REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF-M'SILA

FACULTE DE TECHNOLOGIE
DEPARTEMENT D'ELECTRONIQUE
N°:..19.. /STLC/ 2021

DOMAINE: SCIENCES ET TECHNOLOGIE
FILIERE: ELECTRONIQUE
OPTION: SYSTEME DE TELECOMMUNICATION

Mémoire présenté pour l'obtention
Du diplôme de Master Académique

Par: BELHADJ Mohammed et BOUREZG Ala Eddine

Intitulé

Evaluation de la robustesse d'attribution en reconnaissance d'auteur

Soutenu publiquement le : 20 / 06 / 2021 devant le jury composé de :

Dr. BENNACER Hamza	Université M'sila	Président
Dr. KHENNOUF Salah	Université M'sila	Encadreur
Pr. SAYOUD Halim	Université USTHB	Co-Encadreur
Dr. OUALI Mohamed Assam	Université M'sila	Examinateur

Année universitaire: 2020/2021

بسم الله الرحمن الرحيم

REMERCIEMENTS

Nous remercions avant tout Allah le tout puissant pour son aide, sa bénédiction et pour tout ce qu'il nous a donné.

Un grand merci à nos encadreurs Dr. KHENNOUF Salah et Pr. SAYOUD Halim à qui nous devons beaucoup, pour leurs attentions, leurs disponibilités, leurs conseils et leurs sympathies.

Nos remerciements vont également aux membres du jury, chacun par son nom, pour avoir accepté de faire partie du jury d'évaluation de ce modeste travail.

Nous remercions tous les enseignants du département d'électronique qui ont contribué à notre formation, ainsi que tous les membres du cadre administratif.

Nous tenons à remercier, enfin, tous ceux qui ont aidés de près ou de loin lors de ce projet de fin d'études.

M. BELHADJ & A. BOUREZG

DEDICACE

Je dédie ce modeste travail à :

Mes chers parents qui m'ont aidé et m'ont encouragé pendant toute ma vie d'étude et d'être ma source de bonheur et de réussite.

Mes chers frères et sœurs.

Mes chers amis(es).

A tous ceux qui m'ont été d'un soutien moral ou matériel

Spécialement : HADJI FARES

Et à tous les collègues de ma promotion.

A. BOUREZG

DEDICACE

Ce travail est dédié à mon père, qui m'a appris que la meilleure connaissance est celle qui s'apprend pour soi-même. Il est également dédié à ma mère, qui m'a appris que même la plus grande tâche peut être accomplie si elle est accomplie étape par étape, et je n'oublie pas de la dédier à mes sœurs et à mon frère qui m'ont soutenu.

Je voudrais dédier ce travail à tous mes amis

Spécialement : HADJI FARES

Et à tous les collègues de ma promotion.

M.BELHADJ

Liste des Abréviations

API : Application Programming Interface.

CT : Catégorisation de Texte.

IA: Intelligence Artificielle.

MCE: Most Common Events.

ML : Machine Learning.

OCR : Optical Recognition Character.

OCR 23CAW: Optical Character Recognition of Contemporary Arab Writers

SVM: Support Vector Machine.

TAA: Taux d'Attribution d'Auteurs

TC : Texte Corrigé

TNC : Texte Non-Corrigé

.

Liste des Tableaux

Liste des figures

Table matières

Chapitre-I : Exploration de données textuelles

Chapitre-II: Approches Proposées

Chapitre-III : Simulation et Evaluation

Introduction générale

Introduction générale

De nombreux textes ont été écrits par des auteurs inconnus ou contestés depuis une longue histoire. Pour cette raison, les chercheurs s'efforcent toujours d'incarner un système robuste pour lutter contre la fraude des auteurs de contrefaçon. Par conséquent, il y a un intérêt croissant pour ce domaine en raison de l'importance de joindre une référence à un texte donné et de l'attribution des textes anonymes à leurs auteurs originaux. Le but est de comparer le texte anonyme avec les textes dont on connaît leurs auteurs.

La lutte contre le plagiat dans la littérature arabe nécessite des efforts de collaboration de plusieurs acteurs pour éliminer ce comportement immoral. Pour cela, plusieurs descripteurs seront utilisés pour modéliser le style de chaque auteur, ainsi que l'utilisation de diverses méthodes de classifications. Une base de données textuelle sera conçue et mis en œuvre pour valider les résultats obtenus.

Le présent travail s'intègre dans le cadre de l'attribution d'auteurs des documents textes. L'originalité de ce travail réside dans la manière dont on introduit ces documents textes à l'ordinateur. Cette opération consiste à scanner les pages contenant des textes, ensuite convertir les textes scannés ont format (.txt) à l'aide d'un logiciel appelé OCR (Optical Character Recognition) en français (Reconnaissance Optique de Caractère).

Les objectifs visés dans ce travail sont organisés comme suit :

- Évaluer l'impact de l'OCR des textes numériques sur la robustesse de notre système d'attribution des auteurs.
- Concevoir une base de données de textes par genre d'auteur et type de texte écrit pour mener des expériences afin d'identifier les auteurs de ces textes.
- Réalisation d'un système d'identification d'auteur basé sur des caractères N-gram comme descripteurs (features).
- Appliquer plusieurs classificateurs à l'étude (linéaire SVM, Burrows Delta…).

Enfin, ce mémoire se compose de trois chapitres comme suit ; Le premier chapitre, explique l'exploration de données textuelles et les tâches liées à ce domaine d'application. Ensuite, nous expliquons le processus de classification des différents textes et nous donnons une idée générale du concept de représentation numérique des textes. Enfin, une explication de l'apprentissage automatique pour l'exploration de texte et les techniques de

classification utilisées pour classer les textes est donnée. Le deuxième chapitre donnera un aperçu sur les méthodes de classification utilisées ainsi que le processus de conversion des textes, et enfin une explication du prétraitement de texte. Le troisième chapitre contient le système en place et les résultats obtenus.

Enfin, cette thèse se termine par une conclusion générale contenant une compilation des travaux menés dans cette thèse et des discussions ainsi que des interprétations possibles des résultats obtenus.

Chapitre-I

Exploration de données

Chapitre-I

Exploration de données textuelles

I.1 Introduction

L'exploration de texte est un nouveau domaine en plein essor qui tente de glaner des informations significatives à partir de textes en langage naturel. Il peut être vaguement caractérisé comme le processus d'analyse de texte pour extraire des informations utiles à des fins particulières. Comparé au type de données stockées dans les bases de données, le texte est non structuré, amorphe et difficile à traiter de manière algorithmique. Néanmoins, dans la culture moderne, le texte est le véhicule le plus courant pour l'échange formel d'informations. Le domaine de l'exploration de texte traite généralement des textes dont la fonction est la communication d'informations factuelles ou d'opinions, et la motivation pour essayer d'extraire automatiquement des informations d'un tel texte est convaincante, même si le succès n'est que partiel.

I.2 Les Facteurs de choix d'un type d'analyse de données textuelles

Les chercheurs se situant dans le courant actuel de recherche en Stratégie qui valorise la dimension langagière et communicationnelle ont bien compris l'importance de se doter d'outils pour l'analyse des discours, outils par ailleurs devenus classiques en sciences humaines et sociales (en linguistique et en sociologie bien sûr, mais aussi en histoire, lettres, droit, médecine …).

Analyser u²n discours relève toujours d'une créativité et d'un bricolage, le profil de l'analyste reste donc une variable importante (discipline d'origine, référentiel théorique, compétences, entourage...). Au-delà de ce premier point, et pour une recherche en Stratégie, le choix d'un outil d'analyse devrait surtout dépendre de trois éléments : les choix méthodologiques, la constitution du corpus, et le moment de l'analyse statistique [1].

Tableau I.1. Les facteurs de choix d'un type d'analyse de données textuelles[w]

	Analyses Lexicales	**Analyses Linguistiques**	**Analyses Cognitives**	**Analyses Thématiques**
Cadre Méthodologique	❖ Exploratoire ❖ Modèle	❖ Exploratoire	❖ Exploratoire	❖ Exploratoire ❖ Modèle
Implication du chercheur	❖ Faible	❖ Forte ❖ Faible	❖ Forte	❖ Forte
Axe temporel	❖ Instantané ❖ Longitudinl	❖ Instantané	❖ Instantané	❖ Instantané ❖ Longitudinal
Objet d'analyse	❖ Un groupe	❖ Un individu		❖ Un projet
Taille du corpus	❖ Importante	❖ Limitée	❖ Limitée	❖ Importante
Lisibilité Corpus	❖ Forte	❖ Forte	❖ Faible	❖ Faible
Homogénéité Corpus	❖ Faible	❖ Forte	❖ Forte	❖ Faible
Structuration langage	❖ Faible	❖ Faible		❖ Forte
Moment de l'analyse statistique	❖ Découverte ex-ante ❖ Contrôle ex-post	❖ Ex-ante	❖ Ex-post	❖ Ex-post

I.3 Exploration de texte (Text-Mining)

L'exploration de texte est la découverte par ordinateur de nouvelles informations auparavant inconnues, en extrayant automatiquement des informations à partir de différentes ressources écrites. Un élément clé est la mise en relation des informations extraites pour former de nouveaux faits ou de nouvelles hypothèses à explorer plus avant par des moyens d'expérimentation plus conventionnels.

L'exploration de texte est différente de ce que nous connaissons dans la recherche sur le Web. Dans la recherche, l'utilisateur recherche généralement quelque chose qui est déjà connu et qui a été écrit par quelqu'un d'autre. Le problème est de mettre de côté tout le matériel qui actuellement ne correspond pas à vos besoins afin de trouver les informations pertinentes. [2]

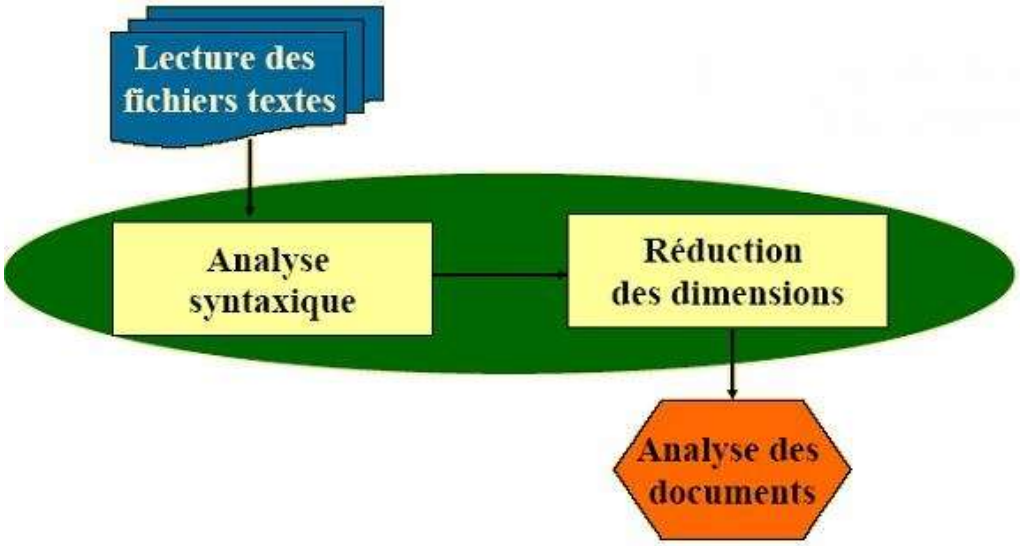


Figure-I.1 : Processus du Text-Mining schéma simplifiée [3].

I.4 Exploration de texte et exploration de données (Text-Mining **and** Data-Mining**)**

Tout comme l'exploration de données peut être vaguement décrite comme la recherche de modèles dans les données, l'exploration de texte consiste à rechercher des modèles dans le texte. Cependant, la similitude superficielle entre les deux caches de réelles différences. L'exploration de données peut être plus complètement caractérisée comme l'extraction d'informations implicites, inconnues auparavant et potentiellement utiles à partir de données. L'information est implicite dans les données d'entrée : elle est cachée, inconnue, et pourrait difficilement être extraite sans recourir à des techniques automatiques de fouille de données. Avec le Text-Mining, cependant, les informations à extraire sont clairement et explicitement énoncées dans le texte. Ce n'est pas du tout caché - la plupart des auteurs s'efforcent de s'exprimer clairement et sans ambiguïté - et, d'un point de vue humain, le seul sens dans lequel il est « inconnu auparavant » est que les restrictions en matière de ressources humaines le rendent impossible pour les gens de lire le texte eux-mêmes. Le problème, bien sûr, est que l'information n'est pas présentée d'une manière qui se prête à un traitement automatique. L'exploration de texte s'efforce de le faire sortir du texte sous une forme adaptée à la consommation par les ordinateurs directement, sans avoir besoin d'un intermédiaire humain.

Bien qu'il y ait une différence claire sur le plan philosophique, du point de vue de l'ordinateur, les problèmes sont assez similaires. Le texte est tout aussi opaque que les données brutes lorsqu'il s'agit d'extraire des informations probablement plus encore

Une autre exigence commune à l'exploration de données et de texte est que les informations extraites doivent être « potentiellement utiles ». Dans un sens, cela signifie actionnable - capable de fournir une base pour des actions à entreprendre automatiquement. Dans le cas de l'exploration de données, cette notion peut être exprimée d'une manière relativement indépendante du domaine : les modèles actionnables sont ceux qui permettent de faire des prédictions non triviales sur de nouvelles données provenant de la même source. Les performances peuvent être mesurées en comptant les réussites et les échecs, des techniques statistiques peuvent être appliquées pour comparer différentes méthodes d'exploration de données sur le même problème, etc. cependant, dans de nombreuses situations de text mining, il est beaucoup plus difficile de caractériser ce que « actionnable » signifie d'une manière indépendante du domaine particulier concerné. Il est donc difficile de trouver des mesures justes et objectives du succès.

Dans de nombreuses applications d'exploration de données, « potentiellement utile » est interprété différemment : la clé du succès est que les informations extraites doivent être compréhensibles dans la mesure où elles aident à expliquer les données. Cela est nécessaire chaque fois que le résultat est destiné à la consommation humaine plutôt qu'à (ou aussi) une base pour une action automatique. Ce critère est moins applicable à l'exploration de texte car, contrairement à l'exploration de données, l'entrée elle-même est compréhensible. L'exploration de texte avec une sortie compréhensible revient à résumer les caractéristiques principales d'un grand corps de texte, qui est un sous-domaine à part entière : la synthèse de texte .[4]

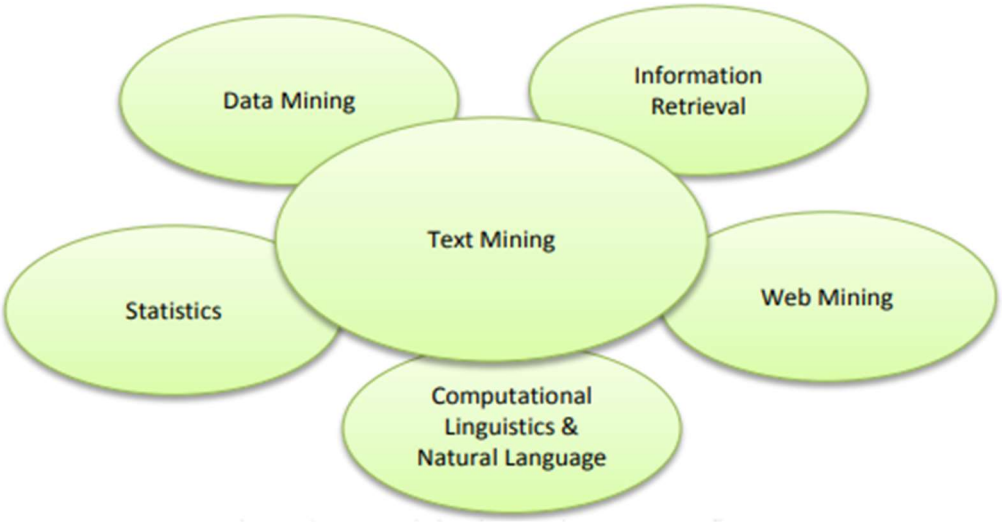


Figure-I.2 : Text-Mining en relation avec d'autres domaines

I.5 Processus du text mining

Trouver des connaissances utiles dans une collection de documents texte implique de nombreuses étapes différentes. Pour les organiser dans un ordre significatif, nous pourrions examiner le processus général d'exploration de texte. Il se compose des étapes suivantes :

I.5.1 Définir le problème

Cette étape est en fait indépendante des actions qui pourraient être entreprises ultérieurement. Ici, le domaine du problème doit être compris et les questions auxquelles il faut répondre, définies.

I.5.2 Collection des données nécessaires

Les sources des textes contenant les informations souhaitées doivent être identifiées et les documents collectés. Les textes peuvent provenir de l'intérieur d'une entreprise (base de données interne ou archive) ou de sources externes, du web par exemple. Dans ce cas, des scrapers Web doivent être fréquemment mis en œuvre pour saisir directement le contenu des pages Web. Alternativement, l'API (Automatique Programming Interface) de certains systèmes Web peut être utilisée pour récupérer les données. Après la récupération, les textes sont stockés afin qu'ils soient prêts pour une analyse plus approfondie.

I.5.3 Caractéristiques de définition

Les caractéristiques qui caractérisent bien les textes et sont adaptées à la tâche donnée doivent être définies. Les caractéristiques sont généralement basées sur le contenu des documents. Une approche très simple, un sac de mots avec pondération d'attributs binaires, prend chaque mot comme une caractéristique booléenne. Sa valeur indique si le mot est ou non dans un document. D'autres méthodes peuvent utiliser des schémas de pondération plus complexes ou des caractéristiques dérivées des mots (mots modifiés, combinaisons de mots, etc.).

I.5.4 Analyse des données

C'est le processus de recherche de modèles dans les données. Selon le type de tâche à résoudre (par exemple, la classification), un modèle ou un algorithme spécifique est sélectionné et ses propriétés et paramètres définis. Ensuite, le modèle peut être appliqué aux données afin de trouver une solution au problème résolu. Pour résoudre un problème

spécifique, plusieurs modèles sont généralement disponibles. Le choix n'est pas explicitement donné à l'avance. Les modèles ont des caractéristiques différentes qui influencent le processus d'exploration de données et son résultat. Le modèle peut être (boîte blanche) ou ne doit pas être (boîte noire) bien interprétable par un humain. Certains modèles ont une complexité de calcul plus élevée que les autres. Selon l'utilisation du modèle, la création rapide peut être préférée à l'application rapide ou vice versa. L'adéquation d'un modèle dépend souvent fortement des données. Le même modèle peut fournir d'excellents résultats pour un ensemble de données alors qu'il peut complètement échouer pour un autre. Ainsi, la sélection d'un modèle approprié, la recherche de la structure appropriée et le réglage des paramètres nécessitent souvent beaucoup d'efforts expérimentaux.

I.5.5 Interprétation des résultats

Ici, certains résultats sont obtenus à partir de l'analyse. Nous devons les examiner attentivement et les relier au problème que nous voulions résoudre. Cette phase peut inclure des étapes de vérification et de validation afin d'augmenter la fiabilité des résultats.

I.6 Apprentissage automatique pour l'exploration de texte

Machine Learning [5] dans le cadre de l'informatique et de l'informatique est l'un des domaines les plus pratiques de l'intelligence artificielle [6] Il a été inspiré par la capacité d'apprendre, c'est-à-dire d'acquérir des connaissances nouvelles ou supplémentaires, qui est l'une des caractéristiques importantes des organismes vivants. L'apprentissage automatique en tant que science se concentre sur la recherche et le développement d'algorithmes, qui peuvent simuler ou émuler les capacités mentales des organismes vivants du point de vue de l'apprentissage.

Les connaissances acquises peuvent ne pas tout traiter complètement – car elles sont de nature générale – mais elles nous permettent de résoudre des problèmes futurs qui ne se sont pas produits fréquemment, sous la même forme et/ou dans le même environnement dans le passé. Si un tel problème futur est similaire à quelque chose qui s'est déjà produit, il n'est souvent pas possible de le résoudre en utilisant les instructions exactes du passé; cependant, une solution (ou procédure) similaire peut, espérons-le, réussir dans la plupart (ou idéalement, tous) les cas qui ne se sont pas encore produits.

L'apprentissage automatique est l'un des outils modernes qui inclut tout ce qui peut être utile. Aujourd'hui, il se compose de dizaines d'algorithmes divers, et la recherche apporte constamment des algorithmes modifiés ou complètement nouveaux, remplaçant parfois les

anciens car le domaine de l'apprentissage automatique est fortement dominé par les besoins pratiques du monde réel qui grandissent dans le temps. En plus des algorithmes et des technologies de l'information pertinentes, l'apprentissage automatique utilise inévitablement les mathématiques, en particulier la théorie des probabilités, les statistiques, la combinatoire, l'analyse mathématique et un ensemble d'autres disciplines essentielles.

Pour résoudre des problèmes à l'aide d'ordinateurs, l'apprentissage automatique diffère de la méthode traditionnelle, qui repose sur l'application de programmes soigneusement mis en œuvre et débogués pour saisir les données. Le programme informatique traditionnel se compose d'ensembles d'instructions qui traitent les données d'entrée et fournissent la sortie à l'aide d'un ensemble de paramètres prédéfinis, qui sont fournis de manière externe. Un exemple est un programme informatique qui recherche une liste prédéfinie de solutions aux problèmes attendus, quelque chose comme un tableau contenant des résultats pour des combinaisons de valeurs d'entrée ou un ensemble d'équations mathématiques. Un tel programme peut être – de manière simplifiée – considéré comme une fonction f(x) = y, qui pour une entrée donnée x renvoie une sortie y, cependant, à condition que f(x) soit connue de manière fiable. Sans connaître une telle fonction, il est nécessaire de l'estimer ou de l'approximer.

Si les conditions et les règles nécessaires sont respectées, le résultat est garanti car il est basé sur des théorèmes et des méthodes mathématiquement prouvés. Malheureusement, s'il y a – souvent même de très petites modifications – des données ou de l'environnement de travail, un tel programme est prédisposé à une erreur fondamentale ou à un dysfonctionnement. Par exemple, les données textuelles analysées peuvent commencer à contenir de nouveaux termes qui n'étaient pas connus lors de la création du programme, ou il devient nécessaire de commencer à analyser ces données dans une autre langue. Dans un tel cas, il est inévitable de retravailler, souvent complètement, ce programme.

L'apprentissage automatique peut utiliser un algorithme implémenté existant, le recycler avec d'autres exemples et appliquer ces connaissances nouvelles ou étendues à la résolution d'un nouveau problème. Un certain défaut pourrait être que cette connaissance n'est pas suffisamment étayée par des preuves mathématiques en apprentissage automatique, le support vient de preuves empiriques et d'heuristiques.

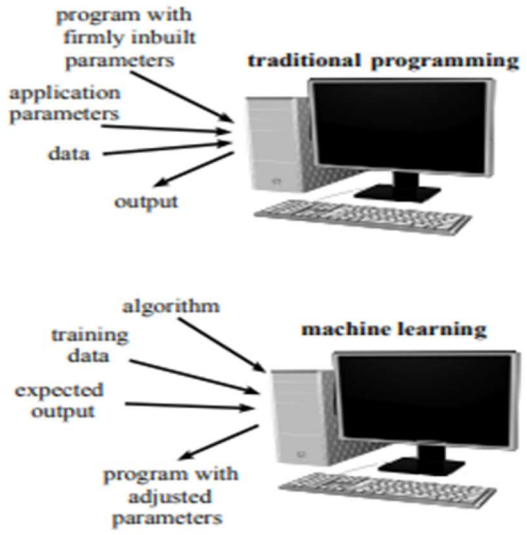


Figure-I.3 : Programmation traditionnelle vs machine Learning – illustration des principes

Une illustration simple de la différence entre les méthodes traditionnelles et l'apprentissage automatique est présentée à la figure 1.3.

Déterminer la structure des modèles et leurs paramètres est l'objectif du processus d'apprentissage, qui utilise (souvent de manière itérative) les échantillons d'apprentissage présentés à cette fin. Le taux de réussite de la phase de formation doit être mesuré par des tests, qui utilisent un autre ensemble d'échantillons qui ne font pas partie du processus de formation.

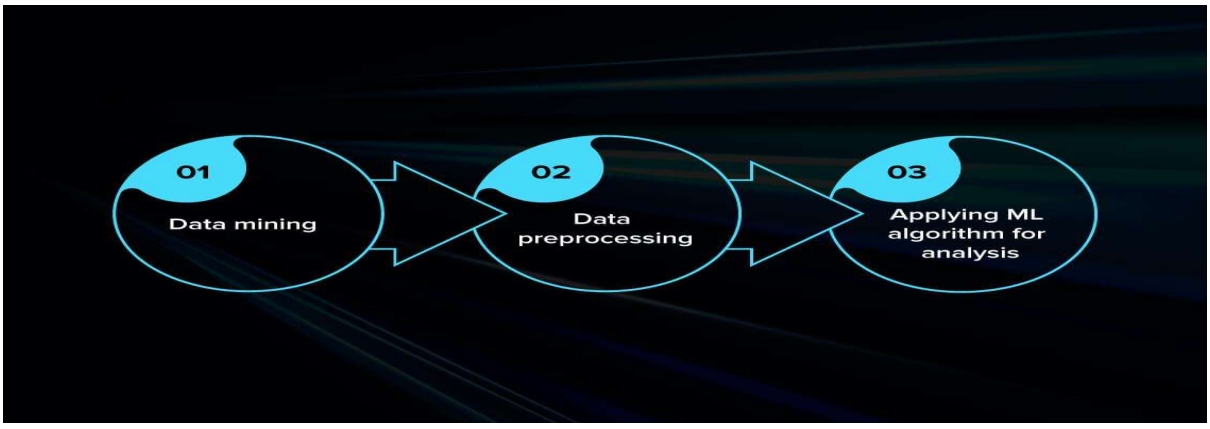


Figure-I.4 : L'interaction de l'exploration de texte avec l'apprentissage automatique

I.7 Catégorisation de texte

La CT est l'une des tâches fondamentales de l'exploration de texte dans l'analyse de données complexes et non structurées qui se préoccupent de « l'attribution de textes en langage naturel à une ou plusieurs catégories prédéfinies en fonction de leur contenu [7] ». Le concept de classification de texte a été anticipé pour la première fois au début des années soixante [8] et il s'est concentré sur l'indexation des revues scientifiques en utilisant le vocabulaire. Dernièrement, ce domaine de recherche a suscité plus d'intérêt en raison de la croissance rapide des documents en ligne contenant des connaissances importantes et utiles. Par conséquent, la classification automatique de texte est devenue l'un des domaines clés pour l'organisation et la gestion des données textuelles Actuellement, de nombreuses applications sont basées sur la catégorisation de texte, notamment : le filtrage de documents, le filtrage de spam, la génération automatique de métadonnées, la classification des ressources Web sous autres [9] [10].

I.7.1 L'exploration de texte arabe (Arabic Text-Mining)

L'arabe est considéré comme l'une des langues les plus parlées dans le monde, et en fait, c'est une langue principale dans les nations arabes ainsi qu'une langue secondaire dans de nombreux autres pays. L'alphabet de la langue se compose de 28 lettres plus un caractère spécial appelé Hamza (ء), qui est. De plus, le sens d'écriture en arabe est de droite à gauche contrairement à l'anglais et au latin. Le style d'écriture des lettres dans un mot varie selon la position de la lettre dans le mot. Ainsi, si les lettres viennent au début, au milieu ou à la fin du mot, la forme des lettres change. Enfin, il existe des signes diacritiques en arabe qui sont des symboles placés au-dessus ou en dessous des lettres pour doubler la lettre dans la prononciation ou pour donner des voyelles courtes [11].

I.7.2 Méthodologie de recherche

La recherche est un processus systématique dans lequel la définition de l'objectif, le contrôle des données et la communication des résultats se déroulent dans des cadres reconnus et associés aux lignes directrices existantes [12]. Le succès de la recherche dépend du choix d'une méthodologie de recherche appropriée. L'utilisation de méthodes qualitatives et quantitatives répond aux objectifs de cette recherche. L'approche quantitative concerne la collecte de données sous forme quantitative qui peuvent faire l'objet d'une analyse quantitative approfondie dans une méthode formelle [13].

Cette stratégie de recherche est appliquée pour analyser les résultats de l'expérimentation à savoir; la précision, le rappel, le taux d'erreur et le nombre de règles générées dérivées de différentes approches de classification. La recherche qualitative implique une évaluation subjective des approches, des opinions et des comportements dans lesquels les résultats sont obtenus soit de type non quantitatif, soit de type non soumis à une analyse quantitative détaillée [13].

Étant donné que l'un des objectifs de cet article est de mener une revue de la littérature sur la classification des textes arabes, l'approche qualitative est considérée comme une approche de recherche appropriée pour mener à bien cette tâche. L'intégration de ces deux approches dans une seule étude est connue sous le nom de méthodologie de recherche mixte [12]. L'importance d'employer ce type d'approche méthodologique est d'atteindre les points forts et de réduire les points faibles des approches de recherche quantitatives et qualitatives.

I.8 Catégorisation des documents textuels

Depuis l'apparition de la catégorisation des documents textuels, plusieurs thématiques de cette dernière sont apparues suite à la diversité de documents disponibles. On peut citer, la catégorisation par langue, par thème, par auteur, par genre, etc [14].

I.8.1 Catégorisation par langue

Elle consiste à reconnaitre la langue d'un texte donné. Ce type de catégorisation est peu abordé par les chercheurs, car il est considéré par certains comme un domaine non difficile et par d'autres comme un problème résolu [15]. Contrairement, l'identification de la langue représente actuellement un défi scientifique quant au traitement des documents multilingues ou de très courts documents (comme les messages Twitter) [16] [14].

I.8.2 Catégorisation par thème

Le deuxième sous-domaine bien connu et dans lequel plusieurs travaux ont été réalisés, c'est la catégorisation par « thème » ou par « sujet » qui est apparue la première fois dans la bibliothéconomie afin d'archiver les documents traitant le même sujet et le même contexte. Par ailleurs, avec l'expansion de l'internet et les différents moyens numériques il est devenu impossible de catégoriser les documents manuellement. De plus, il est parfois difficile de distinguer le thème d'un document/texte vu qu'il aborde plusieurs thèmes (ex. texte qui aborde la relation entre la politique et l'économie) [14].

I.8.3 Catégorisation par auteur

La catégorisation des documents textuels par auteur est l'une des tâches les plus abordées dans ce domaine, où elle représente un vrai défi scientifique. D'après Stamatatos [17], le célèbre physicien Mendenhall était le pionnier de la stylométrie qui étudia les pièces de Shakespeare en 1887 (reconnaitre le vrai auteur des pièces). Mendenhall utilisa les distributions des fréquences des mots de différentes longueurs afin d'identifier l'auteur. D'autre part, on trouve l'un des travaux les plus anciens et influents dans l'identification de l'auteur, c'est le travail de Mosteller sur l'identification de l'auteur des douze articles « Federalist Papers » [18].

I.8.4 Autres catégorisations des documents textuels

Outre que ces trois types de catégorisations, on distingue trois autres sous-catégories qui sont peu ou rarement abordés par les études de recherches ; elles constituent également la catégorisation par « genre » qui consiste à identifier le genre du document (poème, article scientifique, etc.).

De plus, la catégorisation par « genre de l'auteur » (ou en anglais Author Gender) et la catégorisation par « tranche d'âge », consistent à identifier respectivement le genre de l'auteur d'un document (masculin ou féminin) et sa tranche d'âge.

Enfin, les deux autres catégories, c'est la catégorisation par « opinion » et par « avis des clients » qui sont de nouvelles disciplines apparues récemment et largement adressées par les chercheurs suite à l'utilisation des réseaux sociaux et le e-marketing. [14].

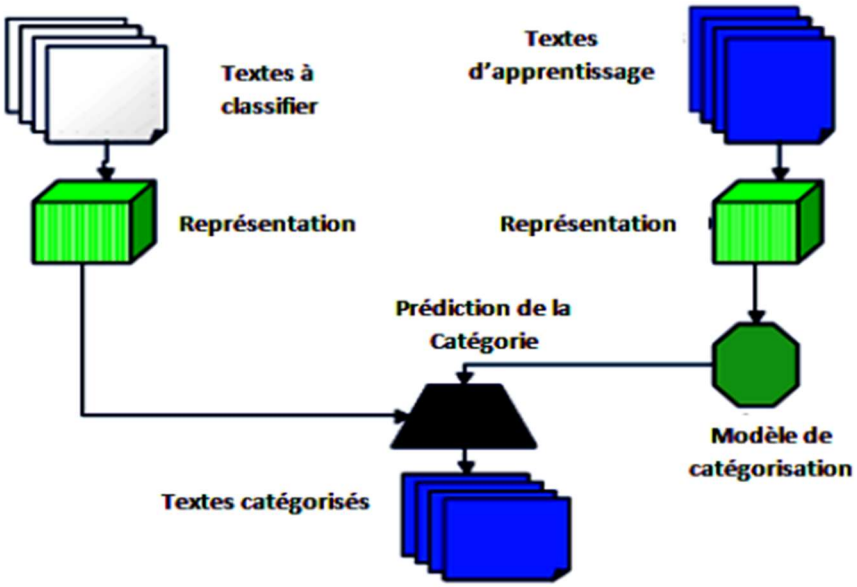


Figure-I.5 Processus de la catégorisation de textes

I.9 Problèmes de la catégorisation de textes

Plusieurs difficultés peuvent s'opposer au processus de la catégorisation de textes. Des problèmes liés à l'apprentissage automatique supervisé comme la subjectivité de la décision prise par les experts, les sur-apprentissages, etc.

Mais aussi des problèmes particuliers liés à la nature des données traitées à savoir des données textuelles comme la polysémie, la redondance, Les variations morphologiques ou même L'homographie, etc...

Dans ce qui suit nous allons signaler les difficultés principales qui s'opposent à la catégorisation de textes [19].

- **Redondance (Synonymie)**

La redondance et la synonymie permettent d'exprimer le même concept par des expressions différentes, plusieurs façons d'exprimer la même chose.

Cette difficulté est liée à la nature des documents traités exprimés en langage naturel contrairement aux données numériques.

Lors d'une représentation vectorielle d'un document, ces termes sont représentés séparément, et les occurrences du concept sont dispersées.

Il est alors important de rassembler ces termes en un groupe sémantique commun.

Pour y remédier, il est alors intéressant de concevoir une ontologie afin de cerner les sens des termes, naturellement, cela engendre des coûts supplémentaires pour sa réalisation.

- **Polysémie (Ambiguïté)**

A la différence des données numériques, les données textuelles sont sémantiquement riches, contrairement des langages informatiques, le langage naturel, autorise des violations des règles grammaticales engendrant plusieurs interprétations d'un même propos.

Un même mot possède, dans différents cas, plus d'un sens et, par conséquent, à cause de la polysémie, les mots seuls sont parfois de mauvais descripteurs.

Le mot livre peut désigner une unité monétaire, ou un bouquin ou le verbe livrer (nom: livraison).

Le mot avocat peut désigner le fruit, le juriste, ou même au sens figuré, la personne qui défend une cause.

Le mot table de cuisine ce n'est pas le même que dans table de multiplication.

Le mot pièce peut correspondre à une pièce de monnaie par exemple, ou à une pièce dans une maison, de même pour pavillon, bloc, glace, etc

- **L'homographie**

Deux mots sont dits homographes s'ils s'écrivent de la même façon sans forcément avoir la même prononciation.

L'homographie est une sorte d'ambigüité supplémentaire. (Ex: avocat en tant que fruit et avocat en tant que juriste).

L'homographie et l'ambiguïté génère du bruit qui va causer une dégradation de précision (indicateur nécessaire pour mesurer la performance du classificateur). Il sera alors préférable d'ôter ces ambigüités.

- **La graphie**

Un terme peut comporter des fautes d'orthographe ou de frappes comme il peut s'écrire de plusieurs manières ou s'écrire avec une majuscule.

Ce qui va peser sur la qualité des résultats. Parce que si un terme est orthographié de deux manières dans le même document (Ghilizane, Relizane), la simple recherche de ce terme avec une seule forme graphique néglige la présence du même terme sous d'autres graphies, ce qui va influencer les résultats puisque les différentes graphies vont être traitées séparément.

- **Les variations morphologiques**

Les conjugaisons, pluriels, influent négativement sur la qualité des résultats puisque les différentes variations morphologiques vont êtes considérées séparément et chaque une va être prise comme un élément à part comme par exemple les trois termes: maître, maîtresse, maîtriser sont traités indépendamment quoi que en réalité cela pivote sur la même idée.

Pour y remédier soit on applique la lemmatisation ou le stemming, à notre texte soit carrément on opte pour une représentation en n-grammes qui peut nous éviter ces prétraitements.

- **Les mots composés**

La non prise en charge des mots composés comme: comme Arc-en-ciel, peut-être, sauve qui peut, etc.

Dont le nombre est très important dans toutes les langues, et traiter le mot Arc en ciel par exemple en étant 3 termes séparés réduit considérablement les performances d'un système de catégorisation néanmoins l'utilisation de la technique des n-grammes pour le codage des textes atténue considérablement ce problème des mots composés.

- **Présence-Absence de termes**

La présence d'un mot dans le texte indique un propos que l'auteur a voulu exprimer, on a donc une relation d'implication entre le mot et le concept associé, quoi que on sait très bien qu'il y a plusieurs façons d'exprimer les mêmes choses.

Dès lors l'absence d'un mot n'implique pas obligatoirement que le concept qui lui est associé est absent du document.

- **Sur-apprentissage**

Le nombre de termes très important et très varié qui ne se répètent dans tous les textes va causer énormément de creux dans le tableau de grande dimension (textes / termes) qui peut provoquer du sur-apprentissage qui s'explique par le fait que le modèle n'arrive pas à bien classer les nouveaux textes, pour tant il l'a bien fait dans la phase d'apprentissage en classant correctement les textes de la base d'apprentissage.

Pour limiter les sur-apprentissages, on doit sélectionner des termes pour réduire la dimensionnalité.

En général le nombre de textes d'apprentissage est limité, c'est pour cela on cherche à agir sur le nombre des termes utilisés en les diminuant, pour éviter ce sur-apprentissage. Sans bien sûr pénaliser le système en supprimant des termes pertinents

- **Subjectivité de la décision**

Après la lecture du texte à classer, l'expert va trancher à quelle (s) catégorie (s) ce texte appartient en se basant sur le contenu sémantique et le contexte du texte et même en consultant d'autres textes préalablement associés à certaines classes, pour valider la décision prise qui ne peut être que subjective.

Les experts humains ne lisent pas de la même manière! Ne réfléchis sent pas de la même manière! Donc ne classent pas de la même manière!

Ainsi un même document peut être classé différemment par deux experts, ou encore un même document peut être classé différemment par le même expert, soumis à deux instants différents.

D'après les expériences: Lorsque deux experts humains doivent déterminer les classes d'une collection de textes, il y a souvent désaccord sur plus de 5% des textes.

Il est donc illusoire de rechercher une catégorisation automatique parfaite.

I.10 Conclusion

Dans ce chapitre, nous avons couvert quelques généralités sur l'exploration de données, nous avons abordé les facteurs de choix du type d'analyse de données textuelles et la relation entre l'exploration de données et d'autres domaines.

D'autre part, nous avons discuté de l'apprentissage automatique pour le Text-Mining qui fait l'objet de recherches contemporaines.

Enfin, la classification de texte est l'étape la plus importante dans l'exploration de données de texte pour de meilleurs résultats


Chapitre-II
Approches Proposées

Chapitre-II

Approches Proposées

II.1 Introduction

Il existe de nombreuses méthodes et méthodologies différentes utilisées pour obtenir un système d'identification d'auteur robuste, ses règles de base étant la mesure de certaines propriétés textuelles, la présentation de la méthodologie de recherche ainsi que les différentes techniques proposées pour l'identification ou l'attribution des auteurs.

II.2 Machine à vecteurs de support (SVM)

L'algorithme SVM est une technique d'apprentissage supervisé basée sur la théorie de l'apprentissage statistique, qui a été développée par vapink1995, récemment considéré comme l'un des outils les plus puissants pour résoudre les problèmes de reconnaissance de formes et de régression dans de nombreuses applications. Il a une bonne capacité de généralisation, une efficacité de calcul et est très robuste en dimensions élevées[20]

II.3 SVM principe de fonctionnement général

II.3.1 Notions de base : Hyperplan, marge et support vecteur

Pour deux classes d'exemples donnés, le but de SVM est de trouver un classificateur qui va séparer les données et maximiser la distance entre ces deux classes. Avec SVM, ce classificateur est un classificateur linéaire appelé hyperplan. Dans le schéma qui suit, on détermine un hyperplan qui sépare les deux ensembles de points] 20].

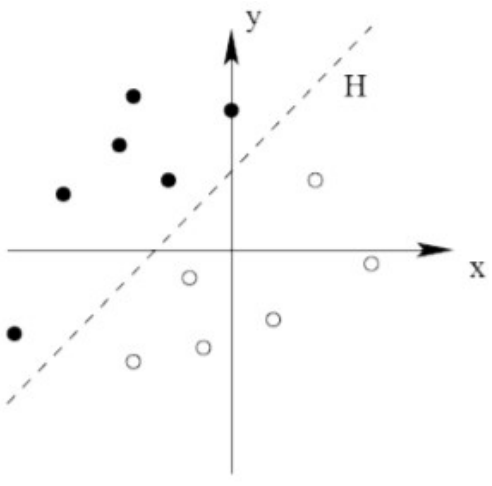


Figure -II.1 : Hyperplan qui sépare les deux ensembles de points

Les points les plus proches, qui seuls sont utilisés pour la détermination de l'hyperplan, sont appelés vecteurs de support

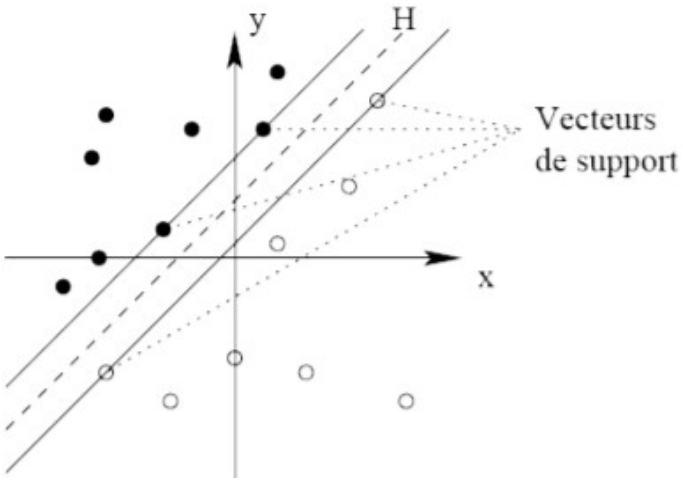


Figure -II.2 : L'hyperplan avec vecteurs de support

Il est évident qu'il existe une multitude d'hyperplan valide mais la propriété remarquable des SVM est que cet hyperplan doit être optimal. Nous allons donc en plus chercher parmi les hyperplans valides, celui qui passe « au milieu » des points des deux classes d'exemples. Intuitivement, cela revient à chercher l'hyperplan le « plus sûr ». En effet, supposons qu'un exemple n'ait pas été décrit parfaitement, une petite variation ne modifiera pas sa classification si sa distance à l'hyperplan est grande. Formellement, cela revient à chercher un hyperplan dont la distance minimale aux exemples d'apprentissage est maximale. On appelle cette distance « marge » entre l'hyperplan et les exemples. L'hyperplan séparateur optimal est celui qui maximise la marge. Comme on cherche à maximiser cette marge, on parlera de séparateurs à vaste marge.

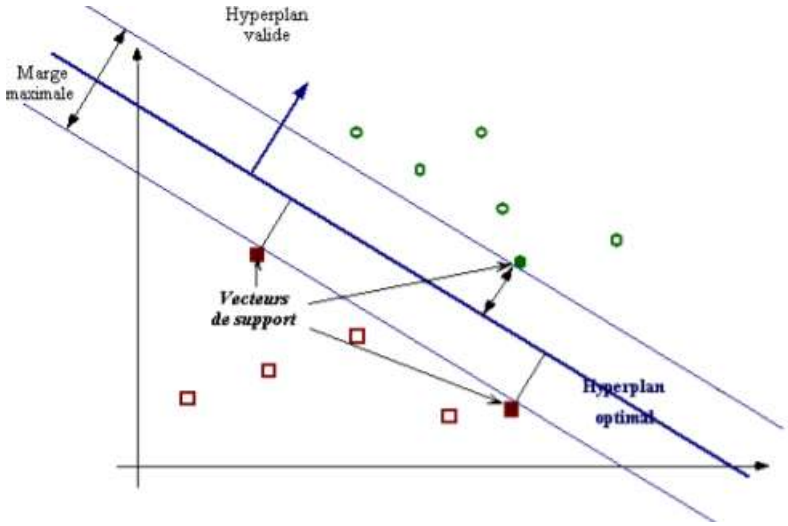


Figure -II.3: SVM- L'hyperplan séparateur optimal et la marge.

II.3.2 Linéarité et non-linéarité

Parmi les modèles des SVM, on constate les cas linéairement séparables et les cas non linéairement séparable. Les premiers sont les plus simple de SVM car ils permettent de trouver facilement le classificateur linéaire. Dans la plupart des problèmes réels il n'y a pas de séparation linéaire possible entre les données, le classificateur de marge maximale ne peut pas être utilisé car il fonctionne seulement si les classes de données d'apprentissage sont linéairement séparables

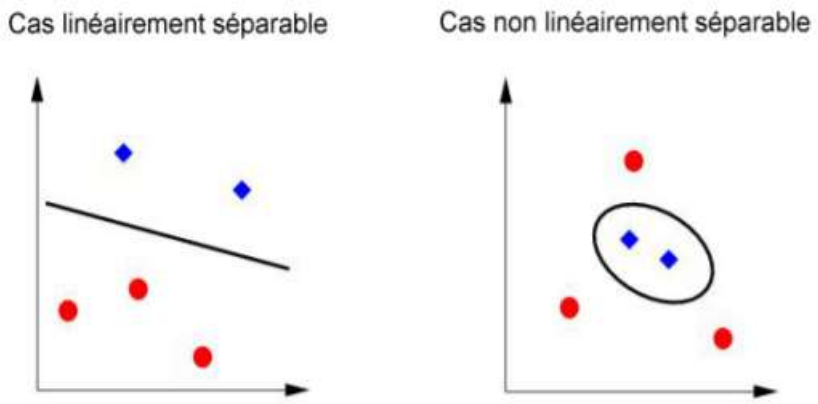
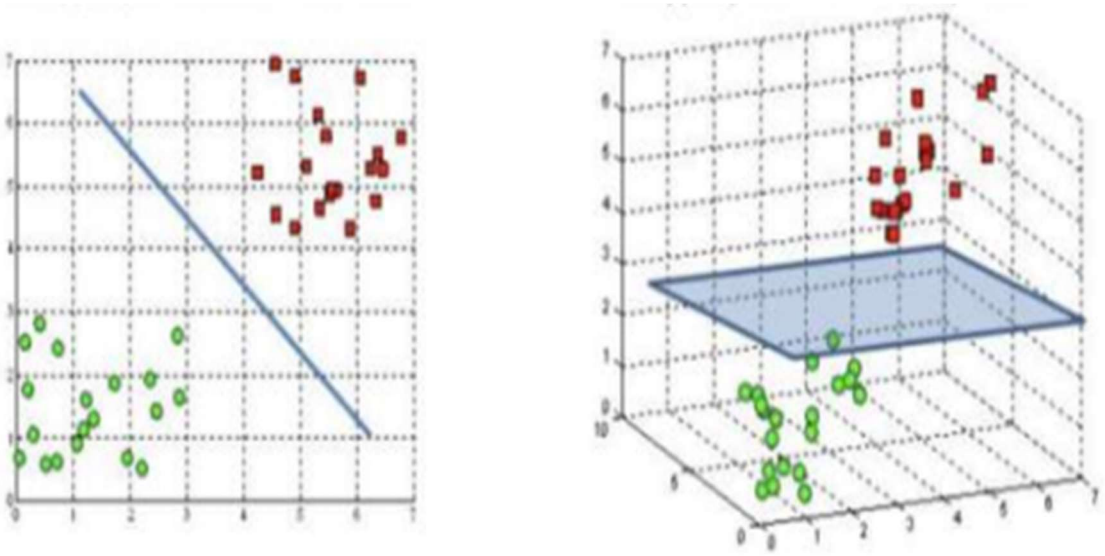


Figure -II.4:les modèles des SVM

II.4 Hyperplans et vecteurs de support

Les hyperplans sont des décisions de limites qui aident à classer les points de données. Les points de données qui tombent sur n'importe quelle partie de l'hyperplan peuvent être dédiés à différentes classes. En outre, la taille de l'hyperplan dépend du nombre de formes. Si le nombre de formes d'entrée est de 2, l'hyperplan n'est qu'une ligne. Si le nombre de formes d'entrée est de 3, alors l'hyperplan devient un plan bidimensionnel. Cela peut être difficile à imaginer si le nombre de formes dépasse 3



un hyperplan en R^3 est une lig un hyperplan en R^2 est un plan

Figure -II.5: Hyperplans dans l'espace d'entités 2D et 3D.

L'algorithme est basé sur les idées de base des machines à vecteurs de support (SVM) et tente de maximiser la marge du classificateur, qui est la distance minimale entre l'hyperplan défini par le classificateur et les points d'entraînement. Il existe une forte motivation théorique pour maximiser la marge d'un classifier qui découle d'un résultat de la théorie de l'apprentissage statistique qui relie une limite sur l'erreur de généralisation d'un classifier à la taille de sa marge. Il a également une interprétation très intuitive: les classificateurs avec des marges plus importantes sont plus robustes.

II.5 Classificateur de marge maximale

La plus courte de ces distances est appelée la distance minimale entre l'hyperplan et l'observation, et elle est appelée marge. Par conséquent, l'hyperplan à marge maximale est l'hyperplan qui a la plus grande marge, c'est-à-dire qui a la plus grande distance entre l'hyperplan et les observations d'entraînement. En utilisant cet hyperplan, nous pouvons classer les données de test[21].

$$Y_i\,(\beta_0 + \beta_1 X_{i1} 4 + \ldots\ldots + \beta_h X_{ih}) > 0$$

B_0: est l'interception

B_1 : Définition du premier axe

B_h: Définition du dernier axe

II.5.1 Marge dure SVM (Hard Margin SVM)

Étant donné un ensemble de données d'apprentissage $\{(xi,yi)\}^{n}_{i=1}$ séparables linéairement, SVM à marge dure cherche le plan affine qui sépare les deux classes avec la marge maximale. Cela revient à résoudre le problème d'optimisation suivant :

$$\Phi^{(n)} \triangleq \min_{w,b} \|w\|^2_2$$

$$\text{s.t. } \forall i \in \{1,\ldots,n\},\ y_i(w^T x_i + b) \geq 1$$

Soit $\hat{w}_H$ et b_H résoudre le problème ci-dessus, alors le classificateur à marge dure appliqué à une observation invisible x est donné par $L_H(x) = \text{sign}\ (\hat{\mathbf{w}}^T_H x + b_H)$.

II.5.2 marge souple SVM (Soft Margin SVM)

Si les données ne sont pas linéairement séparables, les contraintes de la SVM à marge dure ne peuvent pas toutes être satisfaites ensemble. En conséquence, le coût du problème d'optimisation des marges dures est infini, puisque le minimum sur un ensemble vide est par convention . Dans de tels paramètres, une alternative consiste à utiliser le SVM à marge souple qui, par construction, tolère que certaines données d'apprentissage soient mal classées, mais paie le coût de chaque observation mal classée en ajoutant une limite supérieure au nombre d'observations d'apprentissage mal classées. Plus formellement, le SVM à marge souple équivaut à résoudre le problème d'optimisation suivant :

$$\Phi \triangleq \min_{w,b} \|w\|^2_2 + \frac{\tilde{\imath}}{P} \sum_{i=1}^{n} \xi i$$

$$\text{s.t. } \forall i \in \{1,\ldots,n\},\ y_i(w^T x_i + b) \geq 1 - \xi i,\ \xi i \geq 0$$

où$\tilde{\imath}$ est un scalaire strictement positif, défini au préalable par l'utilisateur, et vise à faire un compromis entre la maximisation de la marge et la minimisation de l'erreur d'apprentissage. À cet égard, un petit $\tilde{\imath}$ tend à mettre davantage l'accent sur la marge tandis qu'un grand pénalise l'erreur d'apprentissage. Soit $\hat{\mathbf{w}}_S$ et b_S résoudre le problème ci-dessus, alors le classificateur SVM à marge souple appliqué à une observation invisible x est donné par $L_S(x) = \text{sign}\ (\hat{\mathbf{w}}^T_S x + b_S)$.

II.6 SVM linéaire

Dans le cas de la classification linéaire, la surface de séparation S est un hyperplan défini par[2] :

$$S = \{x : h(x) = w^T x + b = 0\}$$

Entrainer un classifieur, avec un ensemble d'apprentissage {(xi, yi),i=1 :M} consiste à trouver le modèle f:

$$f(x_i) = \text{signe}(w^T xi + b) = yi,\ i = 1,\ldots, M$$

Ce qui est équivalent à :

$$yi\ (w^T x_i + b) > 0,\ i = 1,\ldots, M.$$

Si un tel hyperplan existe, alors l'ensemble d'apprentissage est linéairement séparable :

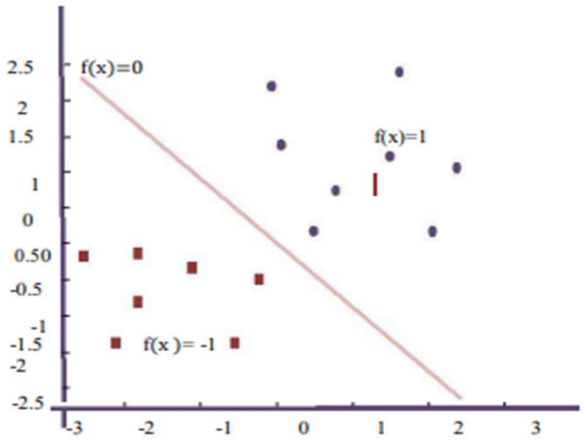


Figure-II.6 : Hyperplan séparateur entre 2 classes

II.7 Méthode Burrows-Delta

Alors que de nombreuses méthodes ont été appliquées au problème de l'attribution automatisée de la paternité, le « Delta la méthode » est une méthode particulièrement simple mais efficace. L'objectif est de déterminer automatiquement, à partir d'un ensemble de données connues documents de formation étiquetés par leurs auteurs, qui est l'auteur le plus probable pour un document de test non étiqueté. La méthode Delta utilise les mots les plus fréquents dans le corpus d'apprentissage comme caractéristiques qu'il utilise pour porter ces jugements. La mesure Delta est définie comme:

La moyenne des différences absolues entre les scores z pour un ensemble de variables de mots dans un groupe de texte donné et les scores z pour le même ensemble de variables de mots dans un texte cible.

Le Delta du document de test est calculé par rapport à chacun des documents de formation, et cet auteur dont le document de formation a un delta minimal et le document de test est choisi pour l'attribution.

Le delta de Burrows peut être considéré avec profit comme une méthode pour classer les candidats à la paternité selon leur probabilité. Soit X et Y sont des vecteurs à n dimensions des fréquences de mots dans deux documents. Notez que le z-score est obtenu en sous trayant la moyenne et en divisant l'écart type. Ensuite, la mesure Delta entre ces documents peut être reformulée [22]

$$\sum_{i=1}^{n} \left| Z(\mathrm{Xi}) - Z(Yi) \right| = \sum_{i=1}^{n} \left| \frac{Xi-\mu}{\sigma i} - \frac{Yi-\mu}{\sigma i} \right| = \sum_{i=1}^{n} \left| \frac{Xi-Yi}{\sigma i} \right|$$

Où x est le classificateur utilisé pour comparer les facteurs d'apprentissage et de test, n est le nombre de classificateurs ou des marqueurs qui seront utilisés dans le test, μ et σ sont respectivement la moyenne et la variance de la population.

Delta est comme une distance mise à l'échelle entre les deux documents. Ce n'est pas la distance ordinaire "à vol d'oiseau", mais plutôt c'est la somme de chaque dimension indépendante. Notez que si nous considérons la moyenne d'une distribution à la place de Yi, cela a une forme similaire à une probabilité de Laplace répartition. Plus précisément, c'est l'exposant du produit des distributions de Laplace indépendantes.

Ainsi, nous supposons que le document individuel avec lequel nous comparons le document de test est une sorte de document moyen pour cet auteur. La prise du z-score correspond à la normalisation dans l'exposant. Donc, dans un sens, Delta mesure la probabilité d'un document étant écrit par un auteur prenant chaque fréquence de mot indépendamment, puis choisissant le document avec la probabilité la plus élevée.

La méthode de Burrows a été améliorée avec succès dans le passé, c'est pourquoi la méthode de type informatique utilise en effet de simples caractéristiques statistiques efficaces.

Le score z utilisé dans le delta dépend de l'écart type. La dépendance moyenne conduit à donner un effet élevé ou poids pour les attributs à haute fréquence, ce qui conduit à obtenir une prédiction et une précision élevées dans les détections de texte.

Nous choisirons un des auteurs (let SAKHAWY) pour illustrer l'algorithme de terrier-delta et choisirons un attribut trio comme fonctionnalité [22]

II.8 Robustesse de recherche utilisé

Notre méthodologie de recherche est basée sur quatre étapes.

La première étape consiste à convertir les textes scannés en textes modifiables à l'aide d'un système OCR

La deuxième étape est consacrée aux opérations de prétraitement des textes obtenus par la conversion OCR afin de les préparés pour l'utilisation dans l'attribution d'auteurs.

Dans la troisième étape, nous extrayons les caractéristiques pertinentes (dans notre cas les caractères n-grammes) et classons les auteurs par genre pour construire un modèle pour chaque auteur.

La quatrième étape est dédiée aux méthodes de classification (Linear-SVM et Burrows-Delta) pour atteindre la robustesse du système et l'identification de l'auteur.

II.8.1 Conversion les textes utilisés

Les textes scannés sont convertis en textes modifiables en utilisant un système de Reconnaissance Optique de Caractères (en anglais : Optical Recognition Character) (OCR). Un texte est une association de caractères appartenant à un alphabet, réunis dans des mots d'un vocabulaire donné. L'OCR doit retrouver ces caractères, les reconnaître d'abord individuellement, puis les valider par reconnaissance lexicale des mots qui les contiennent.

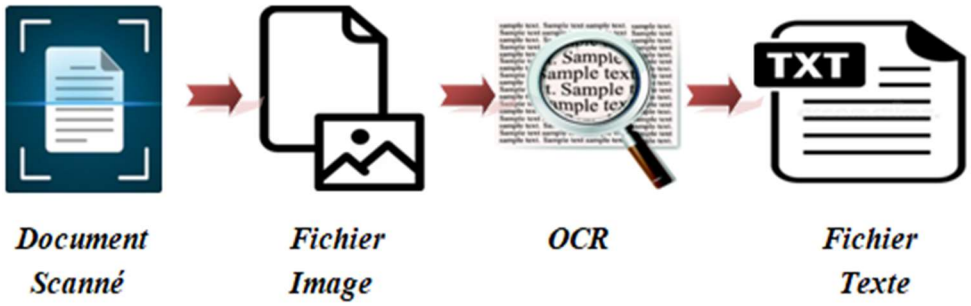


Figure-II.7 : Conversion des textes scannés en textes modifiables à l'aide d'un OCR.

II.8.2 Prétraitement des textes obtenus par la conversion OCR

Les textes convertis à l'aide d'un OCR comportent deux types d'erreurs ; caractères insignifiants ou (bruits) (caractères spéciaux, des chiffres, etc.) et caractères incorrects. Les caractères insignifiants sont des caractères qui n'ont pas un sens bien défini dans la langue arabe et qui apparaissent par erreur dans les textes convertis à l'aide d'un système OCR (i.e. ", %, &, £, *, #, $, 0, 1, …, 9, etc.). Or, les caractères incorrects sont des caractères qui sont mal convertis ou convertis par erreur en autres caractères que les vrais caractères (حconverti en خ ou ج ou encore ق converti en ف .(En conséquence, le prétraitement appliqué dans cette phase consiste à :

- ♣ Supprimer les caractères insignifiants.
- ♣ Supprimer les caractères français et anglais.
- ♣ Supprimer les diacritiques arabes.
- ♣ Supprimer les multiples espaces.

Pour les caractères incorrects sont laissés afin de tester la robustesse de notre méthode proposée pour l'attribution d'auteurs.

Tableau-II.1 : Liste des caractères insignifiants pour la stylométrie

N°	Caractères	Noms
01	1. الأم الشابة	Les titres
02	244- 119- (19)	Les références (numéros dans le texte et en bas de page)
03	0 1 2 3 4 5 6 7 8 9	Chiffres en Français
04	٠١٢٣٤٥٦٧٨٩	Chiffres en Arabe
05	+ - x / = %⁒<>	Symboles mathématiques
06	‘ ’ "" « » “ ”	Guillemets
07	() [ ] { }	Parenthèse, Crochet,
08	, ; ?. ! : … ، ؛ ؟	Ponctuation
09	* # @	Etoile, dièse, arobas
10	§ †| _–	Caractères spéciales
11	€ $ £	Euro, Dollar, Livre
12	a b c … y z	Lettre français minuscules
13	A B C … Y Z	Lettre français majuscules
14	»ومماطداhdipeftelgame/mk	Les sites web publicitaires

II.8.3 Extraction des caractéristiques

La notion de N-grammes de caractères a été utilisée de manière fréquente dans l'identification de la langue ou dans l'analyse de corpus oraux. L'utilisation de profils de fréquence N-gramme, qui est une tranche de N caractères d'une chaîne de caractères, est un moyen simple et fiable de classification des documents dans un large éventail de tâches de catégorisation.

Dans les recherches récentes, cette notion est utilisée pour l'acquisition et l'extraction des connaissances dans les corpus. De nombreux travaux, tel que [23], utilisent les N-grammes de caractères comme méthode de représentation de documents d'un corpus pour la classification. L'ensemble des N-grammes de caractères est le résultat du déplacement d'une fenêtre de N cases sur le texte. Ce déplacement s'effectue par étapes, et chaque étape correspondant à un caractère. Ensuite les fréquences des N-grammes de caractères sont calculées. Ces descripteurs sont indépendants de la langue employée dans le corpus. Il n'est pas nécessaire d'utiliser des dictionnaires, ni de segmenter les documents en mots.

Les N-grammes sont tolérants aux fautes d'orthographe et aux déformations causées lors de la reconnaissance de documents (système OCR). Lorsqu'un document est reconnu à l'aide du système OCR il y a souvent une part non négligeable de bruit. Par exemple, il est possible que le mot "feuille" soit lu "teuille". Un système fondé sur les N-grammes prendra en compte les autres N-grammes comme "eui", "uil", etc.

Dans quelques travaux les N-grammes de caractères sont appliqués pour la classification de petits documents tels que les polluriels (SPAM), courriers électroniques, SMS. D'autres travaux utilisent les N-grammes pour la classification de langues complexes.

Les types de caractéristiques qui ont été proposées et utilisées dans ce travail sont N-grammes (avec N= 3, 4, 5, 6,7) comme illustré ci-dessous:

- ♣ Caractères tri -grammes (n=3),
- ♣ Caractères tétra-grammes (n=4),
- ♣ Caractères penta-grammes (n=5),
- ♣ Caractères hexa-grammes (n=6),
- ♣ Caractères hepta-grammes (n=7).

Pour utiliser ces caractéristiques, une liste de tous les mots est extraite du texte, puis les caractères n-grammes de chaque mot sont pris (figure-2.2), ainsi un profil de caractères n-grammes est créé (contenant les caractères n-grammes et leurs fréquences).

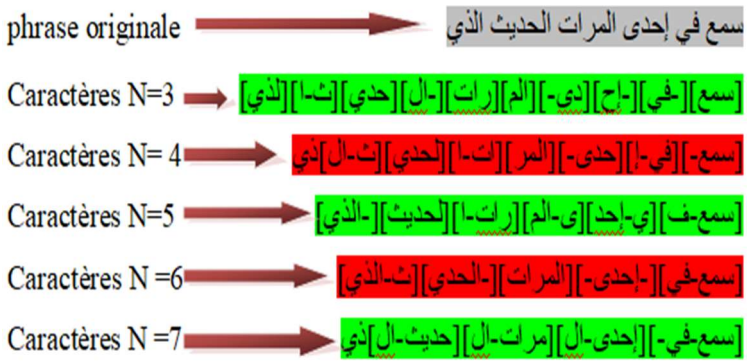


Figure-II.8 : Exemple d'extraction des caractères N-grammes d'un texte.

II.9 Conclusion

Dans ce chapitre, nous avons présenté les différentes approches suivies pour obtenir un système robuste. Les méthodologies présentées se composent de trois phases principales, la première phase étant le prétraitement approprié des textes utilisés pour désigner les auteurs; La seconde est d'extraire les caractéristiques appropriées à ce type de problème (attribution d'auteurs); La troisième étape concerne les méthodes, algorithmes et méthodes de classification sélectionnés (identification de l'auteur).

Enfin, les méthodes choisies pour l'étude changent selon les textes et leur style

Chapitre-III

Expériences et résultats

Chapitre-III

Expériences et Résultats

III .1 Introduction

Dans ce chapitre nous allons exposer les séries d'expériences d'attribution d'auteur effectuées sur notre corpus qui est composé de 23 auteurs dont chacun a écrit 6 textes. Ces textes, numéroté de 1 à 6 et qui ont été obtenus après une opération de Reconnaissance Optique de Caractères (OCR), sont classés en deux classes; textes corrigés et textes non-corrigés selon le type de prétraitement appliqué.

Ces textes ont fait l'objet d'une série d'expériences pour voir l'effet d'acquisition optique sur l'attribution d'auteurs. Par la suite, les résultats obtenus ont été examinés et discutés et des interprétations et des conclusions objectives ont été donnés.

III .2 Corpus d'évaluation

III.2.1 Description du Corpus

L'évaluation expérimentale occupe une place importante dans la classification des textes. A l'aide des corpus de tests, nous pouvons voir l'effet d'acquisition de documents textuels sur l'attribution d'auteurs. Cependant, les études en attribution d'auteur des textes obtenus après une opération OCR disposent d'un nombre relativement restreint de corpus, encore moins pour les textes corrigés, et non corrigés.

De plus, le nombre d'auteurs possibles demeure aussi limité car il s'avère difficile de trouver un nombre important de candidats potentiels respectant des contraintes multiples (même période et langue, cultures proches, thèmes similaires, et volume d'apprentissage important).

Pour cette raison nous avons décidé de construire notre propre corpus qu'on a appelé "**O**ptical **C**aractère **R**ecognition of **23C**ontemporary**A**rabe**W**riters"(**OCR23CAW**).

III.2.2 Constituants du Corpus

Le corpus que nous avons conçu contient 23 écrivains arabes contemporains (8 féminins et 15 masculins) qui sont : Assia_Djebar, Latifa_Zayyat, Ghada_Saman, Houda_ Barakat, Kolite_Sohil, Nazik_Malaika, May_Ziada, Nawel_Saadawi, Djebran_Khalil, Mahmoud_Akad, Hanna_mineh, Abdelkader El -Mazini, Edwaral_Kharat, Youssef_Idriss, Mustapha al-Manfalouti, Ibrahim_Sonallah, Mostapha Sadak_Rafie, Taha_Hocin, Haidar-Haidar, Toufik_Hakim, Mikhail_Naimy, Najib_Mahmoud et najib_mahfoud

On choisit un livre pour chaque auteur, puis on fait extraire aléatoirement un certain nombre de pages contenant le nombre de mots choisi. Pour chaque auteur on sélectionne 6 texte set on les classe en deux catégories; textes corrigés et textes non-corrigés.

Les textes utilisés pour l'opération d'apprentissage (qui sont les textes numéros; 3, 4, 5 et 6 pour chaque auteurs) sont tous de la première catégorie (textes corrigés). Cependant, chacun des textes utilisés pour l'opération de test (qui sont les textes numéros 1 et 2 pour chaque auteurs) on trouve les deux types de textes (c'est-à-dire on a texte-1-corrigé et texte-1-non-corrigé et de même pour le texte-2).

Les textes considérés ont été pris à partir des romans de ces écrivains. Les détails des informations sur les écrivains et les textes de notre corpus sont donnés dans les tableaux suivants :

Tableau-III.1 : Récapitulatif du Corpus (Ecrivains féminins)

Ecrivains	Pays de Naissance	Période	Nbre de livres	Langue	Textes	Nbre de mots	Utilisation
Assia Djebar	Algérie	1936-2015	26livres	AR/FR	Asia-1	2130	Test
					Asia-2	1956	Test
					Asia-3	2408	Apprentissage
					Asia-4	2361	Apprentissage
					Asia-5	2161	Apprentissage
					Asia-6	2510	Apprentissage
Nazik al-Mala'ika	Irak	1923-2007	25-livres	AR	Nazik-1	3211	Test
					Nazik-2	1682	Test
					Nazik-3	1035	Apprentissage
					Nazik-4	886	Apprentissage
					Nazik-5	1015	Apprentissage
					Nazik-6	840	Apprentissage
Ghada al- Saman	Syrie	1942 – à ce jour	46 livres	AR	Ghada-1	3088	Test
					Ghada-2	1825	Test
					Ghada-3	2960	Apprentissage
					Ghada-4	1697	Apprentissage
					Ghada-5	3273	Apprentissage
					Ghada-6	3151	Apprentissage
Houda_ Barakat	Liban	1952 – à ce jour	12 livres	AR	Houda-1	2098	Test
					Houda-2	1984	Test
					Houda-3	2073	Apprentissage
					Houda-4	2116	Apprentissage
					Houda-5	2069	Apprentissage
					Houda-6	1929	Apprentissage
Koulite_ Sohil	Syrie	1931 – à ce jour	29livres	AR	Koulit-1	2329	Test
					Koulit-2	1608	Test
					Koulit-3	1995	Apprentissage
					Koulit-4	1465	Apprentissage
					Koulit-5	1664	Apprentissage
					Koulit-6	1969	Apprentissage
Latifa al-zayyat	Egypte	1923-1996	12livres	AR	Latifa-1	2209	Test
					Latifa-2	1865	Test
					Latifa-3	2232	Apprentissage
					Latifa-4	1983	Apprentissage
					Latifa-5	1993	Apprentissage
					Latifa-6	1753	Apprentissage
May-ziada	Palestine	1886 – 1941	19 livres	FR-EN-ES-IT	May-1	2180	Test
					May-2	1961	Test
					May-3	2149	Apprentissage
					May-4	1946	Apprentissage
					May-5	1978	Apprentissage
					May-6	2258	Apprentissage
Nawel El Saadawi	Egypte	1931- 2021	34 livres	AR	Nawal-1	2050	Test
					Nawal-2	2007	Test
					Nawal-3	1980	Apprentissage
					Nawal-4	2027	Apprentissage
					Nawal-5	3305	Apprentissage
					Nawal-6	2604	Apprentissage

Tableau-III.2: Récapitulatif du Corpus (Ecrivains masculins)

Ecrivains	Pays de Naissance	Période	Nombre de livres	Langue	Textes	Nombre de mots / texte	Utilisation
Najib_mahfoud	Egypte	1911-2006	49 livers	AR	Najib -1	2753	test
					Najib -2	3117	test
					Najib -3	2497	Apprentissage
					Najib -4	2426	Apprentissage
					Najib -5	2638	Apprentissage
					Najib -6	2042	Apprentissage
Edwaral_ Kharat	Egypte	1926-2015	30 livers	AR	Kharat-1	3484	test
					Kharat-2	3486	test
					Kharat-3	3836	Apprentissage
					Kharat-4	3840	Apprentissage
					Kharat-5	3528	Apprentissage
					Kharat-6	3734	Apprentissage
Abdelkader El_Mazini	Egypte	1889 - 1949	19 livres	AR-EN	Mazini-1	2697	test
					Mazini-2	2019	test
					Mazini-3	2088	Apprentissage
					Mazini-4	2055	Apprentissage
					Mazini-5	2077	Apprentissage
					Mazini-6	1971	Apprentissage
Djebran -Khalil	Liban	1883-1931	163 livers	AR-EN-FR	Djebran-1	2306	test
					Djebran-2	2514	test
					Djebran-3	2457	Apprentissage
					Djebran-4	2328	Apprentissage
					Djebran-5	1875	Apprentissage
					Djebran-6	708	Apprentissage
Haidar_ Haidar	Syrie	1936- à ce jour	40 livers	AR	Haider-1	1883	test
					Haider-2	2279	test
					Haider-3	1986	Apprentissage
					Haider-4	2185	Apprentissage
					Haider-5	2412	Apprentissage
					Haider-6	2098	Apprentissage
Hanna_mineh	Syrie	1924-2018	21 livers	AR	Mina-1	1766	test
					Mina-2	1882	test
					Mina-3	1649	Apprentissage
					Mina-4	1577	Apprentissage
					Mina-5	1945	Apprentissage
					Mina-6	1681	Apprentissage
Ibrahim _Sonallah	Egypte	1937- à ce jour	47 livers	AR	Ibrahim-1	3285	test
					Ibrahim-2	3207	test
					Ibrahim-3	3294	Apprentissage
					Ibrahim-4	3628	Apprentissage
					Ibrahim-5	3503	Apprentissage
					Ibrahim-6	3370	Apprentissage
Abbas Mahmoud El Akad	Egypte	1889 - 1964	689 livers	AR	Akad-1	3294	test
					Akad-1	1965	test
					Akad-1	2305	Apprentissage
					Akad-1	1986	Apprentissage
					Akad-1	2535	Apprentissage
					Akad-1	2064	Apprentissage

Tableau-III.2: Récapitulatif du Corpus (Ecrivains masculins) (Suite)

Ecrivains	Pays de Naissance	Période	Nombre de livres	Langue	Textes	Nombre de mots / texte	Utilisation
Mikhail _Naimy	Liban	1889-1988	76 livers	AR	Mikhail-1	2212	test
					Mikhail-2	2238	test
					Mikhail-3	1685	Apprentissage
					Mikhail-4	2194	Apprentissage
					Mikhail-5	2121	Apprentissage
					Mikhail-6	2007	Apprentissage
Nadjib Mahmoud	Egypte	1905-1993	110	AR	Nadjib-1	1960	test
					Nadjib-2	1962	test
					Nadjib-3	1934	Apprentissage
					Nadjib-4	1912	Apprentissage
					Nadjib-5	1844	Apprentissage
					Nadjib-6	1868	Apprentissage
Sadek al-Rafei	Egypte	1880-1937	167	AR	Rafei-1	3261	test
					Rafei-2	1932	test
					Rafei-3	2847	Apprentissage
					Rafei-4	2007	Apprentissage
					Rafei-5	1938	Apprentissage
					Rafei-6	2563	Apprentissage
Taha Hussein	Egypte	1889-1973	381	AR-FR-LAT	Taha-1	2888	test
					Taha-2	1761	test
					Taha-3	1983	Apprentissage
					Taha-4	1884	Apprentissage
					Taha-5	2089	Apprentissage
					Taha-6	2145	Apprentissage
Toufik _Hakim	Egypte	1898-1987	232	AR	Toufik-1	1608	test
					Toufik-2	1578	test
					Toufik-3	3095	Apprentissage
					Toufik-4	1673	Apprentissage
					Toufik-5	1627	Apprentissage
					Toufik-6	1633	Apprentissage
Mustapha al-Manfalouti	Egypte	1876-1924	107	FR	lotfi-1	2691	test
					lotfi-2	2384	test
					lotfi-3	2779	Apprentissage
					lotfi-4	2809	Apprentissage
					lotfi-5	1921	Apprentissage
					lotfi-6	2672	Apprentissage
Youssef Idriss	Egypte	1927-1991	108	AR	Idriss-1	3630	test
					Idriss-2	3306	test
					Idriss-3	3724	Apprentissage
					Idriss-4	3439	Apprentissage
					Idriss-5	2737	Apprentissage
					Idriss-6	3599	Apprentissage

III.2.3 Préparation des documents du corpus

Les documents du corpus doivent être préparés avant leur utilisation pour l'attribution de leurs véritables auteurs. La phase de préparation se résume en opérations pour préparer ce texte :

- ✓ Scanner les pages choisis et les enregistrées en format (.jpeg)
- ✓ Convertir ces images en fichier Word (.txt) à l'aide d'un OCR.
- ✓ Faire les opérations de prétraitement mentionnées dans chapitre précédent.
- ✓ Les documents textes obtenus sont enregistrés sous forme UTF-8 (Encodage basé sur l'Unicode qui peut être codé sur 4 octets).

En général, on a utilisé l'encodage UTF-8 pour encoder tous les textes du corpus, car ce dernier couvre un vaste nombre de caractères, et qui est implicitement capable d'encoder la majorité des langues vu qu'il est encodé sur 4 octets. En revanche, l'utilisation de cet encodage est payée en termes de temps de calcul et en termes de mémoire.

Par la suite, le corpus est divisé en deux sous-ensembles selon la règle (2/3 et 1/3) appliquée dans les bases de données;

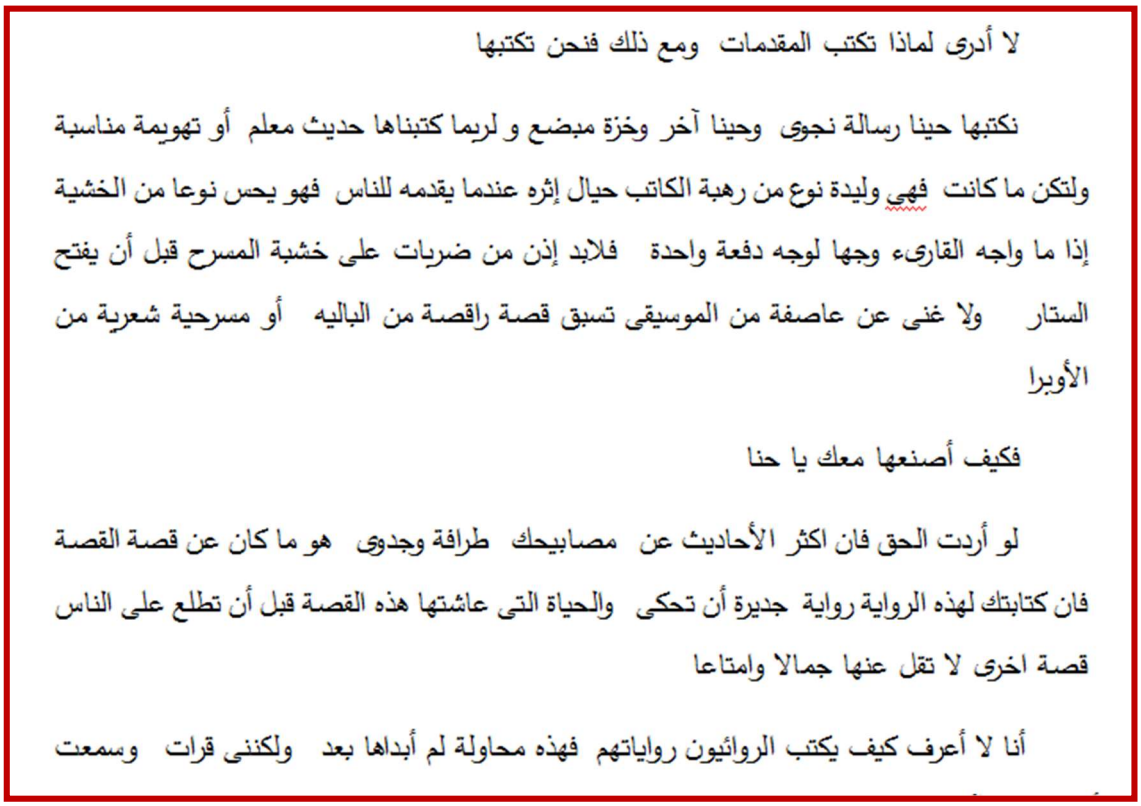
- ❖ Ensemble d'apprentissage constitué des textes (numéros 3, 4, 5 et 6 pour chaque auteur) de chaque catégorie (corrigés et non corrigés).
- ❖ Ensemble de test constitué des textes (numéros 1 et 2 pour chaque auteur) de chaque catégorie (corrigés et non corrigés).

Au totale, le corpus contient 276 textes divisés comme suit; 184 textes pour l'apprentissage (corrigés, non-corrigés) et 92 textes pour le test (corrigés, non-corrigés).

La correction est faite manuellement afin d'obtenir le même texte original

III.2.4 Exemples de textes Word obtenus après une opération OCR

Après le processus de scandes documents en PDF, les résultats obtenus sont considérés comme des documents modifiables (format Word) pour corriger les erreurs, ajouter ou supprimer tout ce qui est supplémentaire. Ci-dessous nous passons en revue des exemples de textes que nous avons obtenus après l'opération OCR afin de les utiliser dans les expériences



لا أدرى لماذا تكتب المقدمات ومع ذلك فنحن تكتبها

نكتبها حينا رسالة نجوى وحينا آخر وخزة مبضع و لربما كتبناها حديث معلم أو تهويمة مناسبة ولتكن ما كانت فهي وليدة نوع من رهبة الكاتب حيال إثره عندما يقدمه للناس فهو يحس نوعا من الخشية إذا ما واجه القارىء وجها لوجه دفعة واحدة فلابد إذن من ضربات على خشبة المسرح قبل أن يفتح الستار ولا غنى عن عاصفة من الموسيقى تسبق قصة راقصة من الباليه أو مسرحية شعرية من الأوبرا

فكيف أصنعها معك يا حنا

لو أردت الحق فان اكثر الأحاديث عن مصابيحك طرافة وجدوى هو ما كان عن قصة القصة فان كتابتك لهذه الرواية رواية جديرة أن تحكى والحياة التى عاشتها هذه القصة قبل أن تطلع على الناس قصة اخرى لا تقل عنها جمالا وامتاعا

أنا لا أعرف كيف يكتب الروائيون رواياتهم فهذه محاولة لم أبداها بعد ولكننى قرات وسمعت

Figure-III.1 : Exemple de texte corrigé.

لا أدرى لماذا تكتب المقدمات . ومع ذلك فنحن تكتبها ٠

نكتبها حينا رسالة نجوى ، وحينا ٢ خر وخزة مبضع . ولربما كتبناها حديث معلم ، أو تهويمة مناسبة ٠ ولتكن ما كانت ٠ فهي وليدة نوع من رهبة الكاتب حيال اثر ه عندما يقدمه للناس. فهو يحس نوعا من الخشية اذا ما واجه القارىء وجها لوجه دفعة واحدة . فلابد اذن من ضربات على خشبة المسرح قبل أن يفتح الستار ٠ ولا غنى عن عاصفة من الموسيقى تسبق قصة راقصةمن الباليه ٤ أو مسرحية شعرية من 1لأوبر ١ ٠

فكيف أصنعها معك يا حنا . ٠

لو أردت الحق فان ١كثر الإحاديث عن « مصابيحك » طرافة وجدوى ، هو ما كان عن قصة القصة ٠ فان كتابتك لهذه الرواية رواية جديرة أن تحكى . والحياة التى عاشتها هذه القصة قبل أن تطلع على الناس قصة اخرى لا تقل عنها جمالا وامتاعا .

٧/١٧١٧١٧٠111138.00111ؤ3 * 10٥31٥12Ñ

أنا لا أعرف كيف يكتب الروائيون رواياتهم . نهذهمحاولةلم ابدأها بعد . ولكننى قرات ، وسمعت ،

Figure-III.2: Exemple de texte non-corrigé

III.3 Expérimentation et résultats obtenus

III.3.1 Protocole expérimental

Dans ce mémoire, la tâche d'attribution d'auteurs est effectuée en utilisant le N-grammes caractère comme caractéristique et deux méthodes de classification : Burrows-Delta et Linear-SVM. Ces techniques ont été utilisées pour évaluer la robustesse de notre système d'attribution des auteurs des documents textes obtenus par une opération OCR. Le Taux d'Attribution d'Auteurs (TAA) est défini par la relation suivante :

$$TAA = \frac{nombredocumentscorrectementattribués}{nombredocumenttesté} \times 100$$

Ce travail expérimental est organisé en quatre séries d'expériences et chaque série comporte plusieurs cas d'applications selon la valeur de N et le sexe d'écrivains.

- Dans la première série, les textes utilisés dans la phase d'apprentissage et les textes utilisés dans la phase de test sont tous les deux des textes corrigés.
- Dans la deuxième série, les textes utilisés dans la phase d'apprentissage sont des textes corrigés et les textes utilisés dans la phase de test sont des textes non-corrigés.
- Dans la troisième série, les textes utilisés dans la phase d'apprentissage sont des textes non-corrigés et les textes utilisés dans la phase de test sont des textes corrigés.
- Dans la quatrième série, les textes utilisés dans la phase d'apprentissage sont des textes non-corrigés et les textes utilisés dans la phase de test sont des textes non-corrigés.

III.3.2 Expériences d'attribution d'auteurs

III.3.2.1 Expérience N°1 : Attribution d'auteurs avec textes d'apprentissage et test corrigés

Cette série d'expériences vise à déterminer le nombre approprié de caractères (N-gramme) pour avoir le meilleur taux TAA, Nous utilisons les textes d'écrivains femmes et hommes pour les deux phases (apprentissage et test). Après le recherche, nous avons choisis d'utiliser la méthode d'analyse des Evénements les Plus Courants « MCE = 300 ». Les résultats obtenus de cette série d'expérience sont présentés dans les tableaux et les figures qui suivent :

Tableau-III.3 : Taux d'attribution d'auteurs pour les textes apprentissage corrigé et test corrigé avec classifieur Linear-SVM

	N=3	N=4	N=5	N=6	N=7
Mâle	94	90	90	90	84
Femelle	82	82	82	94	88
Mâle /femelle	77	87	85	85	85

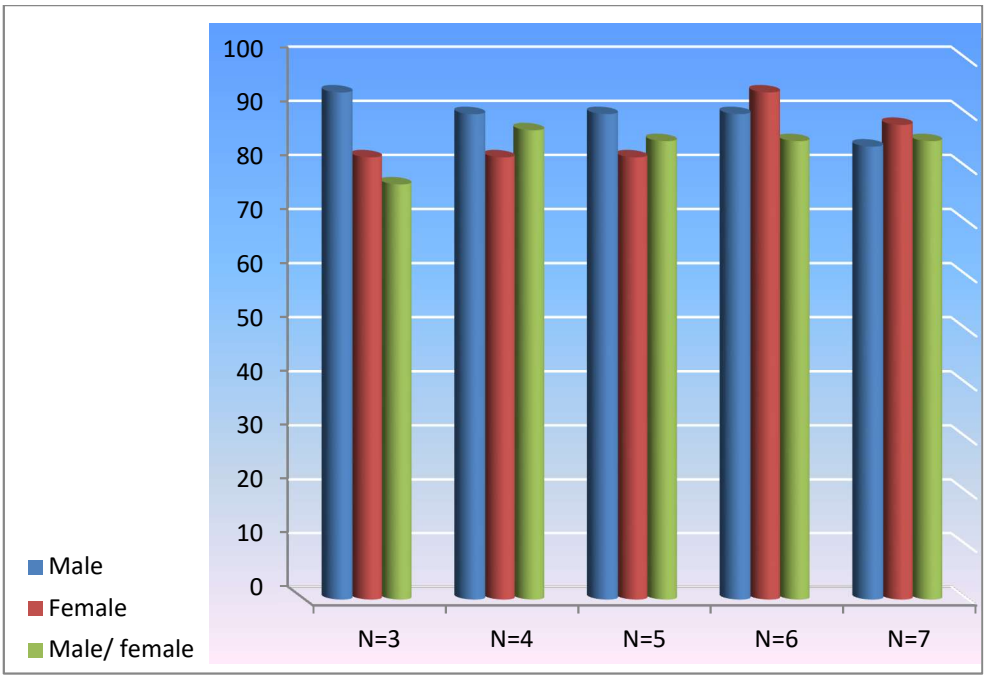


Figure-III.3 : Taux d'attribution d'auteurs pour classifieur Linear-SVM

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 84-94% pour les écrivains mâles, 82-94% pour les écrivains femelle et 77-87% pour les deux (mâle /femelle). Pour les résultats des auteurs mâles on a la meilleure valeur de TAA =94% correspond à N=3. Pour les résultats des auteurs femelles on a la meilleure valeur de TAA =94% correspond à N=6. Pour les résultats mâle / femelle on a la meilleure valeur de TAA =87% correspond à N=4.

Tableau-III.4 : Taux d'attribution d'auteurs pour les textes apprentissage corrigé et test corrigé avec classifieur Burrows-Delta

	N=3	N=4	N=5	N=6	N=7
Mâle	90	87	84	80	80
Femelle	82	88	88	88	82
Mâle / femelle	87	83	79	79	80

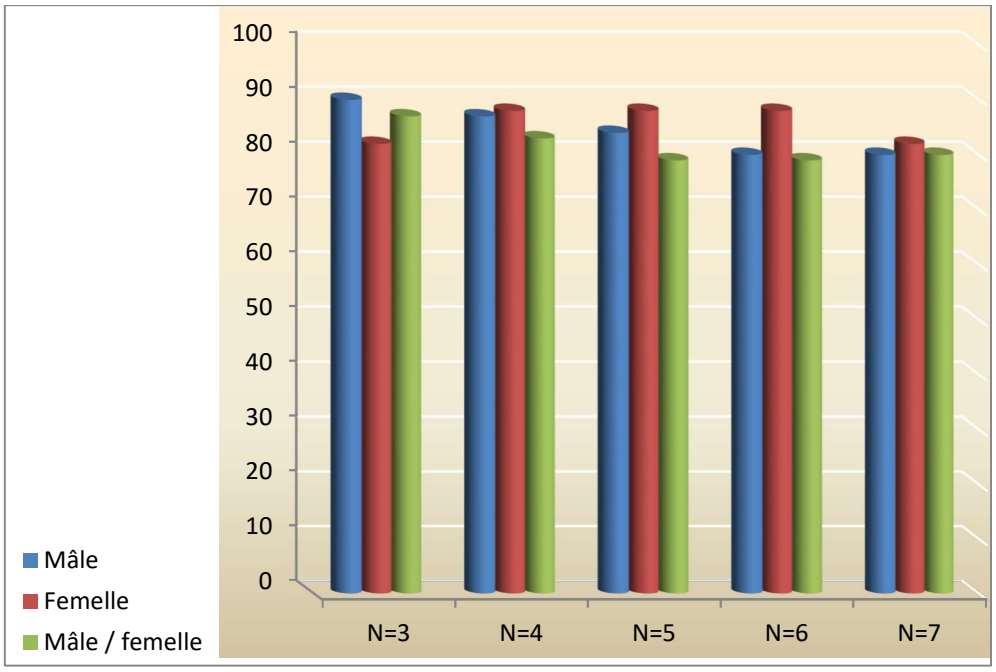


Figure-III.4 : Taux d'attribution d'auteurs pour le classifieur Burrows-Delta

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 80-90% pour mâle, 82-88% pour femelle et 79-87% pour les deux (mâle/femelle). Pour les résultats mâles on a le meilleur taux de TAA = 90% correspond à N=3. Pour les résultats des auteurs femelles on a la meilleure valeur de TAA =88% correspond à N= [4, 5, 6]. Pour les résultats des auteurs mâles/ femelle on a la meilleure valeur de TAA =87% correspond à N=3.

On constate que le meilleur classifieur pour TAA du les textes apprentissage corrigé et test corrigé est Linear-SVM pour les trois cas (mâle, femelle, mâle / femelle).

III.3.2.2 Expérience N°2 : Attribution d'auteurs avec textes d'apprentissage corrigé et textes de test non-corrigé

Cette série d'expériences vise à déterminer le nombre approprié de caractères (N-gramme) pour avoir le meilleur taux TAA, Nous utilisons les textes d'écrivains femmes et hommes pour les deux phases (apprentissage et test). Après le recherche, nous avons choisis d'utiliser la méthode d'analyse des Evénements les Plus Courants « MCE = 300 ». Les résultats obtenus de cette série d'expérience sont présentés dans les tableaux et les figures qui suivent :

Tableau-III.5 : Taux d'attribution d'auteurs pour les textes apprentissage corrigé et test non-corrigé avec classifieur Linear-SVM

	N=3	N=4	N=5	N=6	N=7
Mâle	77	87	80	77	74
Femelle	70	75	75	88	75
Mâle / femelle	72	72	72	72	72

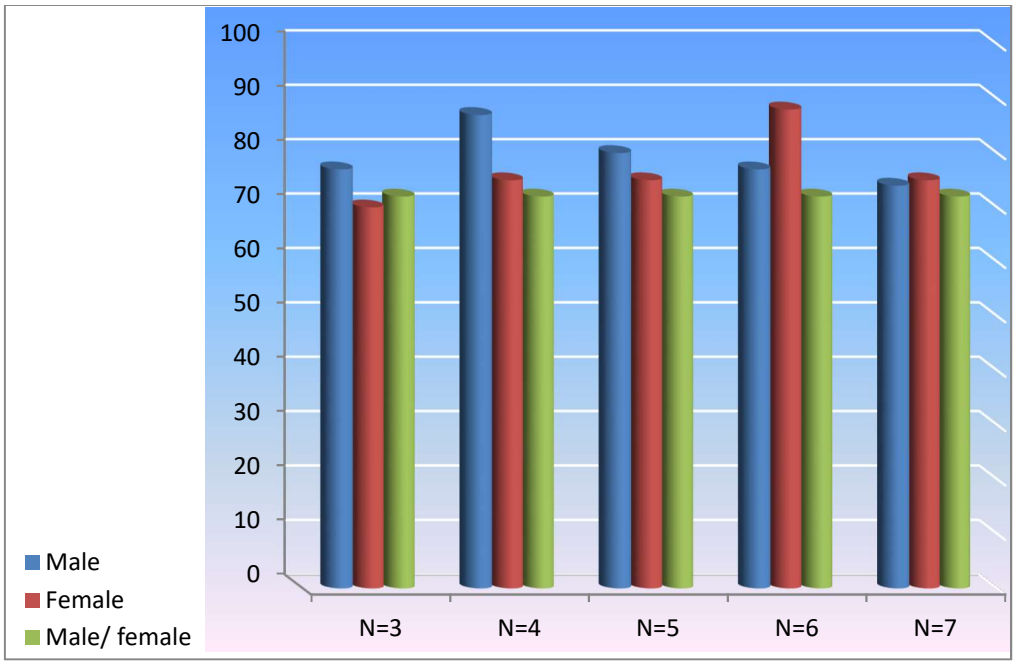


Figure-III.5 : Taux d'attribution d'auteurs pour classifieur Linear-SVM

D'après les résultats obtenus, on peut voir clairement quelle taux TAA de cette expérience est entre 74-87% pour mâle ,70-88% pour Femelle et 72% pour les deux (Mâle

/femelle). Pour les résultats des auteurs mâles on a la meilleure valeur de TAA =87% correspond à N=4

Pour les résultats des auteurs femelles on a la meilleure valeur de TAA =88% correspond à N=6.Pour les résultats mâle / femelle on a la meilleure valeur de TAA =72% correspond à tous les N

Tableau-III.6: Taux d'attribution d'auteurs pour les textes apprentissage corrigé et test non-corrigé avec classifieur Burrows-Delta

	N=3	**N=4**	**N=5**	**N=6**	**N=7**
Mâle	70	77	80	80	77
Femelle	49	50	75	69	82
Mâle / femelle	85	83	78	76	69

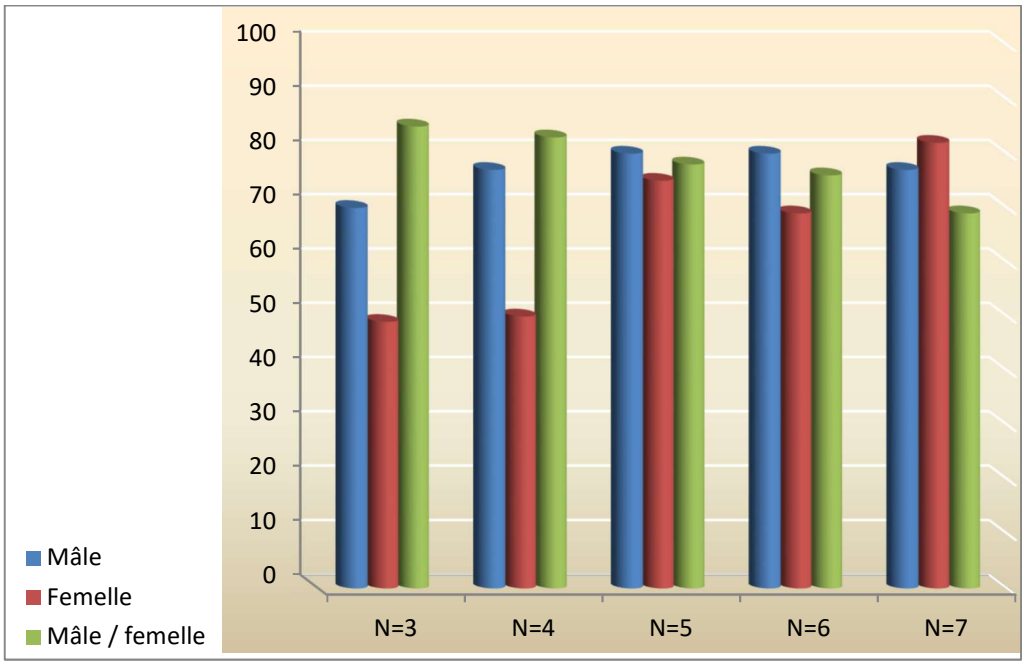


Figure-III.6 : Taux d'attribution d'auteurs pour classifieur Burrows-Delta

D'après les résultats obtenus, on peut voir clairement quelle taux TAA de cette expérience est entre 70-80% pour mâle,49-82% pour femelle et 69-85% pour les deux (mâle /femelle). Pour les résultats mâle on a le meilleure valeur de TAA =80% correspond à N=[5,6].Pour les résultats femelle on a le meilleure valeur de TAA =82% correspond à N= 7.Pour les résultats mâle/ femelle on a le meilleure valeur de TAA =85% correspond à N=3.

On constate que le meilleur classifieur pour les textes apprentissage corrigé et test non-corrigé est Linear-SVM pour les cas (male, femelle) et Burrows-Delta pour le cas (male/femelle). Ceci peut être expliqué par la présence des erreurs de conversion dans le texte résultant d'une opération OCR qui a subi un prétraitement non complet (les textes non-corrigé)

III.3.2.3 Expérience N°3 : Attribution d'auteurs avec textes apprentissage non-corrigé et test corrigé

Cette série d'expériences vise à déterminer le nombre approprié de caractères (N-gramme) pour avoir le meilleur taux TAA, Nous utilisons les textes d'écrivains femmes et hommes pour les deux phases (apprentissage et test). Après le recherche, nous avons choisis d'utiliser la méthode d'analyse des Evénements les Plus Courants « MCE = 300 ». Les résultats obtenus de cette série d'expérience sont présentés dans les tableaux et les figures qui suivent :

Tableau-III.7 : Taux d'attribution d'auteurs pour les textes apprentissage non-corrigé et test corrigé avec classifieur Linear-SVM

	N=3	**N=4**	**N=5**	**N=6**	**N=7**
Male	60	60	64	67	74
Femelle	58	58	75	75	75
Male/ femelle	58	60	64	70	70

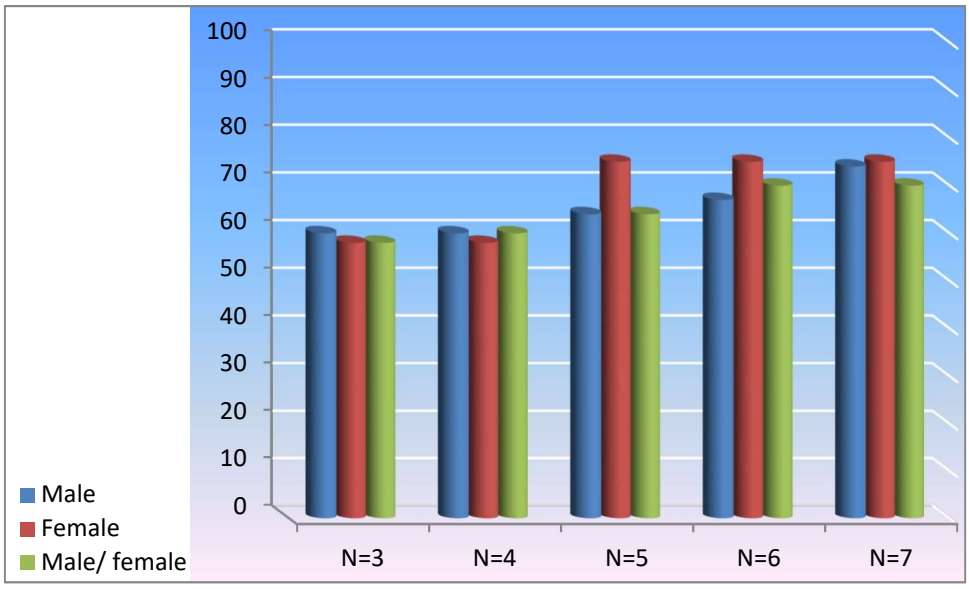


Figure-III.7 : Taux d'attribution d'auteurs pour Classifieur Linear-SVM

D'après les résultats obtenus, on peut voir clairement quelle taux TAA de cette expérience est entre 60-74% pour mâle, 58-75% pour femelle et 58-70% pour les deux (male/femelle). Pour les résultats des auteurs mâles on a la meilleure valeur de TAA =74% correspond à N=7.Pour les résultats des auteurs femelles on a la meilleure valeur de TAA =75% correspond à N=[5,6,7]. Pour les résultats mâle/ femelle on a la meilleure valeur de TAA =70% correspond à N=[6,7]

Tableau-III.8 : Taux d'attribution d'auteurs pour les textes apprentissage non-corrigé et test corrigé avec classifieur Burrows Delta

	N=3	**N=4**	**N=5**	**N=6**	**N=7**
Male	70	80	77	74	70
Femelle	63	82	82	82	75
Male/ femelle	70	83	79	74	74

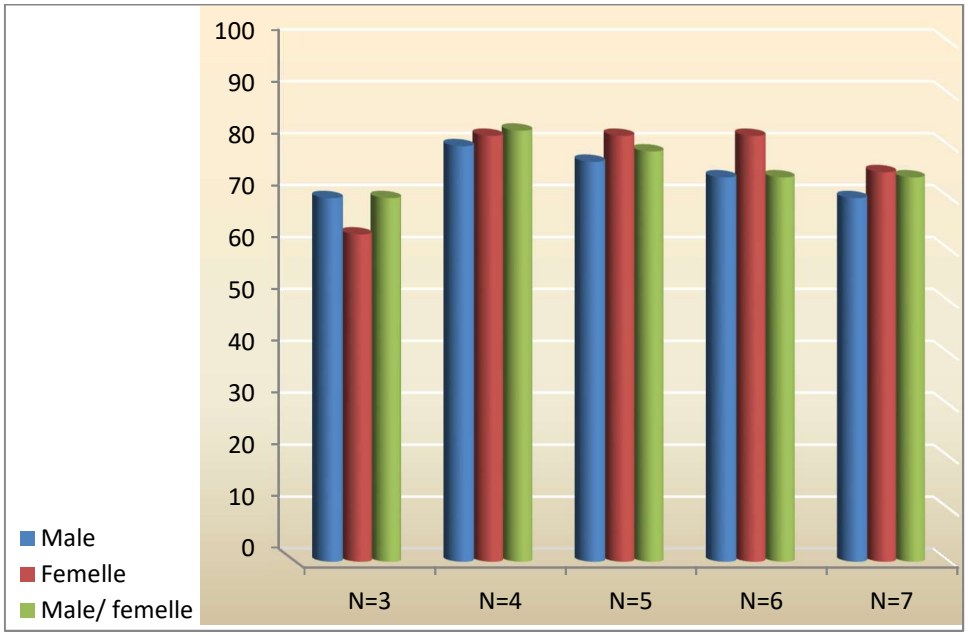


Figure-III.8 : Taux d'attribution d'auteurs pour classifieur Burrows Delta

D'après les résultats obtenus, on peut voir clairement quelle taux TAA de cette expérience est entre 70-80% pour mâle,63-82% pour femelle et 70-83% pour les deux (mâle /femelle).Pour les résultats mâle on a le meilleure valeur de TAA =80% correspond à N=4.Pour les résultats femelle on a le meilleure valeur de TAA =82% correspond à N= [4,5,6].Pour les résultats mâle / femelle on a le meilleure valeur de TAA =83% correspond à N=4

On constat que le meilleur classifieur pour les textes apprentissage non-corrigé et test corrigé est Burrows Delta pour les trois cas (mâle, femelle, mâle /femelle). Ceci peut être expliqué par la présence des erreurs de conversion dans le texte résultant d'une opération OCR qui a subi un prétraitement non complet (les textes non-corrigé)

III.3.2.4 Expérience N°4 : Attribution d'auteurs avec textes apprentissage non-corrigé et test non-corrigé

Cette série d'expériences vise à déterminer le nombre approprié de caractères (N-gramme) pour avoir le meilleur taux TAA, Nous utilisons les textes d'écrivains femmes et hommes pour les deux phases (apprentissage et test). Après le recherche, nous avons choisis d'utiliser la méthode d'analyse des Evénements les Plus Courants « MCE = 300 ». Les résultats obtenus de cette série d'expérience sont présentés dans les tableaux et les figures qui suivent :

Tableau-III.9 : Taux d'attribution d'auteurs pour les textes apprentissage non- corrigé et test non-corrigé avec Classifieur Linear-SVM

	N=3	N=4	N=5	N=6	N=7
Mâle	94	94	94	94	94
Femelle	75	75	75	75	75
Mâle/ femelle	87	90	87	87	87

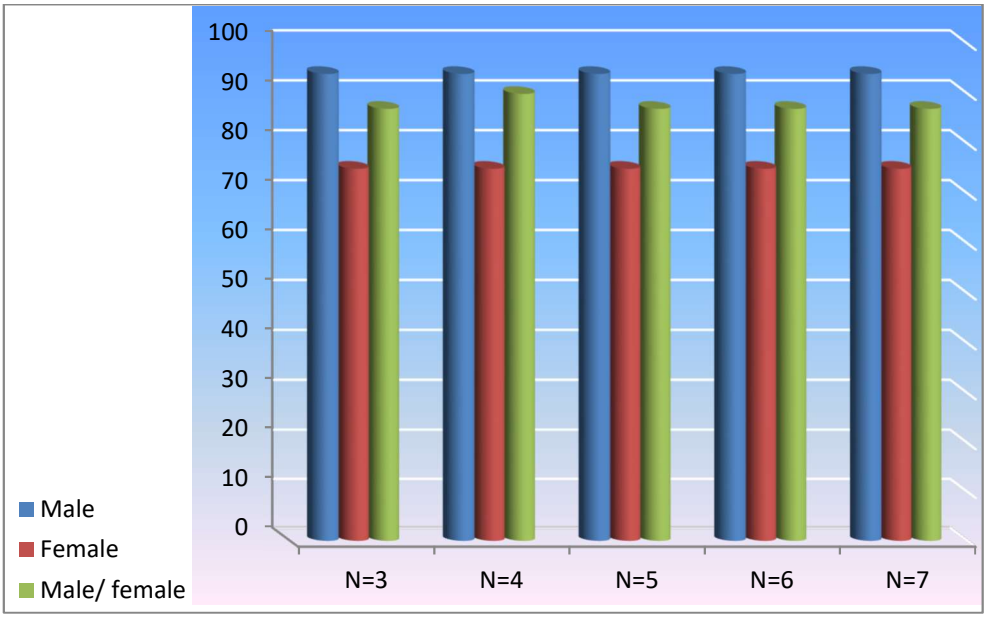


Figure-III.9 : Taux d'attribution d'auteurs pour Classifieur Linear-SVM

D'après les résultats obtenus, on peut voir clairement quelle taux TAA de cette expérience est entre94% pour mâle,75% pour femelle et 87-90% pour les deux (mâle /femelle). Pour les résultats des auteurs mâles on a la meilleure valeur de TAA =94% correspond à tous les N. Pour les résultats des auteurs femelles on a la meilleure valeur de TAA =75% correspond à tous les N. Pour les résultats mâle/ femelle on a la meilleure valeur de TAA =90% correspond à N=4.

Tableau-III.10: Taux d'attribution d'auteurs pour les textes apprentissage non- corrigé et test non-corrigé avec classifieur Burrows Delta

	N=3	**N=4**	**N=5**	**N=6**	**N=7**
male	90	90	90	90	90
female	75	75	75	75	75
Male/ female	87	85	80	83	83

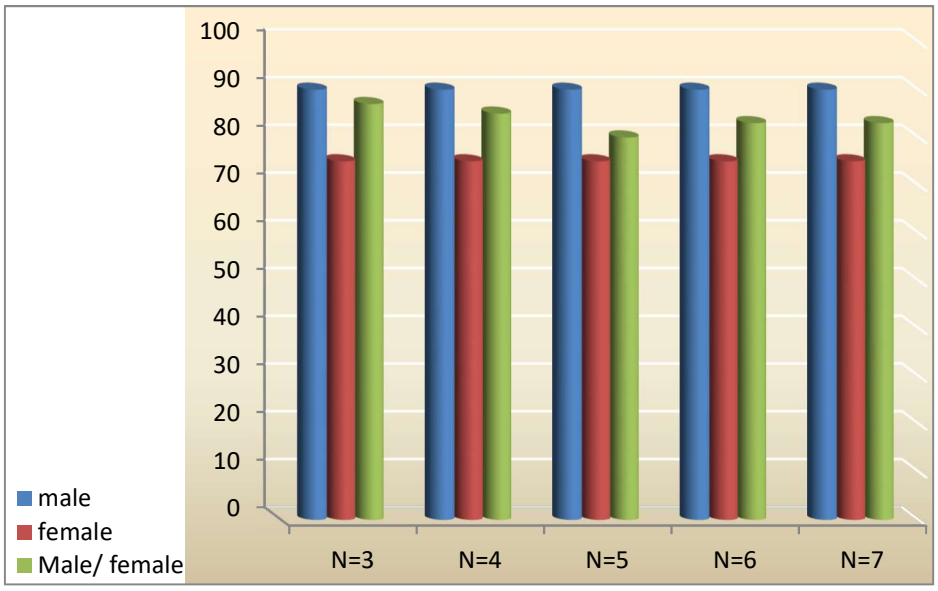


Figure-III.10 : Taux d'attribution d'auteurs pour Classifieur Burrows Delta

D'après les résultats obtenus, on peut voir clairement quelle taux TAA de cette expérience est entre 90% pour Male ,75% pour Femelle et 80-87% pour les deux (male/femelle). Pour les résultats mâles on a la meilleure valeur de TAA =90% correspond à tous les N. Pour les résultats des auteurs femelles on a la meilleure valeur de TAA =75% correspond à tous les N. Pour les résultats des auteurs mâles / femelles on a la meilleure valeur de TAA =87% correspond à N=3.

On constate que le meilleur classifieur pour les textes apprentissage non- corrigé et test non- corrigé est Linear-SVM pour les trois cas (mâle, femelle, mâle/ femelle)

III.4 Robustesse de notre système

D'après les séries d'expériences, nous avons obtenus les résultats des meilleurs TAA représenté dans les tableaux suivants :

Tableau-III.11: Meilleur Taux d'attribution d'auteurs avec classifieur Linear-SVM

Type de textes	Femelle	Mâle	Mâle/Fem
TC TC	94	94	87
TC TNC	88	87	72

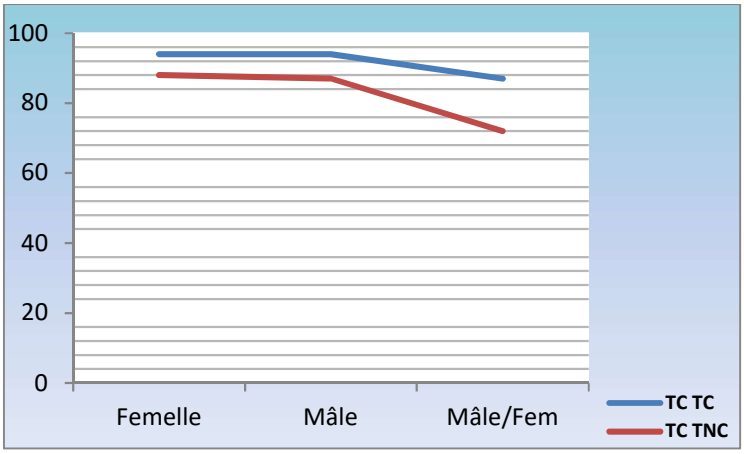


Figure-III.11 : TAA pour le classifieur Linear-SVM

Tableau-III.12: Meilleur Taux d'attribution d'auteurs avec classifieur Burrows Delta

Type de textes	Femelle	Mâle	Mâle/Fem
TC-TC	88	90	87
TC TNC	82	80	85

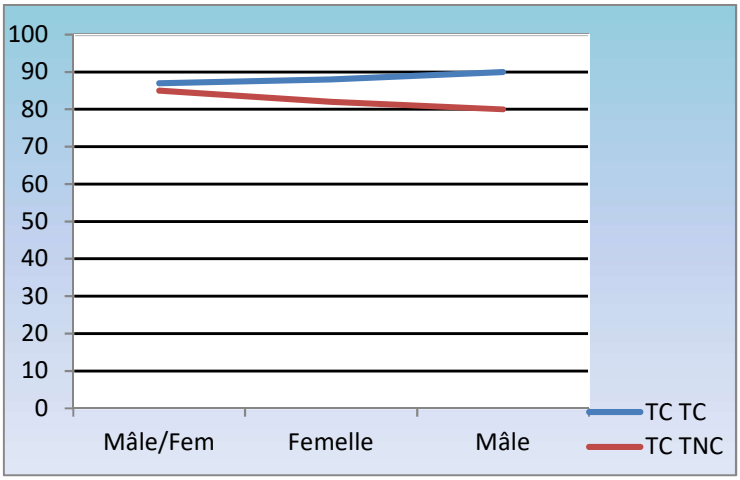


Figure-III.12 : TAA pour le classifieur Burrows Delta

D’après les résultats obtenus, on constate une légère différence entre les résultats des TAA pour les expériences de test en utilisant les différents types de textes (TC et TNC) avec les deux classifieurs (Linear-SVM, Burrows Delta). Ceci, nous donne une idée très claire sur la robustesse contre le bruit (caractères spéciales).

III.5 Conclusion

Dans ce chapitre nous avons effectué des expériences d’attribution d’auteur des documents textes arabes obtenus après une opération OCR avec plusieurs degrés d’erreurs. L’évaluation expérimentale a été réalisé en utilisant une base de données qu’on conçut pour cette fin.

Le sexe d’écrivains de texte et le caractère N Gram influent considérablement sur le taux d’attribution TAA. Le changement de type de textes apprentissage et test influence sur TAA. Dans ce chapitre nous avons effectuaient des expériences d’attribution d’auteur des documents textes arabes

Conclusion Générale

Conclusion générale

❖ **Travail réalisé**

Le sujet traité dans ce travail est lié à l'évaluation de la robustesse de l'attribution dans la reconnaissance des auteurs, la base de données que nous avons conçue pour mener nos expériences est constituée de 23 auteurs arabes, pour chaque auteur nous avons pris 6 textes qui sont convertis en utilisant le processus OCR. L'objectif visait était d'étudier le style des auteurs afin de trouver le véritable auteur, et d'appliquer des propriétés telles que les N-grammes et les classifieur (Linear-SVM et Burrows-Delta).

Nous avons classé les auteurs par sexe pour étudier leur style d'écriture, ensuite nous avons traités leurs textes (corrigés et non corrigés des erreurs d'OCR) en utilisant les deux classificateurs (Linear-SVM et Burrows-Delta) afin d'évaluer la robustesse de notre système, en utilisant les descripteurs de caractères N-grammes (N=3, 4, 5, 6, 7) comme caractéristique pour compléter le processus d'attribution d'auteurs.

❖ **Résultats obtenus**

Les résultats obtenus étaient très encourageants compte tenu des limitations liées à la taille des textes sélectionnés, et au nombre limité des ouvrages utilisés. Nous avons constaté que les valeurs des TAA pour les auteurs féminins sont plus robustes que celles obtenus pour les auteurs masculins. Le meilleur Taux d'Attribution d'Auteurs masculins et féminins (94%) est obtenu par le classifieur Linear-SVM en utilisant les textes corrigés. Ce taux diminue légèrement en utilisant les textes non-corrigés, d'où on peut conclure la robustesse de notre système. On constate, aussi, que les taux d'attributions des auteurs féminins sont légèrement plus robustes contre le bruit (textes non-corrigés) que les taux d'attribution des auteurs masculins (i.e. TAA = 94% en utilisant les TC, TAA= 88% en utilisant les TNC).

❖ **Perspectives suggérées**

Dans le cadre de développement futur de ce travail, on suggère en perspectives de compléter le travail avec les taches suivantes :

- Utilisation des textes sans correction avec le (Deep Learning) après avoir élargir la base de données.
- Tester l'attribution d'auteur en utilisant les textes de petite taille (nombre de mots limité).

Références Bibliographiques et Webographiques

Références Bibliographiques et Webographiques

Références bibliographique :

[1] Florence RODHAIN, Maître de Conférences Université Montpellier 2 CEROM Montpellier-Management 23000 Avenue des Moulins, 3185 Montpellier

[2] Marti Hearst SIMS,UC Berkeley hearst@sims.berkeley.edu October 17, 2003.

[4] Hearst, M. (2003). What is text mining. SIMS, UC Berkeley, Vol. 5

[5] T. M. Mitchell. Machine Learning. WCB/McGraw-Hill, 1997.

[6] P. Larranaga, D. Atienza, J. Diaz-Rozo, A. Ogbechie, C. E. PuertoSantana, and C. Bielza. Industrial Applications of Machine Learning. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, Boca Raton, FL, 2018.

[7] Dumais, S., Platt, J., Sahami, M. & Heckerman, D. (1998) 'Inductive learning algorithms and representations for text categorization', ACM , pp.148 – 155

[8] Feldman, R. & Sanger, J., (2007) the Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, NY: Cambridge University Press

[9] Sebastiani, F. (2002) 'Machine learning in automated text categorization' ACM Publication: ACM Computing Surveys. Vol. 3(1) PP.1-47

[10] Applied Mathematical Sciences, Vol. 6, 2012, no. 81, 4033 – 4046

[11] Duwairi, R. (2007) 'Arabic Text Categorization', International Arab Journal of Information Technology, Vol.4, No.2.

[12] Williams, C. (2007) 'Research Methods', Journal of Business & Economic Research, Vol. 5, No. 3 65

[13] Kothari, C. R. (2009) Research Methodology : Methods And Techniques, New Age Publication

[14] MENASRI, R., & YAKOUBI, M. (2020). Etude et analyse des effets d'acquisition optique à l'aide d'un OCR des textes arabes sur l'attribution d'auteurs (Doctoral dissertation, Univ M'sila).

[15] F. Xia_, C. Lewis et W. D. Lewis, The Problems of Language Identification within Hugely Multilingual Data Sets, Proceedings of LREC, Malta, May 17-23, 2010.

[16] T. Baldwin et M. Lui, Language Identification: The Long and the Short of the Matter, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, Los Angeles, California, June 2010, pp. 229–237

[17] E. Stamatatos, A Survey of Modern Authorship Attribution Methods, Journal of the AmericaSociety for Information Science and Technology, Vol. 60, Issue 3, March, 2009, pp. 538-556.

[18] F. Mosteller et D.L. Wallace, Inference in an Authorship Problem: A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed Federalist Papers, Journal of the American Statistical Association, Vol. 58, No. 302, pp. 275-309, 1963.

[19] BENYAHIA, A. (2019). Etude et analyse sur les performances des techniques d'identification d'auteurs à partir des documents écrits et des documents transcrits (Doctoral dissertation, UNIVERSITE MOHAMED BOUDIAF-M'SILA).

[20] Mohamadally Hasan Fomani Boris BD Web, ISTY3 Versailles St Quentin, France 16 janvier 2006

[22] Abdul-Razzaq, A. A., & Mustafa, T. K. (2014). Burrows-Delta method fitness for Arabic text authorship Stylometric detection. IJCSMC, 3, 69-78.

[23] W. B. Cavnar et J. M. Trenkle, n-gram based text categorization, Proceedings of SDAIR'94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, 1994, pp. 161-175.

Références webographique :

[3] https://touriaelouahabi.wordpress.com/text-mining/definition-du-text-mining/

[21] http://dspace.univ-msila.dz:8080//xmlui/handle/123456789/21045

ملخص

إن البحث في مجال إسناد المؤلف مشكل قديم جدا، وذلك لكثرة النصوص المجهولة والتطور التكنولوجي لمختلف وسائل الاتصال الرقمي الذي سهل عملية انتحال وسرقة الأعمال النصية من المؤلفين الأصليين. وقد نتج عن ذلك صعوبة بالغة في البحث واستخراج المعلومات من هذه الأعمال للتعرف على كتابها الأصليين.

في هذا العمل قمنا بدراسة أسلوب مجموعة من الأدباء المعاصرين العرب من خلال بعض مؤلفاتهم بغرض إنشاء نماذج لهم لإسناد نصوص أدبية مجهولة لأصحابها الأصليين. لأجل ذلك حاولنا أن ننشأ نظاما فعالا، حيث قمنا بتكوين قاعدة بيانات من النصوص المتحصل عليها باستعمال برنامج التعرف الضوئي على الحروف (OCR)، كما اقترحنا مصنفين (Linear-SVM, Burrows-Delta) واستخدمنا خاصية (N-gramme).

الكلمات المفتاحية: التعرف على الكاتب، إسناد المؤلف، اللغة العربية، التعرف الضوئي على الحروف.

Résumé

La recherche dans le domaine de l'attribution d'auteur est un problème très ancien, en raison du grand nombre de textes inconnus et du développement technologique de divers moyens de communication numérique, qui ont facilité le processus de plagiat et de vol des œuvres textuelles des auteurs originaux. Cela a entraîné une grande difficulté à rechercher et à extraire des informations de ces œuvres pour identifier leurs auteurs originaux.

Dans ce travail, nous étudions le style d'un groupe d'écrivains arabes contemporains à travers certaines de leurs œuvres afin de créer des modèles pour qu'ils attribuent des textes littéraires inconnus à leurs propriétaires d'origine. Pour cela, nous avons essayé de créer un système efficace, où nous avons créé une base de données de textes obtenus à l'aide du programme OCR, et suggéré deux classificateurs (Linéaire-SVM, Burrows-Delta) et utilisé la fonction (N-gramme).

Mots clés : Reconnaissance d'auteur, Attribution de l'auteur, Langue Arabe, Reconnaissance Optique des Caractères (OCR).

Abstract

Research in the field of copyright attribution is a very old problem, due to the large number of unknown texts and the technological development of various means of digital communication, which have facilitated the process of plagiarism and theft of textual works. from the original authors. This has resulted in great difficulty in finding and extracting information from these works to identify their original authors.

In this work, we study the style of a group of contemporary Arab writers through some of their works in order to create models for them to attribute unknown literary texts to their original owners. For this, we tried to create an efficient system, where we created a database of texts obtained using the OCR program, and suggested two classifiers (Linear-SVM, Burrows-Delta) and used the function (N - gram).

Keywords: Author recognition, Author attribution, Arabic language, Optical Character Recognition (OCR).