Université Saad DAHLAB - Blida 1

Faculté des sciences

Département d'Informatique

Mémoire présenté par :

BENMILOUD Meriem

DJEBRANI Aymen

Pour l'obtention du diplôme de Master

Domaine : Mathématique et Informatique

Filière : Informatique

Spécialité : Traitement Automatique de la Langue

Sujet :

Détection de tendances des réseaux sociaux en utilisant les techniques du TALN

Soutenu le :26-10-2020 devant le jury composé de :

Mme. N. BENBLIDIA	Universitéde Blida 1	Présidente
Mlle. H. YKHLEF	Universitéde Blida 1	Examinatrice
Mme. H. ALIANE	CERIST	Encadreur
Mme. M. MEZZI	Universitéde Blida 1	Promotrice

Résumé

Nombreuses chercheuses et chercheurs ont recours à Twitter dans leur activité scientifique. Inévitablement, ces courts messages jouent aujourd'hui un rôle dans la dissémination de la science, Bien que Twitter fournisse une liste des sujets les plus populaires tweetés connus sous le nom de sujets tendance en temps réel sauf que la précision de la détection de tendances n'est pas d'une précision élever.

Pour résoudre ce problème, on a recours aux méthodes d'apprentissage automatique ; nous classons les sujets tendance sur Twitteren 20 catégories dans le domaine scientifique ; Nous expérimentons 2 approches de classification automatique pour classerles thèmes et obtenir la tendance ; l'approche bien connue du sac de mots pour la classification des textes et classification basée sur le réseau. Lors de la classification textuelle, nous construisons des vecteurs de mots avec poids tf-idf qui sont classer à l'aide de l'algorithme multinomial Naïve Bayes. Et une méthode de classification basée sur le réseau, en utilisant un algorithme personnalisé.

Expérimenté sur une base de données extraite par le billet de l'API twitter montrant qu'une précision de classification allant jusqu'à 75% et 85% peut être réalisée en utilisant la classification basée sur le texte et sur le réseau modélisation respectivement.

Mots clés : Réseaux sociaux, Twitter, classification des sujets, sujets tendances

Abstract

Manyresearchers and researchers use Twitter in theirscientificactivity. Inevitably, thesecurrent messages play a role in the dissemination of science today, Although Twitter provides a list of the mostpopular topics to tweet known as Trending Topics and trending in real time except that the accuracy of trend detection is not high.

To solve thisproblem, wecategorize the trending topics on Twitter into 20 scientificcategories. We are experimentingwith 2 approaches for the classification of themes and get the trend;Well-knownword bag approach for text classification and network-based classification. In the textual classification method, weconstructwordvectorswithtrending topic definition and tweets and, and tf-idfweights are used to classify topics using a Naive Bayes multinomial algorithm. And network-based classification method, using a custom algorithm.

Experimenting on an annotated twitter API databaserequires classification accuracy of up to 75% and 85% can be achievedusingtext-based classification and network modelling respectively.

Keywords: Social networks, Twitter, subject classification, trending subjects

ملخص

يلجأ العديد من الباحثين والباحثات إلى التويتر خلال مختف نشاطاتهم العلمية. لا محالة، هذه الرسائل القصيرة تلعب اليوم دورا في نشر العلم. على الرغم من أن التويتر يوفر قائمة بالمواضيع الأكثر شيوعا، والمعروفة باسم المواضيع الشائعة في الوقت الحالي، إلا أن دقة الكشف عنها ليست عالية الدقة.

لحل هذه المشكلة، قمنا بتصنيف المواضيع الشائعة على التويتر إلى 20 فئة علمية. واعتمدنا في تصنيفنا والحصول على الاتجاهات لهذه المواضيع على منهجيتين: المنهجية المعروفة بحقيبة الكلمات لتصنيف النصوص ومنهجية التصنيف المعتمدة على الشبكة. في طريقة التصنيف النصي، شكلنا سلسلة كلمات مع التعريف بموضوعها الشائع والتغريدات الخاصة بها والوزن TF-IDF المستخدم لتصنيف هذه المواضيع بالاستعانة بخوارزمية Naïve Bayes متعدد الحدود.والطريقة المعتمدة على الشبكةباستخدام خوارزمية مخصصة.

المشروحة على قاعدة البيانات المعتمدة من طرف منصة تويتر تبين دقة في التصنيف تصل إلى 75% و85يمكن أن تحدث باستخدام التصنيف المعتمد على النصوص وعلى نمذجة الشبكة على التوالي.

الكلمات المفتاحية:الشبكات الاجتماعية،تويتر،تصنيف الموضوع،المواضيع الشائعة

Dédicaces

Je dédicace ce modeste travail aux être qui me sont les plus chers :

***A** mes très chers parents et surtout à mon pére*
***A** mesfréres sabri, et anis*
***A** mes sœurs lilia naima et kenza*
***A** ma meilleur ami katia*
***A** toute ma famille sans exceptions*

Benmiloud meriem

***A** mes très chers parents*
***A** mon frére youcef et*
***A** ma sœur imene*
***A** ma tante ghania qui a toujours était présente a mes cotés quoi qu'il arrive*
***A** toute ma famille sans exceptions*

Djebrani aymen

***A** mes amis et camarades de la promotion TALN, et tous ceux qui m'ont aidé.*

Remerciement

J’exprime toute ma reconnaissance et gratitude à l’administration et à l’ensemble du corps enseignant de l’Université Virtuelle pour leurs efforts à nous garantir la continuité et l’aboutissement de ce programme de Master.

Je tiens à remercier aussi et chaleureusement mes encadreurs Mme. M. MEZZI et Mme H.ALIANE de m’avoir permis de mener ce travail .

Mes remerciements vont également aux, Membres du jury.

Je remercie mes chers parents benmiloud abdelkader et faiza et ma meilleur ami katia d’être toujours présent a coté de moi et me soutenir et m’encourager a faire mieux .

Je remercie aussi mes parents djebrani amirouche et fadila et ma tante ghania pour ses encouragement et soutiens.

Je remercie enfin tous ceux qui, d’une manière ou d’une autre, ont contribué à la réussite de ce travail et qui n’ont pas pu être cités ici.

Table des matières

Liste des figures

Liste des tableaux

Liste des acronymes

AA*: Approche apprentissage*

API *:Application programming interface*

bib*:bibliothèque*

BSD*:Berkley software distriution*

CRUD*:Create read update Delete*

CSS*:Cascading Style Sheets*

CSV *: Comma-separated values*

DEV*:Developement*

DF*:Data Frame*

doc *:document*

DSN*:Dérection de sujet non supervisé*

DSS*:Détection de sujet supervisé*

FCM *:Fuzzy C-means*

FTR*: Fenetre temporelle*

HDF5*: Hierachical data format 5*

html *:hypertexte markup language*

IA *: intelligence artificielle*

idf*:inverse terms frequency*

KNN *: K nearst neighbors*

LIBSVM*:Library for support machine vector*

LLSF *: Linear least square fit*

LPI *:Locality Prescing indexing*

LSI *: Latent smantic indexing*

ML*:Machine Learning*

MTV*: Model view template.*

MVC*: Model view controller*

NB *: Naive Bayes*

NBM *:Naive bayes multinomial*

NLP *:naturel language proccesing*

NLTK *: Natural language toolkit*

ORM*: Object relational mapping*

Liste des acronymes

PDF *: Portable document format*

PHP*:Hypertext Preprocessor*

RF*:Random Forest*

RSN*: Réseau social numérique*

SP *:Sentence probability*

SQL*Structured Query Language*

SVM *support vector machine*

SVM-L *:Support vector machines avec linéaire*

TALN *: traitement automatiqye du langage naturel*

TC *:Texte classification (classification du texte)*

TDT I*:Tendance detection tweete idonésiens*

TF*:Term frequency*

tf-idf *:terme fréquence – inverse fréquence des documents*

URL*: Uniform Resource Locator*

W3C*:World Wide Web Consortium*

Web*: World English Bible*

XML*:Extensible markup language*

Introduction

Contexte globale

Twitter est utilisé comme support d'informations en temps réel et qui présente de nombreuses opportunités de recherche en traitement du langage naturel (TALN) et en apprentissage automatique. Les sujets tendance sont censées représenter les « sujets de conversation » populaires. Un sous-ensemble du problème plus large connu sous le nom de détection et suivi de sujets (TDT), la popularité et la croissance de Twitter présentent certains défis pour les applications de la TALN et de l'apprentissage automatique. Nombre de chercheurs ont recours à Twitter dans leurs activités scientifiques. Inévitablement, ces courts messages jouent aujourd'hui un rôle dans la dissémination de la science, elle avance en partie grâce aux critiques et expositions à de nouvelles idées : débats en ligne, échanges de point de vue, de références.

Problématique et objectifs

Les restrictions de longueur des messages créent des conventions syntaxiques et structurelles qui n'apparaissent pas dans les corpus plus traditionnels, et la taille du réseau Twitter produit un corpus dynamique en constante évolution. De plus, il y a beaucoup de contenu sur Twitter qui serait classé comme sans importance pour un observateur extérieur, consistant en des informations personnelles ou du spam, qui doivent être filtrés afin d'identifier avec précision les éléments du corpus qui sont pertinents pour le Twitter dans son ensemble, et pourraient donc être considérés comme des sujets de tendance potentiels. Notre défi dans cette thèseest de détecter et identifier les sujets tendances. Dans notre travail nos proposant des méthodes d'utilisation des techniques de TALN sur les données de Twitter pour catégoriser et identifier les sujets d'actualité.Notre objectif est d'aider les utilisateurs qui rechercher des informations sur Twitter concernant les sujets scientifique les plus discuter et ne garder que le plus petitsous-ensemble de sujets tendance pour faciliter la recherche d'information.

Organisation du mémoire

Notre mémoire est divisé principalement en quatre chapitres.

Dans le premier chapitre, nous focalisons sur l'état de l'art de l'analyse des données des réseaux sociaux, notamment les dérivés de cette discipline. Le second chapitre est consacré à l'apprentissage automatique et aux algorithmes de classification. Dans

le troisième chapitre, nous présentons en détailles travaux intérieurs faits sur ce sujet. Le quatrième chapitre présente la modélisation de notre système l'expérimentionset la discussion des résultats obtenus. Enfin, nous mettons une conclusion et quelques perspectives.

Chapitre IRéseauxsociaux et analyse de leurs données

I.1 Introduction

Les technologies de l'information et de la communication, ont réalisé une véritable révolution dans nos manières d'être, de penser et d'agir. Le premier réseau social « Classmates.com » a été créé en 1995 par Randy Conrads[1].Diverses plateformes ont émergé de là on parle de l'évolution des technologies liées au Web 2.0, si on se fie à la définition simple du dictionnaire [2] « Site Internet qui permet aux internautes de se créer une page personnelle afin de partager et d'échanger des informations, des photos ou des vidéos avec leur communauté d'amis et leur réseau de connaissances.
Le nombre de travaux traitant de médias ou réseaux sociaux parfois appelés analyse néo-structurale, a presque décuplé tous les 10 ans comme on peut le constater sur la figure -1-.
Dans ce chapitre nous allons survoler brièvement et généralement le monde des réseaux sociaux ainsi que leurs différents types tout en mettant au clair leur rôle incontournable dans la recherche scientifique mais aussi les différentes analyses sur leurs données et leur utilisation.

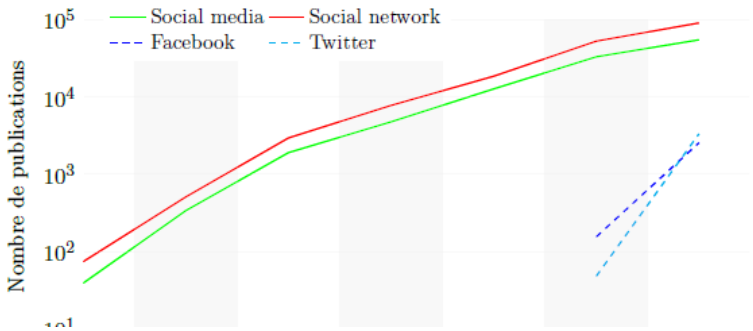


Figure 1Nombre de publications mentionnant les termes « Social media », « social network », « Facebook » et « Twitter » [1].

I.2 Définition des réseaux sociaux

Les réseaux sociaux sont, par définition, de nouveaux moyens de diffusion d'informations intégrant des dimensions à la fois technologiques, virales et sociales, permettant de créer et de diffuser tout type de contenu dans un réseau numérique ouvert ou fermé. Ils peuvent prendre diverses formes, celle d'un blog, d'un réseau social comme Myspace, ou d'une plate-forme collaborative comme Wikipédia Statistiquement parlant, sur les 7,7 milliards d'êtres humains, les mobinautes (5,1 milliards, 67%) sont désormais plus nombreux que les internautes (4,45 milliards,

[1]Randy Conrads :http://engineering.oregonstate.edu/randal-conrads-2003-academy-distinguished-engineers

58%) [3]. Ces chiffres montrent la proportion et l'impact qu'ont ces plateformes dans la vie sociale et économique de notre société.

Les services qu'ils offrent sont principalement basés sur le web, dont la fonctionnalité principale est de connecter des personnes ou des entités, d'où cette définition : « Un réseau social est un ensemble d'individus, organisations ou entités entretenant des relations sociales fondées sur l'amitié, le travail collaboratif et l'échange d'informations, on peut les décrire comme des ensembles finis d'acteurs et les relations définies entre ces acteurs ».

I.3 Types des réseaux sociaux

Les réseaux sociaux peuvent être classés comme le montre le tableau ci-dessous :

Catégories	Types	Exemples	Description
Réseaux sociaux	Professionnels	LinkedIn	Un réseau professionnel international qui permet la mise en relation entre professionnels.
		Viadeo	Permet de construire et de gérer son réseau professionnel.
	Généralistes	Facebook	Permet différents types d'échanges avec sa communauté d'amis.
		Twitter	Permet d'échanger des messages avec d'autres internautes avec une limite de caractères.
		Myspace	Site interactif qui offre à ses abonnés de multiples services combinant blog, espace personnel, espace communautaire.
	Médias de partages	YouTube	Grand espace de disposition de vidéos.
		Dailymotion	Espace ou on peut télécharger, partager et regarder des vidéos.
		Instagrame	Permet de partager, d'éditer des photos et vidéos courte avec son cercle d'amis mais aussi de partager des stories d'une validité de 24 heures

Tableau 1- Les types de réseaux sociaux [4.]

Il existe aussi une typologisation des réseaux sociaux en rapport avec l'évolution et l'apparition ou la fonctionnalité ou encore le point de vue des chercheurs cela reste l'angle de vue qu'on peut associer aux réseaux sociaux selon l'analyse et les recherches qu'on fait. Dans le cadre de ce projet, on s'intéresse à deux types en particulier[5] :

I.3.1 Typologisation des réseaux sociaux numérique selon la fonctionnalité:

- Networking : est un réseau social à usage exclusivement professionnel, orienté sur la mise en valeur et les échanges professionnels de ses membres.
- Bloglike[2] : dont l'usage est strictement privé, cette catégorie permet le partage de la vie quotidienne d'une certaine classe de gens dont l'âge varie entre 12 et 17 ans.
- Spécialisés[3] : Ce type de réseaux sociaux propose des relations beaucoup plus spécifiques et qui pour certains s'apparentent à des communautés d'intérêts.
- Micro-blogging[4] : désigne l'activité de création de contenus courts sur des réseaux sociaux de type Twitter « chat public instantané ».
- Fourre-tout[5] : ce sont les inclassables qui se servent du collaboratif ou du participatif pour alimenter leurs services.

I.3.2 Typologisation selon le point de vue des chercheurs

- Réseaux sociaux de socialisation[6] : Ce type de réseau aide les utilisateurs à trouver une information ou des ressources. Les membres peuvent soit lire les propositions mises en avant en page d'accueil, ou bien utiliser la navigation sociale en lisant les informations postées ou recommandées par leurs amis, ou bien pour certains, recourir plusieurs objectifs [4]
- Réseaux sociaux de réseautage[7]: Utilisés davantage pour trouver de nouveaux contacts et peuvent servir à entrer en connexions avec des personnes

[2]Bloglike. https://www.instagram.com/?hl=fr
[3] Spécialisés. https://fr.linkedin.com/

[4]Micro-blogging. https://twitter.com/login?lang=fr

[5]Fourre-tout. https://www.4chan.org/

[6]Réseaux sociaux de socialisation.https://www.facebook.com/

[7]Réseaux sociaux de réseautage.https://twitter.com/login?lang=fr

inconnues auparavant comme c'est le cas de LinkedIn ou Viadeo, site de réseautage à caractère professionnel [4].

- Réseaux sociaux de navigation [8]: Ce type de réseaux permet aux utilisateurs de trouver une information ou des ressources. Autrement dit, nous trouvons des listes de contacts, listes permettant l'accès à l'information et aux ressources associés à ceux-ci. Les membres peuvent soit lire les propositions mises en avant en page d'accueil, soit utiliser la navigation sociale en lisant les informations postées ou recommandées par leurs amis [4].

I.4 Les réseaux sociaux et la recherche scientifique

Les réseaux sociaux s'imposent progressivement comme une composante à part entière du profil d'un universitaire ou autre personne dans le domaine scientifique. L'image du chercheur ermite, rétif aux nouvelles technologies et volontairement détaché de tout lien électronique est en train progressivement de s'effacer. Deux influences y concourent : d'une part, le développement de réseaux sociaux spécifiquement dédiés aux chercheurs universitaires, comme Researchgate[9] ; ACADEMIA.EDUC[10] ou encore Social Sciences Research Network[11] ; et d'autre part, une demande grandissante de la présence d'universitaires en dehors de leur milieu professionnel. Un canal d'expression fort obligeant à sortir de son laboratoire et de sa recherche pour se confronter à un public élargi (pairs, financeurs, grand public …)

Les nouveaux médias et réseaux sociaux améliorent la connectivité entre les chercheurs, ingénieurs, doctorants, post-doctorants et étudiants en leur permettent d'avoir une information rapide sur les événements scientifiques et de partager directement des connaissances, d'échanger, de débattre de la formulation de questions et d'hypothèses à la diffusion des résultats de recherche un moyen de

[8]Réseaux sociaux de navigation. https://www.reddit.com/r/readit/

[9]Researchgate.https://www.researchgate.net/

[10]ACADEMIA.EDUC.https://www.academia.edu/

[11]Social Sciences Research Network.https://www.ssrn.com/index.cfm/en/

diffuser facilement les résultats de la recherche, néanmoins on se devait de souligner que les contenus échangés concernent surtout le partage de publications, rarement de données de la science, de savoir-faire ou de conseils.

I.5 Analyse des donnés

I.5.1 Définition

L'analyse des réseaux sociaux trouve ses origines théoriques dans les travaux des mathématiciens sur les graphes Un réseau social représente un système d'entités en interaction [6]. « On le modélisera comme un graphe G = (S, A) où S est un ensemble d'entités (les sommets ou nœuds du graphe) et A est l'ensemble des arcs (ou connexions) représentant les interactions entre ces sommets. »

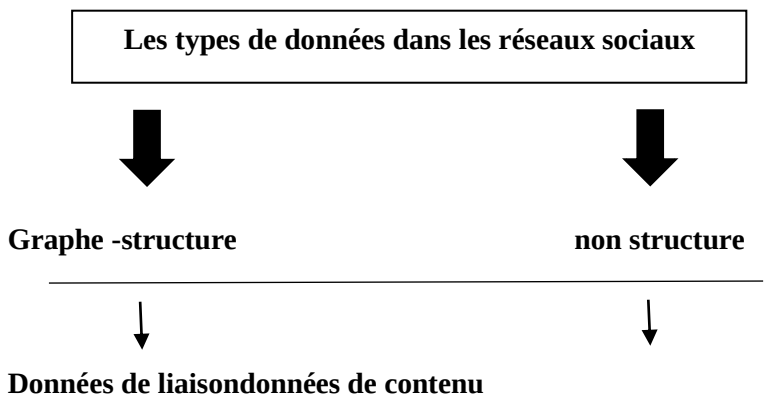
Figure 2- schéma illustrant graphiquement un réseau social.

Après s'être éloigné de l'analyse de simples petits graphiques et les propriétés des nœuds individuels on en vient à considérer des propriétés à grande échelle des graphiques, la nécessité de nouveaux outils et techniques d'analyse et de récolte de données sont inévitables, et de là on trouve différents types d'analyses, de méthodes et de fonctionnalités pour contrer ce défi.

I.6 L'analyse des données des réseaux sociaux

L'analyse des réseaux sociaux est étudiée par le biais d'analyses de Big Data, le défi majeur est de déterminer la nature et la structure de ces données et comment les analyser ce qui est le cas après la collecte dans les réseaux sociaux qui peuvent être structurées et non structurées Ceux-ci impliquent aussi soit une analyse statique ou dynamique dans la première elle peut se faire en mode batch « lot de données ». Inversement, l'analyse dynamique, plus complexe, englobe des données en

streaming qui évolue dans le temps à un rythme élevé. L'analyse dynamique est souvent zone d'interactions entre les entités alors que l'analyse statique traite de propriétés comme la connectivité densité, degré, diamètre et distance géodésique. Le schéma suivant résume les types de données et les approches et méthodes d'analyses correspondantes menées.



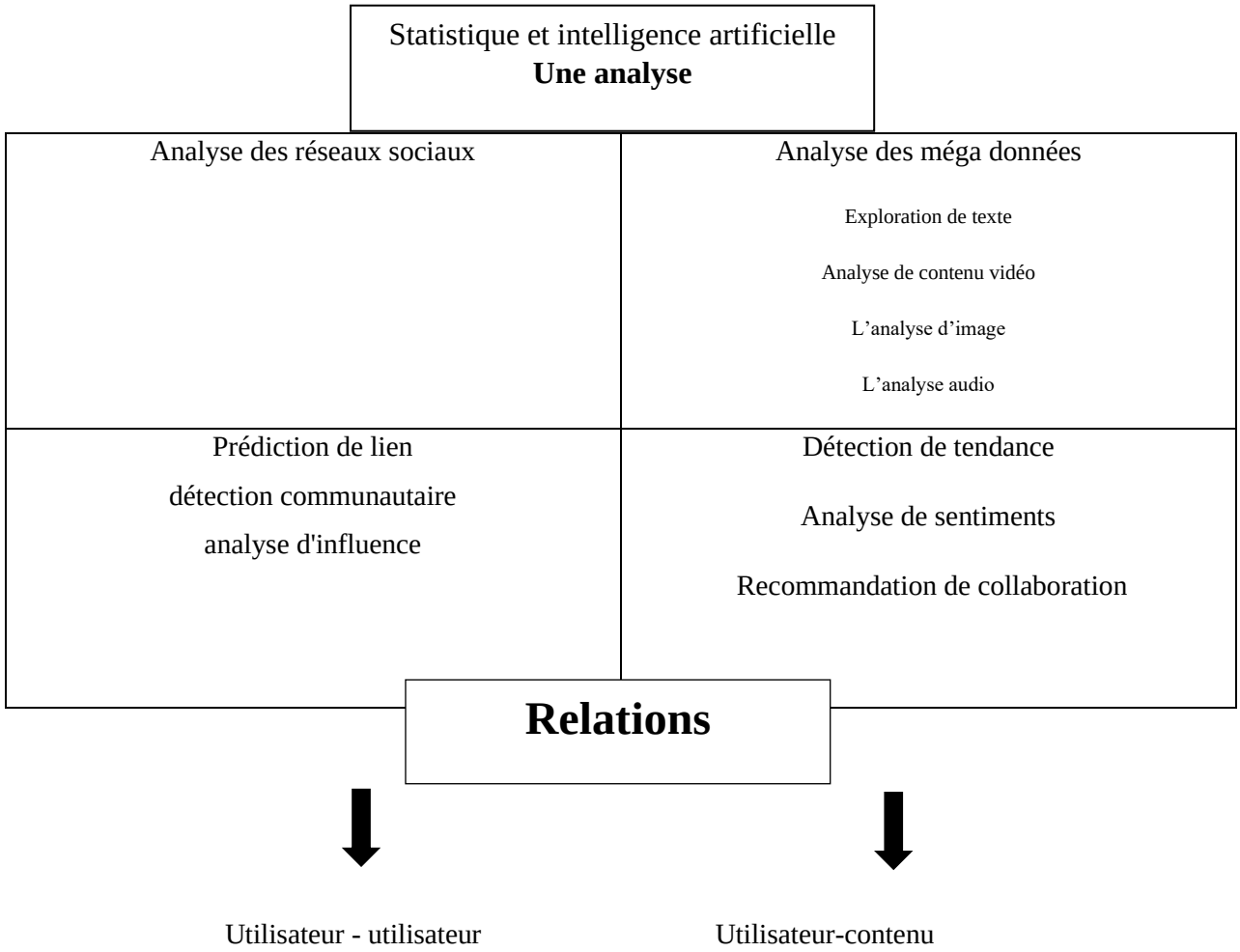


Figure 3- Types de données et analyse.

L'analyse des réseaux sociaux et les approches d'exploration de contenu suivent les principes interdisciplinaires d'Intelligence artificielle (IA), statistiques et domaines connexes.

L'exploration de texte qui est l'un des sujets principaux de notre projet extrait des modèles de données textuelles par le biais de la récupération d'informations, du texte synthétisé et fait un traitement du langage naturel (TALN) qui est une interaction entre l'intelligence artificielle et la linguistique.

I.7 Analyse des réseaux sociaux

I.7.1 Analyse d 'influence

Mesurer la dominance des nœuds, quantifier la force des connexions et découvrir les modèles de diffusion de l'influence. Un sujet de recherche essentiel qui consiste à identifier des utilisateurs [8] « expérimentés » ou « dignes de confiance » qui peuvent être des précurseurs, car leurs messages d'opinion peuvent se propager

rapidement et largement dans le réseau, ce qui leur permet d'influencer d'autres utilisateurs Dans le contexte des microblogs, plusieurs indicateurs ont été discutés pour mesurer l'influence et crédibilité d'un utilisateur[9]: par exemple mentionner l'influence, suivre l'influence et retweeter sont des éléments distincts.

I.7.2 Analyse des liens

L'analyse des liens est utilisée pour évaluer les connexions entre les nœuds. Comprendre la formation et l'évolution de ces connexions dans les réseaux sociaux nécessite des données longitudinales sur les deux interactions sociales et affiliations partagées [10]. L'exploration de liens est généralement associée à l'exploration de texte et peut être utilisée pour la classification, prédiction, clustering ou découverte de règles d'association. Elle est applicable en collaboration avec des systèmes de recommandation pour identifier un groupe d'amis ayant des intérêts similaires [11].

I.7.3 Détections communautaires

Les communautés constituent un aspect important des réseaux elles permettent l'exploration des plateformes et des données de prédictions de connexions qui ne sont pas encore observées [12]. La détection communautaire est essentiellement un problème de clustering de données, où le but est d'affecter chaque nœud à une communauté ou un cluster d'une manière raisonnable. L'analyse peut être classée en matière de dimension comme suite :

- analyse statique : « quelles sont les communautés au temps T ?»,
- analyse temporelle : « comment s'est formée cette communauté ?».
- analyse prédictive : « comment une communauté va-t-elle grandir ?».

I.8 Analyse de Big Data et exploration de texte

I.8.1 Big Data

Le Big Data est un processus de collecte, de gestion et d'analyse de grandes quantités de données pour générer des connaissances et exposer des modèles cachés. Plusieurs défis doivent être relevés, notamment la collecte, le stockage et traitement des données, ainsi que l'acquisition des connaissances précieuses en les analysant et en les visualisant de la meilleure façon possible. L'utilisation d'un ensemble de stockage et les hautes performances des calculs est appelée cluster « une grappe de serveurs sur un réseau, appelé ferme ou grille de calcul » [13], nous pouvons évaluer les méga données concernant leurs quantités et leurs vitesses. De cette façon, le

principe principal de l'analyse des méga données est la nécessité de hautes performances dans les calculs.

I.8.2 Big Data et exploration de texte

L'exploration de texte dans l'analyse des méga données [15] est en train de devenir un outil puissant pour exploiter l'intelligence de données textuelles non structurées en les analysants pour extraire de nouvelles connaissances. L'exploration de texte contient cinq étapes clé qui sont [36] :

1. Collecte de document texte
2. Prétraitement du texte : (Tokenisation, Suppression des mots vide, stemming,(Transformation de texte « Modèle d'espace vectoriel et sac de mots… »)
3. Techniques d'exploration de texte : Il existe différents types de techniques (synthèse, classification, catégorisation, clustering, … etc.)
4. Analysez le texte
5. Découverte des connaissances

I.9 Conclusion

Après avoir présenté brièvement les réseaux sociaux leurs catégories et les caractéristiques de données dans les réseaux sociaux et méthodes et d'approches d'analyse des données. Nous pouvons conclure qu'ils représentent des outils de communication incontournables. L'écosystème des médias sociaux s'organise autour de quatre grands usages : la publication, le partage, la discussion, le réseautage

Nous avons aussi pu voir l'importance des réseaux sociaux au sein de la communauté scientifique, il permet la création de débat et le partage d'informations importantes entre les chercheurs. Ce qui montre l'importance de l'analyse de ce contenu.

Le monde des réseaux sociaux propose toute une panoplie de recherche est d'études dans plusieurs disciplines qui ne cessent de s'étendre et se diversifier selon le besoin.

Dans le prochain chapitre nous verrons une introduction à l'apprentissage automatique en qualification avec les algorithmes les plus performants.

II. Chapitre IIApprentissage automatique et algorithme declassification

II.1 Introduction

Le Machine Learning, aussi appelé apprentissage automatique en français, est une forme d'intelligence artificielle permettant aux ordinateurs d'apprendre sans avoir été programmés explicitement. Le Machine Learning est une méthode d'analyse de données permettant d'automatiser le développement de modèle analytique. Par le biais d'algorithmes capables d'apprendre de manière itérative[38], le Machine Learning permet aux ordinateurs de découvrir des « insights » cachés sans être programmés pour savoir où les chercher.

Dans ce chapitre on va voire une partie de l'intelligence artificielle qu'est l'apprentissage automatique voir les différentes méthodes du « ML » les étapes à faire pour réussir un projet et pour relier ça a notre projet on va étudier vaguement les algorithmes de classification.

II.2 Les principales méthodes de Machine Learning

Les deux méthodes de Machine Learning(ML) les plus couramment utilisées sont l'apprentissage supervisé et le non supervisé[39]**.**

II.2.1 Apprentissage supervisé :

Ils sont entraînés à l'aide d'exemples étiquetés.

L'algorithme reçoit un ensemble d'inputs ainsi que les outputs corrects correspondants, et apprend en comparant les outputs avec les résultats corrects attendus pour détecter les erreurs. Il modifie ensuite son modèle en fonction. Les méthodes comme la classification, la régression, et prédiction permettent à l'apprentissage supervisé d'utiliser des patterns pour prédire la valeur d'une étiquette ou d'une donnée additionnelle sans étiquette.

II.2.2 Apprentissage non supervisé :

Est utilisé pour les données qui n'ont pas d'étiquettes historiques**.** Le système ne connaît pas la réponse correcte, et l'algorithme doit comprendre par lui-même ce qui lui est présenté. L'objectif est d'explorer les données et de trouver une structure

II.3 Algorithmes utilisés :

Ce sont, dans ce domaine :

- Les machines à vecteur de support .
- Le boosting .

- les réseaux de neurones, dont les méthodes d'apprentissage profondes (deeplearning en anglais) pour un apprentissage supervisé ou non-supervisé.
- la méthode des k plus proches voisins pour un apprentissage supervisé.
- les arbres de décisions, méthodes à l'origine des Random Forest, par extension également du boosting (notamment xgboost).
- les méthodes statistiques comme le modèle de mixture gaussienne .
- La régression logistique .
- L'analyse discriminante linéaire .
- Les génétiques et la programmation génétique.

Ces méthodes sont souvent combinées pour obtenir diverses variantes d'apprentissage. L'utilisation de tel ou tel algorithme dépend fortement de la tâche à résoudre (classification, estimation de valeurs…).

II.4 Étapes d'un projet d'apprentissage automatique

L'apprentissage automatique ne se résume pas à un ensemble d'algorithmes mais suit une succession d'étapes.

1. **L'acquisition de données :** l'algorithme se nourrissant des données en entrée, c'est une étape importante. Il en va de la réussite du projet, de récolter des données pertinentes et en quantité suffisante.
2. **La préparation et le nettoyage de la donnée :** les données recueillies doivent être retouchées avant utilisation. En effet, afin d'être compris par l'algorithme, Plusieurs techniques telles que la visualisation de données, la transformation de données ou encore la normalisation sont alors employées.
3. **La création du modèle :** choisir l'algorithme approprié
4. **L'évaluation :** une fois l'algorithme d'apprentissage automatique entraîné sur un premier jeu de donnée, on l'évalue sur un deuxième ensemble de données afin de vérifier que le modèle ne fasse pas de sur apprentissage.

II.5 PROCESSUS DE CLASSIFICATION DU TEXTE

Les étapes de classification du texte (TC) sont reparties comme suit [50] :

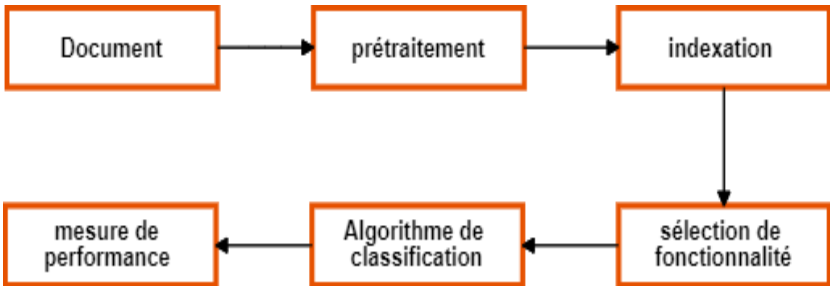


Figure 4-processus de classification des documents.

II.5.1 Collection de documents

Il s'agit de la première étape du processus de classification dans lequel nous collectons les différents types (format) de document comme html, PDF, doc, contenu Web, etc.

II.5.2 Prétraitement

La première étape de prétraitement qui est utilisée pour présenter les documents texte en mots clairs. Les documents préparés pour la prochaine étape de classification de texte sont représentés par une grande quantité de fonctionnalités.

Les mesures prise sont :

Tokenisation : un document est traité comme une chaîne, puis partitionné en une liste de jetons.

Suppression de mots vides : des mots vides tels que « le », « a », «et », etc.

Mot de racine : application de l'algorithme de racine qui convertit une forme de mot différente en forme canonique. Cette étape est le processus de fusion des jetons avec leur forme racine.

II.5.3 Indexation

Le document doit êtretransformé de la version texte intégral en vecteur de document le plus couramment utilisé, la représentation du document est appelée modèle d'espace vectoriel (SMART) [40]

La dimensionnalité de la représentation, perte de corrélation avecles mots adjacents et la perte de la relation sémantique qui existe entre les termes d'un document. Pour surmonter ces problèmes, des méthodes de pondération des termes sont utilisées pour attribuer des pondérations appropriées au terme comme indiqué dans la matrice suivante :

$$\begin{pmatrix} T_1 & T_2 & \dots & T_{at}c_i \\ D_1 & w_{11}w_{21} & \dots & w_{t1}c_1 \\ D_2 & w_{12}w_{22} & \dots & w_{t2}c_2 \\ \vdots & \vdots & & \vdots \\ D_n & w_{1n}w_{2n} & \dots & w_{tn}C_n \end{pmatrix}$$

Où chaque entrée représente l'occurrence du mot dans le document, où « wtn » est le poidsdu mot « i » dans le document « n ». Puisque chaque mot n'apparaît normalement pas dans chaque document.

Il existe plusieurs façons de déterminer le poids « w11 ». Comme la pondération booléenne, « tf-idf », entropie etc.

Où tf-idf = tf *idf /

$Tf = \frac{n}{N}$ où n=Nombre d'apparition du terme t dans le document et

N =Nombre total de termes dans le document

$$idf = log \frac{|D|}{|\{d_j : t_i \epsilon d_j\}|}$$

|D|=nombre total de documents dans le corpus ;

$\{d_j : t_i \epsilon d_j\}$=nombre de documents où le terme t_i apparaît

D'autres méthodes diverses sontprésentées dans [41] comme :

- une représentation d'ontologie d'un document pour garder la sémantiquerelation entre les termes d'un document
- une séquence de symboles (octet, caractère ouWord) appelés N-Grams, qui sont extraits d'une longue chaîne dans un document.
- Latent Semantic Indexation (LSI) qui préserve les caractéristiques représentatives d'un document.
- LocalityPresercingindexing (LPI), découvre la structure sémantique locale d'un document. Mais n'estpas efficace en temps et en mémoire
- une nouvelle représentation pour modéliser les documents Web estproposée. Les balises HTML sont utilisées pour créer la représentation du document Web.

II.5.4 Évaluations de la performance

Les évaluations des classificateurs de texte sont généralementmenées de manière expérimentale plutôt qu'analytique. Elle se base généralement sur l'évaluation de l'efficacité d'unclassificateur, c'est-à-dire sa capacité à prendre les bonnes décisions de catégorisation.

Pour obtenir des estimations de précision et de rappel par rapport à l'ensemble de la catégorie, certaines autres mesures sont égalementutilisées comme seuil de rentabilité, mesure F, [40].

II.6 Algorithme de classification

II.6.1 Algorithme de Rocchio

L'algorithme d'apprentissage de Rocchio [42] a été conçu à l'origine pour utiliser le retour d'information sur la pertinence lors de l'interrogation de bases de données en texte intégral, créer un vecteur prototype pour chaque classement utilisant un ensemble de documents d'apprentissage, c'est-à-dire le vecteur moyen sur tous les vecteurs de documents d'apprentissage qui appartiennent à la classe « ci » , et calculer la similitude entre le document de test et chacun des vecteurs prototypes, qui attribuent le document de test à la classe avec une similitude maximale.

Ci = α * centroïde ci - β * centroïde ~ ci.

$$Q1 = \alpha Q0 + \beta\left(\frac{1}{|D_R|}\sum_{d\epsilon D_R} d\right) - \gamma\left(\frac{1}{|D_n|}\sum_{d'\epsilon D_n} d'\right)$$

Avec :

- D_Rl'ensemble des docs marqués pertinents par l'utilisateur
- D_nl'ensemble des docs marqué non pertinents par l'utilisateur
- $\alpha \geq \beta \geq \gamma$
- Valeur possibles $= 1\ \beta = 0.4\ \gamma = 0.2$, ou même 1 1 0

II.6.2 K-NN

Le classificateur K-NN est un algorithme d'apprentissage [43] basé sur une distance ou une similarité de fonction pour les paires d'observations, telles que la distance euclidienne ou la mesure de similarité cosinus

Cette méthode est essayée pour de nombreuses applications [44] En raison de son efficacité, non paramétrique etpropriétés faciles à mettre en œuvre, mais le temps de classification est long et il est difficile de trouver la valeur optimale de k. Le meilleur choix de k dépend des données généralement

Un bon k peut être sélectionné par diverses techniques heuristiques. Pour pallier cet inconvénient on peut modifier le KNN traditionnel avec différentes valeurs K pour différentes classes plutôt qu'une valeur fixe pour toutes les classes [45].

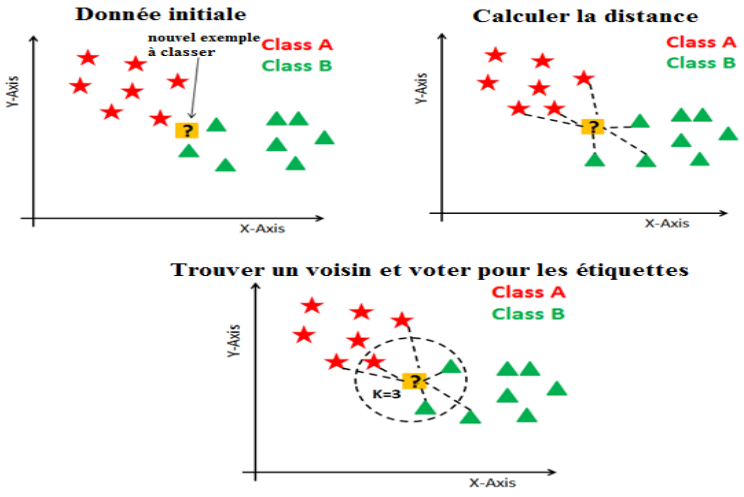


Figure 5- Classification par Knn.

II.6.3 Bayes naïves

La méthode de bayes naïf est une sorte de classificateur de module [46] sous probabilité et classe a priori connues (probabilité conditionnelle). L'idée de base est de calculer la probabilité d'appartenance du document Dà la classe C. Deux modèles d'événement sont présents [47] [48] [49] en tant que multivariéBernoulli et modèle multinomial. Sur ces modèles, le modèle multinomial est plus approprié lorsque la base de données est volumineuse, la base de ces méthodes est l'équation suivante :$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

II.6.4 Arbre de décision

L'algorithme se compose de nœuds internes d'arbre étiqueté par terme, les branches qui s'en éloignent sont étiquetées par test sur le poids, et le nœud feuille est étiquettes par classe correspondantes[51].

L'arbre peut classer le document figure -6- en parcourant la structure de la requête de la racine jusqu'à ce qu'il atteigne une certaine feuille, ce qui représente le but de la classification dudocument. La plupart des données d'entraînement ne rentreront pas dans la construction de l'arbre de décision de la mémoire[52].

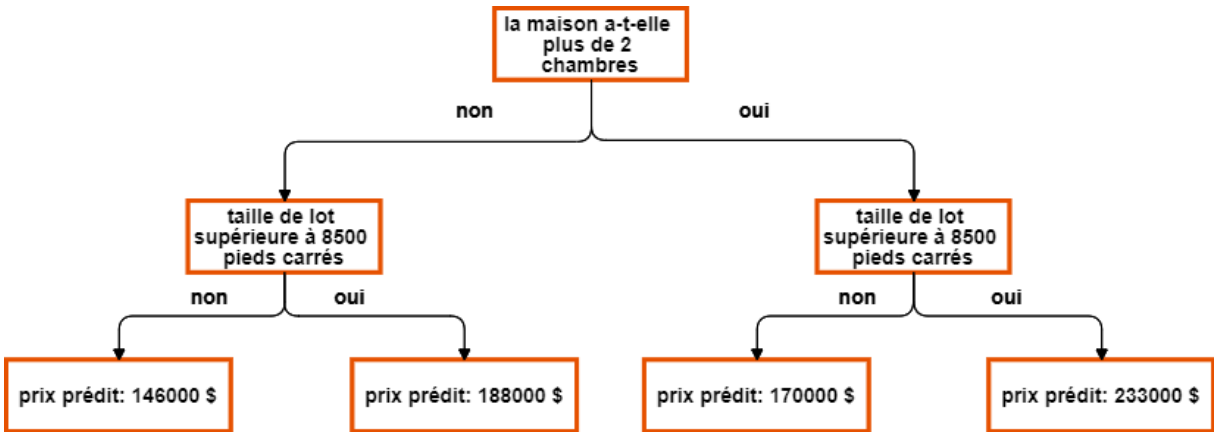


Figure 6-Exemple d'un arbre de décision.

II.6.5 SVM

Le SVM (Support vector machine) [53]a besoin d'un ensemble d'entraînement positif et négatif, ce qui est rare pour une autre méthode de classification.

Ces ensembles de formation positifs et négatifs sont nécessaires pour que le SVM recherche la surface de décision qui sépare le mieux les données positives des données négatives dans l'espace dimensionnel « n », appelé hyper plan[54].

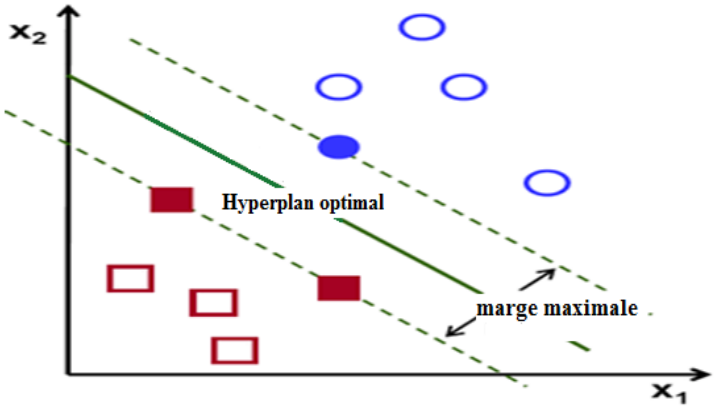


Figure 7-SVM classificateur.

II.6.6 Réseau neuronal

Un classificateur de réseau neuronal est un réseau d'unités, où les unités d'entrée représentent généralement des termes, l'unité de sortie représente la catégorie.

Pour classer un document de test, ses pondérations de terme sontassignées aux unités d'entrée ; l'activation de ces unités se propage vers l'avant à travers le réseau, et la valeur que les unités de sortie prennent en conséquence détermine la décision de catégorisation [56].

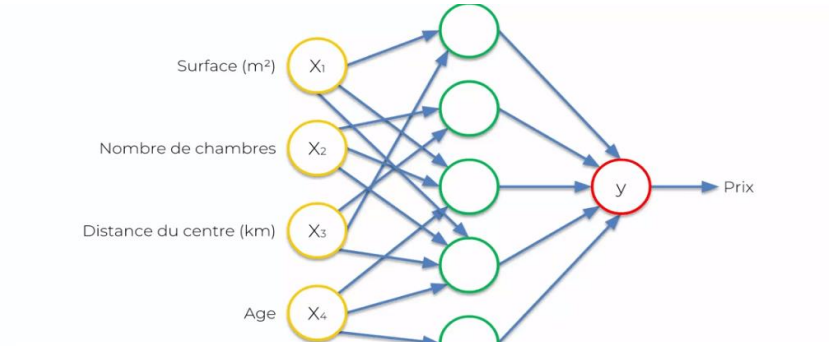


Figure 8-Exemple d'un réseau de neuronal.

II.6.7 LLSF

LLSF signifie Linear Least Squares Fit, c'est une approche de cartographie développée par Yang [57].

Les données d'apprentissage sont représentées sous la forme de paires de vecteurs d'entrée / sortie où le vecteur d'entrée est un document dans le modèle d'espace vectoriel conventionnel (composé de mots avec des poids), et le vecteur de sortisse compose de catégories (avec des poids binaires) du document correspondant. Fondamentalement, cette méthode est utilisée pour la recherche d'informations.

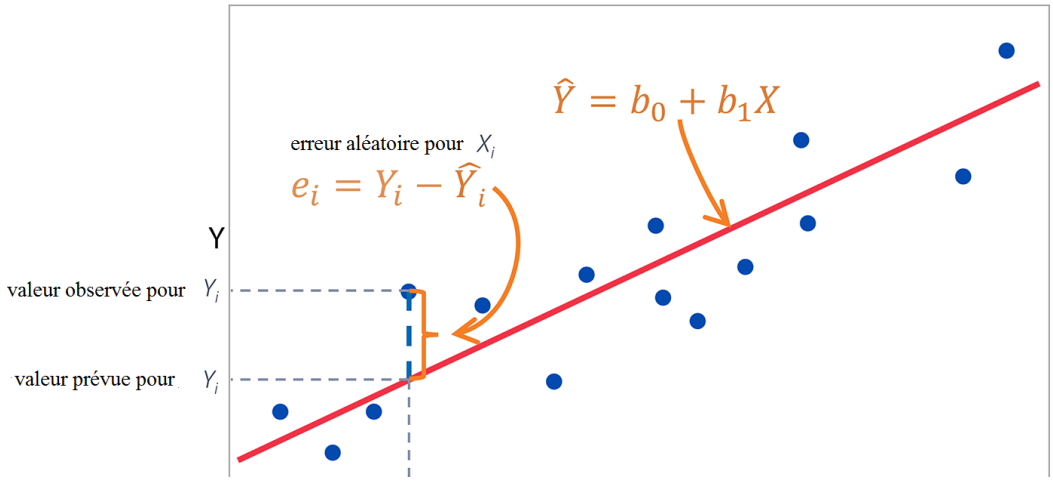


Figure 9-Classification par LLSF.

II.6.8 Vote

Cet algorithme est basé sur la méthode des comités de classification et est basé sur l'idée qu'une tâche donnéequi nécessite des connaissances d'opinion d'expert [58].

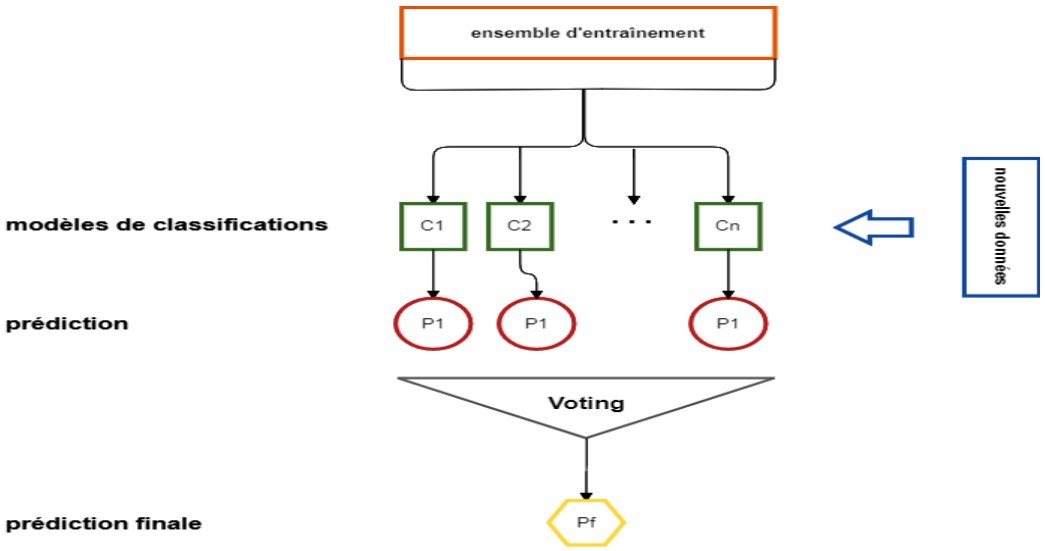


Figure 10-Fonctionnement de classification par Vote.

II.7 OBSERVATIONS COMPARATIVES

Les performances d'un algorithme de classification sont grandement affectées par la qualité de la source de données.

Les caractéristiques non pertinentes et redondantes des données augmentent non seulement le coût du processus d'extraction, mais aussiréduisent la qualité du résultat dans certains cas.

Chaque algorithme a ses propres avantages et inconvénients tels que décrits dans le tableau suivant :

Technique	Apprentissage	Avantage	Inconvénient
SVM	Supervisé	-très haute robustesse -moins de sur-	-incapable de la classification multi class

		apprentissage -Robuste au bruit	-couteux en calcule performance lente
Naïve bayes	Supervisé	-rapide dans la tâche d'entrainement et de classification -pas sensible aux caractéristiques non pertinentes	-suppose l'indépendance de la caractéristique. -moins précis que svm.
K-NN	Supervisé	-Entrainement très rapide. -Simple et facile à comprendre -gère bien les grands ensembles de donnés	-Biaisée par la valeur de k. -Haute complexité de calcul. -se laisse facilement tremper par des attributs non pertinents.
Arbre de décision	Supervisé	-Compréhensibilité du modèle.	-Instabilité de l'algorithme. -Sur-Apprentissages si l'arbre est trop profond.
Réseau de neuronal	Non Supervisé	-Bonne performance en pratique. -Théorème de l'approximation universelle.	-Incompréhensibilité du modèle. -Beaucoup de problème à optimiser.
Rocchio	Supervisé	- Facile à implanter. - Efficace pour des catégorisations ou un texte ne peut appartenir qu'à une seule catégorie	- n'est pas très efficace quand un texte peut appartenir à plusieurs catégories. - certains documents du corpus d'apprentissage appartenant à une catégorie Ci initialement ne seraient pas classés dans Ci par le classificateur.
LLSF	Supervisé	- outil principal de modélisation de processus en raison de son efficacité et	- limitations dans les formes que les modèles linéaires peuvent prendre sur

		de son exhaustivité. - bons résultats obtenus avec des ensembles de données relativement petits.	de longues distances, possiblement de mauvaises propriétés d'extrapolation et sensibilité aux valeurs aberrantes.

Tableau 2- Quelque avantages et inconvénient des algorithmes de classification

II.8 CONCLUSIONS

Les techniques de traitement du langage (NLP), d'exploration de données et d'apprentissage automatique fonctionnent un ensemble de données pour classer et découvrir automatiquement les modèles des différents types de documents.

La classification de texte (TC) est une partie importante de l'exploration de texte, elle a pour but de manuellement définir un ensemble de règles logiques qui convertissent les connaissances d'experts sur la façon de classer les documents sous l'ensemble de donné.

Dans le prochain chapitre nous verrons les travaux intérieurs liés à notre thème et connaitre les performances et les résultats obtenu pour chaque travail.

III. Chapitre IIITravaux dans le domaine de détection des tendances et classification des sujets

III.1 Introduction

Le suivi de la détection des sujets nécessite la réponse automatique de quoi ? Quand ? Où ? Et par qui ? Sont les sujets / événements / tendances populaires. Jusqu'à présent, aucune méthode ne répondait à toutes ces questions efficacement [23].la TDT est généralement utilisée pour détecter les comportements émergents ou suspects et pour une meilleure compréhension des préoccupations de la société [24]. La détection des événements repose principalement sur des techniques d'apprentissage automatique « apprentissage artificiel ou apprentissage statistique est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'apprendre à partir de données ».

Lorsqu'on parle des tendances, l'apprentissage non supervisé « désigne la situation d'apprentissage automatique où les données ne sont pas étiquetées. Il s'agit donc de découvrir les structures sous-jacentes à ces données non étiquetées. » Est préféré tandis que détection d'événements spécifiques repose principalement sur un apprentissage supervisé « est une tâche d'apprentissage automatique consistant à apprendre une fonction de prédiction à partir d'exemples annotés. » La détection des tendances est une tâche très liée à la détection des événements et est couramment appliquée aux réseaux sociaux. Une tendance utile et un outil d'analyse qui a été utilisé dans différentes disciplines, récemment. Il est clair que Twitter est devenu le lieu commun du TDT car il est considéré comme un réseau d'information en plus d'un réseau social.

Dans ce chapitre et grâce à Plusieurs études (Sutton et al. 2008 ; Kwak et al. 2010 ; Becker et al. 2011 ; Jianshu et Bu-Sung, 2011 ; Ozdikis et al. 2012a ; Ozdikis et al. 2012b) qui montre que Twitter est une source intarissable pour dégager les informations et les données. Tous ces facteurs nous ont encouragés à utiliser Twitter pour réaliser notre objectif en plus des travaux fait sur la détection des tendances.

III.2 La plateforme twitter

Twitter se démarque comme la plate-forme par excellence pour accéder ouvertement aux mises à jour en temps réel sur les dernières nouvelles et les événements en cours. Le public qui l'utilise est généralement jeune de 15 à 34 ans. Avec plus de 500 millions d'utilisateurs, C'est une plateforme de micro-blogging, vos posts sont limités à 160 caractères et symboles compris pour faire un message. Twitter voit un flux quotidien de plus de 400 millions de messages courts appelés tweets [25]. Ces tweets incluent des conversations en tête-à-tête et des bavardages englobant toutes sortes d'informations.

Avec des fonctionnalités diverses comme suite [26] :

- **Follow :** « Suivre » quelqu'un signifie qu'un utilisateur va avoir dans sa chronologie personnelle les tweets d'autres personnes (actualiser et mise à jour).
- **Mentions d'un utilisateur :** lorsqu'un utilisateur mentionne un autre utilisateur dans son tweet, un signe @ est placé avant le correspondant au nom d'utilisateur, par exemple @username,
- **Réponses :** lorsqu'un utilisateur souhaite répondre à un tweet antérieur, il place le @username pour mentionner un utilisateur
- **Retweets :** un retweet est considéré comme un repartage d'un tweet publié par un autre utilisateur peut être partagé à plusieurs niveaux.
- **Hashtags :** similaires aux balises sur les systèmes de balisage les hashtags inclus dans un tweeta tendance à regrouper des tweets dans des conversations ou à représenter les principaux termes du tweet, généralement appelés sujets ou intérêts communs d'une communauté. Un hashtag est différencié du reste des termes du tweet par exemple, #hashtag.

III.3 La tendance sur twitter

L'un des phénomènes attrayants du service de micro blogage est le fait que certaines occurrences de nouvelles d'un large intérêt dans les communautés produit une augmentation soudaine des mentions en temps réel au fur et à mesure de leur déroulement , les utilisateurs donnent lieu à des tendances lorsqu'ils partagent des intérêts communs, qui peuvent être déclenchés par différents raisons dans la plupart

du temps ils se manifeste par un #hashtag qui produit ce qu'on appelle une tendance sociale. Les tendances ne comprennent qu'un ensemble de termes qui sont mentionnés de façon remarquable, Quatre types de tendances sont distingués sur twitter selon l'étude d'ArkaitzZubiaga, Damiano Spina [26] :

Nouvelles : les dernières nouvelles ont tendance à arriver sur Twitter très tôt, ce qui a même montré à maintes reprises que les nouvelles passent d'abords par twitter avant que les médias ne le signalent

Événements en cours : un autre type de sujet tendance à propos d'un événement en cours au fur et à mesure qu'il se déroule par exemple la coupe d'Afrique

Mêmes : une partie des sujets de tendances étaient déclenchés par des idées virales un individu ou une organisation, qui était généralement assez populaire pour pouvoir diffuser quelque chose de drôles ou attrayant à grande échelle.

Commémoratives : le dernier type de sujet tendance et probablement le moins fréquent, était celui produite par une commémoration d'une certaine personne ou d'un événement dont on se souvient dans une journée donnée.

III.4 Travaux dans le domaine de détection des tendances et classification des sujets

On va diriger notre travail bibliographique vers deux chemins donc Les travaux vont se concentrer sur la détection de sujets émergents et leur classification a différent niveau et fonctionnalité. Les différences entre ces travaux sont nombreuses a commencé par le public de ce dernier est différent puisque certains travaux visent à aider les scientifiques des données alors que d'autre visent à informer l'utilisateur final ; certains renvoient un ensemble de documents comme tendances et autres un ensemble de mots clés et certaine ce concentre sur la détection de concept spécifique tandis que d'autres sont des outils de domaine ouvert ; et il y'a les classement par types et d'autre part sujet quelques outils supporte la visualisation tandis que d'autres fournissent des composants d'analyse supplémentaires et enfin certains perçoivent la détection le classement en temps réel et d'autres en batch . À cause de ces différences entre les travaux, il est difficile de les comparer car il n'existe pas de référence ou mesure de la qualité.

Les deux tableaux suivants présentent une catégorisation des travaux en fonction de l'année de leur création, des types de service qu'ils offrent, les approches et techniques qu'ils utilisent.

III.5 2.1. Détection des tendances

Le tableau suivant résume et catégorise les travaux les plus importants faits sur la détection de tendance sur tweeter :

Outil	**TwitterMonitor[28]**	**Cloud4Trends[29]**	**TweCom [27]**
Année	2010	2012	2013
Le Service de détection	Sujets généraux	Tendances locales	Propagation des tendances
L 'approche	Fonctionnalité	Fonctionnalité / document	Fonctionnalité
Techniques	-Basé sur le clustering de co-occurrentes	-Basé sur le clustering de co-occurrentes	-Ontologies sémantiques -Genio et RuleGen -Hiérarchique Regroupement -Règle d'association Exploitation minière
En temps réel	Oui	Oui	non
DSN/ DSS	Oui/non	Oui/non	Oui/non
RSN	Twitter	Twitter	Twitter

Tableau 3-Comparaison des différents outils de détection des tendances.

Outil	**Politwi [30]**	**Sociopédia [31]**	**TDT FTR [32]**	**TDT AA [33]**	**TDT I [34]**
Année	2014	2015	2016	2017	2018
Le Service de détection	Sujets politiques Allemand et leur polarité des sentiments	Événements spécifiques	Événements en fenêtres horaires réglable	Détection de sujets tendances en les classant par sujet	Tendance des sujets détection des tweets indonésiens
L 'approche	Fonctionnalité	/	En fonction des caractéristiques	/	document
Techniques	Analyses statistiques	- Ontologies Sémantique - Une analyse Statistique	-Analyse de complexité -Clustering basé sur cooccurrences	L'analyse du texte des médias Fuzzy C-means le cluster-classificateur basé sur naïve bays	-Clustering basé sur cooccurrences
En temps réel	Oui	Non	Oui	Non	non
DSN/ DSS	Oui/non	Oui/non	Oui/non	Oui/non	
RSN	Twitter	Twitter	Twitter	Twitter	twitter

Tableau 4-Travaux réalisés sur la détection des tendances.

- Le champ 1« Année » : fait référence à l'année de création de l'outil.

- Le champ 2 « le service de détection » : illustre le service fourni par l'outil.

- Le champ 3 « L 'approche » : il y'a deux approches principales relier a des caractéristiques temporelles :

-Fonctionnalité-Pivot

-Document-Pivot

Et il y'a aussi les approches hybrides ;
la premier, Fonctionnalité utilise généralement la "burst détection", c'est la détection de mots clés surgis de façon très rapide et en nombre, burst voulant dire éruption, comme Twitter qui présente les tendance en utilisant la fréquence d'utilisation des termes, quand un terme "érupte" il est présenté dans les tendances et pour Le document-pivot est basé sur la similarité textuelle, comme par exemple dans les moteurs de recherche, l'utilisation de certain terme ou certaines phrases similaires de façon nombreuse fait remonter le termes dans les tendances
- Le champ 4 « Techniques » : montre les techniques d'analyse spécifiques utilisées pour développer l'outil.
- Le champ 5 « En temps réel » : indique si l'outil relève le défi de la détection de sujets en temps réel. Dans
Analyse des tendances il s'agit d'une exigence fortement souhaitée.
- Le champ 6 « DSN » et « DSS » : représentent respectivement la détection de sujets non supervisés et supervisés
- Le champ 7 « RSN » : désigne le réseau social numérique pour lequel chaque outil est conçu et testé.

III.5.1 Twitter Monitor

Est l'un des premiers travaux dans le domaine de la détection de sujets émergents sur Twitter [28]. Les chercheurs proposent des algorithmes d'éclatement et de clustering qui sont implémentés dans le noyau application. L'analyse des tendances est effectuée en identifiant les mots clés en rafale (recherche uniquement les mots clés les plus utiliser) ou des mots clés qui sont souvent rencontrés dans les mêmes tweets avec le basé sur les co-occurrentes de mots clés. Plus précisément, étant donné les mots-clés groupés en sous-ensembles disjoints {Kti}, une tendance est identifiée par un seul sous-ensemble Kti, où Kt représente un ensemble de mots-clés éclatés

calculés à chaque instant t et k ∈Kt et Kti est le sous-ensemble de Kt . Concernant le clustering, l'historique de quelques minutes des tweets est récupéré pour chaque burst et les mots clés qui coexistent dans un nombre relativement important de tweets récents sont placés dans le même groupe. Le système applique les caractéristiques contextuelles des tweets pour fournir une description de chaque tendance et une interface utilisateur interactive, où un utilisateur peut classer et soumettre leur propre description, est également disponible.

III.5.2 Cloud4trends

Détecte également les tendances via l'exploitation de la fréquence des mots clés TF-IDF et spécifiquement en attribuant plus de poids aux termes dans les titres et les balises des articles [29]. au lieu d'appliquer une méthode basée sur un seuil fixe qui se définit comme inactive après une période de temps prédéfinie, il observe dynamiquement le taux de mise à jour des clusters et peut identifier les tendances à leur apogée et détecter les sujets qui ne sont plus tendance. Cloud4Trends collecte et regroupe séparément les tweets qui se rapportent à une zone géographique et prend en compte l'emplacement physique de l'utilisateur respectif. La collecte et le traitement des flux pour les différentes zones géographiques simultanément offrent une véritable analyse rapide. La collecte des données à partir de trois sources différentes à savoir les tweets, les blogs et les tweets étendus et la traite dans le Cloud à l'aide du paradigme MapReduce.

III.5.3 TweCom

Est un cadre d'exploration de données pour étudier les tendances les plus pertinentes en termes de propagation du contenu [27]. Il extrait les tweets avec un robot « ad hoc » et fournit des relations / règles sur le contenu et le contexte. Pour générer des taxonomies à partir du contenu des publications et du contexte, des fonctions (spatio-temporelles) de regroupement et d'agrégation hiérarchiques ont été utilisées. Pour chaque groupe on a un mot clé caractérisé par la valeur TF-IDF la plus élevée. L'outil extrait les relations entre les tweets via l'exploration de règles d'association généralisées. Ce dernier est utilisé lorsque la sémantique est requise. Une règle d'association est une implication où X et Y sont des ensembles d'articles, alors qu'en association généralisée, A et B sont des ensembles d'articles généralisés disjoints, à savoir attributs en commun. L'extraction des règles d'association généralisées s'effectue au moyen d'un processus :

Étape1 : extraction fréquente d'ensembles d'articles généralisés via l'algorithme Genio

Étape2 : la règle génération des ensembles d'éléments fréquents extraits via l'algorithme RuleGen. Ce dernier appartient à l'Algorithmes CART et déterminer les relations statistiques entre de nombreuses couches de données

Étape 3 : un arbre de décision binaire. Le classement et la sélection des règles les plus utiles sont limités par le schéma de règle (c'est-à-dire les attributs qui doivent apparaître dans le corps ou l'en-tête de règle), ou une règle spécifique éléments d'intérêt.

Les analystes peuvent ensuite appliquer des requêtes détaillées ou cumulées pour étudier l'évolution temporelle et la répartition géographique de termes spécifiques. Notez que le clustering hiérarchique produit un ensemble de grappes imbriquées organisées comme un arbre, appelé dendrogramme, sur les données et dans ce cas, il est utilisé pour découvrir les relations hiérarchiques entre les mots clés. Les chercheurs utilisent l'approche agglomérative où chaque observation commence dans son propre cluster, et des paires de clusters sont fusionnées au fur et à mesure hiérarchiquement.

III.5.4 Politwi

Est un outil disponible sur Twitter [30], sur le site Web et sur l'application smartphone pour détecter les toptweets politiques allemandes qui sont filtré par heures et jours. L'idée de base de l'approche TDT consiste à comparer le nombre actuel de tweets avec hashtag au nombre de tweets de la période précédente en tenant compte de l'écart type et en utilisant la distribution gaussienne. Pour la mesure, un sujet de tendance se caractérise par une apparition beaucoup plus élevée dans les périodes précédentes. Le graphe des relations contient les mots les plus fréquemment rencontrés à des moments précis ; construit avec chaque hashtag (nœud) pour être entouré par des liens avec des mots connectés (nœud) utilisés dans le contexte actuel avec une polarité prédite pour chacun. et peut être utilisé pour étendre les bases de connaissances existantes pour répondre à des questions telles que « Quelle polarité est actif, sujet à venir, toujours dans un contexte politique ? ».

III.5.5 Sociopedia

Est un système différent pour analyser les sujets des médias sociaux [31]. Il construit automatiquement une ontologie sémantique basée sur un mot-clé donné. Les nœuds

en ontologie sont des entités extraites des tweets les plus recherchés et les relations sont déduites par le biais de documents connexes sur Wikipédia et DBpedia. Le système comprend une analyse de récapitulation des requêtes, une détection de comparaison estconstruite grâce à leur distribution de fréquence des modèles de mots et également une analyse de sentiment menée à travers le lexique AFINN. Pour mieux illustrer ça, la présence du mot (versus ou comparer a) peut indiquer une comparaison et la présence de 5W1H (what (quoi), where(ou), who(qui), why (pourquoi) whether (si), how (comment)) est un indicateur d'une requête.

III.5.6 TDT FTR (Détection sur Twitter des Événement basé sur une fenêtre temporelle [32])

Étudient le problème de la détection d'événements, en temps réglable à travers des fenêtres. Leur système permet aux scientifiques des données de savoir comment un événement brûlant, est arrivé et s'est développé au cours des 120 dernières minutes, et ce qui s'est passé au cours des 60, 30 et 10 dernières minutes. Pour détecter les événements, ils utilisent des uni-grammes comme termes pour chaque nouveau tweet, effectuant un travail sur n-grammes à la fois pour avoir plus d'efficacité et d'efficience. Ils détectent les événements grâce à la détection des anomalies, à savoir qu'ils traitent chaque nouveau tweet et stockent leurs statistiques (nombre de retweets, nombre de tweets par minute, nombre d'utilisateurs et nombre d'utilisateurs qui les retweets) et identifier les termes anormaux la fin de chaque fenêtre de temps. Le clustering est basé sur les co-occurrentes et la sélection est sur la base des clusters les mieux classés. Ils conçoivent une structure de données d'arbre ST pour prendre en charge la fenêtre de temps réglable.

III.5.7 TDT AA (Détection de sujets tendances à l'aide de l'approche d'apprentissage automatique [33])

Dans ce travail, les chercheurs utilisent des techniques basées sur l'analyse du texte des médias. Trois choses sont nécessaires d'abord les données de twitter, création d'un modèle de sujet et évaluation du texte. Des techniques de transformation sont appliquées pour améliorer la qualité des données brutes regrouper à partir de comptes d'utilisant Twitter grâce à Hadoop, apachStrom « est une sorte d'utilisateur tiers une bibliothèque configurée à l'aide de Hadoop cette API aide à pomper les tweets de la base de données twitter à tout système de fichiers Hadoop local et téléchargé dans le répertoire HDFS à l'aide d'un simple fichier texte. »

Et twitter API « c'est une bonne source de données pour l'analyse de texte et sujet tendance pour capturer les tweets en direct. ».Le modèle est conçu de telle manière que les données peuvent être utilisées en ligne et hors ligne. Après vient le prétraitement qui est une action clé apprentissage automatique

- Supprimez les mots vides
- Supprimez les caractères spéciaux
- Calcul du terme fréquence

  TF= (nombre total de fois qu'un mot apparait)/ (Quantité totale de mots dans le tweetsélectionné)
- Probabilité de formation de phrases

  SP= (Mot apparait dans le nombre de phrase)/ (Phrases totales dans les tweets téléchargés)

Un seuil est utilisé comme longueur de 20 processus rend les tweets de longueur régulière.

Après ces étapes on vient aux fonctionnalités calculées l'algorithme FCM Une approche non superviser basé sur les concepts K-means pour partitionner l'ensemble de données en grappes. L'algorithme FCM est une méthode de regroupement « douce » dans laquelle les objets sont assignés aux clusters avec un certain degré de croyance. Par conséquent, un objet peut appartenir à plus d'un cluster avec différents degrés de croyance.

Pseudocode FCM :

Entrée : Compte tenu de l'ensemble de données, définissez le nombre souhaité de Clusters c, le paramètre flou m (une constante> 1), et la condition d'arrêt, initialiser la partition floue matrice et définissez stop = false.

Étape 1 : Faire :

Étape 2 : Calculer les centroïdes de cluster et l'objectif de la valeur J.

Étape 3 : calculez les valeurs d'appartenance stockées dans la matrice.

Étape 4 : Si la valeur de J entre des itérations consécutives est inférieur à la condition d'arrêt, alors stop = true.

Étape 5 : Pendant (! stop)

Sortie : une liste de centres de cluster c et une matrice de partition sont produits.

Une approche non supervisé basé sur les concepts K-means pour partitionner l'ensemble de données en grappes. L'algorithme FCM est une méthode de

regroupement « douce » dans laquelle les objets sont assignés aux clusters avec un certain degré de croyance. Par conséquent, un objet peut appartenir à plus d'un cluster avec différents degrés de croyance.

Pseudocode FCM :

Entrée : Compte tenu de l'ensemble de données, définissez le nombre souhaité de Clusters c, le paramètre flou m (une constante> 1), et la condition d'arrêt, initialiser la partition floue matrice et définissez stop = false.

Étape 1 : Faire :

Étape 2 : Calculer les centroïdes de cluster et l'objectif de la valeur J.

Étape 3 : calculez les valeurs d'appartenance stockées dans la matrice.

Étape 4 : Si la valeur de J entre des itérations consécutives est inférieur à la condition d'arrêt, alors stop = true.

Étape 5 : Pendant (! stop)

Sortie : une liste de centres de cluster c et une matrice de partition sont produits.

Crée des groupes similaires de tweets, selon les domaines sélectionnés initialement. Pour trouver les sujets d'actualité enfin le cluster-classificateur applique La règle algorithmique de classification Naïve Bayes C'est un classificateur probabiliste Il est basé sur des modèles de probabilité et incorpore des hypothèses d'indépendance solides dans un cadre d'apprentissage très superviser $P\left(\frac{H}{X}\right) = \frac{P(\left(\frac{X}{H}\right)P(H)}{P(X)}$

Où, X est un tuple de données et H est une hypothèse d'analyse.

. Les deux modèles sont utilisés pour améliorer la performance et précision de l'analyse des données. La mise en œuvre du concept proposé est donnée en utilisant la technologie développée en JAVA.

III.5.8 TDT I (Tendance des sujets détection des tweets indonésiens à l'aide BN-grammes et Doc-p [34])

Un travail qui présente deux méthodes pour détecter les tendances de sujets de tweets Les méthodes comparées sont, la méthode de pivot de document et les BN-grammes Les types de n-grammes utilisés dans cette étude sont uni-grammes jusqu'à six-grammes et se compose de trois étapes Étape 1. Calcul de DF-IDF Pour chaque n-gramme extrait de la collection de tweets, son DF-IDF « Un n-gramme est un mot généralisé composé de n-grammes consécutifs »

$$DF - IDF_t = \frac{df_i + 1}{log\left(\frac{\sum_{j=i}^{t} df_{i=j}}{t} + 1\right) + 1}.boost$$

Oùdf i est la fréquence des occurrences de n-grammes dans certains tweets à intervalle de temps i, df i-j est la fréquence des occurrences de n-grammes dans certains tweets dans les créneaux horaires ij précédents, et t est le nombre de toutes les fenêtres de temps. Le score de boost est le score de certains termes qui peuvent être classés « une personne, un lieu ou une organisation »

Étape 2. Cluster N-grammes

La fusion de quelques n-grammes pour devenir des grappes donne plus de faits, des informations complètes et fiables sur le sujet tendance.la fusion des n-grammes est effectuée en utilisant un regroupement hiérarchique Les N-grammes sont classés en grappes selon leur distance

$$d(g_1, g_2) = 1 - \frac{A}{min\{B, C\}}$$

Où est la distance entre n-grammes g 1 et g 2, A est le nombre de tweets contenant n-grammes g 1 ou g 2, et B et C sont le nombre de tweets contenant n-grammes g 1 et n-grammes g 2, respectivement.

Étape 3. Classement des sujets

Chaque cluster représente un sujet ou un événement qui se produit dans des médias sociaux. Les clusters sont classés en fonction de leur score de DF-IDF t. Le cluster qui contient n-grammes avec le haut Le score le plus élevé de DF-IDF t représente le sujet le plus discuté

, qui est une méthode de pivot de fonctionnalité, pour cela ils ont utilisé Jeux de données de six ensembles de données, à savoir P1, P2, P3, P4, P5 et P6, comprenant chacun 6 630, 21 306, 74 790, 5327, 807 et 2527.Le deuxième objectif est de décrire un sujet tendance en détail car ils apparaissent sur Twitter avec des hashtags et des mots clés ; et ça ne fait que les rendre plus difficiles à comprendre. Par conséquent, les recherche ont été menées pour faire un résumé des groupes de hashtags pour générer des informations plus détaillées sur Twitter ,en utilisant la combinaison du TF-IDF et du renforcement des phrases visaient à fournir une description complète d'un sujet tendance généré sur la base de trois étapes principales, à savoir la catégorisation des sujets (en utilisant le cosinus similitude), l'extraction de phrases (en utilisant TF-sommaire et hybride TF-IDF), et le clustering de phrases (en utilisant TF-IDF et la méthode de la distance).

Les expériences montrent que la détection des sujets tendances est plus précise lors de l'utilisation de BN-grammes que Doc-p dans les trois ensembles de données. Cependant, pour les mots clés précision, Doc-p est meilleur que BN-grammes

III.5.1 Classification

Le tableau suivant résume et catégorise les travaux les plus importants faits sur la classification :

Outil	**Classification en temps réel [26]**	**Prédire la popularité des tendances en arabe [34]**	**FuzzyFingerprints [35]**
Année	2014	2014	2014
Le Service de classification	Classer les toptweet en type de tendance	Classer chaque tweet avec l'article de presse correspondant pour ensuite prédire la popularité de l'information en arabe	Classer les tweet avec leur thème approprié
L 'approche	one-against-all	-arbre de décision -Naïve Bayes NB -W-JRIP	L'algorithme Fingerprints
Techniques	-Analyse comportementale en ligne - cooccurrences	lightstemming, N-grammes et sac de mots	- Ontologies Sémantique - Une analyse Statistique
En temps réel	Oui	Non	Non
DSN/ DSS	Non/oui	Oui/oui	Non/oui
RSN	Twitter	Twitter	Twitter

Tableau 5-travaux sur la classification en exploration de texte.

III.5.1.1 Classification en temps réel des tendances Twitter

Dans ces travaux les chercheurs essayent de classifier les types de déclencheurs qui déclenchent des tendances sur Twitter [34], suivant ses quatre types : nouvelles,

événements en cours, mêmes et commémoratifs et en temps réel. Les schémas sociaux observés sur le comportement des utilisateurs en termes de diffusion de l'information varieront. Ils ont défini un ensemble de 15 caractéristiques sociales pour aider à capturer ce comportement social en plus de l'analyse textuelle. Et analyserons ces caractéristiques et identifierons les modèles de comportement afin de catégoriser avec précision les tendances en temps réel. Ces 15 ensembles seront divisés en deux

- Calculer avec le nombre moyen d'occurrences moyenne calculée comme la moyenne arithmétique est le résultat de la division du nombre d'occurrences correspondant $AM(f)_t = \frac{1}{|T|}\sum_{i=1}^{|T|} f_i$Où AM (f) t est la moyenne arithmétique de la caractéristique f pour le sujet de tendance t, | T | est le nombre de tweets dans le sujet tendance, et f i est la valeur de la fonction f pour le tweet i.

De fonctionnalités dans les tweets correspondant à une tendance, 10 fonctionnalités différentes qui sont reposent par cette méthode de calcul Retweets (profondeur), Retweets (ratio), Hashtags, Longueur, Exclamations, Questions, Liens, répétition Sujet, Réponses. Sauf la vitesse de propagation: $\boldsymbol{AM(sv)_t = \frac{|T|}{\Delta t}}$Où Δt = est le nombre total de secondes du premier au dernier tweet dans les sujets de tendances.

- Calcule La diversité et la variation: $H'(f)_t = -\sum_{j=1}^{s}(p_{jt} ln\, p_{jt})$; $p_{jt} = \frac{n_{jt}}{N}$Où H ′ (f) t est l'indice de Shannon de la caractéristique f pour le sujet de tendance t, n jt est la population de la valeur j, S est le nombre de valeurs différentes, N est la population totale et p jt est la probabilité observée de la valeur j.

De la fonctionnalité tout au long d'un sujet de tendance. Cette méthode est utilisée pour les 5 fonctionnalités restante c'est-à-dire diversité des utilisateurs, la diversité utilisateur Retweeté, diversité Hashtag,la diversité de la Langue et du vocabulaire.

Cette étude a permis de comprendre si et pourquoi les fonctionnalités utilisées discriminer les types de sujets tendances. Au-dessus de ça on ajoute une analyse de texte dans les tweets Qui donnera un aperçu des termes les plus utilisés dans chaque type de sujet tendance, c'est-à-dire le vocabulaire le plus fréquemment utilisé dans chaque type. Ils sont passés par la suppression des mots vides et la normalisation Après cela, le calcule de TF (fréquence du terme) de chaque mot pour chaque type de ce processus a produit un sac de mots pour chaque type de sujet tendance comme par exemple les sujets de tendances commémoratives, les mots de félicitations

ressortent « joyeux et anniversaire », ainsi que des mots liés au temps « jour et années ».

III.5.1.2 Classification automatique des sujets tendances

Pour ce travail, Étant une tache multi-classe elle va s'appuyer spécifiquement sur une méthode de combinaison binaire one-against-all « un contre tous ». Au lieu de considérer le problème multi-classe comme une tâche unique, un contre tous le divise en plusieurs petites combinaisons binaires. Pour un problème avec k classes, one-against-all définit k classificateurs différents. Dans la phase de formation, chacun des k classificateurs apprend un modèle pour séparer une classe du reste k-1. Ce modèle crée un hyperplan pour séparer une classe du reste. Dans cette tâche, où 4 classes sont définies, les classificateurs suivants sont créés : 1 vs 2-3-4, 2 vs1-3-4, 3 vs 1-2-4 et 4 vs 1-2-3. Dans le processus de catégorisation, chaque classificateur fournit une sortie pour chaque tendance, qui fait référence à la marge i.e. la distance à l'hyperplan comme valeur de fiabilité. Le classificateur maximise la sortie finale prédite par le système.

Twitter utilise un algorithme qui affiche la liste des 10 premiers termes les plus discutés dans un moment précis. Cette liste des 10 principaux termes qui est également disponible via son API de recherche qui s'applique à tous les tweets publics publiés, grâces a ça ils peuvent obtenir et collecté jusqu'à 1 500 des tweets les plus récents un total de 1 036 sujets de tendances uniques. Ces tendances comprennent un total de 567 452 tweets de 348 757 utilisateurs différents, Quatre personnes, étant des utilisateurs quotidiens de Twitter, ont joué le processus d'annotation, attribuer la catégorie la mieux adaptée.

III.5.1.3 Prédire la popularité des tendances en arabe sur Twitter

Le but de cette étude est de construire un modèle qui peut prédire la popularité des articles de presse sur Twitter en classant leurs fonctionnalités [35], en utilisant trois algorithmes pour l'exploration de données : arbre de décision, Naïve Bayes NB et règle basé sur W-JRIP. Quatre approches ont été utilisées pour comparer les caractéristiques extraites : light stemming, N-grammes et sac de mots. Cette étude élargit la littérature limitée sur la classification des textes en arabe et fournit un

modèle prédictif qui peut aider les organisations de presse améliorer leur contenu en ligne. Cette recherche comportait deux phases principales :

La formation et les tests. Le plus gros travaille se concentre sur le jeu de donnée, la collecte des articles de différentes sources site web d'information en les traitant et classant par catégorie, après ça crée une base teste en étiquetant les tweets relatifs à l'article collecté. Ils ont appliqué les algorithmes d'exploration de données et L'arbre de décision à la plus haute performance

III.5.1.4 Détecter le sujet d'un Tweet dans un grand nombre Des tendances Twitter portugais

TwitterTopicFuzzyFingerprints est Une technique consiste à décider si un tweet donné est lié à un sujet #hashtag donné pour le contenu de microblogs Fondamentalement[36], cela peut être classé comme un problème de classification de texte avec un nombre inconnu et important de catégories cette méthode est inspirée d'une autre méthode de classification de livres ou d'articles a des auteurs connus en utilisant l'analogie des empreintes digitales en partant du fait que chaque auteur a un style d'écriture donné en donnant une importance d'établir le parallèle entre le contexte de la propriété d'auteur et Tweet détection de sujet .L'algorithme lui-même et diviser en quatre étapes :

Etape 1 : Rassembler les k « la fréquence des mots dans tous les textes connus de chaque auteur connu » de la même manière l'auteur devient un #hashtag donc Pour chaque tweet, il reconnaît l'existence du # et ajoute chaque mot du tweet à une table #topic en rajoutant son compteur d'occurrences.

Etape 2 : Construire l'empreinte digitale en appliquant une fonction de fuzzifying sur la liste des top- k. Le flou l'empreinte digitale est basée non seulement sur la valeur de fréquence d'occurrence sur les tweet mais aussi en raison de la petite taille de chaque tweet, ses mots doivent être aussi uniques que possible afin de faire les empreintes digitales qui se distingue parmi les différents sujets pour cela la technique de fréquence de document inverse (idf):

$idf = log\frac{N}{n_i}$Où N devient la taille de la bibliothèque d'empreintes digitales du sujet (c.-à-d. le nombre total de sujets), et n i devient le nombre de # sujets où le mot est présent a était adaptation, visant à réduire l'importance des termes fréquents qui sont communs à plusieurs sujets, tels que « suivre », « RT » et « like »

Etape 3 : Effectuez les mêmes calculs pour le texte identifié puis comparez les résultats obtenus avec toutes les empreintes floues d'auteur disponibles. Les plus similaires sont choisies et le texte est attribué à l'auteur de l'empreinte digitale. La même approche que la méthode originale, et utilisez avec l'équation d'adhésion l'équation d'adhésion :$T2S2(\Phi, T) = \frac{\sum_v \mu_\Phi(v):v\varepsilon\ (\Phi \cap T)}{\sum_{i=0}^{j} \mu_\Phi(w_{\mathrm{i}})}$

La fonction T2S2 est une fonction améliorée selon les paramètres en main Φ est l'empreinte digitale #topic, T est l'ensemble des mots du tweet (prétraité), μ Φ (v) est le degré d'appartenance du mot v dans l'empreinte digitale du sujet, et j est le nombre des caractéristiques du tweet.

Twitter api méthode « GET Trends / place » de Twitter DEV, offre le top 10 tendances des sujets du moment dans un lieu donné, avec ces données La technique a était comparé avec deux techniques populaires de classification de texte, (SVM) et (KNN). Les résultats préliminaires montrent qu'il surpasse les autres deux techniques, tout en étant beaucoup plus rapide, ce qui en fait une solution pour le traitement à la volée des flux de Big Data.

III.6 Conclusion

Après cette synthèse faite sur les travaux connexesdans notre travail ,nous constatons que la première étapes c'est-à-dire la collecte est le prétraitement des données et très importante pour le travail et les résultats obtenus ; on constate aussi qu'à chaque objectif et chaque perspectif un algorithme qui sera plus performant que d'autre (supervisé , non supervisé……etc.) mais dans la majorité des cas l'hybridation et la personnalisation des algorithmes nous permet d'obtenir de meilleures performances. Dans le chapitre suivant on va modéliser et expliquer en détail notre travail avec les résultats et les outils utiliser.

IV. Chapitre IVDétection de tendances des réseaux sociaux en utilisant les techniques du TALN

IV.1 Introduction

Les réseaux sociaux connaissent une explosion en termes de volume de données et en fonction du nombre d'utilisateurs à travers le monde suite au progrès du web et des capacités de stockage et d'échanges internet. Plusieurs recherches ont montré que les données publiées par les internautes sur les sites de médias sociaux, notamment Twitter, reflètent presque en temps réel l'intérêt du public.

L'utilisation quotidienne des réseaux sociaux tel que Twitter a changé l'image du web 2.0 et lui a donné une nouvelle dimension et de nouveaux défis le nôtre sera détailler en passant par la modélisation jusqu'à l'implémentation dans le chapitre qui va suivre.

IV.2 Architecture globale du système

Nous commençons par collecter les données dont nous avons besoin puis appliquer un prétraitement pour les nettoyer après ça, les classer par catégorie et sous-catégories tout en mettant en évidences les sujets tendances pour chaque catégorie et la tendance globale du moment en utilisant les algorithmes de classification les plus performant le schéma suivant donne un aperçu sur l'architecture générale de votre système.

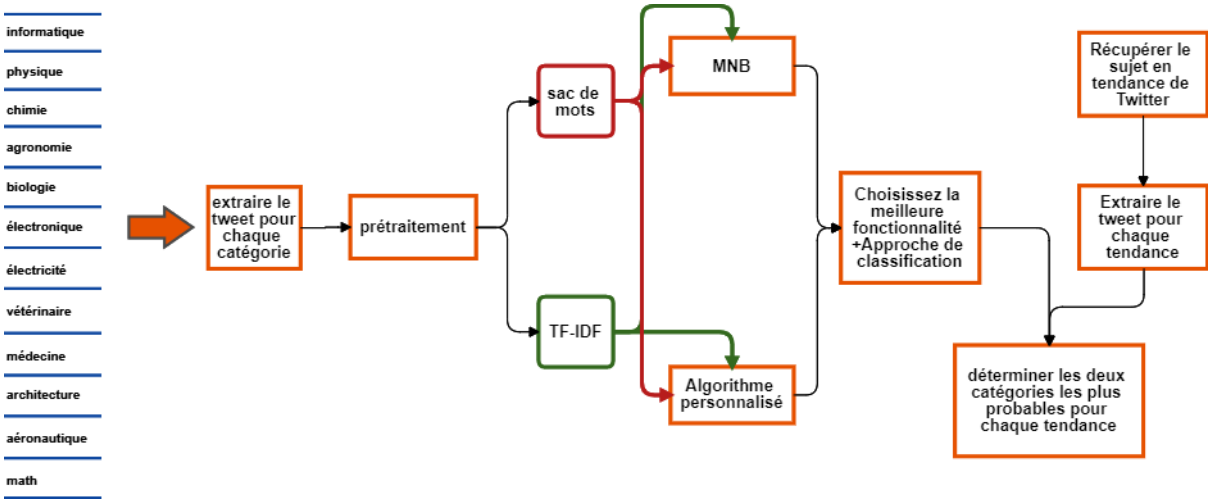


Figure 11-Architicture du travail.

IV.3 Description du Dataset

Twitter permet d'interagir avec ses tweets de données et plusieurs attributs sur les tweets en utilisant les API Twitter. Ils ne sont accessibles que via des requêtes authentifiées, L'accès aux APITwitter est également limité à un nombre spécifique de demandes dans une fenêtre de temps appelée limite de débit. Ces limites s'appliquent aussi bien au niveau de l'utilisateur individuel qu'au niveau de l'application.

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

Figure 12-Utilisation de Tweepy.

IV.3.1 Caractéristiques des données :

Cetteétape nous permet d'obtenir les données d'apprentissage pertinentes à notre système. Le corpus est constitué d'un ensemble de tweets homogènes sur le fond (se reliant tous au thème de la science et de la recherche) et la forme (même format des tweets) en se basant sur l'API Twitter en langue française et anglaise en utilisant l'outil python Tweepy.

```
for tweet in tweepy.Cursor(api.search, q='science', count=30,
                           lang="fr",
                           since="2020-01-01").items():
    print(tweet.created_at, tweet.text)
```

Figure 13-Extraction des Tweets grâce à tweepy.

Pour effectuer des expérimentations et évaluations nécessaires, nous commençons par collecter un corpus composé de plusieurs tweets. Ce corpus a été téléchargé par l'API Twitter en utilisant des requêtes en anglais et en français qui contiennent des mots et des hashtag reliés au domaine de la recherche. La collection regroupe 20 catégories différentes comme l'informatique, la médecine, la chimie, …

Nous avons récupéré environs 54 505 tweets entre le 15 mars 2020 et le 2 aout 2020

```
# Utilisation de Pandas pour catégoriser les différentes informations du Tweet
db_tweets = pd.DataFrame(columns=['username', 'acctdesc', 'location', 'following',
                                  'followers', 'totaltweets', 'tweetcreatedAt',
                                  'retweetcount', 'text', 'hashtags'])
```

Figure 14-Utilisation de Pandas pour catégoriser les informations extraites.

Dans le cadre de notre travail, d'autres informations hormis le texte du tweet sont nécessaires, tel que le nombre de Followers, la localisation et le nombre de retweets

du Tweet en question. Nous avons donc utilisé Pandas dans le but d'annoter et de classifier les différentes informations reliées au Tweets.

Par la suite, une étape de nettoyage des données a été prévue pour les rendre plus pertinentes lors du traitement, cette dernière consiste en l'élimination des différentes parties des tweets qui ne sont pas forcément nécessaires à notre travail et en la catégorisation des autres parties en différentes classes. Lestweets contiennent en moyenne 12,4 mots, le tweet le plus long en contenant 37 après cette étape.

Pour finir, nous avons mis le résultat de ces étapes dans un fichier sous format CSV avec différentes colonnes. Chacune d'elles contient une catégorie d'informations comme le démontre la figure ci-dessous. A noter que les tweets sont séparés dans différents fichiers selon leur langue de base et leur thématique.

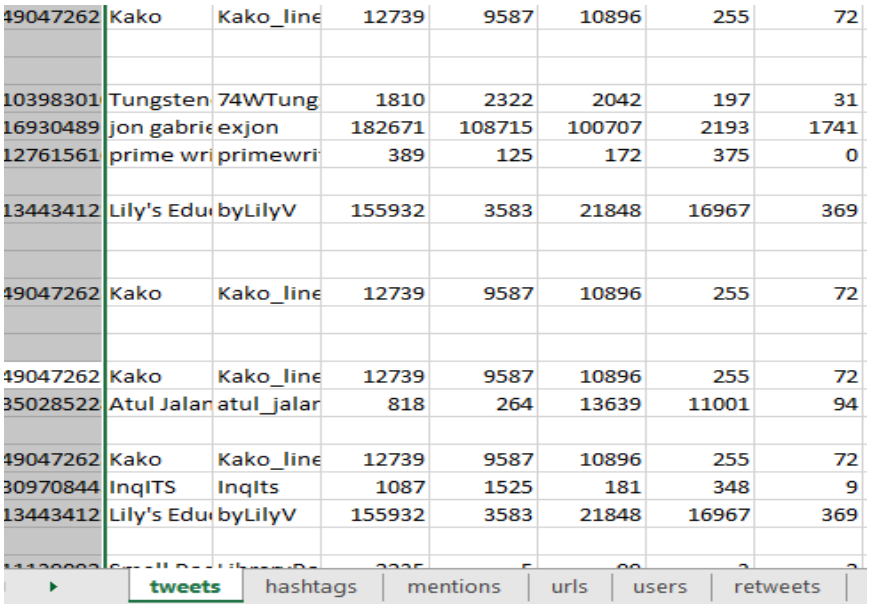


Figure 15-liste des tweets dans une dataset format csv.

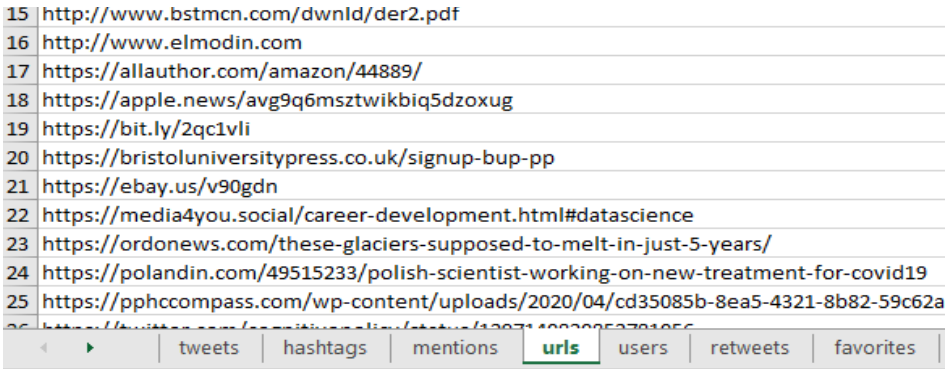


Figure 16-Liste des URLs .

IV.4 Prétraitement

IV.4.1 Problèmes avec les données

Il y a toujours des problèmes avec les données du monde réel. Ceux-ci sont mieux illustrés dans la figure 17. La nature et la gravité des problèmes dépendent de nombreuses raisons qui échappent parfois au contrôle opérateurs humains. Notre inquiétude tient aux effets de ces problèmes sur les résultats de l'analyse des données, l'objectif étant soit de rectifier les problèmes de données à l'avance, soit de reconnaître les effets des problèmes de données sur les résultats. Les problèmes de données peuvent être classés en trois groupes de : trop de données, trop peu de données et données fracturées.

	Problèmes de données dans le monde réel	
Trop de données	Trop peu de données	Données fracturées

-Données corrompues et bruyantes - Extraction de caractéristiques - Données non pertinentes - Très grandes tailles - Numérique / Symbolique	- Attributs manquants - Attribut manquant valeur - Petite quantité de données	- Données incompatibles - Plusieurs sources de données - Données provenant de plusieurs niveaux de granularité

	Techniques de prétraitement des données	
Transformation des données	Collecte d[illegible]mations	Génération de nouvelles informations
- La fusion des données - Commande de données - Edition des données - Modélisation du bruit	- Visualisation de données - Élimination des données - Sélection des données - analyse des principaux composants - Échantillonnage des données	- Ingénierie des données / nouvelles fonctionnalités - Analyse des séries chronologiques - La fusion des données - Simulation / création d'attributs non mesurables - Analyse dimensionnelle - Induction constructive

L'analyse des données
- Découverte des connaissances - Prédiction- Modélisation (processus / système) - Effectuer des tâches intelligentes (aide à la décision, planification des processus, surveillance, diagnostic, etc.). - Construire des systèmes basés sur la connaissance

Figure 17Problèmes avec les données et les techniques de prétraitement des données.

On trouve dans plusieurs cas, des imperfections du texte qui empêchent un bon traitement de son contenu ainsi qu'une analyse fiable. La phase de prétraitement a pour but, de réduire le nombre de ces imperfections pour un résultat meilleur.

Cette étape comprend toutes les actions entreprises avant le début du processus d'analyse des données. Il estessentiellementsujet d'une transformation T qui transforme lesdonnées brutes du monde réel X_{ib} en un ensemble de nouvelles donnéesvecteursY_{ij}tels que :

Y_{ij} préserve les "informations précieuses" dans $X_{ib}Y_{ij}$élimine au moins l'un des problèmes dans X_{ik} et Y_{ij} est plus utile que X_{ik}. Dans la relation ci-dessus :

i = 1, ..., n où n = nombre d'objets,

j = 1, « m où m = nombre d'entités après prétraitement

k = 1, ..., I où I = nombre d'attributs / caractéristiques avant prétraitement, et en général,

M i = I.

IV.4.2 Nettoyage et normalisation des données

Le prétraitement du texte comprend les étapes suivantes :

IV.4.2.1 Tokenisation

Cette étape permet la segmentation du texte initial en unités linguistiques manipulables appelées « token ». Le but de ce procédé est de séparer les unités de base d'un texte qui se prêteront à une analyse pointue par la suite pour le rendre interprétable par une machine.

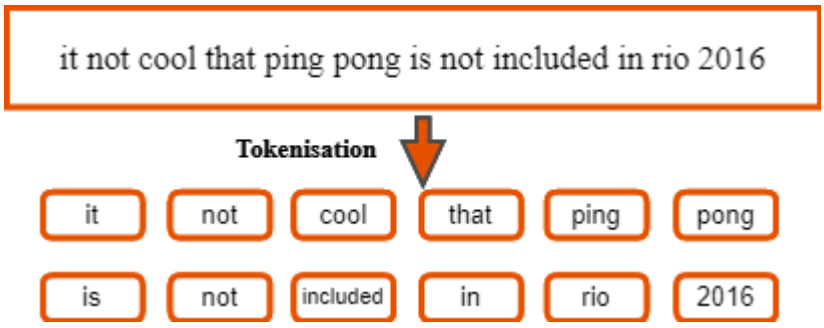


Figure 18-Exemple sur l'étape de tokenisation.

IV.4.2.2 Pos-tagger

Le part-of-speech tag, ou "taggage" selon les parties du discours, est un travail de labellisation qui nous permet de renvoyer tous les mots d'un énoncé à l'ensemble grammatical auquel ils appartiennent. C'est un processus d'étiquetage

morphosyntaxique au niveau du mot qui s'inscrit dans une démarche de linguistique informatique [59].

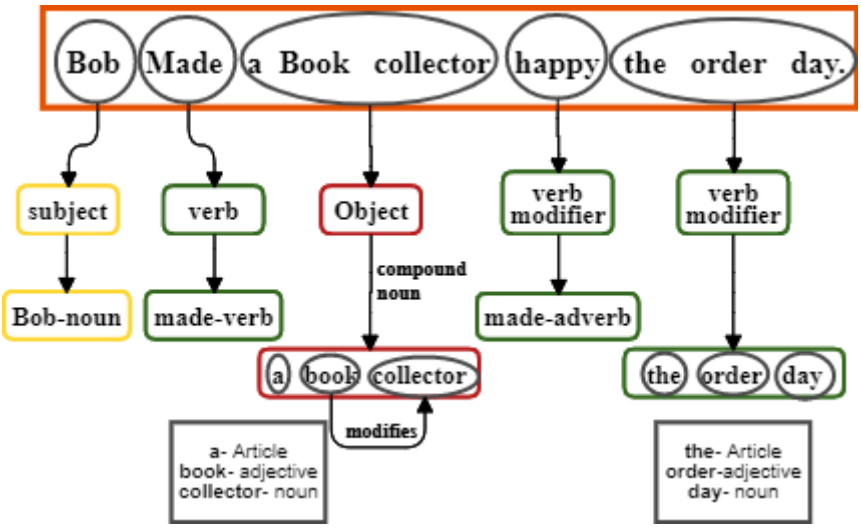


Figure 19-Exemple sur l'étape de Pos-tagger.

IV.4.2.3 La normalisation

L'étape de normalisation et construction du dictionnaire nous permet de ne pas prendre en compte les détails de type ponctuation, majuscules, conjugaison, etc.

Elle consiste à la suppression de ce qu'on appelle en anglais les stopwords, qui sont des mots très courants dans la langue étudiée ("et", "à", "le"... en français), ces derniers n'apportent pas de valeur informative pour la compréhension du "sens" d'un tweet. Ils sont très fréquents et ralentissent l'analyse du texte. Il existe dans la librairie NLTK une liste par défaut des stopwords dans plusieurs langues, notamment pour le français que nous utiliserons pour compléter cette étape.

IV.4.2.4 La lemmatisation

La lemmatisation qui consiste à représenter les mots (ou « lemmes ») sous leur forme canonique. Par exemple pour un verbe, ce sera son infinitif. Pour un nom, son masculin singulier. La lemmatisation nécessite l'utilisation d'un dictionnaire.

IV.4.2.5 Stemming

La racinisation(ou stemming en anglais) consiste à ne conserver que la racine des mots étudiés. L'idée étant de supprimer les suffixes, préfixes et autres des mots afin de ne conserver que leur origine. C'est un procédé plus simple que la lemmatisation et plus rapide à effectuer puisqu'on tronque les mots essentiellement[60].

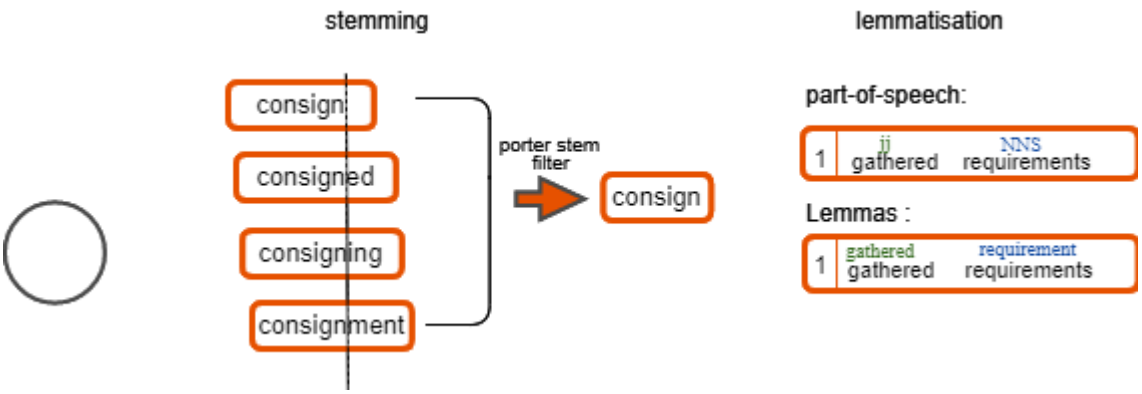


Figure 20-Exemple de Stemming et Lemmatisation.

IV.4.2.6 Décomposition des Hashtags

Un hashtag commence toujours par le caractère « # » ; ce qui permet de le repérer très rapidement dans l'analyse des données (lors de la tokenisation). Ces hashtags créent des problèmes durant l'analyse linguistique. Des hashtags qui sont en relation avec le sujet même. Cela, permet aux autres utilisateurs de trouver plus facilement le tweet posté. Pour le sujet développement web nous avons découvert plusieurshashtags sur le même sujet tels que : #Web_Development, #WebDEV, #webcoding. Ces trois Hashtags ont une relation avec les mots Web, Développent, Coding. Nous voulons extraire les mots qui composent ces hashtags afin de les relier avec des mots appartenant aux catégories déjà fixées au préalable comme suit :

#Web_Development -> (Web, Développent)

#WebDEV -> (Web, Développer)

#Webcoding -> (web, coding)

Ce traitement va nous aider à extraire le plus de mots possibles à partir des hashtags et lesfaire relier avec les autres mots récupérés du tweet. Les sujets discutés dans les tweets, ont généralement symbolisés par des hashtags qui sont le plus souvent des mots composés collés ensemble ou des sujets parlés dans le texte des tweets.

Pour faire le lien entre les deux, nous avons eu besoin de faire cette segmentation.

L'algorithme de la segmentation proposé (voir Algorithme 1) est récursif et traite les hashtags dans la direction de la lecture de texte, c'est-à-dire de gauche vers la droite. Ceci nous permet de décomposer le problème en deux parties principales, comme suit :

1. Détecter des mots en se basant sur la majuscule, utilisée pour marquer le début de chaque mot. Des mots qui commencent parunemajuscule sont collés ensemble, puis les séparer à l'aide des expressions régulières. Par exemple

l'hashtag # BackendWebDev construit àl'aide de trois mots collés ensemble et chaque mot commence par une majuscule.

2. Détecter des mots utilisant une délimitation avec des caractères spéciaux ou par des chiffres. Ce problème a été réglé en cherchant les mots qui sont séparés par des caractères non alphabétiques ou un nombre. Exemple #3Novices,#Dev_Media à l'aide des expressions régulières qui détectent les caractères non alphabétiques, ou bien les chiffres dans le hashtag.

IV.4.2.7 La méthode de pondération

Avant de passer à la phase de classification, il faudra représenter les documents de façon à ce qu'ils puissent être traités automatiquement par un classifieur donné. La plupart des approches se basent sur la représentation vectorielle des documents. Le principe consiste à représenter chaque document de la collection comme un point de l'espace, autrement dit, un vecteur de coordonnées dans l'espace vectoriel. Les coordonnées correspondent en fait aux descripteurs composant le document.

Pour représenter les tweets sous forme vectorielle, nous avons utilisé une méthode automatique qui se décompose en deux phases. Une fois le corpus acquis, il sera représenté de manière vectorielle. Chaque tweet sera considéré comme un sac des mots.

Dans cette représentation dite « Saltonienne », un traitement préalable consistera à éliminer les mots inutiles (préposition, mots vides, etc.). Chaque mot présent dans le corpus représentera une dimension dans l'espace vectoriel sur lequel nous nous appuierons pour effectuer la représentation.

Deux types de représentations peuvent alors être effectuées : une représentation fréquentielle (nombre d'occurrences des mots dans chaque tweet) et la mesure TF-IDF.

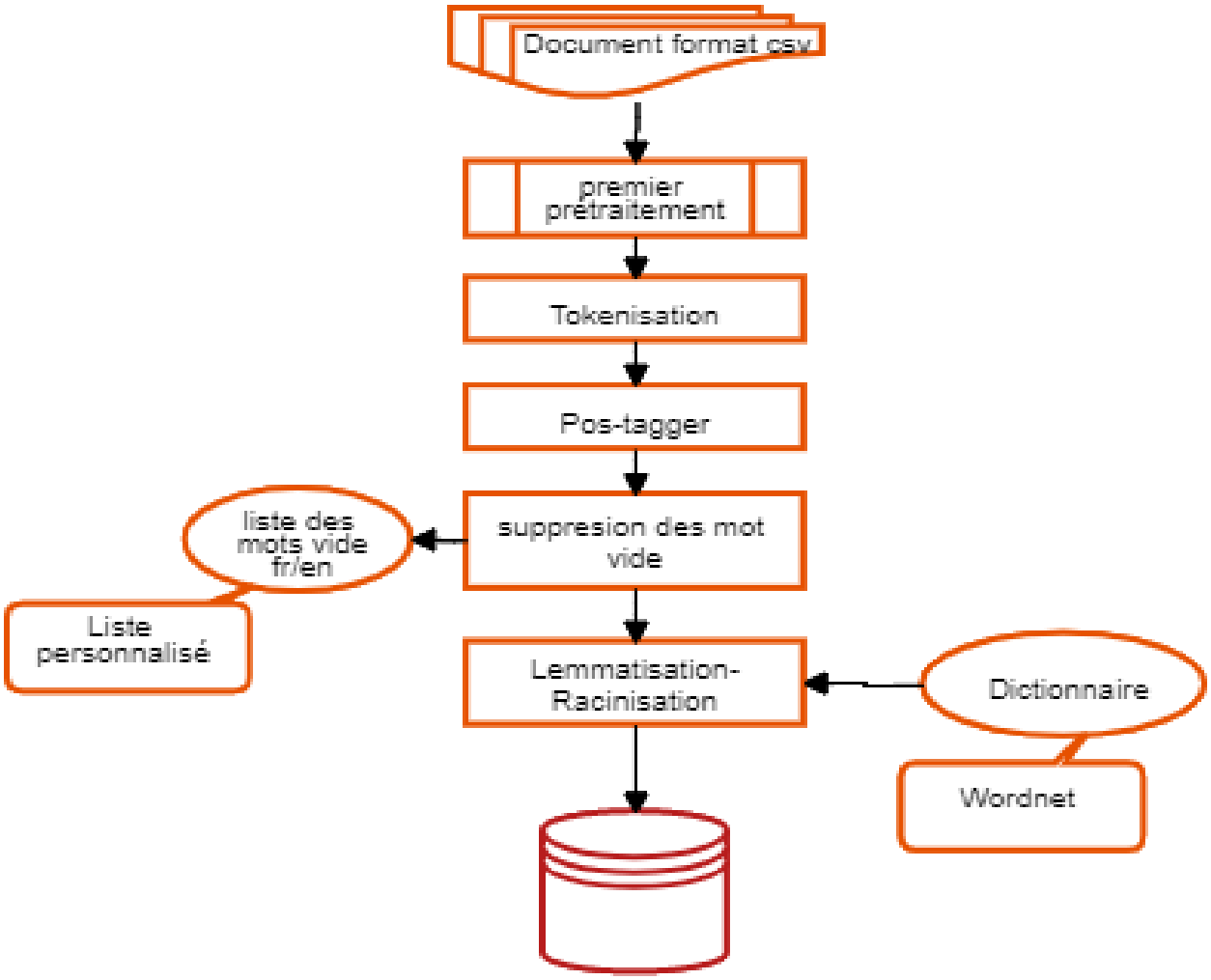


Figure 21-Architucture de notre travail concernant l'étape de prétraitement

IV.5 Annotation

Le système de classification proposé se compose de quatre étapes : collecte de données, catégorisation, Modélisation données et apprentissage automatique. Dans nos expériences,

Nous allons utiliser deux méthodes de modélisation de données : la modélisation de données textuelles ; et modélisation de données basée sur le réseau

On a commencé par élaborer 20catégories(biologie, médecine, informatique, mathématique, vétérinaire, physique , chimie , géologie ,architecture , agronomie, électronique , mécanique , aéronautique , philosophie , économie , sociologie , science politique , psychologie ,énergétique , pharmacie) du domaine scientifique pour classifier les tweets en plus des catégories on a défini chaque catégorie avec une liste de mots d'environ 1000 mots dans le tableau suivant un échantillon des catégories et mots affilier.

biologie	Médecine	informatique	mathématique	physique
Eau distillée Distilled water Eau iodée Iodine water Eau oxygénée, Peroxyde d'hydrogène Hydrogenperoxide Ebullition Boiling Ecaille Echange	Acarbose Achromatopsie Acidose métabolique Acné Acouphène Addictologie Adénopathie Adontie Adrénaline Alcalose métabolique	Applette Backup Bandepassante Batch Bit Boot Bug CPU Cracker Cluster Configuration Customisation	baque abélianiser Abélien Abondant Abscisse Absolu Absorbant Additif Additivité Adjacence Affecter Affine	Absorption Analyse Anion Anticyclone Apogée Aqueux Atmosphère Atome Attraction Azote Balance Baromètre

Exchange	Alcoolisme	Driver	Affixe	Calendrier
Echinodermes	Algodystrophie	Pilote	Aire	Carbone
Echinoderms	Allergie	Dump	Aleph	Carbonique
Ectoblaste Embryologie	Alopécie	En ligne	Algèbre	Catalyseur
Ectoblaste Embryology	Amblyopie	Online	Image	Cation
Ectoderme	Aménorrhée	End-of-file	Imaginaire	Célérité
Ectoplasme	Androgènes	Fin de fichier	Impair	Chlore
Allostérique	Anémie	Framework cadriciel	Imparité	Chromatographie
Efférent	Anesthésie	Firmwaremicroprogramme	Impliquer	Chrome
Effector	Anévrisme	Freeware gratuiciel	Impropre	Clepsydre
Allostericeffector	Bacille	Geek	Incalculable	Combustion
Efferent	Bactérie	Informaticien bricoleur	Inclinaison	Constellation
Effect	Balanite	GroupwareLogiciel	Inclus	Corrosif
Elasmobranches	Barbiturique	Hacker	Inclusion	Décantation
Elasticity	Béribéri	Implémentation	Ligne	Densité
Elastique	Bile	Interface	Linéariser	Diffraction
Elastic	Bioéthique	logging	Lisse	Dilatation
Electriquement neutre	Boulimie	Matériel informatique	Lisette	Dilution
Electricallyneutral	Bradypnée	Hardware	Littéral	Dimension
Electrode	Bronchite aiguë	Middleware	Log	Dioxygène
	Bronchite chronique	Logiciel médiateur	Logarithme	Distillation
	Calcul rénal	Migration	Logicien	Effervescence
		Multitâche -	Logicienne	Electrolyse
			Logistique	

Elément	Cancer	multitasking.	Loi	Electron
Ellipse	Carie dentaire	Open source	Astroïde	Electro-magnétique
Elliptique, ellipsoïde	Césarienne	Protocole ordinateurs	Maximum	Elément
Elodée	Chancre mou	Polling scrutation	N-uple	Emission
Elongation	Chlamydiose	Reset	N-uplet	Equinoxe
Water weed	Choléra	Reboot	Nabla	Etoile
Entoblaste	Chorée de Huntington	Requête	n-aire	Extraction
Endoplasme	Cirrhose	Reverse engineering	Naturel	Faisceau
Endosperme	Clonage thérapeutique	Scrolling	Nécessaire	Fer
Endosperm	Colectomie		Négatif	Filtration
Endothélium	Colite		Négligeable	Force
Endothelium	Colique		Népérien	Fréquence
Engainant	Collapsus		Neutre	Fusio
Sheathing	Coma		n-ième	Galaxie
Ensheathing	Immunologie		niladique	Gnomonique
Entomophile	Infarctus		Nilpotent	Gravitation
Entonnoir	Infection		Noethérien	Halogène
	Laryngectomie		obélus	Hectopascal
			Oblong	Héliocentrique
			octave	Homogène

Tableau 6-Un échantillon des catégories et mots affilier.

IV.6 Modélisation des données

Le schéma suivant résume les étapes avant la modélisation des données textuelle

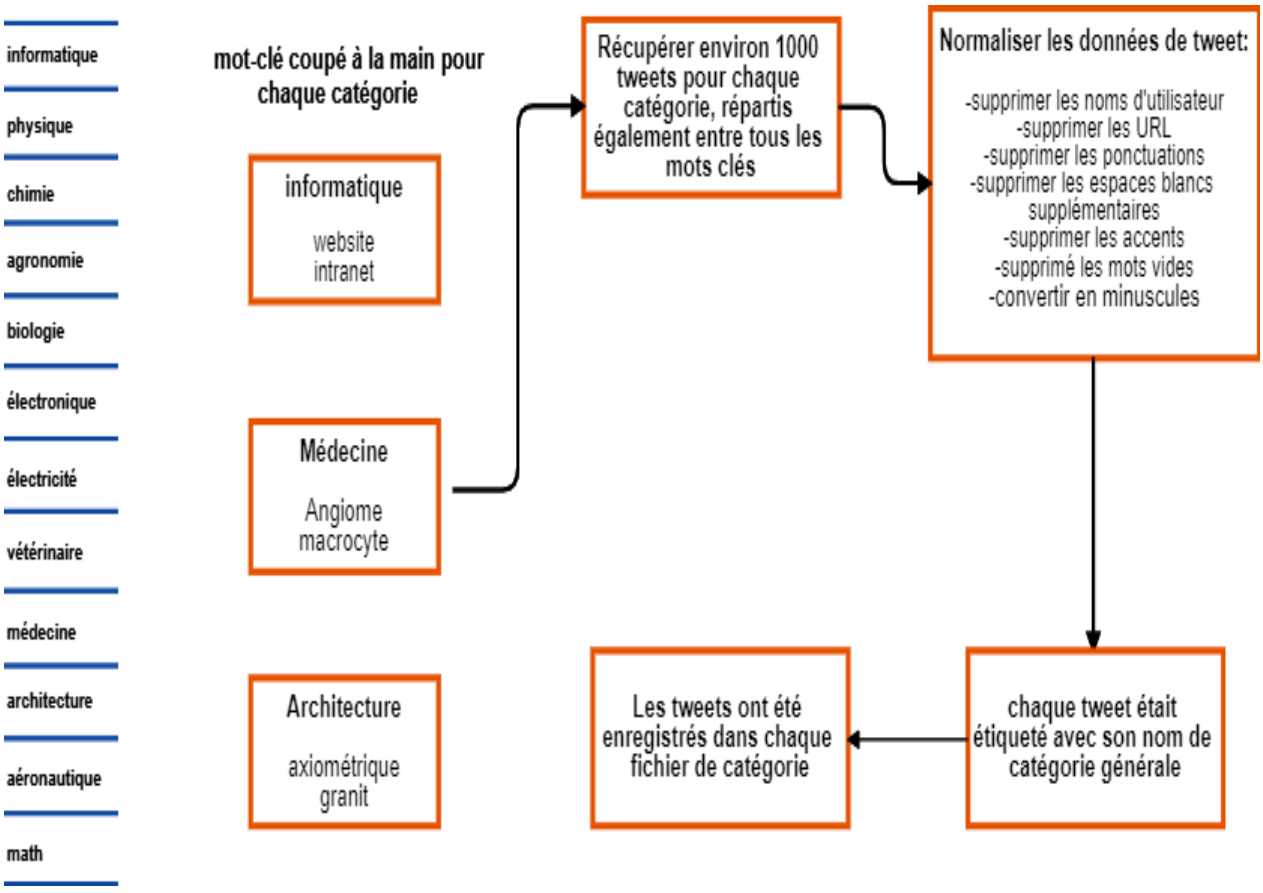


Figure 22-Architecture de modélisation des données.

IV.6.1 Modélisation de données basée sur du texte :

Cette étape consiste en l'utilisation d'un algorithme d'apprentissage supervisé afin de classifier les tweets selon les catégories définies au préalable. Pour se faire, nous avons pris 70% de notre corpus pour effectuer la phase d'apprentissage. Chaque tweet figurant dans le corpus d'apprentissage a été annoté manuellement et affecté à une catégorie parmi les 20 catégories existantes.

En guise de corpus de test, nous nous sommes servis des 30% restants de notre corpus.

Plusieurs classifieurs d'apprentissages automatiques ont été utilisés lors de nos tests pour en choisir le plus performant et le plus optimal pour notre tâche. Suite à nos expérimentations, nous avons conclus que l'algorithme Naïve Bayes Multinomial (NBM) est le plus approprié à utiliser dans notre démarche (Voir graphe de précision dans la section suivante).

IV.6.2 Modélisation de données basée sur le réseau :

Dans le but de réussir à obtenir les tweets les plus tendances dans une catégorie donnée après la classification des tweetseffectuée dans l'étape précédente, nous nous sommes servis des métas-datas obtenus lors de l'extraction des tweets pour arriver à nos fins.

Un tweet tendance est un tweet partagé et liké par plusieurs individus qui partagent un intérêt commun pour le sujet du tweet. Nous nous sommes basés sur le nombre de Likes et de Retweets de chaque tweet appartenant à la même catégorie pour les classé par ordre de popularité.

Une fois cette opération effectuée nous obtiendrons le sujet tendance de chacune des 20 catégories définies au début de notre travail.

Nous allons par la suite nous baser sur les hashtags du tweet le plus populaire de chaque catégorie pour essayer de déterminer le sujet tendance de cette dernière. Car comme nous l'avons déjà mentionné, les hashtags sont des indicateurs qui permettent de comprendre le sujet global d'un tweet donné.

Nous allons utiliser wordNet sur Les hashtags traités au préalable dans une étape précédente, pour regrouper les mots constituant les hashtags qui ont le même sens pour éviter toute répétions non nécessaire.

Une fois cette tâche effectuée, nous obtiendrons donc le sujet le plus tendance de chaque catégorie.

IV.7 Résultat et expérimentation

Dans cette section, nous allons discuter des résultats obtenus après les tests effectués sur notre corpus de test avec plusieurs classifieurs supervisés en utilisant l'outil python Scikit-learn, qui est en soit un outil permettant l'accès à toutes sortes de fonctions et algorithmes en relation avec lamachine Learning.

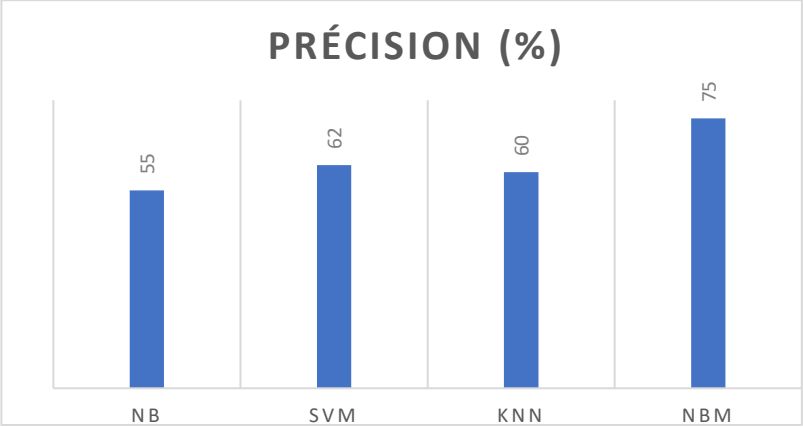


Figure 23-Représentation graphique de la précision par rapport des différents classificateurs.

Utilisation de Naïve Bayes Multinomial (NBM), Naïve Bayes(NB) et Support Vector Machines (SVM-L) et KNN ; nous constatons que la précision de la classification est en fonction du nombre de tweets et des termes fréquents.

Le modèle NB offre toujours une précision inférieure par rapport à NBMmodèle car il modélise le nombre de mots et ajuste le calcul sous-jacent.

SVM-L fonctionne légèrement mieux que NB mais a une précision inférieure par rapport à NBM. Si seulement la définition de tendance est utilisée, quel que soit le plus fréquent terme, la précision est beaucoup plus faible pour les trois classes. Comparés à l'utilisation de la définition des tendances et des tweets. Le les résultats expérimentaux suggèrent que le classificateur NBM utilisant du texte à partir de la définition de tendance, 100 tweets et un maximum de 1000 jetons de mots par catégorie donnent la meilleure précision de 75%.

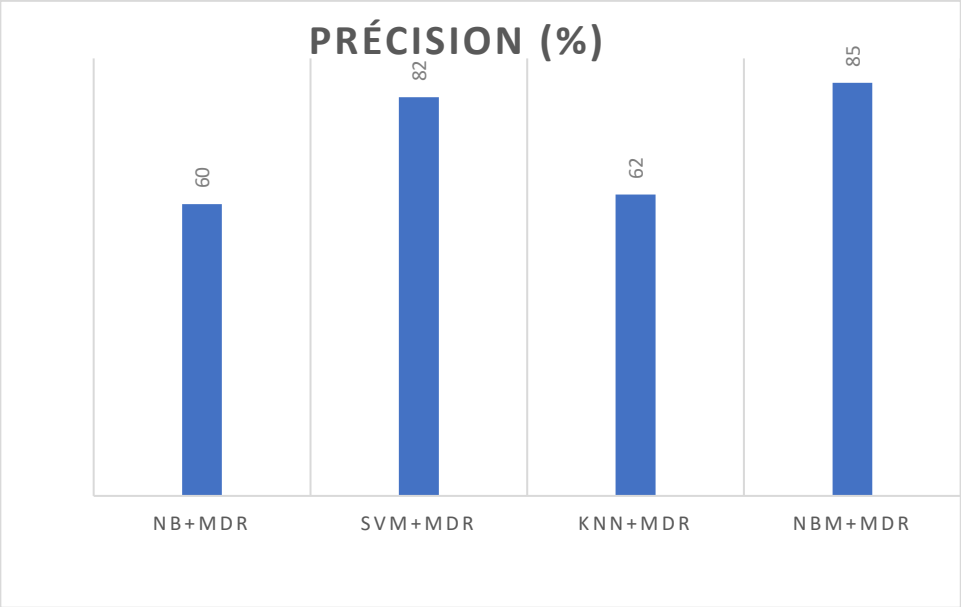


Figure 24-Reppresentation graphique de la précision de chaque classificateur en modélisation réseau.

Figure -23- présente la comparaison de la précision de classification en utilisant différents classificateurs pour la classification textuelle et celle basée sur le réseau. Clairement, le classificateur de notre algorithme donne la meilleure classification précision (85,96%) suivi de Machine à vecteurs de support (82,28%), k-NearestNeighbor (62,349%). Le classificateur utiliséatteint Précisionbeaucoup supérieure à celle de des autres classificateurs. La précision de 85,96% est très bonne compte tenude notre travail qui consiste à ce que nous catégorisons les sujets

en 20 classes. En notre connaissances, le nombre de classes utilisées dans notre expérienceest beaucoup plus grande que le nombre de classes utilisées dans tous lestravaux de recherche antérieurs (la classification en deux classes est la plus commun).

IV.8 Implémentation

IV.8.1 Ressources utilisées

Dans le cadre de la réalisation de notre projet, nous avons utilisé plusieurs outils pour faciliter basés sur le langage Python.

IV.8.1.1 Python

Python est le langage de programmation open source le plus employé par les informaticiens. Ce langage s'est propulsé en tête de la gestion d'infrastructure, d'analyse de données ou dans le domaine du développement de logiciels. En effet, parmi ses qualités, Python permet notamment aux développeurs de se concentrer sur ce qu'ils font plutôt que sur la manière dont ils le font.

Les principales utilisations de Python par les développeurs sont :

- La programmation d'applications
- La création de services web
- La génération de code
- La méta programmation.

Techniquement, ce langage servira surtout pour le Scripting et l'automatisation (interaction avec les navigateurs web)

IV.8.1.2 Django

Django est un cadre de développement web open source en Python. Il a pour but de rendre le développement web 2.0 simple et rapide

Django est un cadre de développement qui s'inspire du principe MVC ou MTV (la vue est gérée par un gabarit) composé de trois parties distinctes :

1. Un langage de gabarits flexible qui permet de générer du HTML, XML ou tout autre format texte ;
2. Un contrôleur fourni sous la forme d'un « remapping » d'URL à base d'expressions rationnelles ;

3. Une API d'accès aux données est automatiquement générée par le cadre compatible CRUD. Inutile d'écrire des requêtes SQL associées à des formulaires, elles sont générées automatiquement par l'ORM.

IV.8.1.3 HTML

L'HTML est un langage informatique utilisé sur l'internet. Ce langage est utilisé pour créer des pages web. L'acronyme signifie HyperText MarkupLanguage, ce qui signifie en français "langage de balisage d'hypertexte". Cette signification porte bien son nom puisqu'effectivement ce langage permet de réaliser de l'hypertexte à base d'une structure de balisage. Est ce qui permet à un créateur de sites Web de gérer la manière dont le contenu de ses pages Web va s'afficher sur un écran, via le navigateur. Il repose sur un système de balises permettant de titrer, sous-titrer, mettre en gras, etc., du texte et d'introduire des éléments interactifs comme des images, des liens, des vidéos... L'HTML est plus facilement compris des robots de crawl des moteurs de recherche que le langage JavaScript, aussi utilisé pour rendre les pages plus interactives.

IV.8.1.4 CSS

Les feuilles de styles (en anglais "Cascading Style Sheets", abrégé CSS) sont un langage qui permet de gérer la présentation d'une page Web. Le langage CSS est une recommandation du World Wide Web Consortium (W3C), au même titre que HTML ou XML. Le but de CSS est séparé la structure d'un document HTML et sa présentation. En effet, avec HTML, on peut définir à la fois la structure (le contenu et la hiérarchie entre les différentes parties d'un document) et la présentation. Avec CSS on peut par exemple définir un ensemble de règles stylistiques communes à toutes les pages d'un site internet. Cela facilite ainsi la modification de la présentation d'un site entier. CSS permet aussi de définir des règles différentes pour chaque support d'affichage (un navigateur classique, une télévision, un support mobile, un lecteur braille...). CSS permet aussi d'améliorer l'accessibilité des documents web.

De plus, CSS ajoute des fonctionnalités nouvelles par rapport à HTML au point de vue du style. En effet, HTML permet une gestion assez sommaire du style des documents.

Autres outils supplémentaires :

IV.8.1.5 NLTK

Natural LanguageToolkit (NLTK) est une boîte-à-outil permettant la création de programmes pour l'analyse de texte. Cet ensemble a été créé à l'origine par Steven Bird et Edward Loper, en relation avec des cours de linguistique informatique à l'Université de Pennsylvanie en 2001. Il existe un manuel d'apprentissage pour cet ensemble titré Natural LanguageProcessingwith Python (en anglais).NLTK est un corpus qui contient différents fonctionnalités qui est de taille plus de 10Go.

Quelque fonctionnalité de NLTK :

- Stop word : parfois, nous avons besoin de « raboter » des éléments inuiles afin que les données soient davantage traduisibles pour l'ordinateur. En NLP, de telles données (des mots, words) sont qualifiées pat stop words. Par conséquent ces mots n'ont aucune signification pour nous, et nous souhaiterions les retirer.

  La libraire NLTK contient quelque mot d'arrêt pour commencer ce traitement.
- Tokénisation : telle que définie dans Wikipédia, est : il s'agit du processus consistant à briser un flux de texte en plusieurs mots, phrases, symboles ou tout autre élément significatifs dénommés signes (tokens) .
- Rechercher : Nous désirerions rechercher (fouiner) le mot. Nous pourrions utiliser la librairie NLTK.

IV.8.1.6 Gensim

est une bibliothèque open-source pour la modélisation de sujets non supervisée et le traitement du langage naturel , utilisant l'apprentissage automatique statistique moderne (Machine Learning).

Gensim inclut des implémentations parallélisées streamées des algorithmes fastText , word2vec et doc2vec, ainsi que l'analyse sémantique latente (LSA, LSI, SVD), la factorisation matricielle non négative (NMF), l' allocation de Dirichlet latente (LDA), tf -idf et projections aléatoires .

Gensim a été utilisé et cité dans plus de 1400 applications commerciales et académiques en 2018, dans un large éventail de disciplines allant de la médecine à

l'analyse des réclamations d'assurance en passant par la recherche de brevets. Le logiciel a été couvert dans plusieurs nouveaux articles, podcasts et interviews.

IV.8.1.7 Textblob

Est une autre bibliothèque NLP extrêmement puissante pour Python. TextBlob est construit sur NLTK et fournit une interface facile à utiliser à la bibliothèque NLTK. Est une bibliothèque (bib) Python (2 et 3) pour le traitement de données textuelles. Il fournit une API simple pour se plonger dans les tâches courantes de traitement du langage naturel (NLP) telles que le balisage d'une partie du discours, l'extraction de phrases nominales, l'analyse des sentiments, la classification, la traduction, etc. quelque caractéristique de texte blob :

- Extraction de phrases nominales
- Analyse des sentiments
- Classification (Naïve Bayes, arbre décisionnel)
- Tokenisation (fractionnement du texte en mots et phrases)
- Fréquences des mots et des phrases
- Analyse
- n-grammes

IV.8.1.8 Spacy

SpaCy est une bibliothèque logicielle Python de traitement automatique des langues développée par Matt Honnibal de l'entreprise Explosion AI.

La bibliothèque SpaCy permet d'effectuer les opérations d'analyse suivantes[3] sur des textes dans plus de 50 langues[3] :

- Tokenization
- Reconnaissance d'entités nommées

IV.8.1.9 Panda

Pandas est une bibliothèque de logiciels sous Python qui permet la manipulation et l'analyse de données. Elle propose en particulier des structures de données et des opérations pour manipuler des tableaux numériques et des séries chronologiques . Il s'agit d'un logiciel libre publié sous la licence BSD .

IV.8.1.10 NumPy

Numpy est une extension du langage de programmation Python, destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux.

Plus précisément, cette bibliothèque logicielle libre et open source fournit de multiples fonctions permettant notamment de créer directement un tableau depuis un fichier ou au contraire de sauvegarder un tableau dans un fichier, et manipuler des vecteurs, matrices et polynômes.

NumPy est la base de SciPy, regroupement de bibliothèques Python autour du calcul scientifique.

IV.8.1.11 Tweepy

Est une bibliothèque Python permettant d'accéder à l'API Twitter. C'est idéal pour une automatisation simple et la création de robots Twitter. Tweepya de nombreuses fonctionnalités. Avec tweepy on peut accéder à API de recherche twitter et meme à API twitterstream.

IV.8.1.12 Scikit-Leran

Scikit-learn est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche comme Inria. Elle comprend notamment des fonctions pour estimer des forêts aléatoires(RF), des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. Elle est conçue pour s'harmoniser avec d'autres bibliothèques libres Python, notamment NumPy et SciPy.

Scikit-learn est écrit en Python, avec quelques algorithmes essentiels écrits en Cython pour optimiser les performances.

IV.8.1.13 SQLite

SQLite est un système de base de données ou une bibliothèque proposant un moteur de base de données relationnelle. Il repose sur une écriture en C, un langage de programmation impératif, et sur une accessibilité via le langage SQL (StructuredQueryLanguage).

SQLite présente la particularité d'être directement intégré aux programmes et dans l'application utilisant sa bibliothèque logicielle alors que ses concurrents comme MySQL reproduisent de leur côté le schéma classique client-serveur.

Avec SQLite, la base de données est intégralement stockée dans un fichier indépendant du logiciel.

Créé au début des années 2000 par D. Richard Hipp, SQLite propose un accès plus rapide aux données, mais aussi plus structuré et avec davantage de sécurité.

À noter que, contrairement à une majorité de systèmes de gestion de base de données (SGBD), SQLite est basé sur un typage dynamique plutôt que sur un typage statique pour le contenu des cellules.

IV.9 Interface et fonctionnalités

Lors de son accès à l'interface. L'utilisateur peut charger son propre dataset dans le but d'obtenir les derniers sujets tendances dans le domaine de la recherche existant dans son corpus à lui. Et visiter l'historique de la dernière classification effectuée dans le but de voir les derniers tweets tendances ainsi que les catégories populaires. Quelques graphes aussi ont été mis à disposition pour pouvoir connaitre les catégories les plus tendance au fil du temps

Les deux figures suivantes donnent un aperçu sur l'interface :

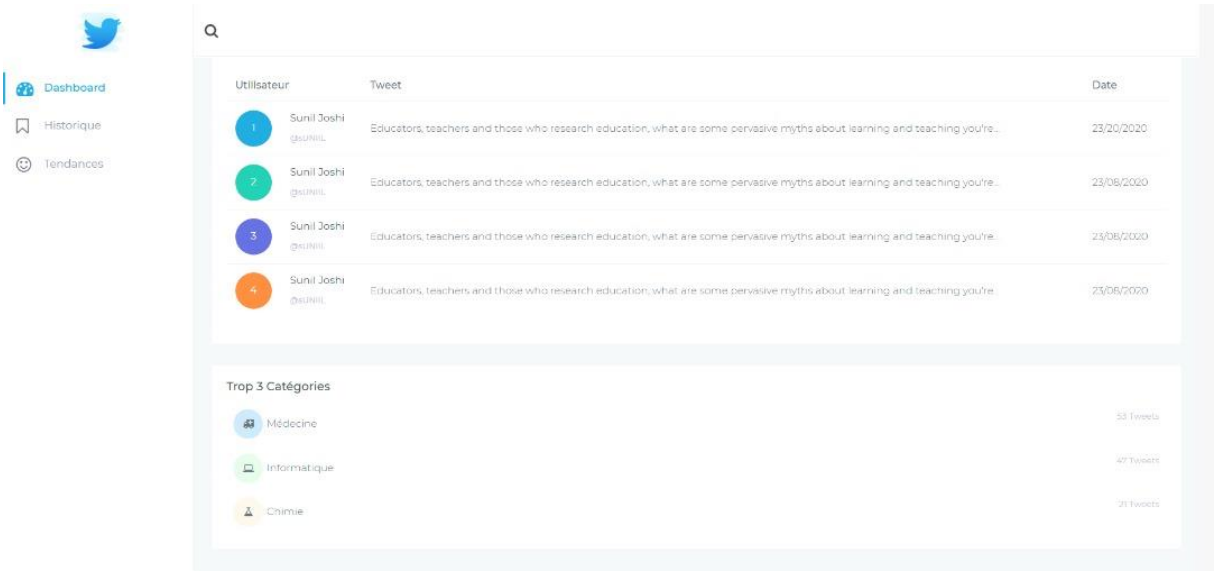


Figure 25-Interface graphique.

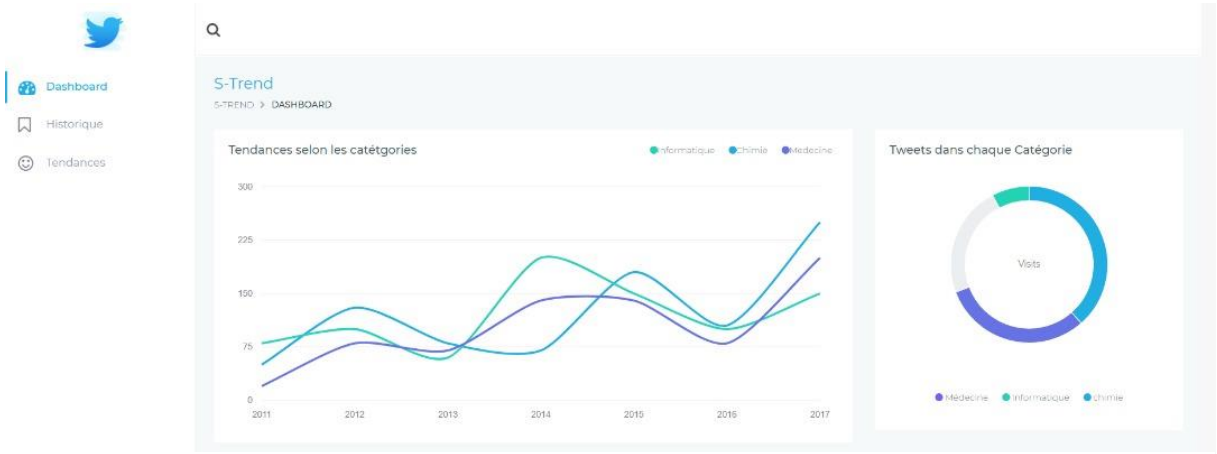


Figure 26- interface graphique.

IV.10 Conclusion

Dans ce chapitre, nous avons décrit brièvement le processus de réalisation de notre projet en spécifiant l'environnement de développement, l'implémentation de la base des données et la démarche suivie pour la réalisation. Nous avons utilisé python pour langage de programmation qui offre beaucoup de fonctionnalités. Après l'extraction des tweets, nous avons appliqués certain nombre d'opération de prétraitement pour le nettoyage. Après le nettoyage et la modélisation de données on a comparé les performances de quatre classificateurs d'apprentissage automatique supervisé pour choisir la meilleur performance et l'amélioré avec un algorithme personnaliser qui se base sur les fonctionnalités de twitter.

Conclusion générale et perspectives

Synthèse

L’objective de ce mémoire est la Détection de tendances des réseaux sociaux en utilisant les techniques du TALN, nous avons utilisé deux schémas de classification différentspour la classification des sujets (scientifique) de tendances Twitter. En plus d'utiliser une classification textuelle, notre contribution clé est l'utilisation la structure du réseau social plutôt que d'utiliser uniquement du texte, qui peuvent être souvent bruyantes données dans le contextedes médias sociaux tels que Twitter en raison de l'utilisation intensive jargon et la limite du nombre de caractères que les utilisateurs sontautoriséssurTwitter. Nos résultats montrent quele classificateur basé sur le réseau a obtenu des performances nettement meilleures que le classificateur basé sur du texte dans notre ensemble de données. Considérant que les tweetsne sont pas aussi grammaticalement structurés que les textes de document normaux,classification textuelle a était réaliser l'aide de Naïve Bayes Multinomial qui a fourni des résultats équitables et peut être mis à profit dans les cas oùnous ne soyons pas en mesure d'effectuer une analyse basée sur le réseau.

Perspectives

Dans nos travaux futurs, nous souhaitons intégrer des classifications utilisant Naïve Bayes Multinomial (NBM) et classification basée sur le réseau. L'idée serait d'intégrer ces deux classificateurs tels que si nous avons les cinq sujets similaires classifié puis utiliser la classification basée sur le réseau sinon utiliser une classification basée sur du texte. Au cours de nos expériences nous avons constaté que certains sujets pouvaient appartenir à plus d'une catégorie. Par exemple, des informations sur l’ADN par exemple relèvent de la biologie, mais aussi de la médecine. Par conséquent, nous voudrions aussi explorer l'utilisation de plusieurs étiquettes dans la catégorisation mais aussi :

- L’enrichissement de notre dictionnaire par plus des mots et expression
- L’enrichissement de Dataset
- L’application d’autres classificateurs et l’utilisation d’autres fonctionnalités.
- L’utilisation des autres configurations telles que bi-gramme, trigramme.

Références bibliographiques

- [1]Jacquesson, A. and Rivier, A., 2005. Bibliothèques et documents numériques. *Concepts, composantes, techniques et enjeux*, *30*, p.235.
- [2]Cerovšek, M., 2018. Quelques défis de classification des expressions idiomatiques du football en français. *Lublin Studies in Modern Languages and Literature*, *42*(4), pp.237-251.
- [3]Thelwall, M., 2009. Social network sites:Users and uses. *Advances in computers*, *76*, pp.19-73
- [4]Thelwall, M., 2009. Social network sites:Users and uses. *Advances in computers*, *76*, pp.19-73.
- [5]Mian, B.S.A., Les Médias Sociaux Numériques outils de Développement Professionnel de l'Enseignant-Chercheur et Chercheur en Afrique.
- [6]Chouchani, N. and Abed, M., 2018. Une approche centrée sur l'utilisateur pour intégrer les acteurs sociaux dans des communautés d'intérêt. In *INFORSID* (p. 149).
- [7]Cagliero, L. and Fiori, A., 2013. Twecom: topic and contextminingfrom twitter. In *The Influence of Technology on Social Network Analysis and Mining* (pp. 75-100). Springer, Vienna.
- [8]Papadopoulos, S., Troncy, R., Mezaris, V., Huet, B. and Kompatsiaris, I., 2011, September. Social Event Detection at MediaEval2011: Challenges, dataset and evaluation. In *MediaEval*.
- [9]Bakshy, E., Rosenn, I., Marlow, C. and Adamic, L., 2012, April. The role of social networks in information diffusion. In *Proceedings of the 21st international conferenceon World Wide Web* (pp. 519-528).
- [10]Kossinets, G. and Watts, D.J., 2006. Empiricalanalysis of an evolving social network. *science*, *311*(5757), pp.88-90.
- [11]Saez-Trumper, D., Comarela, G., Almeida, V., Baeza-Yates, R. and Benevenuto, F., 2012, August. Findingtrendsetters in information networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledgediscovery and data mining* (pp. 1014-1022).
- [12]Macnish, K., 2014. The Ethics of Social Networks and Mining. In *Encyclopedia of Social Network Analysis and Mining*. Springer.
- [13] Bianca, B.L., 2018. The User BehaviorAnalysisBased on Text Messages UsingParafac and Block TermDecomposition. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, *9*(10), pp.55-60.
- [14]Tsai, C.W., Lai, C.F., Chao, H.C. and Vasilakos, A.V., 2015. Big data analytics: a survey. *Journal of Big data*, *2*(1), pp.1-32.

- [15]Poria, S., Gelbukh, A., Cambria, E., Yang, P., Hussain, A. and Durrani, T., 2012, October. MergingSenticNet and WordNet-Affect emotionlists for sentiment analysis. In *2012 IEEE 11th International Conference on SignalProcessing* (Vol. 2, pp. 1251-1255). IEEE.
- [16]Baccianella, S., Esuli, A. and Sebastiani, F., 2010, May. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec* (Vol. 10, No. 2010, pp. 2200-2204).
- [17] Cambria, E., Hussain, A., Havasi, C. and Eckl, C., 2010. Senticcomputing: Exploitation of commonsense for the development of emotion-sensitive systems. In *Development of Multimodal Interfaces: Active Listening and Synchrony* (pp. 148-156). Springer, Berlin, Heidelberg.
- [18] Deng, L. and Wiebe, J., 2015. Mpqa 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 conference of the North American chapter of the association for computationallinguistics:humanlanguage technologies* (pp. 1323-1328).
- [19]Ammar, C. and Mohamed, B., 2011. Amélioration des Environnements de CSCL par les System de Recommandation à Base de L'awareness.
- [20] Chen, H., Cui, X. and Jin, H., 2016. Top-k followeerecommendation over microblogging systems by exploiting diverse information sources. *Future Generation Computer Systems*, *55*, pp.534-543.
- [21] Davis, E. and Marcus, G., 2015. Commonsensereasoning and commonsenseknowledge in artificial intelligence. *Communications of the ACM*, *58*(9), pp.92-103.
- [22]Panagiotou, N., Katakis, I. and Gunopulos, D., 2016. Detectingevents in online social networks:Definitions, trends and challenges. In *Solving Large Scale Learning Tasks. Challenges and Algorithms* (pp. 42-84). Springer, Cham.
- [23]Vakali, A., Giatsoglou, M. and Antaris, S., 2012, April. Social networking trends and dynamicsdetection via a cloud-basedframework design. In *Proceedings of the 21st International Conferenceon World Wide Web* (pp. 1213-1220).
- [24]Milioris, D., 2018. Topic Detection and Classification in Social Networks.
- [25]Zubiaga, A., Spina, D., Martínez, R. and Fresno, V., 2015. Real-time classification of twitter trends. *Journal of the Association for Information Science and Technology*, *66*(3), pp.462-473.
- [26]Cagliero, L. and Fiori, A., 2013. Twecom: topic and contextminingfrom twitter. In *The Influence of Technology on Social Network Analysis and Mining* (pp. 75-100). Springer, Vienna.
- [27]Chowdhury, S.R., Imran, M., Asghar, M.R., Amer-Yahia, S. and Castillo, C., 2013, May. Tweet4act:Using incident-specific profiles for classifyingcrisis-related messages. In *ISCRAM*.

- [28]Vakali, A., Giatsoglou, M. and Antaris, S., 2012, April. Social networking trends and dynamicsdetection via a cloud-basedframework design. In *Proceedings of the 21st International Conferenceon World Wide Web* (pp. 1213-1220).
- [29]Rill, S., Reinel, D., Scheidt, J. and Zicari, R.V., 2014. Politwi:Earlydetection of emergingpolitical topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-BasedSystems*, *69*, pp.24-33.
- [30]Kaushik, R., Chandra, S.A., Mallya, D., Chaitanya, J.N.V.K. and Kamath, S.S., 2016. Sociopedia: an interactive system for eventdetection and trend analysis for twitter data. In *Proceedings of 3rd International Conferenceon Advanced Computing, Networking and Informatics* (pp. 63-70). Springer, New Delhi.
- [31]Wang, Q., She, J., Song, T., Tong, Y., Chen, L. and Xu, K., 2016, June. Adjustable Time-Window-Based Event Detection on Twitter. In *International Conference on Web-Age Information Management* (pp. 265-278). Springer, Cham.
- [32]Gour, P. and Joshi, S., 2017. Trending Topics Detectionusing Machine Learning Approach. *International Journal of Engineering and Management Research (IJEMR)*, *7*(3), pp.796-801.
- [33]Winarko, E. and Pulungan, R., 2019. Trending topics detection of Indonesian tweets using BN-grams and Doc-p. *Journal of King SaudUniversity-Computer and Information Sciences*, *31*(2), pp.266-274.
- [34]Rsheed, N.A. and Khan, M.B., 2014, September. Predicting the popularity of trendingarabic news on twitter. In *Proceedings of the 6th International Conferenceon Management of Emergent Digital EcoSystems* (pp. 15-19).
- [35]Rosa, H., Carvalho, J.P. and Batista, F., 2014. Detecting a tweet's topic within a large number of Portuguese Twitter trends. In *3rd Symposium on Languages, Applications and Technologies*. SchlossDagstuhl-Leibniz-ZentrumfuerInformatik.
- [36]Rosa, H., Carvalho, J.P. and Batista, F., 2014. Detecting a tweet's topic within a large number of Portuguese Twitter trends. In *3rd Symposium on Languages, Applications and Technologies*. SchlossDagstuhl-Leibniz-ZentrumfuerInformatik.
- [37]Lata, S. and Loar, M.R., 2018. Text clustering and classification techniques using data mining. *Int. J. on Futur. Revolut. Comput. Sci. Commun. Eng*, *4*(4), pp.859-864.
- [38] MICHIE, Donald, SPIEGELHALTER, David J., TAYLOR, C. C., *et al.* Machine learning. *Neural and Statistical Classification*, 1994, vol. 13, no 1994, p. 1-298.
- [39] WASKE, Björn, BENEDIKTSSON, Jon Atli, ÁRNASON, Kolbeinn, *et al.* Mapping of hyperspectral AVIRIS data using machine-

learningalgorithms. *Canadian Journal of RemoteSensing*, 2009, vol. 35, no sup1, p. S106-S116.

- [40] KjerstiAas and Line Eikvil "TextCategorization: A Survey" Report No. 941. ISBN 82-539-0425-8. ,June, 1999.
- [41] B S Harish, D S Guru, S Manjunath "Representation and Classification of TextDocuments:ABriefReview" IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition"RTIPPR, 2010.
- [42] Hein Ragas Cornelis H.A. Koster, "Four text classification algorithmscompared on a Dutch corpus"SIGIR1998: 369-370 1998.
- [43] GongdeGuo, Hui Wang, David Bell, Yaxin Bi and Kieran Greer, "KNN Model-BasedApproachinClassification", Proc. ODBASE pp- 986 – 996, 2003
- [44] EijiAramaki and KengoMiyo, "Patient status classification by usingrulebased sentence extraction and bm25-knn based classifier", Proc. of i2b2 AMIA workshop, 2006.
- [45] MuhammedMiah, "Improved k-NN Algorithm for Text Classification", Department of Computer Science and Engineering University of Texas at Arlington, TX, USA.
- [46] SHI Yong-feng, ZHAO, "Comparison of textcategorizationalgorithm", Wuhan university Journal of
- natural sciences. 2004.
- [47] D. Lewis, "Naive Bayes at Forty: The Independence Assumption in Information Retrieval", Proc.ECML-98, 10th European Conf. Machine 1998.
- [48] Vidhya. K.AG.Aghila, "A Survey of Naïve Bayes Machine Learning approach in Text Document Classification", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7,2010.
- [49] McCallum, A. and Nigam K., "A Comparison of Event Models for Naive Bayes TextClassification".AAAI/ ICML -98 Workshop on Learning for TextCategorization
- [50] KORDE, Vandana et MAHENDER, C. Namrata. Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*, 2012, vol. 3, no 2, p. 85.
- [51] Mnish Mehta, Rakeshagrwal" SLIQ: A Fast Scalable Classifier for Data Mining" 1996.
- [52] PeeraponVateekul and MiroslavKubat, "Fast Induction of Multiple DecisionTrees in TextCategorizationFrom Large Scale,Imbalanced,
- [53] Joachims, T. "Textcategorizationwith support vectormachines:learningwithmany relevant features". In Proceedings of ECML-98, 10th EuropeanConference on Machine Learning (Chemnitz,DE), pp. 137–142 1998.

- [54] Loubes, J. M. and van de Geer, S "Support vector machines and the Bayes rule in classification", Data miningknowledge and discovery 6 259-275.2002
- [56]HANSEN, Lars Kai et SALAMON, Peter. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 1990, vol. 12, no 10, p. 993-1001.
- [57]Yiming Yang And Christopher G. Chute Mayo Cllnic "An Example-Based Mapping Method For TextCategorization And Retrieval" ACM Transactions On Information Systems, Vol. 12, No 3, Pages 252-277, July 1994
- [58]Larkey, L. S. and Croft, W. B. "Combiningclassifiers in textcategorization". In Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval (Zurich, CH, 1996), pp. 289–297 1996
- [59] STRAKA, Milan et STRAKOVÁ, Jana. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In : *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. 2017. p. 88-99.
- [60] GOYAL, Amit, DAUMÉ III, Hal, et VENKATASUBRAMANIAN, Suresh. Streaming for large scale NLP: Language modeling. In : *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2009. p. 512-520.