UNIVERSITE KASDI MERBAH OUARGLA
Faculté des Nouvelles Technologies de l'Information et de la Communication
Département d'Informatique et de Technologies de l'Information

Mémoire

MASTER ACADEMIQUE
Domaine : Mathématiques et Informatique
Filière : Informatique Industrielle

Présenté par :
BOUHNIK Nour elhouda
RADI Hanane

Thème

Classification textuelle des articles de presse à l'aide d'une approche d'apprentissage automatique

Devant le jury

Mr. AYADI Ossama	MCA	UKM Ouargla	Président
Mr. KHALDI Bilel	MCA	UKM Ouargla	Examinateur
Mr. MERABTI Hocine	MCA	UKM Ouargla	Encadreur

Année universitaire : 2021 /2022

Résumé

La catégorisation ou la classification automatique des documents est une approche vitale pour la gestion et le traitement des documents non structurés générés à partir des médias sociaux sous forme numérique. Elle prend plus en plus d'importance en raison du grand nombre de documents texte constamment ajoutés sur Web. Les techniques d'apprentissage automatique ML et de traitement du langage naturel NLP ont été largement utilisées pour extraire des informations utiles à partir des documents non structurés disponibles au format numérique. Dans ce travail, nous allons utiliser cinq algorithmes d'apprentissage automatique pour identifier et classifier avec précision les articles de BBC_News en différentes catégories. Ces algorithmes seront testés, analysés et comparés entre eux pour arriver finalement à une conclusion. Les résultats expérimentaux montrent que le model de classification de BBC_News donne des résultats satisfaisants avec la plus part des algorithmes testé sur la base de données. L'algorithme avec le score le plus haut est qualifié de meilleur algorithme d'apprentissage automatique pour l'ensemble de données utilisé.

Mots clés: Machine Learning algorithms, Text Mining, Natural Language Processing, Logistic regression, Random Forest, SVM, KNN, Naïf Bayes, Features Extraction, Text Classification.

Abstract

Automatic document categorization or classification is a vital approach for managing and processing unstructured documents generated from social media in digital form. It is gaining increasing importance due to the large number of text documents that are constantly being added to the web. Machine Learning ML and Natural Language processing NLP techniques have been widely used to extract useful information from unstructured documents available in digital format. In this work, we will use five machine learning algorithms to accurately identify and classify BBC_News articles into different categories. These algorithms will be tested, analyzed and compared with each other to finally reach to a conclusion. The experimental results show that the BBC_News classification model gives satisfactory results, with most of these algorithms tested, on the dataset. The algorithm with the highest score is referred to as the best machine learning algorithm for the dataset used.

Key words: Machine Learning algorithms, Text Mining, Natural Language Processing, Logistic regression, Random Forest, SVM, KNN, Naïf Bayes, Features Extraction, Text Classification.

ملخص

يعد التصنيف الأوتوماتيكي للوثائق نهجًا حيويًا لإدارة ومعالجة المستندات الرقمية الغير المهيكلة و التي تنبثق من خلال وسائل التواصل الاجتماعي. يكتسب هذا التصنيف أهمية متزايدة بسبب العدد الكبير من المستندات النصية التي يتم إضافتها باستمرار إلى الويب. تم استخدام تقنيات التعلم الآلي ML ومعالجة اللغة الطبيعية NLP على نطاق واسع لاستخراج معلومات مفيدة المستندات الرقمية الغير المهيكلة. في هذا العمل، سنقوم باستخدام خمس خوارزميات للتعلم الآلي لتحديد مقالات BBC_News وتصنيفها بدقة إلى فئات مختلفة. إذا، سيتم اختبار هذه الخوارزميات وتحليلها ومقارنتها مع بعضها البعض للوصول من أجل استخلاص نتيجة في النهاية. تظهر النتائج التجريبية أن نموذج تصنيف BBC_News يعطي نتائج مرضية ، بتطبيق معظم هذه الخوارزميات على مجموعة البيانات المستخدمة. يشار إلى الخوارزمية التي حصلت على أعلى الدرجات على أنها أفضل خوارزمية لتعلم الآلة بالنسبة لمجموعة البيانات هاته.

الكلمات المفتاحية: خوارزميات التعلم الآلي ، استخلاص النصوص ، معالجة اللغة الطبيعية ، الانحدار اللوجستي ، الغابة العشوائية ، SVM، KNN، NBC ، استخراج الميزات ، تصنيف النص.

Table des matières

Chapitre II : Classification automatique des articles d'actualité

Chapitre III : Implémentation et réalisation

Liste des tableaux

Liste des figures

Introduction générale

La catégorisation ou la classification automatique des documents est une approche vitale pour la gestion et le traitement des documents non structurés générés à partir des médias sociaux sous forme numérique. Elle prend plus en plus d'importance en raison du grand nombre de documents texte constamment ajoutés sur Web. Les techniques d'apprentissage automatique (ML) et de traitement du langage naturel (NLP) ont été largement utilisées pour extraire des informations utiles à partir des documents non structurés disponibles au format numérique.

Avec la croissance rapide des documents textuels électroniques générés chaque jour sur Internet, la classification des textes a gagné en importance au cours des dernières années. La classification de texte, également connue sous le nom de catégorisation de texte, est le processus d'attribution d'étiquettes de classe à un document texte en fonction de son contenu. La classification de texte a été utilisée avec succès dans des domaines tels que : la détection de topic, le filtrage des spams, la classification des articles de presse (news), la classification des pages Web, la reconnaissance des auteurs et l'analyse des sentiments.

Les news n'étaient pas facilement accessibles jusqu'au début du 21^{e} siècle, mais aujourd'hui, ils sont facilement disponibles sur Internet, car les gens suivent les actualités mondiale avec plus d'intérêt. Par conséquent, la classification des articles de news est maintenant un domaine difficile dans les approches de text mining.

La catégorisation d'un article de presse est effectuée sur la base de leur contenu informatif qui est déterminé par les mots qui constituent les gros titres des articles. La fonction principale de cette approche est de définir le contenu d'un document en seulement quelques mots. Cet étiquetage permettra notamment de faciliter des recherches détaillées en filtrant sur un corpus de documents textuels électronique, et de prendre de décisions.

Problématique et objectifs

Compte tenu de l'énorme expansion d'un système de classement automatisé des articles de presse, son développement pose naturellement un certain nombre de défis. Ces problèmes sont principalement liés au très grand volume de données à classer, ce qui conduit au problème

de la complexité. En plus de ces données de nature textuelle, toutes les problématiques liées à leurs représentations, à la pondération de leurs contenus, à leurs sémantiques rendent difficile le processus de détection des sujets traites [1].

Dans ce travail, on va utiliser cinq techniques d'apprentissage automatique pour identifier et classifier avec précision les articles de presse en différentes catégories. La fonction principale de ces approches est de définir le contenu d'un article en seulement quelques mots. Cet étiquetage permettra notamment de faciliter des recherches détaillées en filtrant sur un corpus de documents textuels électronique et de prendre de décisions.

L'objectif de notre travail consiste à faire une comparaison entre les cinq techniques de classification supervisée, à savoir, le KNN, le SVM, l'Arbre de décision, la régression logistique et la technique de Naïve Bayes, en les appliquant à un ensemble de document textuel appelé BBC-News. A l'issus de ce travail, on va analyser les résultats afin d'évaluer les performances de chaque technique sur l'ensemble de données utilisé.

Organisation du mémoire.

Le présent mémoire s'articule autour de trois chapitres comme suit :

- **Le premier chapitre** est consacré aux domaines de medias sociaux et de l'exploitation de données à partir de médias sociaux.
- **Le deuxième chapitre** est consacré à la présentation des articles d'actualité, et les méthodes d'apprentissage automatique utilisées pour catégoriser ces articles.
- **Le troisième chapitre** présente la conception et l'implémentation des modèles proposés avec une discussion et une comparaison des résultats obtenus.

Ce travail s'achève par une conclusion générale et quelques perspectives.

1. Introduction

Pour toute recherche, la connaissance des concepts du domaine étudié devient une étape indispensable avant toute démarche de conception ou de développement. En ce sens, ce premier chapitre sera consacré à présenter les concepts de base de notre domaine d'étude ainsi que tous les concepts qui s'y rapportent. Il s'agira donc principalement de visionner le domaine des médias sociaux en général et l'exploitation de données textes à partir de ces médias en particulier.

2. Medias Sociaux (Socials Media)

Depuis l'avènement d'Internet, les médias se sont révélés être l'un des outils les plus importants pour transmettre des informations et des nouvelles à un large public. Cette méthode a évolué et est passée par plusieurs étapes. Au début, au cours de laquelle des informations ont été échangées avec des peintures et des livres. Mais maintenant, de nouvelles méthodes et technologies ont été conçues, la diffusion de l'information a révolutionné le domaine de l'échange d'informations et de la communication, où l'échange de messages avec les auditeurs vise à accroître leur compréhension et leur prise de conscience et à influencer leurs perceptions.

2.1 Définition

La définition de medias sociaux se trouve dans le dictionnaire Larousse : Le terme média désigne tout moyen de distribution, d'édition ou de transmission d'œuvres, de documents ou de messages écrits, visuels, sonores ou audiovisuels(tels que la radio, la télévision, le cinéma, Internet, presse, télécommunications, …etc.) [2].

Les médias sociaux désignent également toutes les technologies qui facilitent les interactions sociales (Facebook, Linkedin, Twitter, etc.), et la création et le partage de contenus (blogs, forums, messageries, wikis, applications mobiles, etc.).

2.2 Historique des médias sociaux

On peut dater l'apparition des medias sociaux à partir de 1990, quand l'Internet a pris son envol. Aujourd'hui, les médias sociaux, qu'ils soient privé ou professionnel, sont devenus indispensables à la vie de tous. Le tableau ci-dessous présente une chronologie de l'évolution de ces médias.

Tableau I.1. L'évolution de principaux médias sociaux [3].

Années	Brevet à	Événements
1973	Dave Wooly et Douglas Brown	Création de Talkomatic à l'Université de l'Illinois. C'était une salle de chat multi-utilisateurs avec des fonctionnalités limitées.
1997	Barry Appelman, Eric Bosco et Jerry Harris	Création de programme de messagerie instantanée AOL.

1999	Jerry Yang et David Filo	Yahoo! Messager.
1999	Microsoft	Création MSN Messenger pour introduire de nouvelles fonctionnalités telles que les appels vidéo.
1999	Brad Fitzpatrick	Création LiveJournal ; plateforme de blogs.
2001	Windows Messenger	Lancement de la version intégrée de MSN Messenger.
2003	Reed Hoffman et sescollègues Thomas Anderson NiklasZenstrom&DaneJanus.Friese	LinkedIn. MySpace Skype.
2004	Mark Zuckerberg	Facebook
2005	/	Youtube
2006	Jack Dorsey	Twitter
2011	Evan Spiegel Bradley Horowitz	Snapchat Google plus
2017	Zhang Yiming	TikTok
2022	Mark Zuckerberg : PDG de Facebook	Meta

2.3 Différents types de médias sociaux

Avant la diffusion des moyens technologiques les plus utilisés à l'heure actuelle, comme Internet et télévision, la presse écrite (journaux et magazine) était le seul moyen d'interagir et de partager des informations avec le public. Aujourd'hui, après le développement extraordinaire et rapide d'Internet, avec un simple clic, nous pouvons parcourir les sites Web et trouver toutes sortes d'informations et de nouvelles en temps réel.

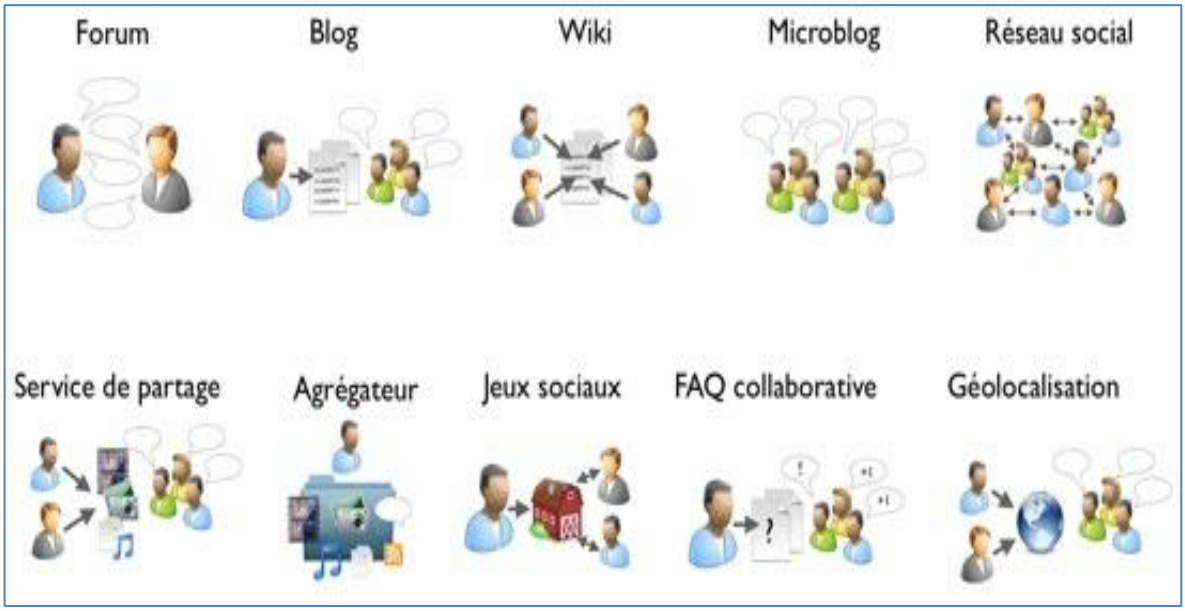

Figure I.1 : Les Différents types de médias sociaux [4].

Il existe plusieurs formes des medias sociaux, notamment [5]:

- ***Les réseaux sociaux*** (grand public comme *Facebook* ou professionnels comme *LinkedIn*) : ces services permettent à des individus de se connecter et d'interagir avec d'autres individus ayant des intérêts communs.
- ***Les sites de micro-blogging :*** ces services permettent à l'utilisateur de publier des prises de parole courtes et limitées en nombre de mots ou des images (comme *Twitter*).
- ***Les réseaux de partage de médias :*** ces services permettent de rechercher et de partager des photos, des vidéos ou d'autres types de médias sur le web. On trouve dans cette catégorie : *YouTube, Instagram, Snapchat,* ...etc.
- ***Les forums de discussion :*** ces services permettent à un groupe de personnes d'échanger des informations, des idées, des points de vue ou des actualités sur un thème précis. Ils fonctionnent sur un principe de questions/réponses (comme par exemple, *Doctissimo* dans le domaine de la santé).
- ***Les plateformes de social bookmarking*** : appelées aussi plateformes de partage de signets. Ces services permettent aux utilisateurs de sauvegarder et de partager ses ressources web (médias, contenus, signets web, …etc.) qu'ils jugent pertinentes (voir par exemple la plateforme *Diigo*).
- ***Les blogs :*** désignent une forme de journal en ligne dans lequel l'utilisateur publie des contenus sur un sujet de son choix et invite ses visiteurs à réagir sous la forme de commentaires. Ils sont hébergés sur des plateformes de blog Ging comme *Blogger* ou *WordPress*.
- ***Les sites participatifs et wikis :***Ces médias à un caractère informatif. Ils désignent un type de site web dynamique dans lequel les utilisateurs partagent leurs connaissances sur un sujet ou trouvent des informations pertinentes. Ils sont utilisés pour relier de nombreux types d'informations y compris les nouvelles, les contenus humoristiques ou les discussions. Ils facilitent la participation démocratique du Web (comme par exemple le site *Wikipédia*).

2.4 Types de continues médiatiques

2.4.1 Contenus d'actualités

Ce contenu peut être présenté périodiquement par le biais de vidéos ou de publications interactives. On trouve plus de détails dans la section 2 du deuxième chapitre.

2.4.2 Contenu pédagogique

Le contenu pédagogique Ou éducatif est une méthode marketing simple. Elle consiste à apporter des réponses précises et gratuites à des questions tout aussi précises. Les questions abordées sont issues d'une analyse type du besoin client de l'éditeur ou de l'administrateur de la plateforme [6], par exemple la plate-forme « Next Thought ».

2.4.3 Contenu inspirant

Un contenu inspirant est motivant, encourageant et une source d'énergie renouvelable comme Les récits sur les marques et comment elles sont parties de zéro [7].

2.4.4 Contenu interactif

Un contenu interactif est un contenu qui vise à attirer l'attention des internautes afin de les motiver à faire quelque chose comme répondre à des questions, s'orienter vers une certaine pensée ou faire des choix ...etc [7], par exemple : les quiz.

2.4.5 Contenu associatif

Le contenu associatif peut être compris comme un groupe qui travaille ensemble et a le désir de se réunir pour partager et faire des choses ensemble, par exemple le contenu que les organisations caritatives publient sur leurs pages et est réalisé collectivement avec des bénévoles.

2.4.6 Contenu promotionnel

Le contenu promotionnel parle d'un produit ou d'un service, en informant les abonnés et les lecteurs dans le but de commercialiser ces produits ou services non seulement pour attirer des abonnés existants en tant que clients, mais également pour gagner de nouveaux abonnés à la recherche de ce produit par exemple « Blog ouvert de Buffer» [8].

2.4.7 Contenu divertissant

Ce type de contenu prend de nombreuses formes similaires aux vidéos d'humours et aux blagues, dont le but est de divertir les abonnés afin qu'ils ne se sentent pas fatigués ou répétitifs et restent en constante interaction. Comme participer à des concours pour gagner des cadeaux

2.5 Impact et conséquences des médias sociaux

Les médias sociaux ont des effets positifs et négatifs sur l'individu et sur la société dans son ensemble. Les avantages sont nombreux, mais les inconvénients aussi, et parfois, un avantage peut devenir un inconvénient. Voici quelques impacts des médias sociaux [8]:

2.5.1 Les avantages des médias sociaux

- Élargir le cercle des relations sociales.
- Réduire les obstacles à la communication.
- Un moyen de former une opinion publique efficace.
- Un moyen efficace de promouvoir.
- Suivez l'actualité mondiale.
- Aider les hommes d'affaires et les entreprises.

2.5.2 Les inconvénients des médias sociaux

- Risques de fraude ou d'usurpation d'identité.
- Faire perdre du temps aux gens.
- pirater la vie privée des gens.
- commettre des crimes contre les utilisateurs.
- Impact sur les relations familiales.
- Violation du système des coutumes et traditions.
- Isolement.
- Faible réussite scolaire chez certain élèves.

2.6 Exploitation de données des médias sociaux

2.6.1 Concept

L'exploration de données des médias sociaux (Social Media Data Mining) fait référence au processus d'extraction de données sociales. Contrairement à l'exploration de données classique, l'exploration de données des médias socaux explore au-delà des bases de données et des systèmes internes d'une entreprise ou d'une société de recherche donnée. Cela implique généralement la collecte, le traitement et l'analyse de données brutes obtenues à partir de plateformes de médias sociaux (telles que Facebook, Instagram, Twitter, TikTok, LinkedIn, YouTube et autres), afin de découvrir des modèles et des tendances significatifs, de tirer des conclusions et de fournir des informations pertinentes et exploitables.

Le Social Media Data Mining récolte divers types de données sociales qui sont soit *accessibles au public* (par exemple, l'âge, le sexe, la profession, l'emplacement géographique, etc.) ou qui sont *générées quotidiennement sur les plateformes de médias sociaux* (par exemple, les commentaires, les likes, les clics, etc.)[9].

En règle générale, les données représentent les attitudes, les relations, le comportement et les sentiments des personnes à l'égard d'un certain sujet, produit ou service. Lors d'une tentative d'optimisation d'un contenu social, de promouvoir une activité en ligne, de découvrir des clients influents ou d'améliorer des stratégies de marketing et d'engagement, il devrait toujours se concentrer sur la collecte des types de données susmentionnés.

2.6.2 Fonctionnement

Généralement, le processus d'extraction de données sociales implique une combinaison de techniques *statistiques*, de *mathématiques* et *d'apprentissage automatique*. La première étape de ce processus consiste à collecter les données sociales provenant de différentes sources de médias sociaux. Toutes ces informations doivent ensuite être traitées avant de passer à l'étape suivante. Une fois les

données collectées et traitées, l'application de diverses techniques d'exploration de données est introduite. Ces techniques permettent d'identifier plus facilement les modèles communs et la corrélation entre les différents points de données dans de grands ensembles de données. Certaines des techniques d'exploration de données des médias sociaux les plus couramment utilisées incluent la *classification*, la détection et le suivi des formes, l'analyse prédictive, l'extraction de mots clés, l'analyse des sentiments et l'analyse du marché et des tendances.

De plus, un certain nombre de solutions logicielles d'exploration de données sont également utilisé pour optimiser le processus d'exploration. Parmi ces solutions logicielles, les approches de l'apprentissage automatique sont les plus connues.

La dernière étape du processus d'exploration consiste à créer une représentation visuelle des informations obtenues à partir de l'ensemble du processus afin de fournir les informations au public ciblé. Cela se fait généralement en utilisant l'analyse des médias sociaux ou divers outils de visualisation de données [9].

2.6.3 Utilisation

En raison des quantités massives de données générées par les utilisateurs qui sont collectées et analysées grâce au Social Media Data Mining, ce processus a trouvé une large utilisation et est de plus en plus reconnue comme un atout inestimable dans de nombreux domaines. Bien qu'il ait été principalement utilisé à des fins commerciales, ce processus est aujourd'hui souvent utilisé par les chercheurs et par les agences gouvernementales [9]

Les entreprises, les hôtels, les détaillants, les compagnies aériennes, les fabricants et même les groupes politiques achètent des ensembles de données à des sociétés d'exploration de données pour les aider à personnaliser l'expérience client, à améliorer les stratégies marketing et la satisfaction du service, et à optimiser leurs activités, en général [9].

Voici quelques exemples de qui et comment le Social Media DataMining est utilisée :

- Certaines de ses principales utilisations dans les entreprises comprennent les campagnes de marketing ciblées, les études de marché, l'aide à la vente, l'analyse prédictive, le marketing d'influence et la surveillance de la réputation de la marque.
- ***Analyse des tendances*** : Les entreprises utilisent le Social Media DataMining pour obtenir des informations précieuses sur les mots-clés, les mentions et les sujets actuellement à la mode sur les plateformes de médias sociaux.
- ***Détection d'événements (cartographie thermique sociale)*** : Cette métrique est d'une grande importance pour les agences et les chercheurs qui utilisent la surveillance des médias sociaux.

- ***Détection de spam social*** : Le Social Media Data Mining permet une détection plus facile des spammeurs et des bots sur les plateformes de médias sociaux comme Instagram et Twitter.
- ***Commerce électronique*** : Le Social Media DataMining est utilisée pour analyser la façon dont les gens parlent des produits.
- ***Médias numériques*** : Le Social Media DataMining s'applique également au domaine des médias numériques. Par exemple, le contenu qui doit être affiché sur un panneau d'affichage numérique particulier peut être décidé en menant un processus de Social Media Data Miningafin de répondre aux préférences ou aux besoins du public.
- ***Blogueurs et influenceurs des médias sociaux*** : Le Social Media DataMining est souvent utilisée par les blogueurs et les influenceurs des médias sociaux pour les aider à analyser les attitudes et les sentiments de leurs abonnés, ce dont ils parlent et ce qu'ils pensent de certains sujets de discussion.
- ***Marques*** : Le Social Media DataMining aide les marques à prendre des décisions importantes, par exemple, lorsqu'elles décident de futurs marchés potentiels.
- ***Fins de recherche*** : Les chercheurs trouvent que l'utilisation des données des médias sociaux dans leur recherche est un atout précieux pour leur travail en raison de la magnitude et la facilité d'accès aux données. Le Social Media DataMining peut être appliquée dans différents domaines de recherche, notamment les sciences sociales, la recherche en santé et la recherche technologique.
- ***Agences gouvernementales*** : Le Social Media DataMining est également de plus en plus utilisée par les agences gouvernementales à des fins d'interventions axées sur le bien-être. Pour ce faire, le Social Media DataMining consiste notamment à suivre les mouvements des résidents lorsqu'ils documentent leurs activités à des endroits marqués tout au long de la journée. De toute évidence le Social Media DataMining peut être un outil puissant qui peut aider à améliorer la vie des résidents et la sécurité des communautés.

3. La Fouille De Texte (Text Mining)

Le text mining désigne l'ensemble des méthodes informatisées utilisées pour extraire et quantifier des informations à partir de documents texte. L'exploration de texte a une définition large en incluant des outils de récupération de données et le nettoyage de texte en ligne La flexibilité du text mining le rend adapté à de nombreuses questions de recherche. Elle peut être utilisée par exemple pour définir l'usage de certains termes dans les archives, pour comprendre l'évolution de la jurisprudence Ces méthodes peuvent être appliquées à tous types de textes, qu'ils soient d'origine numérique (articles en ligne, textes postés sur les réseaux sociaux réseaux) ou qu'ils soient informatisés par Chercheur (archives, retranscriptions d'entretiens).Cette formation a pour but de permettre à chacun d'acquérir les

bases lui permettant de construire et nettoyer une base de données textuelle et d'effectuer plusieurs types d'analyses statistiques.

3.1 Concepts :

Le text mining est un processus d'extraction de connaissances, nécessaire pour transformer des documents texte non structurés en informations structurées précieuses, en mettant en œuvre des techniques statistiques ou d'apprentissage automatique [10].

3.2 Techniques liées au Text Mining

On peut dire que le text mining est un ensemble de chemins extraits de l'intelligence artificielle, qui combine de nombreux domaines linguistiques, sémantiques, statistiques et informatiques. Il s'apparente à d'autres domaines avec qui elle est très complémentaire : le traitement automatique de la langue naturelle (TALN), la recherche documentaire (RI) et l'extraction de l'information (EI).

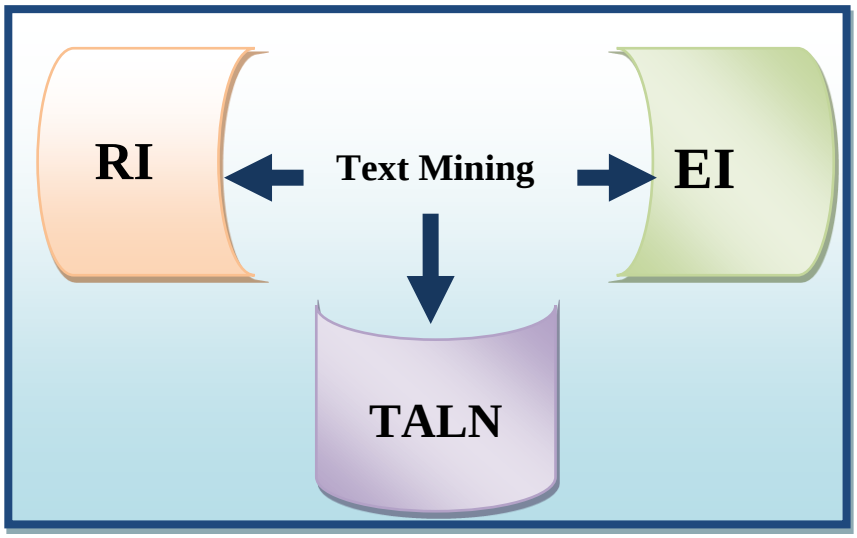

Figure I.2 : Techniques les plus célèbres utilisées au Text Mining.

3.2.1 Le traitement automatique de la langue naturelle(TALN)

Le traitement automatique du langage naturel (TALN), Automatic Natural Language Processing(ANLP), ou Automatic Language Processing (TAL), communément appelé NLP (Natural Language Processing) est un domaine interdisciplinaire qui combine la linguistique, l'informatique et l'intelligence artificielle, qui vise à créer un langage naturel outils de traitement pour diverses applications [11].

3.2.2 La recherche d'information (RI)

La recherche d'informations (RI) fait référence au processus d'extraction de modèles pertinents et associés basés sur un ensemble spécifique de mots ou de phrases. Dans cette technique d'exploration de texte, les systèmes IR utilisent différents algorithmes pour suivre et surveiller les comportements

des utilisateurs et découvrir les données pertinentes en conséquence. Les moteurs de recherche Google et Yahoo sont les deux systèmes IR les plus connus [12].

3.2.3 L'extraction d'information (EI)

Ce sont les techniques de text mining les plus populaires. L'échange d'informations fait référence au processus d'extraction d'informations significatives à partir de gros morceaux de données textuelles. Cette technique de fouille de texte se concentre sur l'identification de l'extraction d'entités et d'attributs et de leurs relations à partir de textes semi-structurés ou non structurés. Quelle que soit l'information extraite, elle est stockée dans une base de données pour un accès et une récupération futurs [13].

4. Conclusion

Le premier chapitre du notre travail considère comme une introduction au sujet de notre recherche où nous avons introduit les concepts de base sur les médias sociaux, les types d'articles les plus importants qui y sont présentés ainsi que leurs utilisations. Ensuite, nous avons abordé de l'exploration de texte et son fonctionnement afin d'arriver aux leurs techniques les plus importantes utilisées et qui sont liées à notre travail.

1. Introduction

La lecture d'articles de presse est essentielle et vitale pour comprendre et suivre les événements émergents et en développement aux niveaux local, étatique et mondial, ainsi que pour comprendre les demandes des citoyens et les opinions des critiques. Grâce à la prolifération des médias sociaux en tant que canaux d'information, les citoyens et les groupes professionnels échangent facilement des nouvelles et des opinions, ajoutant plus de nouvelles à découvrir et à traiter. Cette étude se concentre sur la façon dont les algorithmes de classification d'apprentissage automatique peuvent aider à classer les actualités en différentes catégories pour atteindre facilement la catégorie d'actualités souhaitée et filtrer les actualités bruyantes et nuisibles à mesure qu'elles arrivent en temps réel.

Dans ce sens, ce chapitre sera principalement consacré à la présentation des articles d'actualité en générale, et les méthodes d'apprentissage automatique utilisées pour catégoriser ces articles en particulier.

2. News d'actualité

2.1 Définition

L'actualité (nouvelle) est par définition des informations liées à des événements récents qui sont diffusées par les médias.

2.2 Critères d'actualité

La présentation des informations dans les médias a des priorités en fonction de l'importance de l'actualité qui ferait que la publication soit montrée en premier et avec les détails, et donc les nouvelles les moins important sont montrées en dernier et avec moins de détails. Cette distinction dans l'importance de l'information se fait par le jugement des journalistes, sur le niveau d'intérêt pour la société et l'importance relative de l'événement. Nous mentionnons le plus important de ces critères dans les points suivants [14] :

2.2.1 Nouveauté de l'actualité

L'un des critères les plus élevés qui déterminent la qualité d'un article est le facteur temps. Il est très important pour déterminer le niveau d'importance de l'actualité, car à chaque instant qui passe, l'information ne sera pas nouvelle, donc le public ne lui donnera pas le niveau d'attention qu'il mérite.

2.2.2 Irrégulier *(*inhabituel*)*

Les nouvelles inhabituelles varient d'un pays à l'autre, par exemple, «le banquet de bœuf » n'est pas une nouvelle importante dans notre société et ne sera pas en tête de l'actualité. Contrairement à ce qui

se passerait dans un pays comme l'Inde, il est inhabituel que la consommation de bœuf soit strictement interdite dans leur croyance.

2.2.3 Contenu intéressant

De nombreux sujets attirent l'attention du public, par exemple, des informations telles que « a découverte d'une nouvelle planète dans notre galaxie» est une actualité et inconnue en mémé temps, mais cette nouvelle n'intéressera pas le public. Alors que des nouvelles comme «Le prix du pétrole baisse ou augmentede 80% » est une actualité qui fera grand bruit car elle affectera directement la situation financière des habitants d'un pays producteur de pétrole.

2.2.4 Exclusivité

Il existe une source pour chaque journal ou chaîne de télévision dans chaque grande entreprise du monde, ces sources leur donnent des informations exclusives. Par exemple, « Des augmentations significatives des salaires des travailleurs », ce genre de nouvelles attirera beaucoup l'attention du public, et il sera donc à la une des journaux ou des chaînes d'information.

2.3 Les domaines d'actualité

Les préférences des utilisateurs pour une telle ou telle catégorie d'articles publiés sur les sites de médias sociaux sont différentes, ce qui crée des disparités d'audience ou de navigation sur les sites. La figure ci-dessous montre les catégories d'articles les plus consultés :

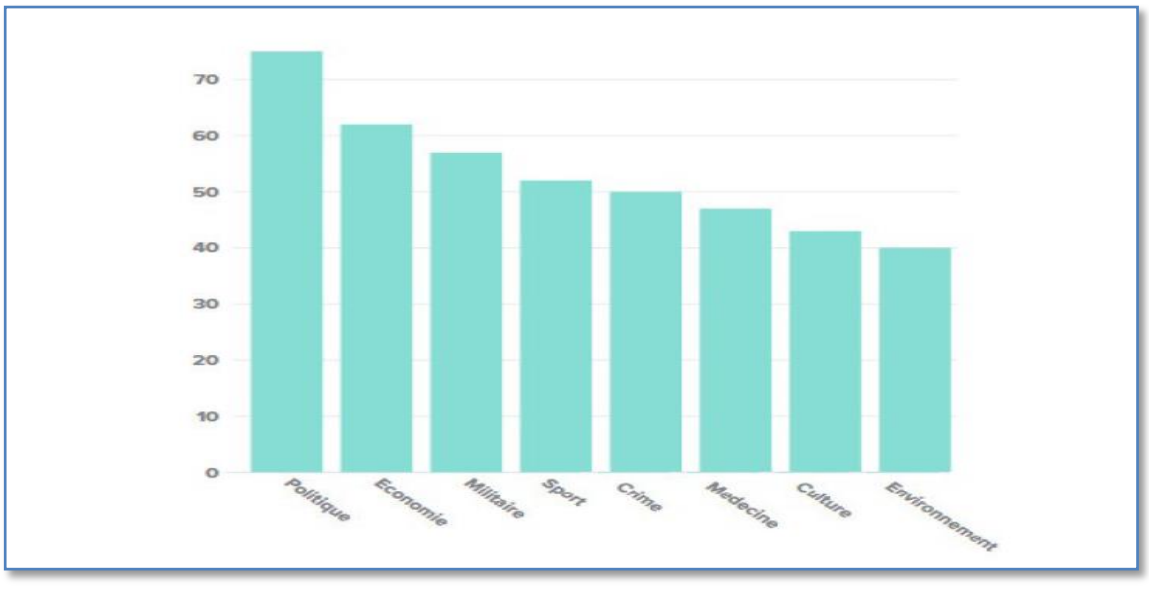

Figure II.1 : Les topiques le plus consultée par le public de masse [14].

2.4 Diffusion des actualités

Récemment, on constate un déplacement de la consommation d'information du public vers Internet, ce qui a réduit le nombre de vus des informations qui sont traditionnellement diffusées. Ainsi, la presse écrite est menacée. Une partie de la migration du public vers les journaux en ligne liés au journalisme

a été le passage au monde numérique. En outre, il existe des systèmes tels que les services de recherche d'actualités, les moteurs de recherche et les notifications d'actualités, qui permettent d'accéder à une gamme de sources d'actualités [15].

Les radiodiffuseurs traditionnels tentent de rivaliser en fournissant des informations de dernière minute sur les évènements importants, ce qui signifie que les actualités d'aujourd'hui ne sont pas fixes au moment du dernier tirage, mais changent constamment.

3. Classification des articles d'actualité

Sur Internet, on suppose que la multiplicité des sources garantit la multiplicité des informations. La couverture des milliers d'articles issus de toutes catégories de sites, l'analyse lexicale des titres et l'identification des sujets d'actualité dévoilent une réalité encore plus contrastée. La grande variété des sujets abordés sur le web se traduit simultanément par une concentration sur quelques sujets majeurs, souvent traités de manière redondante jusque dans leur formulation. Ce traitement repose principalement sur les méthodes de classification automatique des articles.

3.1 Méthodes de classification

Les méthodes de classification visent à déterminer à quelle classe appartiennent les objets spécifiés dans leur description. Il existe deux types d'approches de classification : l'approche supervisée et l'approche non supervisée (clustering) [16].

- ***Classification supervisée.*** Dans la classification supervisée, un échantillon représentatif de l'ensemble des formes à reconnaître est fourni au module d'apprentissage. Chaque forme est étiquetée par un opérateur appelé superviseur ou professeur. Cette étiquette permet d'indiquer au module d'apprentissage la classe dans laquelle le professeur souhaite que la forme soit rangée. Cette phase d'apprentissage consiste à analyser les ressemblances entre les éléments d'une même classe et les dissemblances entre les éléments de classes différentes pour en déduire la meilleure partition de l'espace des représentations. Les paramètres décrivant cette partition sont stockés dans une table d'apprentissage à laquelle le module de décision se référera ensuite pour classer les formes qui lui sont présentées.
- ***Classification non-supervisée***. Dans la classification non-supervisée (ou clustering), il n'y a aucun superviseur explicite et le système forme des clusters (ou des groupements) des exemples d'entrée (formes non étiquetées). L'étape de la classification va se charger d'identifier automatiquement les formes appartenant à une même classe.

3.2 Processus de classification de texte

La classification de texte consiste à créer un classificateur à base d'un corpus de documents étiquetés. Ce corpus est partitionné en deux ensemble : ensemble d'apprentissage et ensemble de test. Tous d'abord, ce classificateur s'entraine avec l'ensemble d'apprentissage, et une fois appris, il teste son efficacité avec l'ensemble test. Dans certain cas, ce processus se termine par une étape de validation avec un ensemble de nouveau documents. Le schéma ci-dessous illustre parfaitement ces différentes étapes

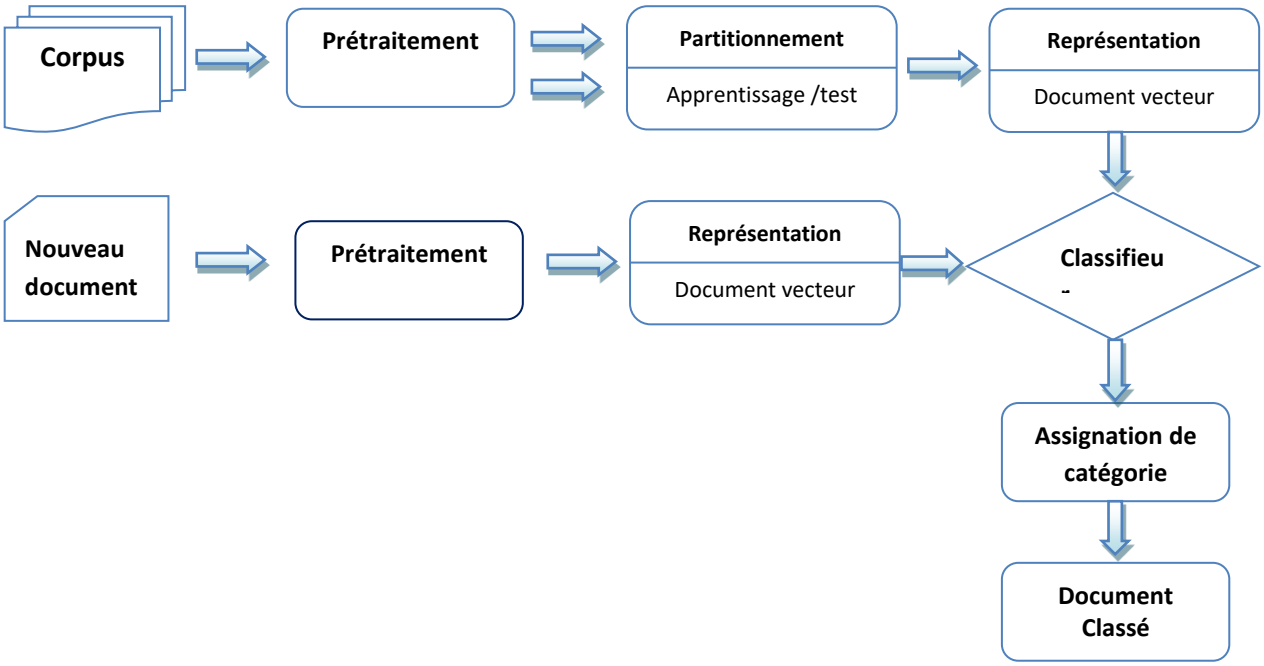

Figure II.3: Processus de classification de documents.

La première étape consiste à transformer le document texte en document vecteur pour ensuite le présenter aux algorithmes de classification. Cette transformation est appelée le « prétraitement ».

3.2.1 Le prétraitement :

Dans cette étape, nous ne présentons quels prétraitements classiques qui apparaissent souvent dans la littérature. Cependant, les traitements utilisés dans notre cas seront détaillés dans le troisième chapitre.

- ***Tokenisation :***

Un token (jeton) est définie comme une séquence de caractères comprise entre deux séparateurs. Les séparateurs étant des blancs, des signes de ponctuation et certains autres caractères tels que des guillemets ou des parenthèses. La Tokénisation consiste à segmenter un document texte en jetons de mots, en séquences de mots ou en de phrases pures, mais travaille souvent sur des mots [17].

- **Mots vides :**

Les mots vides tels que: les articles, les prépositions, les déterminants, les adverbes...etc., comme « la, dans, cet, ici, car, » dans la langue française et « the, in, and, after » dans la langue anglaise, sont présents dans tous les textes du corpus. Ils représentent environ 30% des mots dans un texte. La présence de ces mots ne fait aucune différence d'un point de vue linguistique et lexical. Cela signifie que leur présence dans tous les textes du corpus les rend non discriminatoires, et donc leurs utilisation dans une tâche de classification se révèle être inutile. D'autre part, leur suppression réduit la dimension du document vecteur, par conséquent, le temps de traitement et le temps d'apprentissage seront considérablement réduits[17].

- **La racinisation (Stemming) et la lemmatisation:**

Pour représenter l'ensemble de texte, chaque mot est considéré comme un descripteur. Cette manière de faire comporte certaines irrégularités, notamment pour les verbes à l'infinitif et les verbes conjugués (acheter, achète, acheté,...), on remarque que ces derniers ont le même sens mais chacun d'eux est considéré comme un des cripteurs à part. La racinisation permet de surmonter ce problème, en ne considérant que la racine de ces mots plutôt que les mots en entier sans se soucier de l'analyse grammaticale [17].

Contrairement à la racinisation, la lemmatisation met en évidence l'analyse grammaticale. Elle ramène les termes à leur forme canonique en mettant les noms au singulier, les adjectifs au masculin singulier, et les verbes conjugués à l'infinitif.

La substitution des mots par leurs racines permet une meilleure représentation et réduit considérablement la dimension des descripteurs (surtout dans le cas de la lemmatisation).

3.2.2 La représentation (poids)

Avant qu'un classificateur puisse être crée et formé à partir d'un corpus de documents donné, il est fortement nécessaire de convertir les documents textes en entrées valides et compréhensibles par les différents algorithmes de classification. Ces entrées valides sont en fait des vecteurs ou des matrices qui définissent le poids (fréquences) de chaque descripteur (mot ou groupe de mots, n-gramme) dans chaque document texte dans lequel elles apparaissent. En revanche, si un tel descripteur n'existe pas dans un tel document texte, alors son poids sera nul [17].

À partir de ces poids, on peut dire qu'un descripteur quelconque est discriminant ou pas (apparait fréquemment ou pas du tout dans le document en question) si son poids est élevé ou pas, respectivement. Mais malheureusement cette méthode de mesure n'est pas toujours correcte, surtout dans des longs documents avec de nombreux paragraphes.

Il existe plusieurs manières pour représenter l'ensemble de descripteurs.

- **Le sac à mots (bag of words)**

Cette méthode consiste à utiliser les mots comme descripteurs. Elle permet de construire un sac de mots (bag of words) à partir de tous les mots qui apparaissent au moins une fois dans le corpus. Les mots sont regroupés en vrac et sont traités indépendamment. Du fait de la présence de certaines anomalies, notamment en ce qui concerne la variation des fréquences par rapport à la longueur du document, la présence des mots composés, et principalement, le fait que l'ordre d'apparition des mots dans les phrases du document n'est pas pris compte, font que cette méthode est loin de répondre à toutes les attentes de la classification de texte [17].

- **Fréquence des termes (TF) :**

TF fais référence à la fréquence (répétition) d'un mot (descripteur) dans un texte donné. C'est un calcul très simple, mais il s'avère efficace et pratique. Il a souvent utilisé en combinaison avec d'autres fréquences. Il existe plusieurs manières de calcul de la TF [17]:

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

$$TF = NT \tag{1}$$

Où NT *:* est le nombre de fois que le terme est apparait dans le texte.

TF relative : c'est le rapport entre le nombre de fois où un terme est apparait dans le texte et le nombre de tous les termes du texte. Cette manière est généralement utilisée pour limiter l'effet de la longueur des textes. En effet, un terme qui apparait 10 fois dans un texte de 400termes n'a pas le même degré de distinction qu'un terme qui apparait le même nombre de fois dans un texte de 30 termes.

$$TF = \frac{NT}{ST} \tag{2}$$

Où *NT* : est le nombre de fois que le terme est apparu dans le texte. ST : est le nombre detous les termes du texte.

TF booléenne : pour juste avoir la présence ou l'absence d'un terme dans le texte.

$$TF = 0 \; ou \; 1 \tag{3}$$

Le principal inconvénient de la *TF* est le fait que le terme peut apparaitre avec une fréquence assez élevée dans tous les documents d'un corpus. Ce terme, dans ce cas, perd toute sa notion de distinction

liée au degré de présence. Pour rectifier ce cas exceptionnel, un autre concept nommé *IDF* (*Inverse documents frequancies*)a été introduit [17].

- **Fréquence documents inverses (IDF) :**

IDF mesure le degré de rareté d'un terme, non pas dans un document, mais danstous les documents d'un corpus. Elle est déterminée par l'équation suivante:

$$IDF = \log(\frac{ND}{DT}) \quad (4)$$

Où *ND* :est le nombre de documents dans le corpus, et *DT* : est le nombre de documents dans lesquels le terme est apparu.

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

$$IDF = \log(ND) \quad (5)$$

Cette pondération seule ne détermine en rien le degré de discrimination d'un terme dans un document car elle est relative au corpus. L'association d'*IDF* avec *TF* donne des résultats importants.

- **TFIDF :**

On a vu que la fréquence d'un terme dans un document texte joue un rôle important dans lecalcul de son degré de discrimination. En outre, la représentation d'un texte ne dépend pas seulement de son contenu, mais aussi au corpus au quel le texte appartient. La rareté d'un terme dans les autres documents du corpus s'avère aussi importante que sa fréquence dans le document en question. La combinaison du principe de fréquence particulière avec du principe de la rareté générale a conduit à la pondération dite *TFIDF*. Cette combinaison est calculée par la formule suivante [15]:

$$TFIDF = TF * \log(\frac{ND}{DT}) \quad (6)$$

Où *TF* est relative ou absolue.

3.2.3 Les Algorithmes :

Il existe plusieurs algorithmes de la classification de texte. Dans ce manuscrit, on s'intéresse uniquement aux cinq algorithmes [21], qu'on va les comparer dans le chapitre 3.

a. Machine à Vecteur de Support (Support Vector Machine SVM)

Le SVM est un classificateur dit linéaire, ce qui signifie que dans le cas idéal, les données (document texte) doivent être linéairement séparables. Le corpus est représenté sous forme d'un espace vectoriel, où chaque document texte est représenté par un point dans ce dernier. La problématique

maintenant est de trouver le meilleur séparateur (ligne, plan ou hyperplan) qui divise le corpus en deux catégories. Alors, le principe est simple : Le SVM vise à séparer les textes en classes à l'aide d'une frontière, de telle façon que la distance entre les différents groupes de textes et la frontière qui les sépare soit maximale. Cette distance est aussi appelée « marge » qui est définie par les points (Vecteurs de support) les plus proches de la frontière. Le SVM est ainsi qualifié de « séparateurs à vaste marge », dont l'objectif principale est de maximiser cette marge, plus elle est grande, meilleurs est le résultat [18].

Dans le cas où les données ne sont pas linéairement séparables, le SVM s'appuie souvent sur l'utilisation des fonctions mathématiques qui permettent de séparer les données en les projetant dans un espace vectoriel de plus grande dimension. Ainsi, la technique de maximisation des marges assure une meilleure robustesse au bruit (erreur) et donc un modèle plus généralisable.

La figure ci-dessous montre comment SVM divise les données dans le cas où elles sont linéairement séparables, de plus, il choisit la meilleure séparation où la marge est maximale.

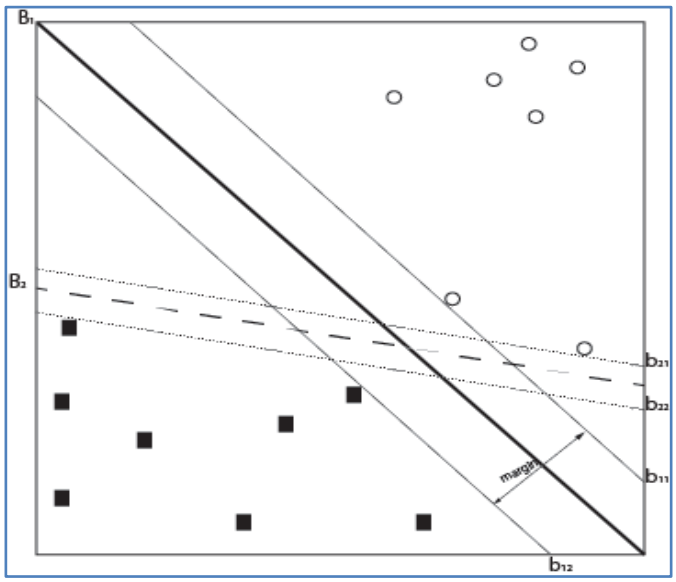

Figure II.4:SVM pour un problème à deux classes. B2: Séparation correcte. B1: séparation optimale [17].

- **Avantages**
  - Le SVM est un classificateur qui se généralise bien avec les nouvelles données, puisque les vecteursde support sont les points les plus proches du séparateur, pas les plus éloignés. Par conséquent, lesvaleurs aberrantes n'affectent en rien le classificateur.
  - Le SVM est flexible dans le traitement des cas non-linéairement séparables.
- **Inconvénients**

Le temps d'apprentissage du SVM est relativement élevé par rapport aux autres algorithmes d'apprentissage machine.

b. Arbre de décision

L'arbre de décision est composé d'un ensemble de nœuds liés par des branches. Il consiste souvent en un nœud central à partir duquel peuvent être tirées plusieurs Data possibles. Les nœuds conduisent à d'autres nœuds qui, à leur tour, ouvrent plusieurs autres possibilités. On distingue trois types de nœuds. Le nœud de décision (représenté par un carré) indique une décision à prendre. le nœud de hasard (représenté par un cercle) met en évidence les probabilités de certaines Data. Le nœud terminal permet d'avoir le résultat final d'un chemin sur l'arbre [17].

L'arbre de décision permet aussi de classer les textes en plusieurs catégories. Il est facile à interpréter et à entraîner. Tout d'abord, il faut mettre dans le nœud racine le descripteur qui distingue le mieux les textes du corpus, pour obtenir de nouveaux nœuds enfants (sous nœuds), et ainsi de suite. On répète le processus pour chaque nœud, jusqu'à ce que la séparation des textes ne soit plus possible ou souhaitable. Finalement, les nœuds feuilles sont constitués d'ensemble de textes de même classe [17].

La figure ci-dessous présente une forme d'un arbre avec des branches multiples.

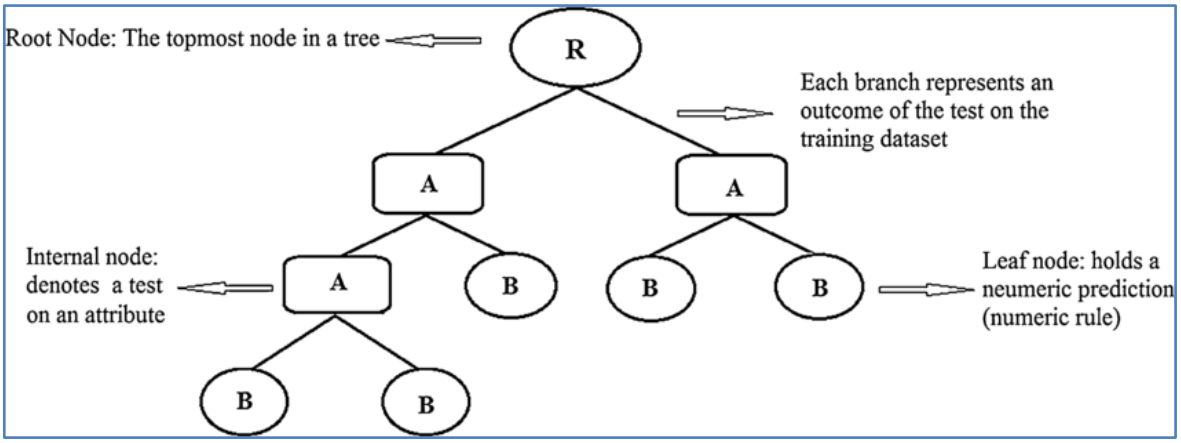

Figure II.5. Arbre de décision [19].

- **Avantages**
  - Les arbres de décision sont construits à partir d'un ensemble de règles très explicites qui permettent bien comprendre les résultats par l'utilisateur.
  - Ils ont généralement besoin de peu de ressources, et leurs temps de d'apprentissage et de test sont relativement courts.

- **Inconvénients :**
  - Lorsque l'arbre devient assez grand, c'est-à-dire qu'il comporte beaucoup de feuilles, il perd leur pouvoir explicatif.

- Quand les sous-nœuds soient directement dépendants du nœud racine, il rend la présence ou l'absence d'un seul descripteur dans un texte cruciale pour le sort de leur classement.
- La modification d'un seul nœud, s'il est proche de la racine, modifie complètement l'arbre.
- La probabilité que l'arbre de décision soit exposé à un sur-apprentissage.

c. Classifieur bayésien naïf (NBC)

La classifieur bayésien naïf (Naïf Bayes Classifier) est un typede classification Bayésienprobabiliste simple basée sur le théorème de Bayesavec une forte indépendance (dite naïve) des hypothèses. En termes simples, il suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques .

$$P(A/B) = P(B/A)\frac{P(A)}{P(B)} \quad (7)$$

➢ **Avantages**

- Le classifieur Naïve Bayes permet d'apprendre rapidement pour un modèle de classification car les calculs de probabilités (les hypothèses d'indépendance des descripteurs) ne sont en fait pas très coûteux.
- Il n'est pas nécessaire de fournir un gros volume de données pendant de la phase d'apprentissage, en effet, la classification est possible même avec un petit jeu de données

➢ **Inconvénients :**

S'il y a une grande corrélation entre les caractéristiques (dans le casdes documents longs c.-à-d. un vocabulaire riche qui favorise les dépendances entre les descripteurs), alors le NBC va donner de mauvaises performances.

d. Régression logistique

La régression logistique est par définition un modèle d'analyse statistique qui consiste à prédire une valeur de données d'après les observations réelles d'un jeu de données. Il permet d'étudier le lien entre une variable principale et plusieurs variables explicatives (prédictives). En effet, elle est un modèle linéaire généralisé qui utilise une fonction logistique comme une fonction de lien entre la variable d'intérêt et les autres variables [17].

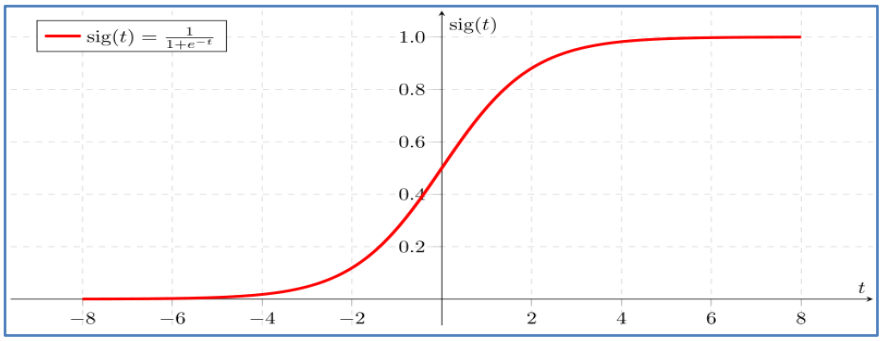

Figure II.6: Exemple d'une fonction logistique [17]

➢ **Avantages**

- La régression logistique est plus facile à mettre en œuvre et à interpréter, et ne nécessite pas de mise à l'échelle des caractéristiques d'entrée.
- Elle est très efficace à entraîner et les résultats qu'elle fournit sont des probabilités prédites bien calibrées.
- Elle fonctionne plus efficacement lorsque les attributs non liés à la variable de sortie et ceux qui sont corrélés, sont omis.
- Elle fonctionne bien avec un ensemble de données volumineux, surtout sicet ensemble est linéairement séparable.

➢ **Inconvénients**

- La régression logistique ne peut pas être utilisée pour résoudre des problèmes non linéaires et, malheureusement, de nombreux systèmes actuels sont non linéaires.
- La vulnérabilité au sur-apprentissage est connue.

e. K-voisins les plus proches (K – Nearest Neighbours KNN)

Le KNN est un algorithme de classification supervisé non linéaire basé sur la prédiction de données en trouvant les similitudes avec les données sous-jacentes. Le principe est de prédire à quelle classe appartient un nouveau point de données de test en identifiant la classe de ses k voisins les plus proches. La sélection de ces k voisins les plus proches est basée sur une mesure de similarité (généralement la distance euclidienne). À partir de ces k voisins, le nombre de points de données dans chaque catégorie est compté et la catégorie avec le plus de voisins est attribuée au nouveau point de données [17].

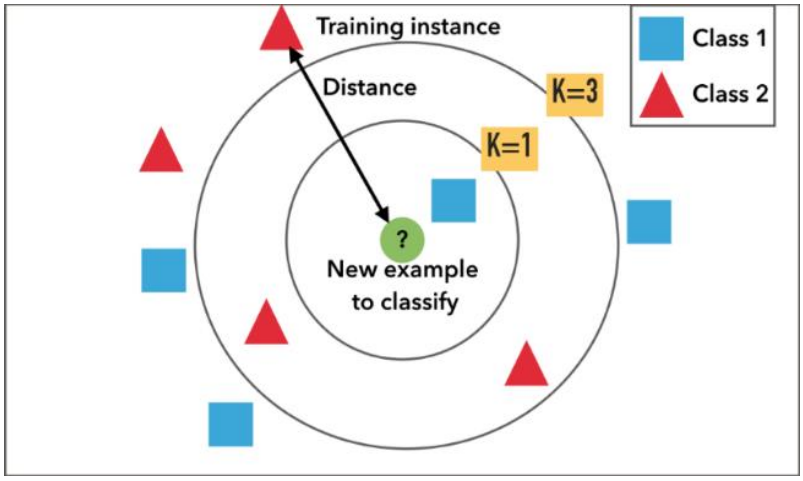

Figure II.7: KNN [17]

➢ **Avantages**

- Le KNN est modèle simple à comprendre, facile à mettre en œuvre, rapide et efficace.
- Il n'existe aucune hypothèse sur les données (linéaires, affines,...).

➢ **Inconvénients**

- L'implémentation de KNN a besoin de choisir manuellement la méthode de calcul de la distance ainsi que le nombre de voisins «k».
- Si le nombre de données d'apprentissage est augmenté, Il devient beaucoup plus lent à mesurer.

3.2.4 Les Scores

Les scores sont également importants pourvoir à quel point un tel classificateur est fiableetse comporte-t-il bien avec de nouvelles données? Plusieurs scores(mesures) existent pour vérifier et estimer avec précision le degré de généralisation d'un tel classifieur, et ainsi refléter lesuccès ou l'échec de ce classifieur.

❖ **La matrice de confusion**

Puisque la performance d'un algorithme d'apprentissage automatique est directement liée à sa capacité à prédire un résultat. L'utilisation de la matrice de confusion est notre meilleure option pour comparer les résultats de l'algorithme avec la réalité. L'apprentissage automatique consiste à alimenter un algorithme avec des données afin qu'il apprenne par lui-même à effectuer une tâche particulière. Dans les problèmes de classification, il prédit les résultats qui doivent être comparés à la réalité pour mesurer le degré de sa performance. On utilise généralement la matrice de confusion, également appelée tableau de contingence. Cela fera non seulement ressortir des prédictions correctes et incorrectes, mais surtout nous donnera une idée du type d'erreurs commises [32].

Table 2.2 : Explication de la matrice de confusion

		Value Prédictif	
		Positive	**Négative**
Value actuel	**Positive**	Vrai positif [VP]	False Négative [FN] (Type II Erreur)
	Négative	False Positive [FP] (Type I Erreur)	Vrai Négative [VN]

Terminologies utilisées dans Confusion Matrice

- Vrai positif → Classe positive qui est prédite comme positive.
- Vrai négatif → classe négative qui est prédite comme négative.
- Faux positif → Classe négative qui est prédite comme positive. [Erreur de type I]
- Faux négatif → Classe positive qui est prédite comme négative. [Erreur de type II]
- Le faux positif (FP) est également connu sous le nom d'erreur de type I.

- Le faux négatif (FN) est également connu sous le nom d'erreur de type II

Notez qu'avec ces quatre paramètres, toutes les autres métriques sont calculées :

❖ **Le rappel :** Le rappel (ou recall en anglais) est un paramètre qui permet de mesurer le nombre de prévisions positives correctes sur le nombre total de données positives.

$$Rappel\ (C) = VP/\ VP + FN \tag{8}$$

❖ **La précision :** La précision est le nombre de documents pertinents retrouvés rapporté au nombre de documents total proposé pour une requête donnée.

$$Précision(C) = VP\ /VP + FP \tag{9}$$

❖ **La F-Mesure :** C'est la moyenne harmonique de la combinaison de la valeur de rappel et de la valeur de précision.

$$F - Mesure = 2 * (Précision * Rappel)/\ Précision + Rappel \tag{10}$$

❖ **Le taux d'erreur et de succès:**

Le taux de succès est le rapport entre les documents bien classés sur le nombre total des documents du corpus.

$$Taux\ de\ succès = VP + VN\ /VP + FP + VN + FN \tag{11}$$

Le taux d'erreur est le rapport entre les documents mal classés sur le nombre total des documents du corpus [31].

$$Taux\ d'erreur = FP + FN/\ VP + FP + VN + FN \tag{12}$$

4. Conclusion

Dans ce chapitre, nous avons présenté les principaux fondements, méthodes et techniques de classification des articles d'actualité. Dans la première partie de ce chapitre, nous avons abordé les articles d'actualité, en nous concentrant sur leurs critères et leurs domaines d'actualité. Dans la deuxième partie, nous avons présenté les principaux étapes suivies pour la réalisation d'un système de classification automatique des articles d'actualité. Nous avons également détaillé chaque étape avec analyse et spécification des différentes méthodes et techniques qui y sont développées et utilisées.

Dans le chapitre suivant, nous décrivons, étape par étape, les solutions que nous avons adopté à cette classification afin d'améliorer et d'obtenir une meilleure précision.

1. Introduction

Le chapitre précédent permet de présenter l'architecture de notre modèle d'apprentissage automatique qui prend en compte l'aspect abstrait de notre travail. Dans ce chapitre, nous chercherons à expérimenter et à évaluer les algorithmes proposés d'apprentissage automatique pour la classification des articles d'actualité.

2. Problématique de la classification des news

Compte tenu de l'énorme expansion d'un système de classement automatisé des articles de presse, son développement pose naturellement un certain nombre de défis. Ces problèmes sont principalement liés au très grand volume de données à classer, ce qui conduit au problème de la complexité. En plus de ces données de nature textuelle, toutes les problématiques liées à leurs représentations, à la pondération de leurs contenus, à leurs sémantiques rendent décile le processus de détection des sujets traites [26].

Ci-dessous une liste de défis auxquels nous pouvons rencontrer dans le contexte de ce travail :

- ***La graphie***

Un terme peut inclure des erreurs d'orthographe ou de frappe car il peut être écrit de plusieurs manières ou écrit en majuscules (Ouargla et Wargla) etce qui affectera la qualité des résultats.

- ***Les mots composés***

Lorsque le système ne prend pas en charge des mots composés (tels que : casse-tête, laissez-passer, après-midi, pomme de terre, etc.) qui sont trop nombreux dans toutes les langues, cela réduit considérablement les performances de ce système de classification. Par exemple, il traite le mot pomme de terre comme étant trois termes distincts.

- ***Redondance***

La redondance permet d'exprimer le même concept à travers diverses expressions différentes, c.-à-d. il existe plusieurs façons d'exprimer la même chose. Par exemple, on peut tout simplement dire « je suis tout à fait d'accord avec toi » ou bien « tu as tout à fait raison » ; « tout à fait » renforce ici la notion d'accord. Cette difficulté de redondance est liée à la nature des documents qui ont été traités en langage naturel contrairement aux données numériques

- ***Ambiguïté***

Contrairement aux données numériques, les données textuelles se caractérisent par leur richesse linguistique car c'est la pensée humaine qui les a conçues et sa logique, ce qui a entraîné une certaine

ambiguïté, par exemple : le mot avocat peut désigner le fruit, le juriste, ou encore au sens figuré, la personne qui défend une cause.

- ***Complexité de l'algorithme d'apprentissage***

Un texte est représenté généralement sous forme de vecteur qui contient les nombres d'apparitions des termes dans ce texte. Cependant, la taille de corpus (le nombre de textes) qu'on va traiter est très important sans oublier la taille de chaque texte (le vocabulaire ou le nombre de termes composant le même texte).Par conséquent, on peut bien imaginer la dimension du vecteur (nombre de textes * nombre de termes) qui sera traité et qui compliquera considérablement la tâche de classification en diminuant la performance du système.

- ***Présence et absence de termes***

Il y a une relation implicite entre le mot et le concept qui lui est associé car la présence d'un mot dans le texte indique un point que l'écrivain a voulu exprimer, même si l'on sait qu'il y a plusieurs manières d'exprimer les mêmes choses, mais l'absence de un mot ne signifie pas nécessairement l'absence du concept qui lui est associé. C'est ce qui nous rend plus prudents dans l'utilisation de techniques d'apprentissage automatique qui consistent à exclure un certain mot

- ***Le temps d'apprentissage***

Il partira en fait du document vectoriel (données d'entrée pour l'algorithme de classification) qui est généré à partir de l'ensemble qui a une dimension incroyable (large). En calculant le nombre de documents dans l'ensemble de documents et le nombre de mots dans chaque document, ce qui au final on aboutit à un tableau contenant tous les mots à traiter avec leurs propres poids, la mise en œuvre de ces étapes occupera une beaucoup de ressources et prendra plus de temps

- ***Le sur-apprentissage (overfitting)***

Les documents vectoriels de grande dimension produisent généralement un sur apprentissage, ce qui signifie que les documents d'apprentissage sont bien triés, mais que le classifieur interagit mal et ne généralise pas aux nouveaux documents. Il existe plusieurs façons de réduire les dimensions pour résoudre ce problème. C'est l'un des points les plus importants sur lesquels nous nous sommes concentrés.

- ***Étiquette***

Ce problème est spécifique à l'apprentissage supervisé, où l'ensemble d'apprentissage est catégorisé par des experts humains. Parce que chaque expert a sa propre évaluation de l'attribution d'une catégorie à un document. Le document peut être classé en deux ou plusieurs catégories différentes, à la discrétion des experts. Il en résulte une percée dans l'objectif d'obtenir un classeur parfait.

3. Formulation du problème

Le problème de classification des articles que nous somme entrain de traiter peut être formulé comme suit :

Input:

- **S** =S_1, S_2,S_3,. . .,S_n. *l'ensemble des sources d'actualités.*
- **C_News**=N_1, N_2,N_3,. . .,N_m. *le corpus de news (l'ensemble des articles).*
- **T**= T_1,T_2,T_3, . . .,Tn. *l'ensemble des termes (mots) existantes dans tout le corpus.*
- *Chaque article (texte)*A_i *sera représenté par un vecteur pondéré.*
  A_i= {A_iT_1, AiT_2, AiT_3,...,A_iT_n}.
  AvecA_iT_j*est le poids (pondération) de mot* T_j *dans l'article* A_i.

Output :

- **Catégories_Articles**= (C_1, C_2, C_3,..., C_k).*Ensemble de groupes d'articles ou chaque groupe* (C_i, $0<i\leq k$) *rassemble un certain nombre d'articles* ***traitant d'un même sujet d'actualité.***

Contrainte :

- **Densité de la catégorie (Intra-classe) :** Minimiser la distance moyenne entre l'article A_i et tous les autres points du classe C_k auquel il appartient.
- **Diversité entre catégories (inter-classes) :** Maximiser la distance moyenne entre l'article A_i qui appartient à une classe C_k et tous les autres points d'une autre classe C_l

4. Environnement de développement

4.1 Environnement logiciel

4.1.1 Langage de programmation

Le langage de programmation utilisé dans notre travail est Python. Selon la communauté TIOBE, Python est élu « langage de l'année » pour la 5ème fois de l'histoire du classement [27].

Il est utilisé pour développer des applications GUI, des sites Web, la science et la visualisation des données et il est fortement recommandé dans l'apprentissage automatique et l'intelligence artificielle.

Nous choisissons Python en raison de ses différentes bibliothèques qui permettent aux développeurs d'effectuer des tâches complexes sans avoir besoin de réécrire de nombreuses lignes dans le code.

Python est un langage de programmation informatique interprété orienté objet. Il est à la fois simple et puissant, il permet d'écrire des scripts très simples et grâce à ses nombreuses bibliothèques, il permet de travailler sur des projets encore plus ambitieux (comme les gros projets de sites internet et les tâches de maintenance).

Python est un langage facile à apprendre et son code est plus lisible, il est donc plus facile à maintenir.Il est parfois plus concis que d'autre langage de programmation, ce qui augmente la productivité du développeur et réduit mécaniquement le nombre d'erreurs.

4.1.2 Environnement (framework)

PyCharm est un Environnement de développement intégré (IDE) Python développé et édité par JetBrains[28].

PyCharm est un IDE complet basé sur la productivité avec : des systèmes d'auto-complétion intelligente, d'analyse de code en temps réel, l'intégration des outils de tests et de debugging ; et l'intégration d'un grand nombre de raccourcis clavier permettant de réaliser presque n'importe quelle tâche rapidement sans lever les mains sur le clavier pour utiliser la souris.Il existe de nombreuses versions de cet IDE, dans notre cas, nous utilisons la version PyCharm2018.3.7.

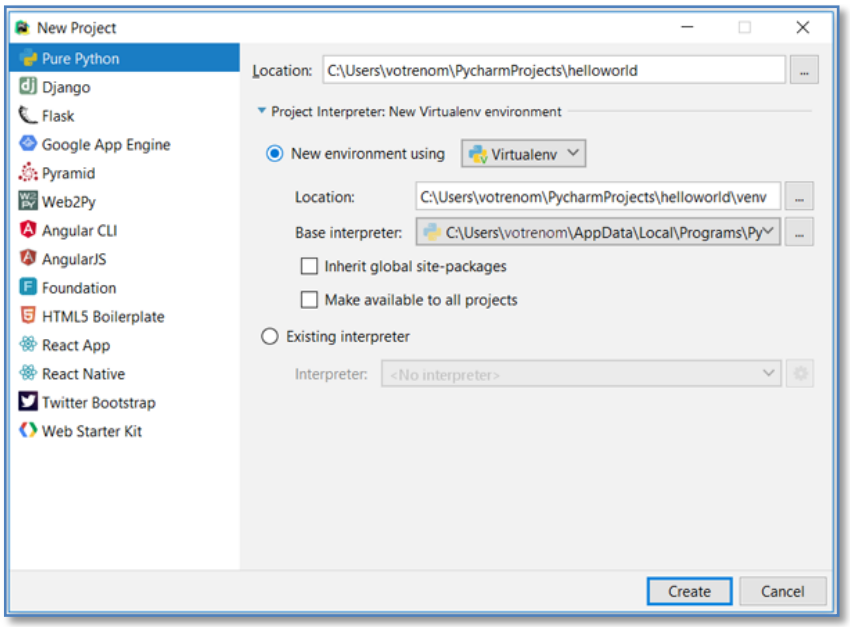

Figure III.1 : Éditeur PyCharm[28].

4.1.3 Bibliothèques

Pour générer des prévisions de catégories d'articles d'actualité à partir des publications sur les médias sociaux, nous avons utilisé des modules intégrés qui nous ont aidés à obtenir une solution standardisée dans notre implémentation, nous en mentionnerons certains dans ce qui suit :

- Pandas : Pandas est une bibliothèque Python de bas niveau construite sur NumPy.
- Numpy : L'arrivée de NumPy a été la clé pour étendre les capacités de Python avec des fonctions mathématiques, sur lesquelles des solutions d'apprentissage automatique seraient construites, car Python n'a pas été développé à l'origine comme un outil de calcul numérique[29].
- Nltk : c'est une plate-forme leader pour la création de programmes Python pour travailler avec des données de langage humain [30].
- Scikit-learn : Scikit-learn était initialement conçu comme une extension tierce à la bibliothèque SciPy(SciPy est construit au-dessus de NumPy et peut opérer sur ses tableaux). Aujourd'hui, c'est une bibliothèque autonome et l'une des plus populaires sur GitHub [29]
- Wordcloud : Une bibliothèque qui permet la visualisation de données textuelles (nuage de mots, également appelé nuage de tags ou liste pondérée) Les mots sont généralement des mots simples et l'importance de chacun est indiquée par la taille ou la couleur de la police [38].
- Matplotlib :Un module de NumPy, SciPy et Matplotlib est destiné à remplacer le besoin d'utiliser le langage statistique propriétaire MATLAB. Ce fait explique pourquoi les fonctions des bibliothèques mentionnées sont similaires à celles de MATLAB [29].
- Gensim : c'est une bibliothèque Python open-source gratuite pour représenter des documents sous forme de vecteurs sémantiques, aussi efficacement (informatiquement) [41].

4.2 Environnement matériel

Dans notre cas, un ordinateur qui fonctionne bien et sous n'importe quel système d'exploitation (Windows, Linux, Mac Os), est suffisant pour mettre en place notre système. Nous avons utilisé donc deux ordinateurs portable HP et DELL doté d'un processeur duale core et I5 respectivement, et ayant 4gb de ram.

5. Architecture du système

Les différentes étapes de notre système peuvent être résumées par le schéma suivant :

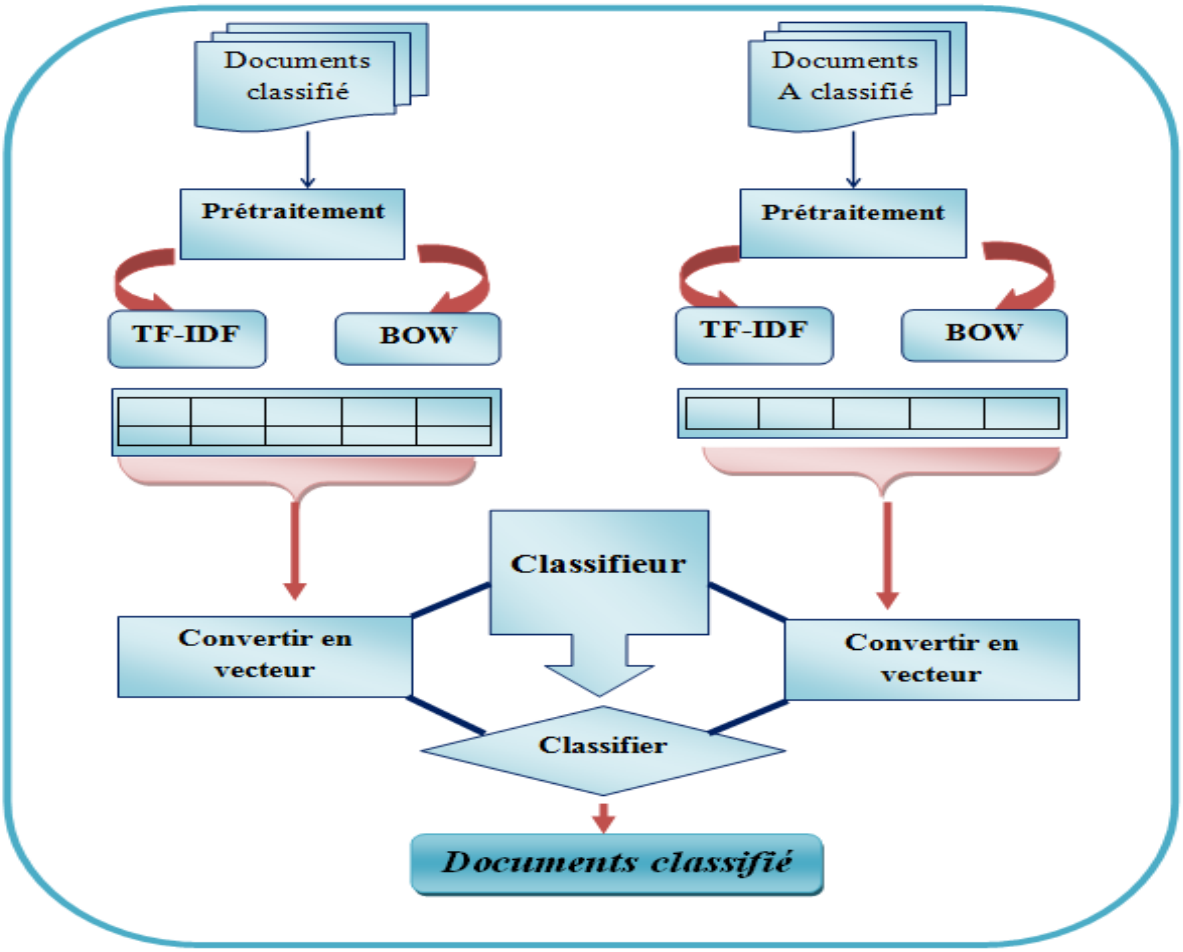

Figure III.2 : Architecture de notre système de classification

5.1 Préparation de données

5.1.1 Collection de données

Nous utiliserons un ensemble de données publiques de la BBC-News. Cet ensemble de données est composé de 2225 articles, chacun étiqueté dans l'une des 5 catégories suivantes : business, entertainment, politics, sport et tech. Cet ensemble a été téléchargé au format CSV depuis la plateforme Kaggle [39] le tableau III.1 présente le nombre de documents utilisés pour chaque catégorie de la collection.

TableauIII.1 : Distribution des documents d'articles dans la collection de données

Catégorie ID	Catégorie	Nombre d'articles
0	Technologies	401
1	Business	510
2	Sports	511
3	Entertainments	386
4	politics	417
Total	05	2225

```
--------------------------------------------------------------------------------
-------------------------1_/ importing Dataset---------------------------------
        category                                                text
0           tech  tv future in the hands of viewers with home th...
1       business  worldcom boss  left books alone  former worldc...
2          sport  tigers wary of farrell  gamble  leicester say ...
3          sport  yeading face newcastle in fa cup premiership s...
4  entertainment  ocean s twelve raids box office ocean s twelve...
```

Figure III.3 : Exemple d'importation de paquet de données

5.1.2 Prétraitement

Les prétraitements appliqués sur un document texte permettent, d'une part, d'éliminer ou de réduire le bruit dans le texte, et d'autre part, de simplifier les traitements ultérieurs. Noter que cette étape ne sert qu'à préparer l'ensemble de documents, afin qu'il soit possible d'extraire les descripteurs et ainsi calculer leur poids. Nos opérations de prétraitements comprennent : la Conversation en minuscules, le Filtrage des nombres, la suppression des signes de ponctuation, la suppression des mots vides,La racinassions (Stemming) et la lemmatisation.

- **La Conversation en minuscules** : utilisé pour mettre toutes les lettres majuscule en minuscules.

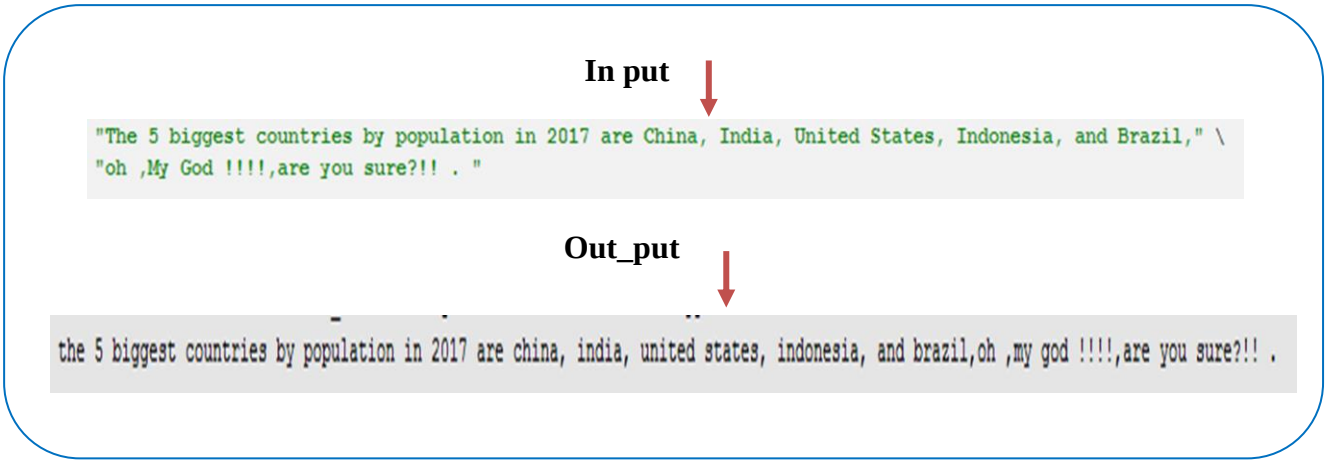

Figure III.4 : Exemple de Convertir tout en minuscules.

- Supprimer tous nombres en texte

Figure III.5 : Exemple de Supprimer nombres en texte.

- **La suppression des signes de ponctuation:** sert à supprimer toutes les balises ponctuations et caractères spéciaux.

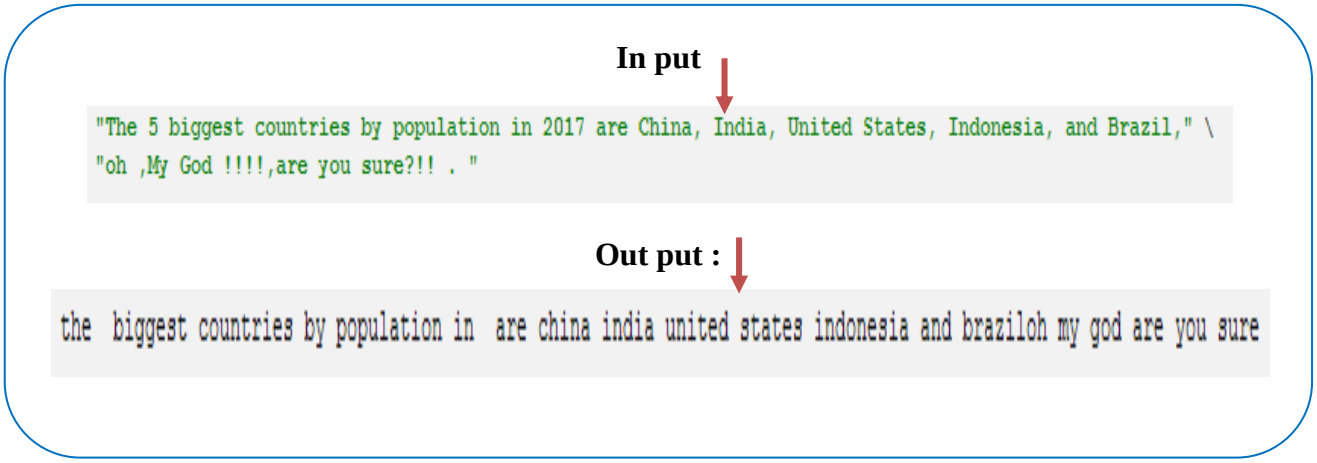

Figure III.6 : Exemple de la suppression des signes de ponctuation.

- **La suppression des mots vides :** utilisé pour supprimer tous les mots vides que nous avons définis dans le deuxième chapitre.

In put

```
"The 5 biggest countries by population in 2017 are China, India, United States, Indonesia, and Brazil," \
"oh ,My God !!!!,are you sure?!! . "
```

Out put

```
['biggest', 'countries', 'population', 'china', 'india', 'united', 'states', 'indonesia', 'braziloh', 'god', 'sure']
```

Figure III.7 : Exemple de la suppression des mots vides.

- **La racinisation et la lemmatisation**:sert à remplacerles mots par leurs racines ou à les ramènerà leurs formes de base (comme expliqué dans le chapitre 2).

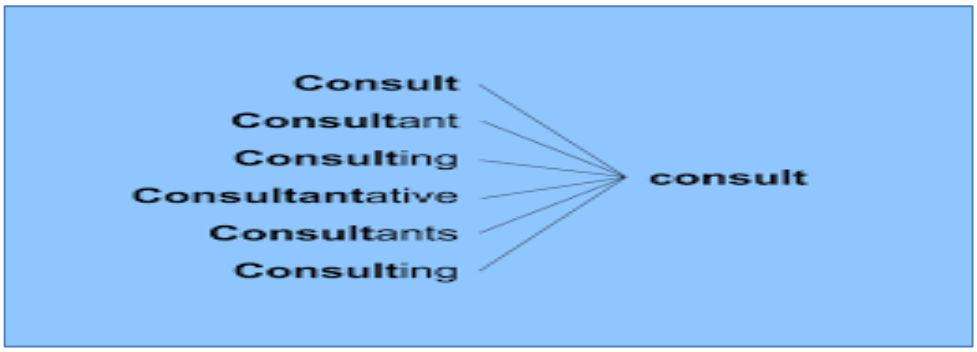

Figure III.8 : Exemple de racinisation [42]

In put

```
print(lemmatizer.lemmatize("I am loving it"))
print(lemmatizer.lemmatize("loving"))
print(lemmatizer.lemmatize("loving", "v"))
print(lemmatize_sentence("I am loving it"))
```

Out put

```
I be love it
```

Figure III.9 : Exemple de lemmatisation.

5.2 Partitionnement des données

C'est la dernière étape avant d'apprendre les algorithmes. Elle consiste à diviser l'ensemble de données en deux ensembles, l'un pour l'entrainement et l'autre pour les tests. Pour cela, nous avons préféré de partitionner notre ensemble en 84% pour l'apprentissage et 16% pour le test.

5.3 Représentation de données

La troisième étape de la préparation de documents consiste à représenter ces documents sous forme d'entrées valides et lisible par machine, ou plus précisément, par lesdifférents algorithmes de classification.Cette représentation est également appelée calcul des pondérations ou de poids. Dans notre travail, nous allons utiliserons la méthode de sac à mots et la méthode de fréquence que nous avons expliqué au deuxième chapitre.

- **La méthode de sac à mots :** permet de grouper tous les termes (descripteurs) qui apparaissent au moins une fois dans tout le corpus.

Tableau III.2 : Exemple de la méthode de sac à mots

Texts									
Doc 1	Mohammed	Lives	In	Algeria					
Doc 2	Ahmed	Lives	In	Egypt					
Doc 3	They	Meet	Once	a	year	In	London		
After cleaning text									
c 1	Mohammed	Lives	Algeria						
Doc 2	Ahmed	Lives	Egypt						
Doc 3	Meet	Once	year	London					
Vectorizing									
	Mohammed	**Lives**	**Algeria**	**Ahmed**	**Egypt**	**Once**	**year**	**London**	**Meet**
Doc 1	1	1	1	0	0	0	0	0	0
Doc 2	0	1	0	1	1	0	0	0	0
Doc 3	0	0	0	0	0	1	1	1	1

- **La méthode de fréquence :** permet d'évaluer l'importance d'un terme (descripteurs) contenu dans un document relativement à une collection de documents. Nous avons utilisé la fréquence relative dans notre travail.

Tableau III.3: Exemple de la méthode de TF_IDF

Texts									
Doc 1	Mohammed	Lives	In	Algeria					
Doc 2	Ahmed	Lives	In	Egypt					
Doc 3	They	Meet	Once	a	year	In	London		
After cleaning text									
Doc 1	Mohammed	Lives	Algeria						
Doc 2	Ahmed	Lives	Egypt						

Doc 3	Meet	Once	year	London					
Vectorizing									
	Mohammed	Lives	Algeria	Ahmed	Egypt	Once	year	London	Meet
Doc 1	1	1	1	0	0	0	0	0	0
Doc 2	0	1	0	1	1	0	0	0	0
Doc 3	0	0	0	0	0	1	1	1	1
TF									
	Mohammed	Lives	Algeria	Ahmed	Egypt	Once	year	London	Meet
Doc 1	0,12	0,12	0,12	0	0	0	0	0	
IDF									
	Mohammed	Lives	Algeria	Ahmed	Egypt	Once	year	London	Meet
Doc 1	0 ,47	0,17	0,47	0	0	0	0	0	0
TF_ IDF									
	Mohammed	Lives	Algeria	Ahmed	Egypt	Once	year	London	Meet
Doc 1	0,05	0,02	0,05	/	/	/	/	/	/

6. Résultats obtenus et discussion

6.1 Modèles d'apprentissage automatique

Une fois que nous avons terminé le prétraitement, calculé les poids (avec les deux méthodes BoW et TFIDF) et divisé le document vectoriel, nous passons maintenant à l'entraînement des classifieurs avec l'ensemble d'apprentissage, puis nous prédisons la catégorie de chaque document à partir de l'ensemble de test. Pour ce faire, nous allons utiliser cinq classifieurslargement utilisé dans la littérature à savoir , l'arbre de décision (Random forest), Machine à Vecteur de Support (SVM) , et l'arbre de décision, le classifieur bayésien naïf (NBC), Régression logistique (Logistic regression) et le K-voisins les plus proches (KNN).

6.2 Résultats avec initialisation par défaut des paramètres

Les algorithmes d'apprentissage automatique proposés nécessitent une initialisation de certains paramètres donnés par l'utilisateur (Tableau III.5). Dans ce contexte, il sera utile de développer une conception expérimentale ou de proposer des techniques afin de découvrir la meilleure combinaison de paramètres et de les fixer automatiquement. Pour cela, nous allons essayer d'obtenir les meilleurs paramètres pour chacun d'algorithme en suggérant une méthode de sélection (appelé hyper paramètres), puis comparer les résultats de chaque algorithme configuré par défaut avec les résultats obtenus par notre suggestion.

Tableau III.5 : Paramètres d'initialisation des Classifieurs.

Classifieurs	**Paramètres**
	N_estimators == représenté le nombre des arbres de forêt

Random forest	Max_depth ==est la profondeur de l'arbre Min_sample_ split == représente les nœuds classe supérieur Min_sample_leaf == sont les feuilles
Logistic regression	C== variable de contrôle Penalty == est la somme es valeurs absolues des éléments du vecteur de poids
KNN	n_neighbors == nombre des voisin dans le processus de vote p_values ==ou valeur de probabilité est, pour un modèle statistique donné
Naive Bayes	Alpha == probabilité a priori de différences classes
SVM	***Gamma*** == Gamma est un hyper paramètre que nous devons définir avant d'entraîner le modèle. Gamma décide de la courbure que nous voulons dans une limite de décision ***C*** == Cout de la violation séparabilité linéaire*(*il indique à l'optimisation SVM dans quelle mesure on souhaite éviter de mal classer chaque exemple d'entrainement).

Tableau III.6 : Initialisation des paramètres**.**

Classifieurs	Paramètres initiaux
Random forest	N_estimators =300 Max_depth = 2 Min_sample_ split =2 Min_sample_leaf =1
Logistic regression	C = 0.001 , penalty= 'l2'
KNN	n_neighbors = 3, p_values = 1
Naive Bayes	Alpha = 0.001
SVM	Gamma= 'auto_deprecated' , C = 1.0

Tableau III.7 : Résultats de Random Forest avec initialisation par défaut des paramètres

Category	Precision		Recall		F1-score		support	
	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW
Technologies	1.00	1.00	0.81	0.64	0.90	0.78	58	58
Business	0.73	0.66	0.96	0.98	0.83	0.79	83	83
Sports	0.66	0.62	0.99	0.99	0.79	0.76	80	80
Entertainments	1.00	1.00	0.38	0.30	0.55	0.47	69	69
politics	0.96	0.98	0.79	0.71	0.87	0.82	66	66
perforance	0.80	0.74	0.80	0.74	0.80	0.74	356	356

Tableau III.8 : Résultats de Logistic regression avec initialisation par défaut des paramètres

Category	Precision		Recall		F1-score		support	
	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW
Technologies	0.00	0.98	0.00	0.95	0.00	0.96	58	58
Business	0.48	0.96	0.98	0.95	0.64	0.96	83	83
Sports	0.43	0.96	1.00	0.99	0.60	0.98	80	80
Entertainments	0.00	0.99	0.00	0.96	0.00	0.97	69	69
politics	0.00	0.94	0.00	0.98	0.00	0.96	66	66
perforance	0.45	0.97	0.45	0.97	0.45	0.97	356	356

Tableau III.9 : Résultats de KNN avec initialisation par défaut des paramètres

Category	Precision		Recall		F1-score		support	
	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW
Technologies	0.00	0.96	0.00	0.45	0.00	0.61	58	58
Business	0.48	0.87	0.98	0.65	0.64	0.74	83	83
Sports	0.43	0.41	1.00	1.00	0.60	0.58	80	80
Entertainments	0.00	0.93	0.00	0.39	0.00	0.55	69	69
politics	0.00	0.95	0.00	0.64	0.00	0.76	66	66
perforance	0.45	0.64	0.45	0.64	0.45	0.64	356	356

Tableau III.10 : Résultats de NB avec initialisation par défaut des paramètres

Category	Precision		Recall		F1-score		support	
	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW
Technologies	0.00	0.96	0.00	0.45	0.00	0.61	58	58
Business	0.48	0.87	0.98	0.65	0.64	0.74	83	83
Sports	0.43	0.41	1.00	1.00	0.60	0.58	80	80
Entertainments	0.00	0.93	0.00	0.39	0.00	0.55	69	69
politics	0.00	0.95	0.00	0.64	0.00	0.76	66	66
perforance	0.45	0.64	0.45	0.64	0.45	0.64	356	356

Tableau III.11 : Résultats de SVM avec initialisation par défaut des paramètres

Category	Precision		Recall		F1-score		support	
	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW
Technologies	0.00	1.00	0.00	0.53	0.00	0.70	58	58
Business	0.48	0.63	0.98	0.94	0.64	0.75	83	83
Sports	0.43	0.54	1.00	1.00	0.60	0.70	80	80
Entertainments	0.00	1.00	0.00	0.14	0.00	0.25	69	69
politics	0.00	0.98	0.00	0.65	0.00	0.78	66	66
perforance	0.45	0.68	0.45	0.68	0.45	0.68	356	356

6.3 Résultats avec initialisation automatique des paramètres

La méthode de sélection de paramètres best basé sur la technique *GridSearch*. Elle prend dans un premier temps une liste de valeurs comme intervalle d'hyper paramètres. Ensuite, le modèle est entrainer pour chaque ensemble d'hyper paramètres, et le résultat est calculé à chaque fois. Enfin, le groupe de paramètres qui atteignent le score le plus élevé celui qui choisit pour chacun des algorithmes.

- ***Random forest***

Le choix du nombre d'arbres à créer et du nombre de variables à utiliser pour chaque section d'un nœud est ce qui explique le réglage et l'optimisation des hyper paramètres de cet algorithme.

Tableau III.12 : Hyper paramètres de Random forest

Classifieurs	Hyper paramètres	
Random forest	**TF_IDF**	**BOW**
	N_estimators = [100,300, 500, 800, 1200] Max_depth = 30 Min_sample_ split =15 Min_sample_leaf =1	
Meilleur paramètres	N_estimators = 300 Max_depth = 30 Min_sample_ split =15 Min_sample_leaf =1	N_estimators = 100 Max_depth = 2 Min_sample_ split =2 Min_sample_leaf =1

Tableau III.13 : Meilleur score de Random Forest

Category	Precision		Recall		F1-score		support	
	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW
Technologies	0.98	0.98	0.97	0.97	0.97	0.97	58	58
Business	0.98	0.96	0.96	0.96	0.96	0.96	83	83
Sports	0.93	0.95	0.96	0.99	0.96	0.97	80	80
Entertainments	0.98	1.00	0.96	0.96	0.96	0.98	69	69
politics	0.95	0.96	0.65	0.97	0.95	0.96	66	66
perforance	0.96	0.97	0.96	0.97	0.96	0.97	356	356

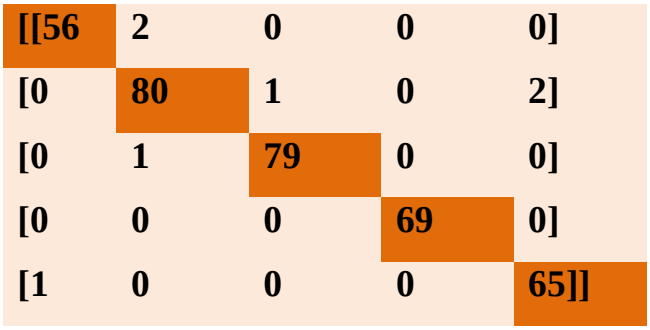

Figure III.10 : Matrice de confusion de classifieur Random forest

Selon le tableau III.13, nous remarquons que la catégorie ‘Technologies’ obtenu la même précision avec la méthode TF-IDF et BoW 0.98 (98%), recall et F1-score ont voir le même pourcentage 0.97 (97%).

Pour la catégorie ‘Business’ obtenu une précision 0.98 (98%) avec TF-IDF et 0.96 (96%) avec BoW. Pour le Recall et F1-score, il obtenu le même pourcentage 0.96 (96%).

la catégorie ‘Sports’ marquait 0.93 (93%) de précision avec la méthode TF-IDF et 0.95 (95%) avec BoW,le Recall était 0.96 (96%) avec TF-IDF et 0.99 (99%) avec BoW. En F1-score, elle marquait 0.96 (96%) avec TF-IDF et 0.98 (98%) avec BoW.

la catégorie ‘Entertainments’ marquait 0.98 (98%) de précision avec la méthode TF-IDF et 1.00 (100%) avec BoW, le Recall était 0.96 (96%) avec TF-IDF aussi avec BoW. En F1-score, elle marquait 0.96 (96%) avec TF-IDF et 0.98 (98%) avec BoW.

la catégorie ‘politics’ achevait 0.95 (95%) de précision avec la méthode TF-IDF et 0.96 (96%) avec BoW, le Recall était 0.65 (65%) avec TF-IDF et 0.97 (97%) avec BoW. En F1-score, elle marquait 0.95 (95%) avec TF-IDF et 0.96 (96%) avec BoW.

En ce qui concerne le support, chaque catégorie a obtenu le même valeur dans les deux méthodes (TF-IDF et BoW) comme suit : Technologies' 58, 'Business'83, 'Sports' 80, 'Entertainments' 69 et 'politics' 66.

Dans la matrice de confusion de l'algorithme *Random forest,* chaque ligne et colonne correspond à une classe, alors que la valeur 56 dans la première ligne indiques que le classifieur est capable de classifier 56 texte bélongue à la catégorie 'Technologie', et la valeur 1 dans la même ligne, indiques que le classifieur le classait dans une autre classe. En simulation sur la deuxième ligne, la valeur 80indiques que le classifieur est capable de classifier 80 texte bélongue à la catégorie 'Business', et que les valeur 1 et 2dans la même ligne sont classées dans une autre classe. dans la troisième ligne, la valeur 79indiques que le classifieur est capable de classifier 79 texte bélongue à la catégorie 'Sport', et la valeur 1 dans la même ligne, indiques que le classifieur le classait dans une autre classe, dans la quatrième ligne, la valeur 66indiques que le classifieur est capable de classifier 66 texte bélongue à la catégorie 'Entertainment', et les valeur 1et 2dans la même ligne, indiquent que le classifieur les classait dans une autre classe, dans la dernière ligne, la valeur 64 indiques que le classifieur est capable de classifier 64 texte bélongue à la catégorie 'Politique', et la valeur 1 dans la même ligne, indiques que le classifieur le classait dans des autre classes qui sont 'Technologie' et 'Business'.

- ***Logistic regression***

La régression logistique ne fournit pas vraiment de coefficient supercritique. Cependant, des pénalités peuvent s'appliquer pour ajuster et augmenter la précision en plus, la pénalité associée à « C » qui va le contrôle

Tableau III.14 : Hyper paramètres de Logistic regression

Classifieurs	**Hyper paramètres**	
Logisticregression	**TF_IDF**	**BOW**
	C =[0.1 , 0.001, 1] penalty= [l2 , l1]	
Meilleur paramètres	C = 1 , penalty= l2	C = 1 , penalty= l2

Tableau III.15 : Meilleur score de Logistic regression

Category	Precision		Recall		F1-score		support	
	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW
Technologies	0.98	0.98	0.97	0.97	0.97	0.97	58	58
Business	0.96	0.96	0.96	0.96	0.96	0.96	83	83
Sports	0.99	0.95	0.99	0.99	0.99	0.97	80	80
Entertainments	1.00	1.00	1.00	0.96	1.00	0.98	69	69
politics	0.97	0.96	0.98	0.97	0.98	0.96	66	66
perforance	0.98	0.97	0.98	0.97	0.98	0.97	356	356

```
[[56  1  1  0  0]
 [ 0 80  1  0  2]
 [ 0  1 79  0  0]
 [ 0  0  2 66  1]
 [ 1  1  0  0 64]]
```

Figure III.11: Matrice de confusion de classifieur Logistic regression

Selon le tableau III.15, nous remarquons que la catégorie ‘Technologies’ obtenu la même précision avec la méthode TF-IDF et BoW 0.98 (98%), recall et F1-score ont voir le même pourcentage 0.97 (97%).

Pour la catégorie ‘Business’ obtenu 0.96 (96%) de une précision avec TF-IDF et aussi avec BoW. le même pourcentage pour le Recall et F1-score 0.96 (96%).

la catégorie ‘Sports’ marquait 0.99 (99%) de précision avec la méthode TF-IDF et 0.95 (95%) avec BoW,le Recall était 0.99 (99%) avec TF-IDF et BoW. En F1-score, elle marquait 0.99 (99%) avec TF-IDF et 0.97 (97%) avec BoW.

la catégorie ‘Entertainments’ marquait 1.00 (100%) de précision avec la méthode TF-IDF aussi avec BoW, le Recall était 1.00 (100%) avec TF-IDF et 0.96 (96%) avec BoW. En F1-score, elle marquait 1.00 (100%) avec TF-IDF et 0.98 (98%) avec BoW.

la catégorie ‘politics’ achevait 0.97 (97%) de précision avec la méthode TF-IDF et 0.96 (96%) avec BoW, le Recall était 0.98 (98%) avec TF-IDF et 0.97 (97%) avec BoW. En F1-score, elle marquait 0.98 (98%) avec TF-IDF et 0.96 (96%) avec BoW.

Dans la matrice de confusion de l'algorithme *Logistic regression*, chaque ligne et colonne correspond à une classe, alors que la valeur 56 dans la première ligne indiques que le classifieur est capable de classifier 56 texte bélongue à la catégorie 'Technologie', et la valeur 2 dans la même ligne, indiques que le classifieur le classait dans une autre classe. En simulation sur la deuxième ligne, la valeur 80 indiques que le classifieur est capable de classifier 80 texte bélongue à la catégorie 'Business', et que les valeur 1 et 2 dans la même ligne sont classées dans une autre classe. dans la troisième ligne, la valeur 79 indiques que le classifieur est capable de classifier 79 texte bélongue à la catégorie 'Sport', et la valeur 1 dans la même ligne, indiques que le classifieur le classait dans une autre classe, dans la quatrième ligne, la valeur 69 indiques que le classifieur est capable de classifier 69texte bélongue à la catégorie 'Entertainment', et les valeur 0 dans la même ligne, indique que le classifieur''uniquement dans cette ligne''estsuccède de classement en 0% erreur. dans la dernière ligne, la valeur 65 indiques que le classifieur est capable de classifier 65 texte bélongue à la catégorie 'Politique', et la valeur 1 dans la même ligne, indiques que le classifieur le classait dans une autre classe qui est 'Technologie'.

- ***KNN***

L'hyper paramètre le plus important pour KNN est le nombre de voisins (n_neighbors).

Tableau III.16 : Hyper paramètres KNN

Classifieurs	**Hyper paramètres**	
KNN	**TF_IDF**	**BOW**
	n_neighbors=[1 :21] p-value == [1,2,5]	
Meilleur paramètres	n_neighbors = 19 p-value == 2	n_neighbors = 3 p-value == 2

Tableau III.17 : Meilleur score de KNN

Category	**Precision**		**Recall**		**F1-score**		**support**	
	TF-IDF	**BoW**	**TF-IDF**	**BoW**	**TF-IDF**	**BoW**	**TF-IDF**	**BoW**
Technologies	0.97	0.98	0.98	0.69	0.97	0.81	58	58
Business	0.99	0.89	0.94	0.70	0.96	0.78	83	83
Sports	0.96	0.69	1.00	0.90	0.98	0.78	80	80
Entertainments	1.00	0.95	0.99	0.61	0.99	0.74	69	69
politics	0.97	0.52	0.98	0.80	0.98	0.63	66	66
perforance	0.98	0.74	0.98	0.74	0.98	0.74	356	356

```
[[57  1  0  0  0]
 [0  80  1  0  2]
 [0   1 79  0  0]
 [0   0  2 68  0]
 [1   0  0  0 65]]
```

Figure III.11: Matrice de confusion de classifieur KNN

Selon le tableau III.17, nous remarquons que la catégorie 'Technologies' obtenu 0.97 (97%) de précision avec la méthode TF-IDF et 0.98 (98%) avec BoW , le recall était 0.98 (98%) avec la méthode TF-IDF et 0.69 (69%) avec BoW, et F1-score étais 0.97 (97%) avec la méthode TF-IDF et 0.81 (81%) avec BoW .

Pour la catégorie 'Business' obtenu 0.99 (99%) de une précision avec TF-IDF et 0.98 (98%) avec BoW. pour le Recall 0.94 (94%) avec TF-IDF et 0.70 (70%) avec BoW et F1-score 0.96 (96%) avec TF-IDF et 0.78 (78%) avec BoW.

La catégorie 'Sports' marquait 0.96 (96%) de précision avec la méthode TF-IDF et 0.69 (69%) avec BoW, le Recall était 1.00 (100%) avec TF-IDF et 0.90 (90%) BoW. En F1-score, elle marquait 0.98 (98%) avec TF-IDF et 0.78 (78%) avec BoW.

La catégorie 'Entertainments' marquait 1.00 (100%) de précision avec la méthode TF-IDF et 0.95 (95%) avec BoW, le Recall était 0.99 (99%) avec TF-IDF et 0.61 (61%) avec BoW. En F1-score, elle marquait 0.99 (99%) avec TF-IDF et 0.74 (74%) avec BoW.

La catégorie 'politics' achevait 0.97 (97%) de précision avec la méthode TF-IDF et 0.52 (52%) avec BoW, le Recall était 0.98 (98%) avec TF-IDF et 0.80 (80%) avec BoW. En F1-score, elle marquait 0.98 (98%) avec TF-IDF et 0.63 (63%) avec BoW.

Dans la matrice de confusion de l'algorithme *KNN*, chaque ligne et colonne correspond à une classe, alors que la valeur 57 dans la première ligne indiques que le classifieur est capable de classifier 57 texte bélongue à la catégorie 'Technologie', et la valeur 1 dans la même ligne, indiques que le classifieur le classait dans une autre classe. En simulation sur la deuxième ligne, la valeur 80 indiques que le classifieur est capable de classifier 80 texte bélongue à la catégorie 'Business', et que les valeur 1 et 2 dans la même ligne sont classées dans une autre classe. dans la troisième ligne, la valeur 79 indiques que le classifieur est capable de classifier 79 texte bélongue à la catégorie 'Sport', et la valeur 1 dans la même ligne, indiques que le classifieur le classait dans une autre classe dans la quatrième

ligne, la valeur 68indiques que le classifieur est capable de classifier 68texte bélongue à la catégorie 'Entertainment', et les valeur 1 dans la même ligne,indique que le classifieur le classait dans une autre classe, dans la dernière ligne, la valeur 65indiques que le classifieur est capable de classifier 65 texte bélongue à la catégorie 'Politique', et la valeur 1 dans la même ligne, indiques que le classifieur le classait dans une autre classe qui est 'Technologie'.

❖ ***Naive Bayes***

Tableau III.18 : Hyper paramètres Naive Bayes

Classifieurs	**Hyper paramètres**	
Naive Bayes	**TF_IDF**	**BOW**
	Alpha = [0.001, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 1]	
Meilleur paramètres	Alpha =0.01	Alpha =0.09

Tableau III.19 : Meilleur score de NB

Category	**Precision**		**Recall**		**F1-score**		**support**	
	TF-IDF	**BoW**	**TF-IDF**	**BoW**	**TF-IDF**	**BoW**	**TF-IDF**	**BoW**
Technologies	0.98	0.98	0.95	1.00	0.96	0.99	58	58
Business	0.96	0.99	0.98	0.96	0.97	0.98	83	83
Sports	0.99	1.00	0.99	0.99	0.99	0.99	80	80
Entertainments	1.00	0.99	1.00	1.00	1.00	0.99	69	69
politics	0.97	0.97	0.98	0.98	0.98	0.98	66	66
perforance	0.98	0.99	0.98	0.99	0.98	0.99	356	356

```
[[58  0  0  0  0]
 [ 0 80  0  1  2]
 [ 0  1 79  0  0]
 [ 0  0  0 69  0]
 [ 1  0  0  0 65]]
```

Figure III.12: Matrice de confusion de classifieur NB

Selon le tableau III.19, nous remarquons que la catégorie 'Technologies' obtenu 0.98 (98%) de précision avec la méthode TF-IDF aussi avec BoW, le recall était 0.95 (95%) avec la méthode TF-IDF et 1.00 (100%) avec BoW, et F1-score étais 0.96 (96%) et 0.99 (99%) avec BoW .

Pour la catégorie ‘Business’ obtenu 0.96 (96%) de une précision avec TF-IDF et 0.99 (99%) avec BoW. pour le Recall 0.98 (98%) avec TF-IDF et 0.96 (96%) avec BoW et F1-score 0.97 (97%) avec TF-IDF et 0.98 (98%) avec BoW.

la catégorie ‘Sports’ marquait 0.99 (99%) de précision avec la méthode TF-IDF et 1.00 (100%) avec BoW, le Recall était 0.99 (99%) avec TF-IDF et aussi avec BoW. En F1-score, elle marquait 0.99 (99%) avec TF-IDF et aussi avec BoW.

la catégorie ‘Entertainments’ marquait 1.00 (100%) de précision avec la méthode TF-IDF et 0.99 (99%) avec BoW, le Recall était 1.00 (100%) avec TF-IDF aussi avec BoW. En F1-score, elle marquait 1.00 (100%) avec TF-IDF et 0.98 (98%) avec BoW.

la catégorie ‘politics’ achevait 0.97 (97%) de précision avec la méthode TF-IDF et aussi avec BoW, le Recall était 0.98 (98%) avec TF-IDF aussi avec BoW. En F1-score, elle marquait 0.98 (98%) avec TF-IDF aussi avec BoW.

Dans la matrice de confusion de l’algorithme *NB,* chaque ligne et colonne correspond à une classe, alors que la valeur 58 dans la première ligne indiques que le classifieur est capable de classifier 58 texte bélongue à la catégorie ‘Technologie’, et les valeurs1 dans la même ligne, indiquent que le classifieur le classait dans d’autreclasse. En simulation sur la deuxième ligne, la valeur 80 indiques que le classifieur est capable de classifier 80 texte bélongue à la catégorie ‘Business’, et que le valeur 1 dans la même ligne est classées dans une autre classe. dans la troisième ligne, la valeur 79 indiques que le classifieur est capable de classifier 79 texte bélongue à la catégorie ‘Sport’, et la valeur 1 dans la même ligne, indiques que le classifieur le classait dans une autre classe. dans la quatrième ligne, la valeur 69 indiques que le classifieur est capable de classifier 69texte bélongue à la catégorie ‘Entertainment’, et les valeur 0 dans la même ligne, indique que le classifieur‘‘uniquement dans cette ligne’’ est succède de classement en 0% erreur. dans la dernière ligne, la valeur 65 indiques que le classifieur est capable de classifier 65 texte bélongue à la catégorie ‘Politique’, et la valeur 1 dans la même ligne, indiques que le classifieur le classait dans une autre classe qui est ‘Technologie’.

- ***SVM***

Pour les SVM on peut voir 2 ou 3 les plus importants (le C et le gamma, et le kernel peut être considéré comme un hyper paramètre dans certains cas)

Tableau III.20 : Hyper paramètres SVM

Classifieurs	Hyper paramètres	
SVM	**TF_IDF**	**BOW**
	Gamma =[0.001, 0.01, 0.1, 1] **C** =[0.1, 1, 10]	
Meilleur choix	**Gamma** = 10 **C** =1	**Gamma** = 10 **C** =1

Tableau III.21 : Meilleur score de SVM

Category	Precision		Recall		F1-score		support	
	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW
Technologies	0.98	0.98	0.98	1.00	0.98	0.99	58	58
Business	0.98	0.99	0.95	0.96	0.96	0.98	83	83
Sports	0.96	1.00	1.00	0.99	0.98	0.99	80	80
Entertainments	0.97	0.99	0.96	1.00	0.96	0.99	69	69
politics	0.95	0.97	0.95	0.98	0.95	0.98	66	66
perforance	0.97	0.99	0.97	0.99	0.97	0.99	356	356

```
[[57  0  0  1  0]
 [ 0 79  1  1  2]
 [ 0  0 80  0  0]
 [ 0  0  2 66  1]
 [ 1  2  0  0 63]]
```

Figure III.13: Matrice de confusion de classifieur SVM

Selon le tableau III.21, nous remarquons que la catégorie ‘Technologies’ obtenu 0.98 (98%) de précision avec la méthode TF-IDF aussi avec BoW, le recall était 0.98 (95%) avec la méthode TF-IDF et 1.00 (100%) avec BoW, et F1-score étais 0.98 (98%) et 0.99 (99%) avec BoW .

Pour la catégorie ‘Business’ obtenu 0.98 (98%) de une précision avec TF-IDF et 0.99 (99%) avec BoW. pour le Recall 0.95 (95%) avec TF-IDF et 0.96 (96%) avec BoW et F1-score 0.95 (95%) avec TF-IDF et 0.98 (98%) avec BoW.

la catégorie ‘Sports’ marquait 0.96 (96%) de précision avec la méthode TF-IDF et 1.00 (100%) avec BoW, le Recall était 1.00 (10%) avec TF-IDF et 0.99 (99%) BoW. En F1-score, elle marquait 0.98 (98%) avec TF-IDF et 0.99 (99%) avec BoW.

la catégorie ‘Entertainments’ marquait 0.97 (97%) de précision avec la méthode TF-IDF et 0.99 (99%) avec BoW, le Recall était 0.96 (96%) avec TF-IDF et 1.00 (100%) avec BoW. En F1-score, elle marquait 0.96 (96%) avec TF-IDF et 1.00 (100%) avec BoW.

la catégorie ‘politics’ achevait 0.95 (95%) de précision avec la méthode TF-IDF et 0.97 (97%) avec BoW, le Recall était 0.95 (95%) avec TF-IDF et 0.98 (98%) avec BoW. En F1-score, elle marquait 0.95(95%) avec TF-IDF et 0.98 (98%) avec BoW.

Dans la matrice de confusion de l’algorithme *SVM*, chaque ligne et colonne correspond à une classe, alors que la valeur 57 dans la première ligne indiques que le classifieur est capable de classifier 57 texte bélongue à la catégorie ‘Technologie’, et la valeur 1 dans la même ligne, indiques que le classifieur le classait dans des autre classes. En simulation sur la deuxième ligne, la valeur 79 indiques que le classifieur est capable de classifier 79 texte bélongue à la catégorie ‘Business’, et que les valeurs1 et 2 dans la même ligne sont classées dans une autre classe. dans la troisième ligne, la valeur 80 indiques que le classifieur est capable de classifier 80 texte bélongue à la catégorie ‘Sport’, et que les valeurs1 et 2 dans la même ligne sont classées dans une autre classe.dans la quatrième ligne, la valeur 66 indiques que le classifieur est capable de classifier 66 texte bélongue à la catégorie ‘Entertainment’, les valeurs1 et 2 dans la même ligne sont classées dans des autre classes, dans la dernière ligne, la valeur 63 indiques que le classifieur est capable de classifier 63 texte bélongue à la catégorie ‘Politique’, et la valeur 1et 2dans la même ligne, indiques que le classifieur les classait dans des autre classes qui est ‘Technologie’et ‘Business’.

Les résultats étaient différents d'un algorithme à l'autre, nous allons essayer de les expliquer à travers les points suivants évoqués :

6.4 Analyse et discussion des résultats :

Selon les méthodes d'extraction de caractéristiques : BOW et TF_IDF, nous avons obtenu deux types de résultats, et chaque type est divisé en deux sections, une avec des paramètres de support et une sans, tous extraits en appliquant les algorithmes de classification spécifiés dans ce travail. Nous nous sommes appuyés sur l'élément de ***l’Accuracy*** performance à titre de comparaison comme le tableau suivant spectacles:

Tableau III.22 : Performance des Classifieurs

	TF_IDF		BOW	
	par défaut	**Hyper paramètres**	**par défaut**	**Hyper paramètres**
Random forest	0,7977	0,9634	0,7443	0.9652
Logisticregression	0,4522	0,9803	0,9662	0.9743
KNN	0,45	0,9775	0,6432	0.7443
Naive Bayes	0,9803	0,9803	0,6432	0.9859
SVM	0,2247	0,9807	0,6797	0.9859

Premièrement, les résultats des cas par défaut étaient relativement faibles, que nous classons par ordre décroissant,NB avec une précision98.08%, Logistic regression avec une précision96.62%, Random forest avec une précision79.77%,SVM avec une précision 67.97%,et KNN avec une précision64.32%.

Après l'optimisation des paramètres, les résultats moyens montrent que le SVM avec une précision de 98,59 % est performant mieux que les quatre autres algorithmes, après NB et Logistic regression avec une précisionde 98,03 %. Vient ensuite KNN avec une précision 97,75 %etRandomf orestavec précision 96,52 % , comme illustré à leTableau III.22.

6.5 Évaluation

Tableau III.23 : Meilleur score des Classifieurs avec Hyper paramètres

	précision	**rappel**	**F1 score**
Random forest	0 .96	0 .96	0 .96
Logistic regression	0.98	0.98	0.98
KNN	0.98	0.98	0.98
Naive Bayes	0.98	0.98	0.98
SVM	0 .99	0 .99	0 .99

Selon le Tableau III.23, on remarque que les résultats montrent la précision, rappel et F1 score meilleurs scores étaient plus fortes proches avec l'algorithme SVM de 99 % de précision. Suivre par les Classifieurs NB, Logistic regression et KNN 98 %. Vient ensuite, Random forest 0.96%.

6.6 Comparaison

Après avoir discuté nos résultats, on va maintenant comparer les cinq algorithmes en fonction des quatre paramètres, à savoir : la précision, l'exactitude, le score F1 et le support.Ces quatre paramètres sont comparés sur un historgamme(bar-graphe) afin d'afficher une comparaison parfaite.Les quatre comparaisons sont présentées ci-dessous comme suit :

6.6.1 Précision

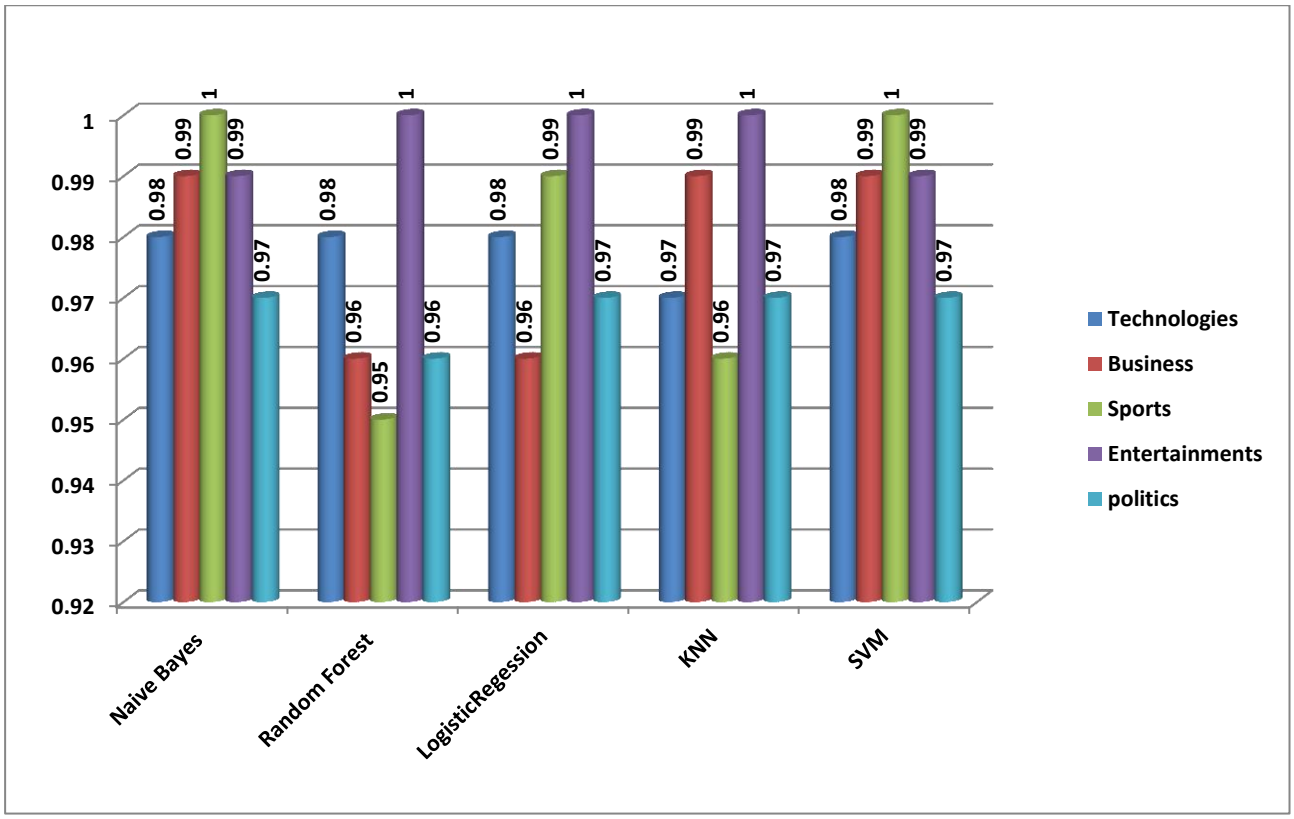

Figure III.14: Catégories versus précision : montrant les variations des classes d'ensembles de données par rapport au changement de précision.

Le graphique obtenu est illustré dans la figure **III.14.**

Dans la catégorie business, la régression logistique a obtenu une précision de 0,96 (moyenne 96%), la forêt aléatoire a une de précision 0,96 (moyenne 96%), le KNN a une précision de 0,99 (moyenne 99%), le NB a une précision de 0,99 (moyenne 99%) et le SVM a une précision de 0,99 (moyenne 99%). Dans la catégorie entertainment, la régression logistique a obtenu une précision de 1,00 (100%), la forêt aléatoire a une précision de 1,00 (100%) , le KNN a une précision de 1,00 (100%), le NB a une précision de 0,99 (99%) et le SVM a une précision de 0,99 (99%). Dans la catégorie politics, la régression logistique a obtenu une précision de 0,97 (97%), la forêt aléatoire a une

précision de 0,96 (96%) , le KNN a une précision de 0,97 (97%), le NB a une précision de 0,97 (97%) etle SVM a une précision de 0,97 (97%). Dans la catégorie sport, la régression logistique a obtenu une précision de 0,99 (99%), la forêt aléatoire a une précision de 0,95 (95%) , le KNN a une précision de 0,96 (96%), le NB a une précision de 1,00 (100%) etle SVM a une précision de 1,00 (100%). Enfin dans la catégorie tech, la régression logistique a obtenu une précision de 0,98 (98%), la forêt aléatoire a une précision de 0,98 (98%) , le KNN a une précision de 0,97 (97%), le NB a une précision de 0,98 (98%) etle SVM a une précision de 0,98 (98%).

b. Performance

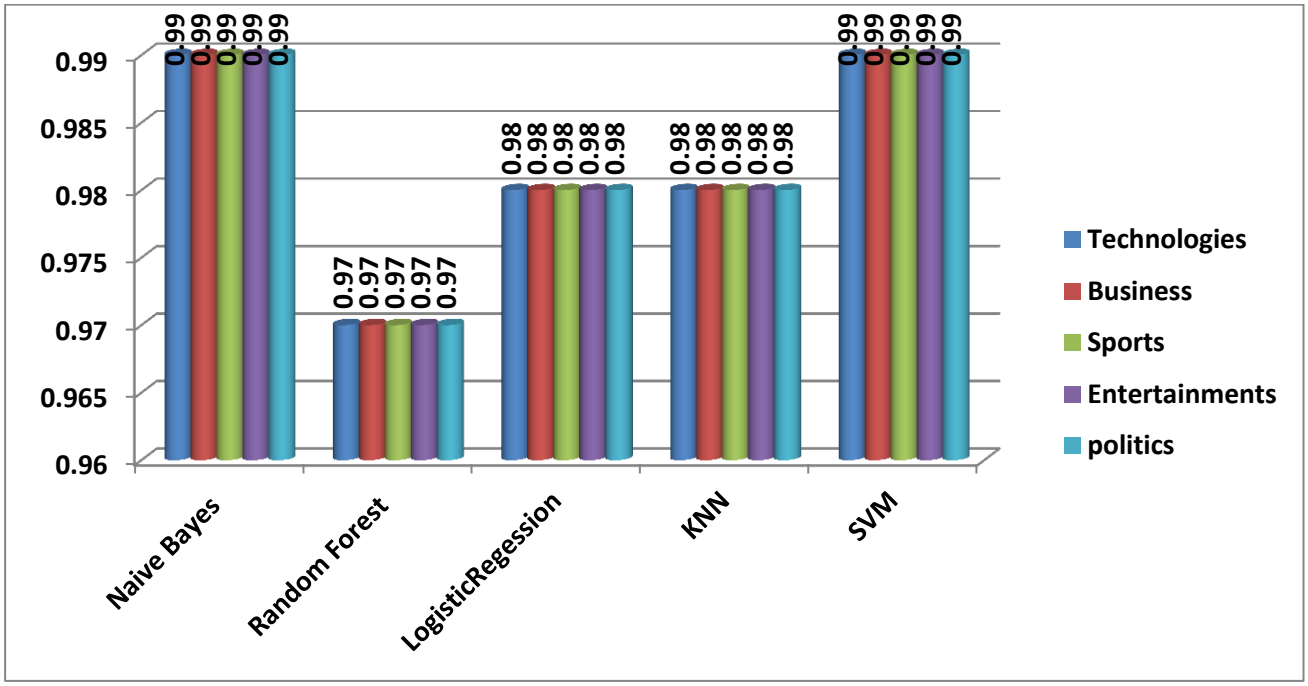

Figure III.15: Catégories versusperformance : montrant les variations des classes d'ensembles de données par rapport au changement de performance.

Le graphique de performance est présentédans la figure **III.15**

Dans la catégorie business, la régression logistique a obtenu une performance de 0,98 (98%), la forêt aléatoire a une performance de 0,97 (97%), le KNN a une performance de 0,98 (98%), le NB a une performance de 0,99 (99%) et le SVM a une performance de 0,99 (99%). Dans la catégoriee ntertainment, la régression logistique a obtenu une performance de 0,98 (98%), la forêt aléatoire a une performance de 0,97 (97%), le KNN a une performance de 0,98 (98%), le NB a une performance de 0,99 (99%) etle SVM a une performance de 0,99 (99%). Dans la catégorie politics, la régression logistique a obtenu une performance de 0,98 (98%), la forêt aléatoire a une performance de 0,97 (97%), le KNN a une performance de 0,98 (98%), le NB a une performance de 0,99 (99%) et le SVM a

une performance de 0,99 (99%). Dans la catégorie sport, la régression logistique a obtenu une performance de 0,98 (98%), la forêt aléatoire a une performance de 0,97 (97%), le KNN a une performance de 0,98 (98%), le NB a une performance de 0,99 (99%) etle SVM a une performance de 0,99 (99%). Enfin dans la catégorietech, la régression logistique a obtenu une performance ion de 0,98 (98%), la forêt aléatoire a une performance de 0,97 (97%), le KNN a une performance de 0,98 (98%), le NB a une performance de 0,99 (99%) etle SVM a une performance de 0,99 (99%).

c. F1-score

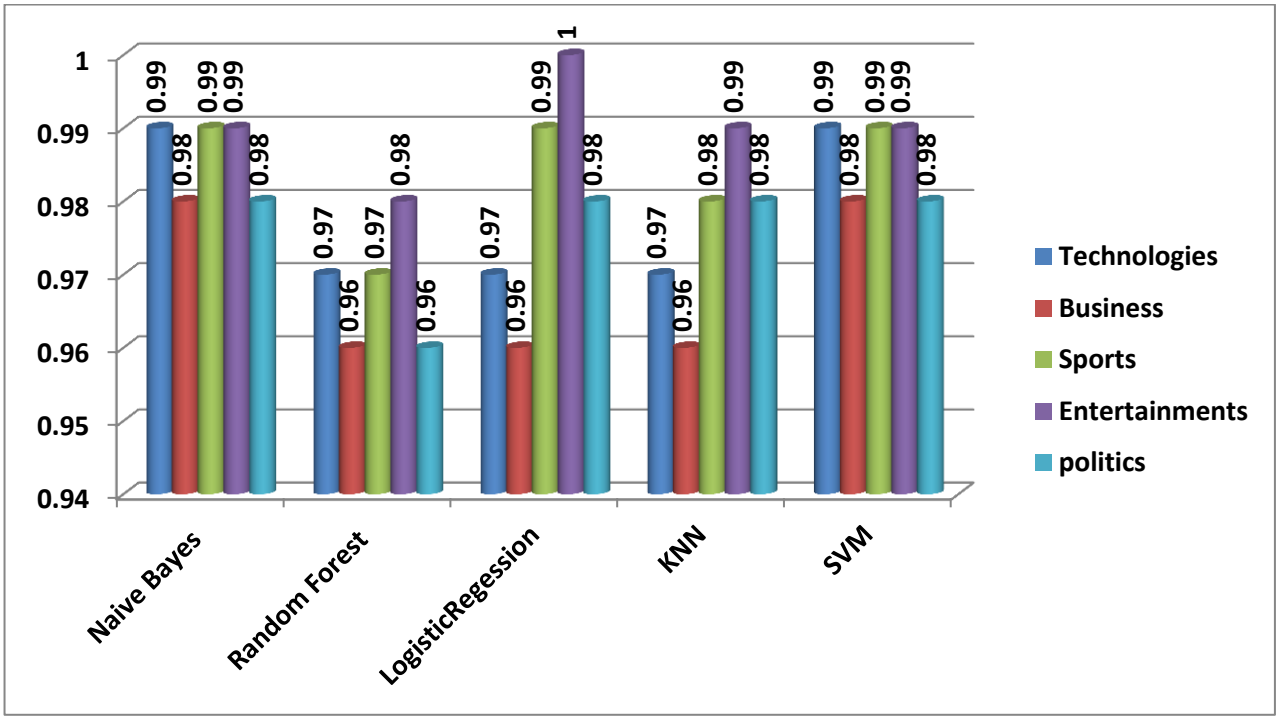

Figure III.16: Catégories versus*F*1-score : montrant les variations des classes d'ensembles de données par rapport au changement de *F*1-score .

Le graphique à barres pour le *F*1-score est montré dans la figure **III.16**.

Dans la catégorie business, la régression logistique a obtenu un *F*1-scorede 0,96 (96%), la forêt aléatoire a un *F*1-scorede 0,96 (96%), le KNN a un*F*1-scorede 0,96 (96%), le NB a un*F*1-scorede 0,98 (98%) etle SVM a un*F*1-scorede 0,98 (98%). Dans la catégorie entertainment, la régression logistique a obtenu un *F*1-scorede 1.00 (100%), la forêt aléatoire a un *F*1-scorede 0,98 (98%) , le KNN a un *F*1-scorede 0,99 (99%), le NB a un *F*1-scorede 0,99 (99%) et le SVM a un *F*1-scorede 0,99 (99%). Dans la catégorie politics, la régression logistique a obtenu un *F*1-scorede 0,98 (98%), la forêt aléatoire a un *F*1-scorede 0,96 (96%) , le KNN a un *F*1-scorede 0,98 (98%), le NB a un *F*1-scorede 0,98 (98%) et le SVM a un *F*1-scorede 0,98 (98%). Dans la catégorie sport, la régression logistique a obtenu un *F*1-scorede 0,99 (99%), la forêt aléatoire un *F*1-scorede 0,97 (97%) , le KNN a un *F*1-scorede 0,98 (98%),

le NB a un *F*1-scorede 0,99 (99%) et le SVM a un *F*1-scorede 0,99 (99%). Enfin dans la catégorie tech, la régression logistique a obtenu un *F*1-scoreion de 0,97 (97%), la forêt aléatoire a un *F*1-scorede 0,97 (97%) , le KNN a un *F*1-scorede 0,97 (97%), le NB a un *F*1-scorede 0,99 (99%) etle SVM a un *F*1-scorede 0,99 (99%).

d. Support

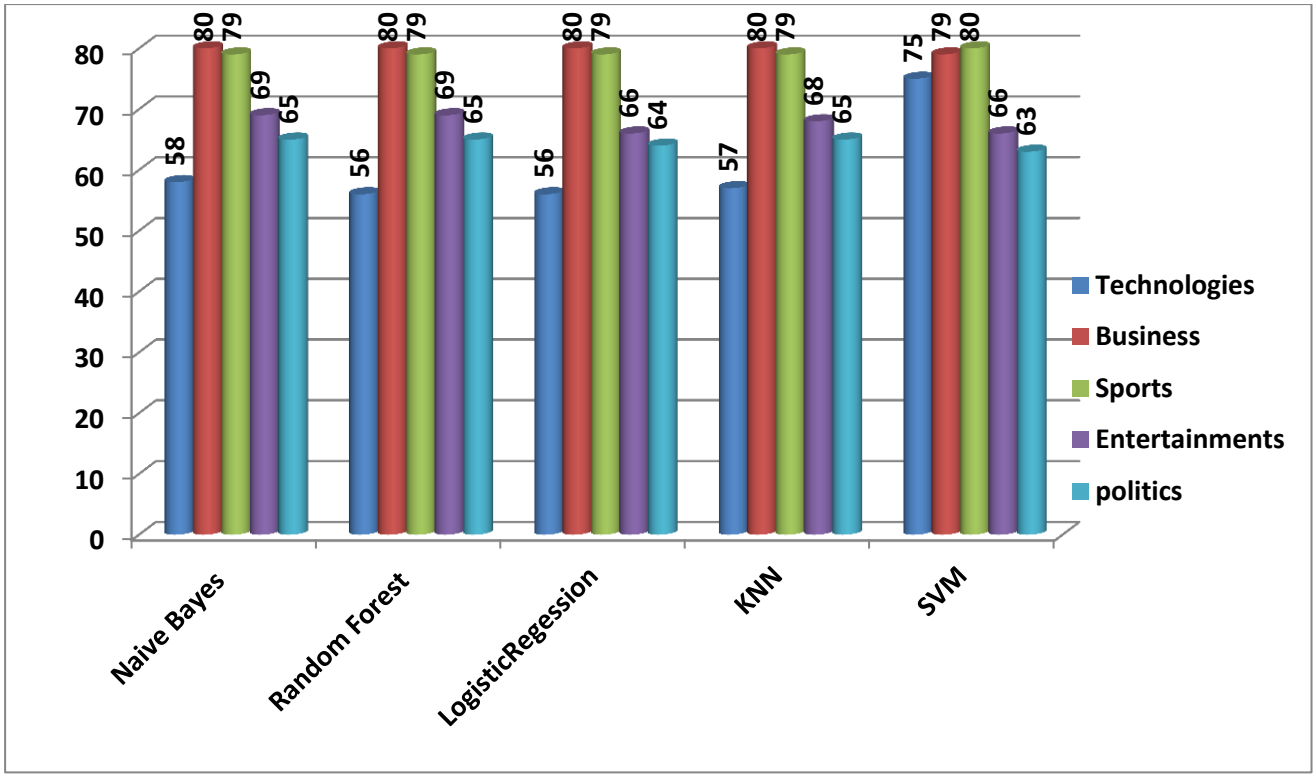

Figure III.17: Catégories versus support: montrant les variations des classes d'ensembles de données par rapport au changement de support.

Pour le paramètre support, l'histogramme obtenu est illustré à la figure **III.17**

Dans la catégorie business, la régression logistique a obtenu un support de 80, la forêt aléatoire a un supportde 80, le KNN a unsupportde 80, le NB a un supportde 80 etle SVM a unsupportde 79. Dans la catégorie entertainment, la régression logistique a obtenu un supportde 66, la forêt aléatoire a un supportde 69, le KNN a un supportde 68, le NB a un supportde 69 etle SVM a un supportde 66. Dans la catégorie politics, la régression logistique a obtenu un supportde 64, la forêt aléatoire a un supportde 65, le KNN a un supportde 65, le NB a un supportde 65 et le SVM a un support de 63. Dans la catégorie sport, la régression logistique a obtenu un support de 79, la forêt aléatoire un supportde 79, le KNN a un supportde 79, le NB a un supportde 79 et le SVM a un supportde 80. Enfin dans la catégorie tech, la régression logistique a obtenu support de 56, le KNN a un supportde 57, le NB a un supportde 58 et le SVM a un support de 75, la forêt aléatoire un supportde56.

7. Conclusion

Dans ce chapitre, nous avons présenté le côté technique de notre application. Nous avons identifié les outils utilisés lors de la mise en place de notre application, en plus de présenter les résultats de l'étude que nous avons appliquée et de les analyser selon l'ensemble d'algorithmes mentionnés ci-dessus.

Conclusion générale et perspectives

Conclusion

Dans ce travail, nous avons essayé de construire un modèle de classification des articles d'actualités, de la base BBC_News, basé sur un ensemble d'algorithmes d'apprentissage automatique. Ce travail propose les algorithmes de la régression logistique, de la forêt aléatoire, du SVM, du NB et de K-plus proche voisin,et décrive chaque aspect du modèle en détail en fournissant les métriques d'évaluation.

Lorsque des algorithmes d'apprentissage automatique sont implémentés sur l'ensemble de données utilisé, le paramètre le plus important qui compte est la performance.Par conséquent, le résultat montre que le classificateur de SVM avec la représentation TF-IDF de caractéristiques atteint la performance la plus élevée (99%) pour l'ensemble de données.Cet algorithme est devenu le classificateur le plus stable dans un petit ensemble de données.Le deuxième meilleur score était obtenu par les classificateurs de KNN, NV et la régression logistique en même temps avec une performancede 98%. Le classificateur avec le moins de performance parmi les cinq étaitla forêt aléatoire avec une performance globale de 92 %.Le classificateur de SVM a donné une performance conforme à celle attendue en termes de tous les paramètres. Par conséquent, lesrésultatsobtenus sont conformément aux attentes.

Perspectives

Bien que notre modèle ait implémenté avec une grande précision et un taux de précision élevé, certains défis peuvent être envisagés pour le développement futur de ce travail :

- L'ensemble de données utilisé ici est entièrement statistique et basé sur du texte. Le problème de l'apprentissage à partir de données textuelles avec un déséquilibre de classe est le problème auquel nous sommes confrontés.
- La classification de texte est confrontée aussi au problème de la grande dimensionnalité de l'espace des caractéristiques, et plus précisément le nombre de vocabulaire. Dans notre cas, il existe plusieurs dizaines de milliers de caractéristiques, bien qu'un nombre

important de ces caractéristiques ne soient pas utilisées pour la tâche de classification de texte même certains d'entre eux peuvent fortement réduire la précision de la classification

- La portée de ces algorithmes peut s'étendre à différents ensembles de données qui peuvent avoir des caractéristiques basées sur des images et des audios POS (Part-Of-Speech). L'utilisation de ceux-ci fournirait un large domaine d'application pour cette recherche.
- Le développement actuel vers l'automatisation peut être considérablement avancé par l'utilisation d'applications de classification de texte. Ceux-ci peuvent altirer directement la facilité d'application en transformant les commandes que nous prononçons, en actions directes par des machines [44].

Références Bibliographiques

[1] Matallah Hocine, Mémoire magister en informatique, « Classification Automatique de Textes Approche Orientée Agent », Université Aboubeker BELKAID Tlemcen,2011.

[2] Site web :'//www.yumens.fr/expertise/smo/reseaux/2020', consulté le (04/ 03/2022)

[3] Site web : ' https://dewzilla.com/a-brief-history-of-social-media', consulté le (06/03/2022)

[4] Site web : 'https://fredcavazza.net/2011/02/06/description-des-differents-types-de-medias-sociaux 09/2011/ ', consulté le (10/03/2022)

[5] Site web : 'https://www.alqiyady.com', consulté le (12/03/2022)

[6] Site web : 'https://www.outsource.be/fr/rp/medias-traditionnels-les-differents-types-de-contenu-expliques/ Consulté (31 janvier 2019)', consulté le (20/03/2022)

[7] Site web : 'https://menaeditors.com/training/ 411-2021-10-28-14-37-04', consulté le (22/03/2022)

[8] Site web : 'https://audreytips.com/glossaire-web/contenu-interactif', consulté le (26/03/2022)

[9] Site web : ' https://buffer.com/resources/open/', consulté le (29/03/2022)

[10] Site web : 'https://mawdoo3.com', consulté le (05/04/2022)

[11] Site web : 'https://whatagraph.com/blog/articles/social-media-data-mining', consulté le (13/04/2022)

[12] Site web : 'https://ia-data-analytics.fr/logiciel-data-mining/text-mining/definition/', consulté le (22/04/2022)

[13] Site web : 'https://www.upgrad.com/blog/', consulté le (27/04/2022)

[14] Site web : 'https://e3arabi.com', consulté le (02/05/2022)

[15] Bechar Amine, Tenfir Nassim, Mémoire de Master en informatique, « Annotation et classification automatique des articles d'actualité », Université Akli Mohand Oulhadj de Bouira, 2020

[16] Mérabti H, Thèse de doctorat en informatique, « Approches bio-inspirées pour la reconnaissance de formes», Université 8 Mai 1945 Guelma, 2016

[17] Lahlou Ouchiha, mémoire de Master en informatique, « classification supervisée de documents étude comparative », Université du Québec en Outaouais, (Janvier 2016)

[18] Songbo Tan , "An effective refinement strategy for KNN text classifier, Expert SystAppl", (2006)

[19] Site web : 'https://JKP', consulté le (05/05/2022)

[20] Stéphane Tufféry, «Data Mining et statistique décisionnelle, L'intelligence des données», (2012)

[21] Site web : 'https://fr.acervolima.com /', consulté le (07/05/2022)

[22] Site web : 'https://datascience.eu/fr/', consulté le (09/05/2022)

[23] Site web: 'https://mrmint.fr/', consulté le (11/05/2022)

[24] Site web : ' https://www.tibco.com/fr/reference-center/ ', consulté le (14/05/2022)

[25] Site web :' /www.analyticsvidhya.com/blog/2021/06/', consulté le (16/05/2022)

[26] Site web :' /https://www.section.io/engineering-education/', consulté le (19/05/2022)

[27] Matallah Hocine, Mémoire magister en informatique, « Classification Automatique de Textes

Approche Orientée Agent », UNIVERSITE ABOU BEKR BELKAID-TLEMCEN, (2011)

[28] Site web : 'https://www.blogdumoderateur.com/2021', consulté le (22/05/2022)

[29] Site web : '/www.jetbrains.com/fr-fr/', consulté le (25/05/2022)

[30] site web : '/https://mobiskill.fr/blog/conseils-emploi-tech/ ', consulté le (27/05/2022)

[31] Site web : 'https://www.nltk.org/ ', consulté le (28/05/2022)

[32] Site web : 'https://datascientest.com/', consulté le (30/05/2022)

[33] Site web :' https://openclassrooms.com/fr/courses/4297211', consulté le (01/06/2022)

[34] BrunoTrstenjak, SasaMikac, DzenanaDonko," KNN with TF-IDF based framework for text categorization" , (2014)

[35] Kanish Shah, Henil Patel, Devanshi Sanghvi, Manan Shah ,"A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification" , (13 /02/2020)

[36] Ben Abdessalam Wahiba, Ben El Fadhl Ahmed "New fuzzy decision tree model for text classification, (28/30/2015)

[37] Abolfazl Nadi, Moradi Hadi, "Increasing the views and reducing the depth in random forest. Expert Syst Appl", (30 /12/ 2019)

[38] Site web : 'https://ieeexplore.ieee.org', consulté le (27/05/2022)

[39] Site web :'/https://www.python-graph-gallery.com/ ', consulté le (02/06/2022)

[40] Site web : 'https://www.kaggle.com/datasets', consulté le (03/06/2022)

[41] Site web :'https://radimrehurek.com/2021', consulté le (06/06/2022)

[42] Site web :'https://devopedia.org ', consulté le (07 /06/2022)

[43] Kizito Nyuytiymbiy ,"Parameters and Hyper parameters in Machine Learning and Deep Learning//",(Dec 30, 2020)

[44] Chen J, Huang H, Tian S, Qu Y (2009) Feature selection for text classification with Naıve Bayes. Expert Syst Appl 36(3–1):5432–5435