "Hama Soltani" en twitter

Utilisation de tweepy

Tweepy prend en charge l'accès à Twitter via l'authentification de base et la méthode plus récente, OAuth. Twitter ayant cessé d'accepter l'authentification de base, OAuth est désormais le seul moyen d'utiliser l'API Twitter.

5 Implémentation

Dans notre implémentation en utilisant le twitter comme source de données et en faire un petit code pour classifier les tweets en plusieurs catégorie pour simuler la phase filtrage par domaine.

Cette implémentation passe sur les étapes suivant

L'authentification : pour permettre nous d'utiliser les données de twitter.

L'acquisition : collecte les tweets

La classification : en utilisant les model (entrainer les modèles d'abord)

En fin la phase extraction des tweets des selon le domaine avec une interface graphique TKinter.

5.1 Environnement de travail

Afin de mener notre expérimentation et évaluation, nous avons utilisé un PC marque LENOVO, équipe d'un processeur multi-coré I5, cadence par une horloge d'une fréquence de2.47GHZ, avec 4 GO Octets de RAM, un disque dur d'une capacité de 500 Giga Octets.platforme Windows 10 professionnel.

5.2 L'Authentification

Twitter est ce qui se passe dans le monde et ce dont les gens parlent en ce moment. Vous pouvez accéder à Twitter via le Web ou votre appareil mobile. Pour partager des informations sur Twitter aussi largement que possible, nous fournissons également aux entreprises, aux développeurs et aux utilisateurs un accès programmatique aux données Twitter via nos API (interfaces de programmation d'applications).

```
class Tweetifier:
    def __init__(self, user, count=10):
        self.consumer_key = "0p4DnO6HaUbKPQbmDUxCSA"
        self.consumer_secret = "V6FYODRVEhE8flhZwoNxuK227CVmBZPRIP4DVl1JgkQ"

        self.access_token = "1728327853-EgIxPDpBtEVyWtaeka7y9EGVIYUP5S8f2jzJwtL"
        self.access_token_secret = "eK82eY6d9t4gB8yWt5t2pXAsrRer0V6Hu92Y3GxsvWkMv"

        self.auth = tweepy.OAuthHandler(self.consumer_key, self.consumer_secret)
        self.auth.set_access_token(self.access_token, self.access_token_secret)
```

Figure 5:API TWITTER

5.3 L'Acquisition

Cette étape pour collecte le maximum des tweets partager à l'aide de l'API twitter

```
def crawl(self):
    try:
        api = tweepy.API(self.auth)
        for status in tweepy.Cursor(api.user_timeline,
                                    id=self.user, retweets=True).items(self.count):
            self.tweets.append(status.text)
    except Exception as e:
        print(e)
```

Figure 6: collecte des tweets

5.4 La Classification

En utilisant les model de l'apprentissage automatique pour classifier les tweet collecte

```
def classify(self):
    for t in self.tweets:
        if is_actionable(t):
            topic = predict_topic(t)
            if self.topic_bucket.get(topic):
                self.topic_bucket[topic].append(t)
            else:
                self.topic_bucket[topic] = [t]
```

Figure 7:classification des tweets

5.5 Quelque interface utilise

Les interfaces sont générer aves le module Tkinter, le meilleur module pour les interfaces graphiques sous python, elle est simple et facile à utiliser.

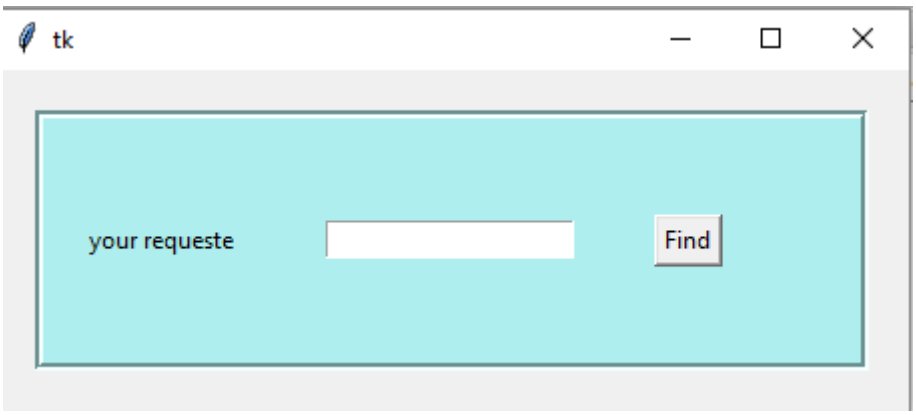


Figure 8 : interface de la détection des domaine

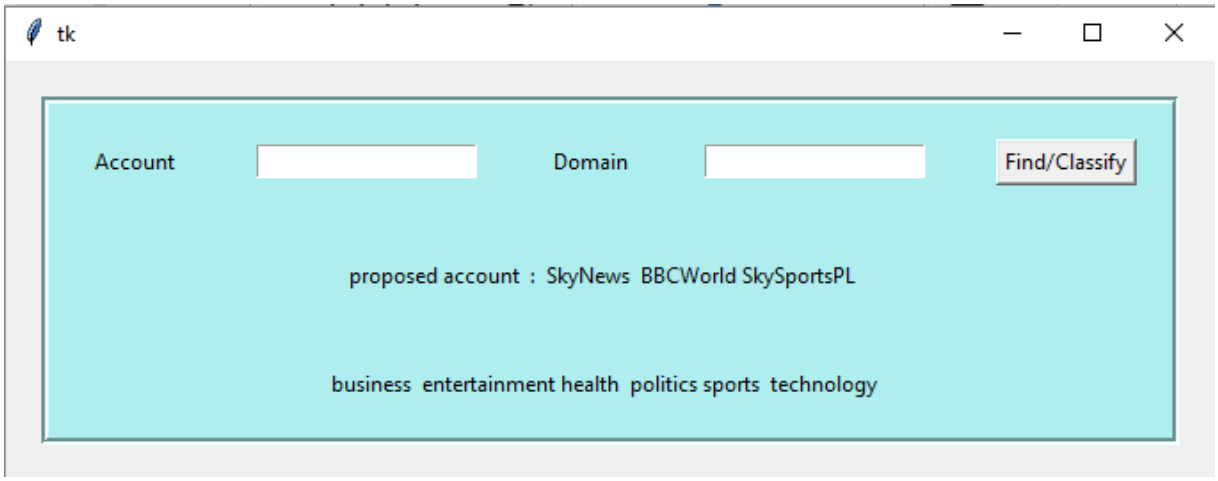


Figure 9:Linterface de la deuxieme phase de la processus propose

6 La conclusion

Au cours de ce chapitre, nous avons utilisé l'environnement anaconada/spyder pour implémenter quelques étapes de base pour notre mprocessus proposée telles que la detection du domaine et filtrage selon le domaine avec une des model de l'apprentissage automatique pour classification des données et en utlisant comme source du données le reseaux sociaux Twitter Ceci est fait par API de twitter, en essayant de montrer l'importance de les deux filtres propose .Ce travail peut toujours être développé pour couvrir toutes les étapes de notre méthode.

Conclusion generale& perspectives

Les réseaux sociaux est un domaine très intéressant à cause les quantités énormes des données et informations partage sur les réseaux sociaux, les réseaux sociaux jouent le rôle de première fournisseur de données dans notre époque.

L'analyse et contrôle de la diffusion de l'information sont devenus domaine de recherche très important et reste un problème ouvert qui renvoie aux questions suivantes : « quelle elle la meilleure technique de d'analyse ? comment on avoir le meilleur résultat ? quelle méthode d'analyse on doit choisir »

Le Big Data est un ensemble de technologies, architectures et outils permettant de capturer, traiter et analyser de large quantités de données hétérogènes et changeants afin d'extraire les informations pertinentes à un cout accessible car les outils classiques ne sont plus inappropriés pour gérer ce type de données.

Dans notre travail, nous avons présenté un modelé pour l'analyse et l'exploitation des informations des réseaux sociaux. Ce travail repose un processus qui définit les modèles d'apprentissage automatique (naïve bayes multinomial, ...)et l'apprentissage approfondi(RNN, LSTM, ...).

Cetravail a consisté à utiliser des modèles d'apprentissage pour l'extraction de données selon un domaine ou un sujet spécifique ensuite en appliquant d'autre modèles à ces données en peut extraire des informations appropriées au contexte de notre requête.

A la suite les résultats que nous avons obtenus, montrent que les technique d'apprentissage automatique et approfondi et les techniques traitement de langage automatique performent pour minier l'espace de recherche et améliorer la qualité des résultats obtenus lors d'une requête

Plusieurs perspectives sont envisageables.

Consiste à mettre en œuvre les modèles d'analyse et extraire les informations à plusieurs réseaux en envisageant des tailles de données importantes.

Nous envisageons de travailler sur les utilisations des profile d'utilisateur pour spécialiser les données qui peut utiliser dans plusieurs domaine comme (la marketing ,la sante,l'éducation,le transport, ...)et améliorer et faciliter la opération de recherche de l'information.

Références

[1] F.Fogelman, E.Viennet ., «L'analyse des réseaux sociaux,» *Bulletin de la société informatique de France*, pp. 25-43, 2016.

[2] MILGRAM, STANLEY, «An Experimental Study of the Small World Problem,» 1967.

[3] M.Haider, A.Gandomi , «Beyond the hype: Big data concepts, methods, and analytics,» *ELSEVIER,* pp. 137-144, 2014.

[4] LAOUAR M, BRADJI L, TABET K, "APPROACH FOR URBAN TERRITORY PLANNING BASED BIG DATA," *Courrier du Savoir,* pp. 45-52, 2017.

[5] Villanustre, Borko Furht · Flavio, Big Data, Switzerland: Springer, 2016.

[6] Na Pandey1, S Rajeshwari, R Shobha,Mrs Mounica , «A Comparison on Hadoop and Spark,» *International Journal of Innovative Research in Computer,* vol. 6, 2018.

[7] Ahmed Oussous, Fatima-Zahra Benjelloun , Ayoub Ait Lahcen , Samir Belfkih, «Big Data technologies: A survey,» *Journal of King Saud University –Computer and Information Sciences,* vol. 30, pp. 431-448, 2018.

[8] H Karau, A Konwinski,P Wendell. M Zaharia, «Introduction to Data Analysis with Spark,» chez *Learning SPARK*, 2015.

[9] A_G. Shoro , T_R Soomro, «Big Data Analysis: Ap Spark Perspective,» *Global Journal of Computer Science and Technology: Software & Data Engineering,* vol. 15, 2015.

[10] Y SAMADI, M ZBAKH ,C Tadonki, «Comparative study between Hadoop and Spark based on Hibench benchmarks,» *2nd International Conference on Cloud Computing Technologies and Applications (CloudTech),* pp. 267-275, 2016.

[11] Jiawei Han, Micheline Kamber,Jian Pei, Data Mining :Concepts and Techniques, AMSTERDAM ,LONDON: Morgan Kaufmann,Elseiver, 2012.

[12] Fatih KAYAALP, Muhammet Sinan BAŞARSLAN, «Open Source Data Mining Programs: A Case Study on R,» *Düzce University Journal of Science & Technology,* vol. 6, pp. 455-468, 2018.

[13] Elangovan, Subedha , Sathishkumar , Ambeth kuma, Data Mining Techniques for Social Network, Advances in Engineering Research, 2018.

[14] K. Harkiran, «A STUDY ON DATA MINING TECHNIQUES AND THEIR AREAS OF APPLICATION,» *International Journal of Recent Trends in Engineering & Research (IJRTER),* vol. 3, pp. 93-95, 2017.

[15] Parneet, Kaur kamaljit et Kaur, «Clustering Techniques in Data Mining For Improving Software Architecture: A Review,» *International Journal of Computer Applications,* vol. 9, pp. 35-39, 2016.

[16] N Jothia, N Abdul Rashidb, WHusainc, «Data Mining in Healthcare – A Review,» *Procedia Computer Science Elsiever,* vol. 72, pp. 306-313, 2015.

[17] C. Mehta, «Basics of Data Mining: A Survey Paper,» *International Journal of Trend in Research and Development,,* vol. 4, pp. 4.-41, 2017.

[18] CLIFTON P, VINCENT L, KATE S, ROSS G, «A Comprehensive Survey of Data Mining-based Fraud Detection Research,» 2013.

[19] Savitha S. Kadiyala,Alok Srivastava,, «Data Mining For Customer Relationship Management,» *International Business & Economics Research Journal,* vol. 1, pp. 61-70.

[20] R.A.Khan ,A Mushtaq,H Kanth, «Data Mining for Marketing,» *Journal of Marketing and Consumer Research,* vol. 9, pp. 17-29, 2015.

[21] Dandan Li, Jing Maa, Zihao Tian, Hengmin Zhub, «An evolutionary game for the diffusion of rumor in complex,» *elseiver,* vol. 433, pp. 51-58, 2015.

[22] X. W. K. G. a. S. Z. Mei Li, «A Survey on Information Diffusion in Online Social,» *information,* 2017.

[23] M. E. J. Newman, «The Structure and Function of,» *Society for Industrial and Applied Mathematics,* vol. 45, pp. 167-256, 2003.

[24] Liang Liu, Bo Qu, Bin Chen, Alan Hanjalic and Huijuan Wang, «Modeling of Information Diffusion on Social Networks with Applications to WeChat,» *RESEARCH,* 2017.

[25] Yu Jin, Wendi Wang , Shiwu Xiao, «An SIRS model with a nonlinear incidence rate,» *Chaos, Solitons and Fractals,ELseiver,* vol. 34, pp. 1482-1497, 2007.

[26] K. X. G. Z. Chao Wang, «A SEIR-based model for virus propagation on SNS,» chez *Fourth International Conference on Emerging Intelligent Data and Web Technologies*, china, 2013.

[27] Xuejun, DING, «Research on propagation model of public opinion topics based on SCIR in microblogging,» *Computer Engineering and Applications,* vol. 51, pp. 20-26, 2015.

[28] Li H, Cui JT, Ma JF, «Social In uence Study in Online Networks,» *JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY,* vol. 30, pp. 184-199, Jan 2015.

[29] Farman Ullah, Sungchang Lee, «Identification of influential nodes based on temporal-aware modeling of multi-hop neighbor interactions for influence spread maximization,» *Physica A,* vol. 486, pp. 968-985, 2017.

[30] K. S. Masahiro Kimura, «Tractable Models for Information Diffusion in,» *Knowledge Discovery in Databases,* vol. 4213, pp. 259-271, 2006.

[31] Shakarian P, Bhatnagar A, Aleali A, Shaabani E & Guo R., «The Independent Cascade and Linear Threshold Models. Diffusion in Social Networks,» *SpringerBriefs in Computer Science,* pp. 35-48, 2015.

[32] Supriya Bhosale, Sucheta Kokate, «Traffic Detection Using Tweets on Twitter Social Network,» *International Journal of Science and Research (IJSR),* vol. 4, 2015.

[33] Tengfei Yang 1,2 , Jibo Xie, Guoqing Li , Naixia Mou , Zhenyu Li , Chuanzhao Tian , «Social Media Big Data Mining and Spatio-Temporal Analysis on Public Emotions for Disaster Mitigation,» *internationale journale of Geo-Information,* vol. 8, 2019.

[34] Chouchani, Nadia, «Une approche de détection des communautés d'intérêt dans les réseaux sociaux :,» *Web. Université de Valenciennes et du Hainaut-Cambresis,* 2018.

[35] Liu, B., «Sentiment analysis and opinion mining,» *Synthesis lectures on human language technologies,* vol. 1, 2012.

[36] Patel, Harsh Thakkar and Dhiren, «Approaches for Sentiment Analysis on Twitter:,» *arXiv e-prints,* 2015.

[37] Kharde, Vishal A, Sonawane, S.S., «Sentiment Analysis of Twitter Data: A Survey of Techniques,» *International Journal of Computer Applications,* vol. 139, 2016.

[38] Liu, Bing, «Sentiment Analysis and opinion mining,» *Morgan &Claypool Publishers,* 2012.

[39] Shepelenko, Olha, «Opinion Mining and Sentiment Analysis using Bayesian and Neural Networks Approaches,» UNIVERSITY OF TARTU, TARTU, 2017.

[40] Walaa Medhat, Ahmed Hassan , Hoda Korashy, «Sentiment analysis algorithms and applications:,» *Ain Shams Engineering Journal,* vol. 5, p. 1093–1113, 2014.

[41] Neha Upadhyay, Angad Singh, «Sentiment Analysis on Twitter by using Machine,» *International Journal for Research in Applied Science & Engineering,* vol. 4, pp. 488-494, 2016.

[42] Nurulhuda Zainuddin, Ali Selamat, «Sentiment Analysis Using Support Vector Machine,» chez *international conferance in computer science,comunication,and control technology*, malisya, 2014.

[43] Megha Joshi, Purvi Prajapati,Ayesha Shaikh,Vishwa Vala, «A Survey on Sentiment Analysis,» *International Journal of Computer Applications,* vol. 163, pp. 34-38, 2017.

[44] MAHAMMED, FATIMA AIT, «APPROCHES D'APPRENTISSAGE AUTOMATIQUE POUR LA DÉTECTION,» UNIVERSITÉ DU QUÉBEC À MONTRÉAL, MONTRÉA, 2018.

[45] Gwanghoon Yoo, Jeesun Nam, «A Hybrid Approach to Sentiment Analysis Enhanced by Sentiment Lexicons and Polarity Shifting Devices,» *The 13th Workshop on Asian Language Resources,* pp. 21-28, 2018.

[46] Devi D., Venkata R K , Mounika K, Sowjanya Swathi N., «Assay: Hybrid Approach for Sentiment,» *Information and Communication Technology for Intelligent Systems. Smart Innovation, Systems and Technologies,Springer,* vol. 106, 2019.

[47] FARZINDAR ATEFEH ,WAEL KHREICH, «A SURVEY OF TECHNIQUES FOR EVENT DETECTION IN TWITTER,» *Computational Intelligence,* vol. 31, pp. 132-164, 2015.

[48] T .SAKAKI,M. OKAZAKI , Y. MATSUO, «Earthquake shakes Twitter users: Real-time event detection by social sensors.,» chez *19th International Conference on World Wide Web,ACM*, New York, 2010.

[49] SANKARANARAYANAN, J., H. SAMET, E. TEITLER D. LIEBERMAN, and J. SPERLING, «TwitterStand: News in Tweets,» New York, 2009.

[50] E. Negre, « Comparaison de textes: quelques approches...,» *hal-00874280,* 2013.

[51] Lina Alfantoukh ,Arjan Durresi, «Techniques for Collecting data in Social Networks,» 2014.

[52] Facebook developper, «Using the Graph API,» chez *https://developers.facebook.com/docs/graph-api/using-graph-api*.

[53] twitter developer, «Twitter libraries,» chez *https://developer.twitter.com/en/docs/developer-utilities/twitter-libraries.html*.

[54] Julien Maitre, Michel Menard, Guillaume Chiron, Alain Bouju, «Utilisation conjointe LDA et Word2Vec dans un contexte d'investigation numérique.,» chez *Extraction et Gestion des Connaissances 2017*, Grenoble, France, jan 2017.

[55] Marwa Naili, Anja Habacha Chaibi, Henda Hajjami Ben Ghezala, «Comparative study of word embedding methods in topic segmentation,» *Procedia Computer Science,* vol. 112, pp. 340-349, 2017.

[56] Mikolov, Tomas & Corrado, G.s & Chen, Kai & Dean, Jeffrey., «Efficient Estimation of Word Representations in Vector Space,» 2013.

[57] Tomas Mikolov,Ilya Sutskever,Kai Chen,Greg Corrado,Jeffrey Dean, «Distributed Representations of Words and Phrases and their Compositionality,» 2013.

[58] Jeffrey Pennington, Richard Socher, Christopher D. Manning, «GloVe: Global Vectors forWord Representation,» 2014.

[59] Jeffrey Pennington, Richard Socher, Christopher D. Manning, «GloVe: Global Vectors forWord Representation,» chez *https://nlp.stanford.edu/projects/glove/*, 2014.

[60] facebook research team, «Word representations,» 2016.

[61] Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, «HDLTex: Hierarchical Deep Learning for Text,» *International Conference on Machine Learning and Applications (ICMLA),* pp. 364-371, 2017.

[62] aishwarya, «Introduction to Recurrent Neural Network,» *https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/* .

[63] Lynda Tamine, Nesrine Zemirli, Wahiba Bahsoun, «Approche statistique pour la définition du profil d'unutilisateur de système de recherche d'information,» *Information Interaction Intelligence,* vol. 7, pp. 5-25, 2007.

[64] Ramiandrisoa, Faneva and Mothe, Josiane, «Profil utilisateur dans les réseaux sociaux : État de l'art.,» chez *Rencontres Jeunes Chercheurs en Recherche d'Information (RJCRI 2017)*, Koria, 2017.

[65] Python Software Foundation. , «Le tutoriel Python,» *https://docs.python.org/fr/3/tutorial/index.html.*

[66] Anaconda Inc, «Anaconda Distribution,» *https://www.anaconda.com.*

[67] anaconda Inc, «spyder,» *https://anaconda.org/anaconda/spyder.*

[68] Wikipedia, «twitter,» *https://en.wikipedia.org/wiki/Twitter.*