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°: .. 12 .. /INST/ 2020

DOMAINE: SCIENCES ET TECHNOLOGIE
FILIERE: ÉLECTRONIQUE
OPTION: INSTUMENTATION

Mémoire présenté pour l'obtention
du diplôme de Master Académique

Par : MENASRI Radja et YAKOUBI Mebarka

Intitulé

Etude et analyse des effets d'acquisition optique à l'aide d'un OCR des textes arabes sur l'attribution d'auteurs

Soutenu publiquement le : 30 /09 / 2020 devant le jury composé de:

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

Année universitaire : 2019 /2020

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 que nous avons trouvés en eux, nous sommes très reconnaissants.

Nous remercions messieurs le chef du département d'électronique et le doyen de la faculté de technologie et tous les enseignants du département d'électronique qui ont contribué à notre formation, ainsi que tous les membres du cadre administratif.

Comme nous tenons tant à remercier l'ensemble des enseignants ayant contribué à notre formation durant les différents cycles primaire, moyen, secondaire et universitaire.

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

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.

Ma chère sœur et son mari et ses petites filles

Mes chers amis(es).

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

Spécialement :INSAF,RIYAD ,BECHAR

Et à tous mes collègues de ma promotion.

Mebarka

DEDICACE

Je dédie ce modeste travail :

A mon père : << Nourddine >>

A ma mère : << Mebarka>>

A mes frères : Seif Eddine, Mouhamed Amine, Ghoulem et Youcef

Ma sœur : Chyma

Ma belle-sœur : Insaf

Ma partenaire et ma chère amie avec qui je partage ce modeste travail : Mebarka

Radjaa

LISTE DES ABREVIATIONS

ALA : Arabe Littéraire Ancienne

ALC : Arabe Littéraire Classique

ASM : Arabe Standard Moderne

SVM : Support Vecteur Machine

MLP: Multi Layer Perceptron

OCR: Reconnaissance Optique de Caractères.

ANN : Artificiel Neural Network

MCD: Manhattan Centroid Distance.

AA : Author Attribution

RNA : Réseaux de Neurones Artificiels

MRS : Minimisation du Risque Structurel

LISTE DES TABLEAUX :

LISTE DES FIGURES:

Table des matières

Introduction générale

INTRODUCTION GENERALE

Notre motivation

L'attribution d'auteur des textes anonymes est l'une des plus vieilles difficultés de la statistique appliquée à la littérature. Il s'agit de rapprocher le texte anonyme à d'autres textes dont les auteurs sont connus et dont on soupçonne qu'ils ont pu participer à leur rédaction. La lutte contre le plagiat dans la littérature arabe nécessite une collaboration des efforts pour se débarrasser de ces comportements débauchés.

Le développement se déroulant dans divers domaines, notamment l'informatique et l'intelligence artificielle, en est venue avec leurs nombreuses propriétés afin de résoudre divers problèmes de l'attribution d'auteurs. L'idée principale derrière l'attribution ou l'identification d'auteurs est fondée sur des statistiques ou des calculs qu'en mesurant certaines caractéristiques textuelles, nous pouvons distinguer les textes écrits par différents auteurs.

Pour cette fin, le présent travail s'intègre dans le cadre de la reconnaissance d'auteurs des textes dont les auteurs sont inconnus, en proposant un système basé d'un côté sur l'intégration générique des outils de reconnaissance de caractères, et de l'autre côté sur l'utilisation des méthodes et techniques de classification.

Nos objectifs

La reconnaissance optique de caractères en Anglais (Optical Character Recognition) abrégée (OCR) est un moyen de conversion des documents images (text image) en documents textes qui peuvent être traités par l'ordinateur en utilisant les différents outils de traitements de texte, avant d'utiliser des méthodes spécifiques qui permettent d'étudier le style de l'auteur et de pouvoir attribuer ces textes à leur auteur réel.

Le travail présenté dans ce mémoire entre dans le cadre générale d'attribution de l'auteurs des textes arabes. La particularité de ces textes est qu'ils ont été obtenus après une opération de reconnaissance de caractères (OCR) appliquée sur des textes scannés (numérisés). Dans cette étude, on s'est fixé les objectifs principaux suivants :

- Conception d'une base de données textuelle pour valider les techniques que nous avons proposées.
- Evaluation de l'influence de l'opération OCR des textes scannés sur la robustesse de notre système d'attribution d'auteurs.
- Conception d'un système d'attribution d'auteurs basé sur les réseaux de neurones, les machines à vecteurs de support et les N-gramme caractéristiques.

Structure de la thèse

Le présent mémoire est organisé autour de trois chapitres comme suit : La reconnaissance optique de caractères ainsi que ces différents types de systèmes, les notions fondamentales de l'exploration de texte, la stylométrie et l'attribution de textes, la langue arabe et ces particularités font l'objet du premier chapitre. Le deuxième chapitre aborde la méthodologie de recherche qui a été adopté dans ce mémoire ainsi que les approches et techniques proposées pour l'attribution d'auteurs des documents textuelles en littérateur arabe. Le troisième et dernier chapitre présente les différentes expérimentations conduites suivi de quelques discussions et interprétations des résultats obtenus. Nous terminons notre mémoire par une conclusion générale, où on récapitule l'ensemble des travaux réalisés et on rappelle les principaux résultats obtenus dans cette étude et on propose des perspectives envisagées pour poursuivre cette recherche.

Chapitre-1

Généralités sur la Reconnaissance Optique des Caractères et l'Exploration de Données Textuelles

CHAPITRE-1

GENERALITES SUR LA RECONNAISSANCE OPTIQUE DES CARACTERES ET L'EXPLORATION DE DONNEES TEXTUELLES

1.1 Introduction

Les outils d'exploration de données traditionnels sont incapables de gérer les données textuelles car il faut du temps et des efforts pour extraire des informations. La sélection de la technique appropriée aide à augmenter la vitesse et diminue le temps et les efforts nécessaires pour extraire des informations précieuses.

L'exploration de texte est un domaine multidisciplinaire basé sur la recherche d'informations, l'exploration de données, l'apprentissage automatique, les statistiques et la linguistique informatique.

Dans ce chapitre, nous présentons des généralités sur la Reconnaissance Optique des Caractères, l'exploration de texte, ensuite la stylométrie. Enfin, la catégorisation des textes écrits en langue Arabe

1.2 Reconnaissance Optique des Caractères (ROC)

La Reconnaissance Optique des Caractères (ROC) (en anglais Optical Character Recognition OCR) est une technologie qui permet de reconnaître automatiquement les caractères à travers un mécanisme optique. Bien que l'OCR ne soit pas en mesure de rivaliser avec la lecture humaine capacités, il peut reconnaître à la fois le texte manuscrit et imprimé.

Les performances de l'OCR, qui dépendent directement de la qualité des documents d'entrée, permettent de convertir les différents types de documents tels que documents papier numérisés, documents PDF des fichiers ou des images capturées par un appareil photo numérique en données consultables [Mit,Ind,Div ,2013].

La reconnaissance de texte imprimé est de transformer ce dernier en une représentation compréhensible par une machine, et facilement reproductible par un traitement de texte, ceci

n'est pas toujours facile du moment que les mots peuvent avoir plusieurs représentations, différentes polices et différents styles de caractères (Gras, Italiques, etc.).

1.2.1 Types de systèmes OCR

On distingue trois types de systèmes OCR, à savoir ; les systèmes de reconnaissances mono-fonte, multi-fonte ou omni-fonte. Un système OCR mono-fonte reconnaît les caractères d'une fonte bien déterminée. Les systèmes OCR multi-fonte permettent la reconnaissance des caractères de plusieurs fontes, toutes ayant en principe fait l'objet d'un apprentissage unique. La phase de reconnaissance est souvent précédée par une phase de normalisation. Les systèmes OCR omni-fonte font abstraction de l'information sur la fonte, dans la mesure où ils sont capables de reconnaître des caractères de n'importe quelle fonte, et n'importe quelle taille [Bou,Ben, 2005].

1.2.2 Les étapes d'un processus OCR

Les principales étapes d'une chaîne de reconnaissance sont présentées ci-après :

- L'acquisition permettant la conversion du document papier sous la forme d'une image numérique (bitmap). Cette étape est importante car elle se préoccupe de la préparation des documents à saisir, du choix et du paramétrage du matériel de saisie (scanner), ainsi que du format de stockage des images.
- Le prétraitement dont le rôle est de préparer l'image du document au traitement. Les opérations de prétraitement sont relatives au redressement de l'image, à la suppression du bruit et de l'information redondante, et enfin à la sélection des zones de traitement utiles.
- La reconnaissance du contenu qui conduit le plus souvent à la reconnaissance du texte et à l'extraction de la structure logique. Ces traitements s'accompagnent le plus souvent d'opérations préparatoires de segmentation en blocs et de classification des médias (graphiques, tableaux, images, etc.).
- La correction des résultats de la reconnaissance en vue de valider l'opération de numérisation. Cette opération peut se faire soit automatiquement par l'utilisation de dictionnaires et de méthodes de correction linguistiques, ou manuellement au travers d'interfaces dédiées.

1.3 Exploration de données et exploration de textes

1.3.1 Exploration de données

L'exploration de données est un processus analytique qui recherche des tendances et des modèles dans des ensembles de données qui révèlent de nouvelles perspectives. Ces nouvelles informations sont des informations implicites, auparavant inconnues et potentiellement utiles.

Les données, qu'elles soient constituées de mots, de chiffres ou des deux, sont stockées dans des bases de données relationnelles. Il peut être utile de considérer ce processus comme une exploration de bases de données ou comme certains l'appellent « Découverte de connaissances dans des bases de données ». L'exploration de données peut être utilisée pour extraire n'importe quelle base de données, pas seulement celles créées à l'aide de l'exploration de texte [Cla,2013].

1.3.2 Exploration de textes

L'exploration de textes est un processus permettant d'extraire des schémas intéressants et significatifs pour explorer les connaissances à partir de sources de données textuelles ou bien transforme le texte en des données qui peuvent être analysées.

Plusieurs techniques d'exploration des textes comme la synthèse, la classification, le regroupement etc., peuvent être appliquées pour extraire des connaissances. Cette approche est essentiellement basée sur des règles et recherche des mots prédéfinis et des modèles de mots à partir desquels le sens peut être déduit.

Une approche différente utilisant des méthodes statistiques devient de plus en plus populaire et les techniques s'améliorent régulièrement. Ces méthodes utilisent la fréquence et l'emplacement des mots pour révéler des concepts. Un exemple simple consisterait à déterminer les termes les plus courants ou les moins courants dans un texte, puis à identifier les autres termes qui se produisent avec ceux-ci. Cette technique a été utilisée pour créer des outils de classification automatique. Dans ce cas, le système est d'abord entrainé à l'aide d'un échantillon de documents dont la classification est connue, ensuite il utilise les modèles qu'il a appris pour classer les nouveaux documents [Tal,Han,Aye,Fat,2016].

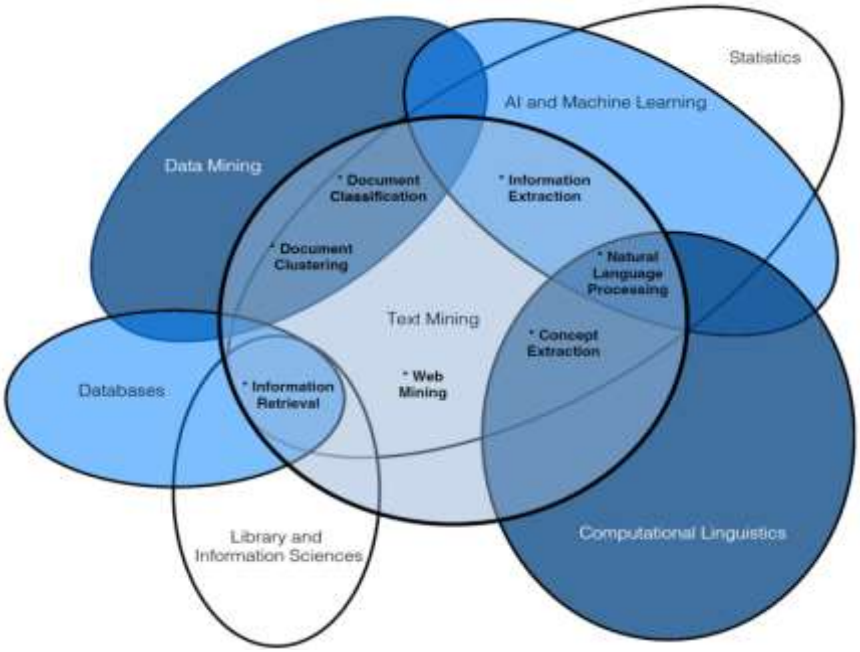


Figure-1.1 L'interaction de l'exploration de texte avec les autres domaine (Diagramme de VENN) [Tal,Han,Aye,Fat,2016]

1.4 La stylométrie

La technologie moderne et plus particulièrement l'informatique permet de nos jours d'analyser la trame stylistique d'un texte lorsque l'identité d'un auteur est contestée. Par l'analyse stylométrique, il est maintenant possible de déterminer avec un haut degré de certitude si telle ou telle personne est l'auteur ou non de l'ouvrage concerné.

1.4.1 Petit historique de la stylométrie

Les premières mentions de la stylométrie pour identifier des auteurs sont apparues en 1851. Mais, compte tenu de la difficulté des mesures à effectuer, les premières études crédibles ont dû attendre l'arrivée des ordinateurs modernes, pour leur précision de comptage et leur traitement à grande vitesse des données. Au début des années 1980, une équipe de chercheurs a travaillé pour affiner et rendre plus performantes les techniques de la stylométrie. Les travaux montrent que les méthodes de comptage et de comparaison entre différents textes ont été grandement améliorées. La stylométrie continue à évoluer vers une fiabilité et une sensibilité toujours plus grandes, elle a atteint un niveau qui permet la mise en œuvre d'une technique de mesure rigoureuse qui donne des réponses fiables dans l'analyse des textes de plusieurs milliers de mots d'un même auteur, en flux libre [Ang].

1.4.2 Définition de la stylométrie

La stylométrie est l'étude quantitative du style littéraire à l'aide de méthodes informatiques de lecture distante. Elle se base sur l'observation faite que chaque auteur a tendance à écrire de façon relativement constante, reconnaissable et unique. C'est un ensemble de techniques à l'intersection de la linguistique et de la statistique, dont le but est d'identifier le style de documents textuels. Le style d'un texte est une caractéristique de son contexte d'écriture au sens large : son auteur, son époque, son « genre », etc…

La stylométrie tente de montrer qu'un texte est écrit dans un style différent d'une collection d'autres textes. Cette différenciation permet donc, dans une certaine mesure, de déterminer si un « anonyme » a été écrit par un auteur précis, et surtout de déterminer si un texte n'a pas été écrit par un auteur précis [Ang].

1.4.3 Caractéristiques utilisées dans la stylométrie

Chaque individu possède son propre vocabulaire, parfois riche, parfois limité. Bien qu'un vocabulaire étendu soit généralement associé à une littérature de qualité, ce n'est pas toujours le cas. Certaines personnes écrivent en phrases courtes, tandis que d'autres préfèrent les phrases complexes comportant plusieurs propositions. Il n'y a pas deux auteurs qui utilisent les points-virgules, les tirets et autres signes de ponctuation exactement de la même façon.

L'identification de l'auteur d'un texte anonyme constitue cependant l'une des applications les plus courantes de la stylométrie. Il est parfois possible de découvrir l'identité de l'auteur d'un texte en mesurant certaines caractéristiques de ce texte, comme la longueur moyenne des phrases ou le rapport entre le nombre d'articles définis et indéfinis [Ang].

Ces mesures sont ensuite comparées avec celles observées dans des textes dont les auteurs sont connus Lorsque l'on parle de trame non-contextuelle, il s'agit de mots qui sont souvent interchangeables ou qui peuvent même être omis sans perte de la signification générale du texte. Ces mots contribuent peu à l'information contextuelle et sont souvent ignorés consciemment, aussi bien par le lecteur que par l'auteur. Ces mots constituent typiquement 20 à 45% du texte total, ce qui permet d'avoir un nombre important de choix statistiques, et plus les mesures statistiques sont nombreuses, plus leurs résultats sont fiables.

1.5 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.

1.5.1 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 [Xia, 2010]. 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) [Bal, 2010].

1.5.2 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).

1.5.3 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 [Sta, 2009], 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 [Mendenhall, 1887]. 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 » [Mos, 1963].

1.5.4 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.

1.6 Attribution d'auteur

L'attribution d'auteur (AA) est le processus visant à identifier la paternité probable d'un document donné, compte tenu d'une collection de documents dont l'auteur est connu. L'attribution de la paternité devient un problème important car le la gamme d'informations anonymes augmente avec une croissance rapide de l'utilisation d'Internet dans le monde entier. Application de la paternité de l'attribution comprend la détection du plagiat, déduire l'auteur de communications inappropriées qui envoyé de manière anonyme ou sous un pseudonyme, ainsi que la résolution de questions historiques paternité peu claire ou contestée [Boz,Bag,Uya 2007].

L'attribution de la paternité est le moyen de déterminer l'auteur d'un texte lorsqu'il n'est pas clair qui l'a écrit. Il est utile lorsque deux personnes ou plus prétendent avoir écrit quelque chose ou quand personne ne veut (ou ne peut) dire qu'elle ou il a écrit la pièce.

1.6.1 L'état de l'art

Plusieurs recherches ont été menées à l'attribution d'auteurs au cours des dernières années. Avec la quantité croissante de documents sur Internet, et comme la plupart des écrits sont anonyme, l'attribution de la paternité devient importante. Les recherches portent sur différentes propriétés des textes. On distingue deux propriétés différentes des textes qui sont utilisés dans classification ; le contenu du texte et le style de l'auteur.

L'analyse statistique du style littéraire complète la bourse littéraire traditionnelle car elle offre un moyen de saisir le caractère souvent insaisissable de l'auteur style en quantifiant certaines de ses caractéristiques. La majorité des études stylométriques utilisent des éléments de langage et la plupart de ces éléments sont à base lexicale.

1.6.2 Les étapes de l'attribution d'auteur

Un processus complet d'attribution de l'auteur consiste en :

- Rassemblement des textes qui sont les observations à classer.
- Un méthode d'extraction de caractéristiques qui calcule les informations numériques ou symboliques issues de ces observations.
- Un système de classification ou de catégorisation qui fait le classement à partir de ces observations.

1.7 Les textes écrits en langue arabe

Grâce à la propagation de l'Islam et la diffusion du Coran, la langue arabe étendu à partir du 7ème siècle. Les recherches pour le traitement automatique de l'arabe ont commencé vers les années 1970. Les premiers travaux concernaient notamment les lexiques et la morphologie. Par ses propriétés morphologiques et syntaxiques la langue arabe est considérée comme une langue difficile à maîtriser dans le domaine du traitement automatique de la langue.

Avec la diffusion de la langue arabe sur le Web et la disponibilité des moyens de manipulation de textes arabes, les travaux de recherche ont abordé des problématiques plus variées comme la syntaxe, la traduction automatique, l'indexation automatique des documents, la recherche d'information, etc…

1.7.1 Particularités de la langue arabe

La langue Arabe fait partie de la famille des langues sémitiques (i.e. Hébreu, Araméen, Babylonien, etc.) en référence au nom de Sem le fils du prophète Noah. Elle est l'une des langues les plus intéressantes au monde, où elle est classée comme la 5 langue la plus parlée au monde en 2014. Cependant, ce classement est basé sur le nombre de personnes ayant l'Arabe comme langue maternelle (3.3% de la population mondiale).

Or, vu que l'Arabe est la langue du saint Coran, alors tous les musulmans maitrisent au moins les bases linguistiques qui permettent de lire le Coran pour prier. Ainsi, et en considérant le nombre de musulmans compté en 2015 (24% de la population mondiale) qui est reporté dans les statistiques faites par le centre de recherche PEW [Pew, 2015], alors on trouve que l'Arabe peut occuper la 1ère ou la 2ème place dans le classement.

En effet, la langue Arabe est l'une des langues les plus riches ; sa richesse réside dans le fait qu'elle est très flexionnelle en comportant un large vocabulaire (approximativement 500 million mots), en outre, elle est dotée d'une morphologie complexe [Sha, 2010].

L'alphabet de la langue arabe contient 28 lettres principales (tableau-1.1) et s'écrit et se lit de droite à gauche. Elle est dotée aussi de quelques autres caractéristiques telles que :

- Chaque lettre peut prendre plusieurs formes différentes dépendantes de sa position dans le mot (tableau-2.2).
- Toutes les lettres s'attachent entre elles sauf six lettres « أ، د، ذ، ر، ز، و » qui ne sont joignables avec aucune lettre à gauche.

Tableau-1.1 : Les lettres de l'alphabet arabe (28 lettres).

N°	Lettre arabe	Lettre correspondante en français	Pronunciation	N°	Lettre arabe	Lettre correspondante en français	Pronunciation
1	ا	a	Alef	**15**	ض	d	Dad
2	ب	b	Ba '	**16**	ط	t	Tah
3	ت	t	Ta'	**17**	ظ	z	Zah
4	ث	th	Tha'	**18**	ع	a	Ayn
5	ج	j	Jim	**19**	غ	gh	Ghayn
6	ح	h	Hha'	**20**	ف	f	Fa
7	خ	kh	Kha'	**21**	ق	q	Qaf
8	د	d	Dal	**22**	ك	k	Kaf
9	ذ	d	Thal	**23**	ل	l	Lam
10	ر	r	Ra	**24**	م	m	Mim
11	ز	z	Zayn	**25**	ن	n	Nun
12	س	s	Sin	**26**	ه	h	Ha
13	ش	sh	Shin	**27**	و	w	Waw
14	ص	s	Sad	**28**	ي	y	Ya

Tableau-2 : Exemple de variation de la lettre « ع » (Ayan).

A la fin		Au milieu	Au début
Non joignable	**Joignable**		
ع	ع	ـعـ	عـ

Une autre spécificité de la langue Arabe qui la rend particulière c'est l'utilisation des diacritiques, ou le Tashkil en Arabe (التشكيل). En effet, le sens d'un mot Arabe est déterminé par ces derniers, or le changement d'une seule diacritique dans le mot peut changer radicalement le sens du mot.

Par exemple, le mot Arabe Na3am (نَعَم) avec la diacritique Fatha au début signifie « oui » en français, contrairement le mot Arabe Ni3am (نِعَم) avec la diacritique Kasra au début signifie « des grâces » en français. Bien que les deux mots soient syntaxiquement semblables, leur sémantique diffère en changeant une seule diacritique.

On note aussi une autre caractéristique, c'est l'existence de deux différentes approches de lemmatisation dans la littérature. En fait, la première approche fait référence au nom Albasrion (البصريّون), elle consiste à dériver les mots Arabes à partir de El-Masdar (المصدر). Tandis que la deuxième fait référence au nom de Kuffien (الكوفيّون), qui est la plus répandue et la plus correcte selon les linguistes Arabes, consiste à dériver les mots Arabes à partir de El-Fiîl El-Maadi (الفعل الماضي) qui correspond au verbe au passé.

1.7.2 Morphologie de la langue arabe

Le lexique arabe comprend trois catégories de mots : verbes, noms et particules. Les verbes et noms sont le plus souvent dérivés d'une racine à trois consonnes radicales. Une famille de mots peut ainsi être générée d'un même concept sémantique à partir d'une seule racine à l'aide de différents schèmes.

Ce phénomène est caractéristique à la morphologie arabe. On dit donc que l'arabe est une langue à racines réelles à partir desquelles on déduit le lexique arabe selon des schèmes qui sont des adjonctions et des manipulations de la racine.

1.7.3 Syntaxe arabe

Étudie la formation correcte des phrases par l'analyse de :

- La position des unités lexicales les unes par rapport aux autres pour déterminer leur ordre,
- Marquage casuel des unités lexicales de la phrase, Ainsi, la fonction syntaxique de chaque unité est déterminée en s'appuyant sur la morphophonologie [Khe, 06].

1.7.4 Les types de la langue arabe

La langue arabe est passée dans l'histoire par plusieurs variétés :

a) L'arabe littéraire ancienne (ALA) : C'est la langue de la poésie préislamique, se retrouve dans un nombre restreint des documents d'aujourd'hui.

b) L'arabe littéraire classique (ALC) : Ce type représente une autre étape de l'évolution de la langue. Il apparaît avec la naissance de l'Islam. Cette arabe évoluée a utilisé les règles de base de la langue du Coran et a ajouté une grammaire considérée comme une norme idéale.

c) L'arabe standard moderne (l'ASM) : une forme un peu différenciée de l'arabe classique constitue la langue écrite de tous les pays arabophones. L'ASM reste la langue de la presse et de la littérature, alors que l'arabe classique appartient au domaine religieux et est pratiqué par les membres du clergé.

1.7.5 Les dialectes arabes

Malgré l'existence d'une langue officielle commune, chaque pays a développé son propre dialecte. On peut regrouper ces dialectes en quatre grandes catégories comme suit [Bou, 08] :

- Dialectes arabes parlés dans la Péninsule Arabique (Golfe, Najd, Yéménite),
- Dialectes arabes parlés dans les pays maghrébins (Algérien, Marocain, Tunisien, Hassaniya de Mauritanie).
- Dialectes parlés au Proche-Orient (Egyptien, Soudanais, Syro-Libano-Palestinien, Irakien).
- La langue Maltaise est également considérée comme un dialecte arabe.

1.8 Conclusion

Dans ce chapitre, nous avons exposé quelques généralités sur l'exploration de données textuelles et la Reconnaissance Optique des Caractères (ROC ou en Anglais OCR), les types de systèmes OCR ainsi que les étapes d'un processus OCR sont abordés. Par la suite, des définitions de l'exploration de données, de l'exploration de textes et la stylométrie ainsi qu'un bref historique de cette dernière sont présentés.

D'autre part, nous avons discuté les différentes formes de la catégorisation des documents textuels et nous avons mis l'accent sur l'attribution d'auteurs. Enfin, nous avons abordé la catégorisation des textes écrits en langue Arabe.

Dans le prochain chapitre, nous allons présenter la méthodologie de recherche ainsi que les techniques utilisées dans ce travail de recherche.

Chapitre-2

Méthodologiede Recherche etTechniques Proposées

CHAPITRE-2

METHODOLOGIEDE RECHERCHE ETTECHNIQUES PROPOSEES

2.1 Introduction

L'idée principale derrière l'attribution d'auteurs basée sur des statistiques ou des calculs est qu'en mesurant certaines caractéristiques textuelles, nous pouvons distinguer les textes écrits par les différents auteurs.

Ce chapitre est consacré à décrire et à présenter la méthodologie de recherche ainsi que les différentes techniques proposées pour l'identification ou l'attribution d'auteurs.

2.2 Méthodes de catégorisation des textes par auteur

Un suivi sur les approches modernes d'identification de l'auteur a été fait par Stamatatos[Sta, 2009], où il souligna les limitations des approchesproposées dans la période des années 80 (jusqu'à 1990). Depuis 1990, et avec l'expansion de l'internet et de divers moyens de communications, les choses ont bien évolué et de nouvelles perspectives sont apparues telles que :

- Vérification de l'auteur (décision si un texte appartient à tel auteur ou non) ;
- Détection de plagiat (vérifier si un texte a été copié d'une autre source) ;
- Caractérisation d'auteur (création d'un profil représentant un auteur ; sexe, âge, etc.) ;
- Détection stylistique : repérer les parties écrites par tel auteur dans un texte écrit en collaboration.

De plus, Stamatatos présenta les différentes caractéristiques stylistiques utilisées dans ce domaine, où on trouve les caractéristiques lexicales (i.e. la longueur des mots, la longueur des phrases, les fréquences des mots, etc.), les caractéristiques basées sur les caractères (i.e. caractères N-grammes, les caractéristiques syntaxiques (i.e. structures des phrases, part-of-speech, etc.) et les caractéristiques sémantiques (i.e. synonymes, mots fonctionnels, etc.). Il présenta également les différentes méthodologies appliquées telles que les méthodes probabilistes, méthodes de compression, méthodes basées sur les n-grammes, méthodes basées sur les modèles d'espace de vecteur, méthodes basées sur les

similarités, méthodes basées sur les modèles de méta-apprentissage et enfin les méthodes hybrides [Sta, 2009].

2.3 Méthodologie de recherche proposée

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 on fait l'extraction des caractéristiques pertinentes (dans notre cas les caractères n-grammes) et construction du modèle de chaque auteur. Enfin, la quatrième étape est dédiée aux méthodes de classification (MLP et SVM) pour réaliser le processus d'identification (figure-2.1).

2.3.1 Conversion des textes scannés

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-2.1. Conversion des textes scannés en textes modifiables à l'aide d'un OCR.

La technologie d'OCR a été appliquée ces dernières années à travers tout le spectre d'industries en train de révolutionner le processus de gestion des documents. Les systèmes

d'OCR ont permis à des documents numérisés de se transformer en documents entièrement consultables avec le contenu du texte qui est reconnu par les ordinateurs [Gaj,All].

Cependant, après plus de deux décennies de recherche sur la numérisation des documents, ces systèmes peuvent encore laisser quelques imperfections pour parvenir à une réédition du document ce qu'il peut être dû aux différents problèmes dont la qualité du document et de l'impression, la discrimination de la forme, le type d'acquisition, les variations des dimensions, le nombre de scripteurs, la taille du vocabulaire, etc.

2.3.2 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 (ح converti 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 demots.

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

Tableau-2.1. Liste des caractères insignifiantspour la stylométrie

N°	Caractères	Noms
01	a b c … y z	Lettre français minuscules
02	A B C … Y Z	Lettre français majuscules
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

2.3.3 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[Cav,1994], 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 expérimentations fondées sur diverses valeurs de N (de 2 à 5-grammes) révèlent de bons résultats avec N = 3 et N = 4. D'autres travaux montrent également de bonnes performances pour la classification de textes du langage Telugu5en se fondant sur les 3-grammes [Lam ,Béc ,Ham, Roc ,2010].

Les types de caractéristiques qui ont été proposées et utilisées dans ce travail sont N-grammes (avec N= 1, 2, 3, 4,5) comme illustré ci-dessous:

- Caractères uni-grammes (n=1),
- Caractères bi-grammes (n=2),
- Caractères tri-grammes (n=3),
- Caractères tétragrammes (n=4),
- Caractères pentagrammes (n=5).

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).

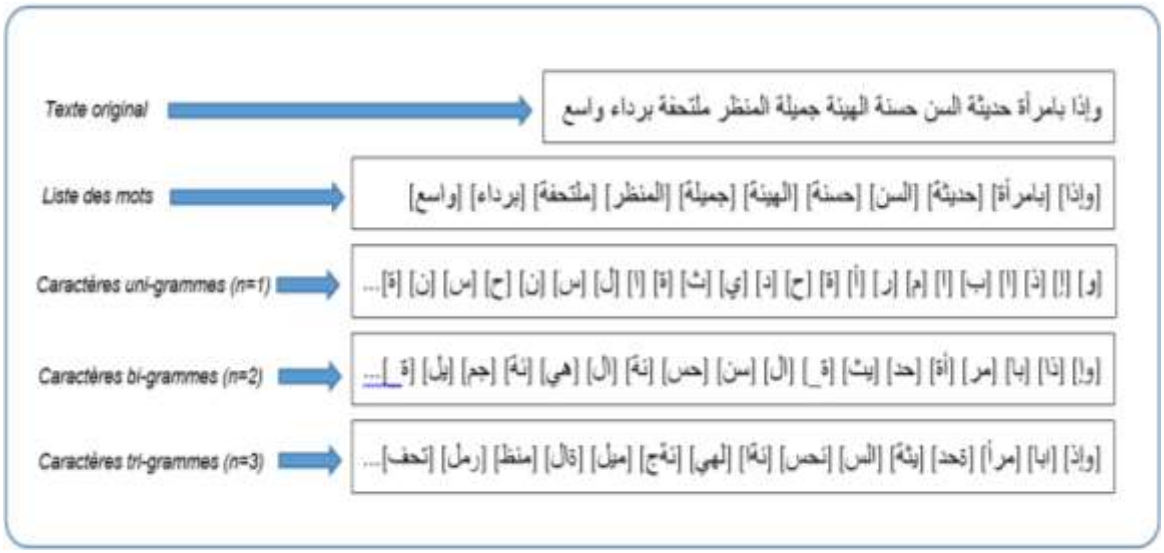


Figure-2.2. Exemple d'extraction des caractères N-grammes d'un texte.

2.3.4 Approches proposées pour l'attribution d'auteurs

2.3.4.1 Approche Neuronale

Sans revenir trop à l'histoire quant aux origines des Réseaux de Neurones Artificiels "RNA" (ou en Anglais Artificiel Neural Network "ANN"), on peut dire que Minsky en 1969 est le premier à avoir proposé un modèle mathématique du neurone comme unité de calcul binaire. Cette approche dite connexionniste(à connaissance répartie), a alors été supplantée par l'approche symbolique qui promouvait les systèmes experts (à connaissance localisée).

Face aux difficultés rencontrées lors de la modélisation des connaissances, l'approche symbolique s'est éteinte avec les années 80, permettant ainsi la relance de l'approche connexionniste, qui s'inspire d'une modélisation du cerveau en mettant l'accent sur les mécanismes d'apprentissage. Les RNA sont des circuits électroniques dont chaqueélément est sensé simuler le fonctionnement d'une cellule élémentaire du cerveau humain qu'est le neurone [MED,2010].

A- Neurone biologique et neurone formel

Depuis le début du vingtième siècle, plusieurs recherches s'intéressent au cerveau humain d'un point de vue microscopique et cellulaire afin de relier les faits du

comportement et les réactions électriques et chimiques qui se produisent à l'intérieur du cerveau.

Le système nerveux est composé de 10^{12}cellules nerveuses, appelées neurones, interconnectés. Bien qu'il existe une grande diversité de neurones, ils fonctionnent tous sur le même schéma. Le neurone se décompose en trois régions principales : Le corps cellulaire, les dendrites et l'axone. (Voir figure-2.3).Chaqueneuronereçoit à son entrée unensembledepotentielsexcitateurs (ouinhibiteurs) par l'intermédiaire des synapses qui le relient aux neurones voisins. Les dendrites calculent une somme pondérée de ces entrées. Une fois le niveau d'activation est obtenu, le neurone génère un potentiel d'action qui se propage le long de l'axone pour arriver aux neurones voisins. De cette manière,leneuroneréalise les cinqfonctions requises à savoir [……]:

- Recevoir des signaux provenant des autres neurones,
- Intégrer ces signaux,
- Engendrer un influx nerveux,
- Conduire ce flux le long de l'axone,
- Transmettre ce flux aux neurones voisins prêts à le recevoir.

Ce modèle biologique simple sert de base au modèle mathématique du neurone artificiel (ou formel). Le neurone artificiel (appelé aussi neurone formel), qui s'inspire amplement du neurone biologique, est une fonction non linéaire à seuil, paramétrée et à valeurs bornées.Par analogie au neurone biologique, le neurone formel est un modèle qui se caractérise par : des signaux d'entrée, un état interne, une fonction d'activation et une sortie [ZEM,2003]. (Voir figure-2.3).

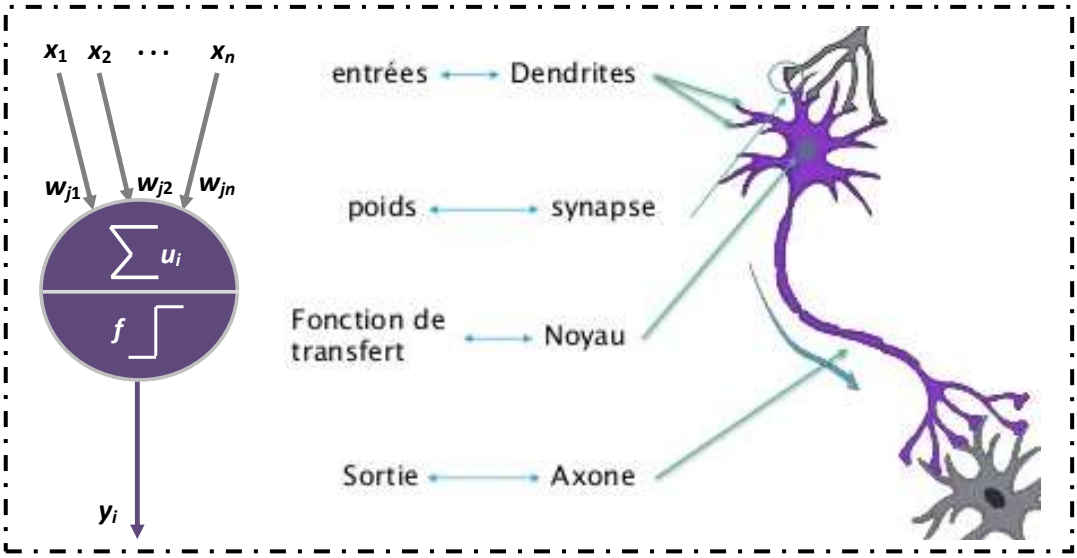


Figure-2.3. Analogie entre neurone biologique et neurone artificiel (formel)

Le neurone formel, représenté graphiquement dans la figure-2.4, réalise une fonction ***f*** de la somme pondérée par les poids synaptiques (w_{ji}) des entrées (x_1,…, x_n). Cette valeur, qui représente l'état interne du neurone u_i, est alors transmise à la fonction d'activation ***f***.

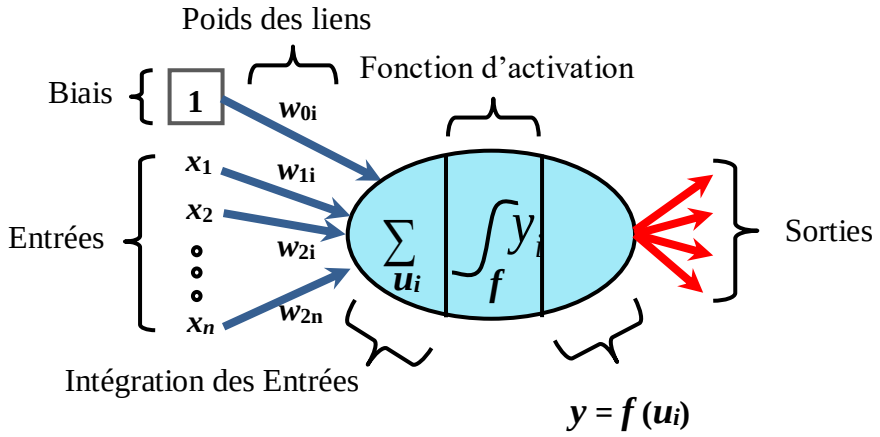


Figure-2.4. Représentation d'un modèle de neurone formel

La sortie y_i, qui représente l'activation du neurone, est donnée par l'équation 2.1 :

$$y_i = f(w_0 + \sum_{i=1}^{n} w_i x_i) = f(w_0 + w^{'} x) \qquad (2.1)$$

f : Fonction d'activation agissant comme une transformation d'une combinaison affine des signaux d'entrée. Les différents types de neurones se distinguent par la nature de leur fonction d'activation.

w_0 : Les biais du neurone.

w_i : Les poids associés à chaque neurone dont les valeurs sont estimées dans la phase d'apprentissage. Ces poids constituent "la mémoire" ou "la connaissance répartie" du réseau de neurone.

B- Types de fonctions d'activation

Un neurone artificiel réalise un produit scalaire entre un vecteur d'entrée *X* et un vecteur de poids *W*, ajoutant un biais *b*. Une fonction *f*, appelée fonction d'activation, définit la valeur de la sortie *y* en termes de niveaux d'activité de ses entrées.

Cette fonction est appliquée sur l'ensemble pour déterminer la réponse (sortie) du neurone : $y = f(wx + b)$. On peut trouver plusieurs types de fonctions d'activations. Les plus utilisées, sont données dans la figure-2.5 suivante [ZEM,2003] :

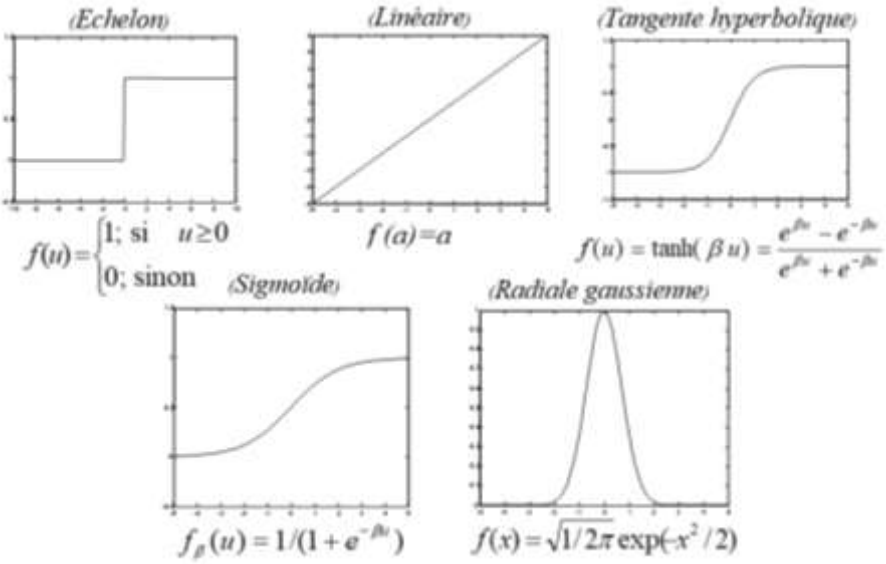


Figure-2.5. Exemples de fonctions d'activation.

Les fonctions d'activation non linéaires différentiables sont plus intéressantes et souvent plus utilisées. Les fonctions, tangente hyperbolique et sigmoïde représentent d'ailleurs un exemple très pratique. Le choix de la fonction d'activation est motivé par des considérations théoriques et pratiques. Les propriétés de cette dernière influent en effet sur celles du réseau de neurone. Il est donc important de bien choisir celle-ci pour obtenir un modèle performant.

C- Apprentissage des RNAs

Parmi les principales caractéristiques des RNAs est leur capacité à apprendre par l'exemple. L'apprentissage des réseaux de neurones est la procédure qui consiste à estimer les paramètres du réseau afin que celui-ci remplisse au mieux la tâche qui lui est affectée. L'étape d'apprentissage est une phase du développement d'un RNA durant laquelle sa conduite du est modifiée jusqu'à l'obtention du réseau désiré [Ouk,2012].

Selon le critère d'apprentissage, les RNAs se divisent en trois catégories ; les réseaux à apprentissage supervisé, les réseaux à apprentissage non supervisé et apprentissage semi supervisé.

- **Apprentissage supervisé (ou à partir d'exemples étiquetés) :** L'objectif de ce type d'apprentissage est de construire un modèle prédictif d'une grandeur numérique qui permet d'apprendre au mieux la fonction inconnue qui génère des données aléatoires, indépendantes et identiquement distribuées et dont nous ne disposons que de quelques exemples. Il s'agit, dans ce cas, d'un problème de classification.
- **Apprentissage non supervisé :** Il vise à apprendre certaines informations sur des données non étiquetées dans un but de les rassembler en des groupes homogènes. Dans ce cas, on présente une entrée au réseau et on le laisse évoluer librement jusqu'à ce qu'il se stabilise. La règle d'apprentissage, n'étant pas fonction du comportement de la sortie du réseau, mais plutôt du comportement local des neurones. Il existe de nombreux cas où on ne possède aucune information sur les classes de l'ensemble d'apprentissage. Il s'agit, dans ce cas, d'un problème de clustering.
- **Apprentissage par renforcement :** L'apprentissage par renforcement, appelé aussi semi supervisé, est seulement une évaluation qualitative de la performance désirée, qui est généralement spécifiée sous forme d'une fonction de coût à minimiser. Cette fonction est directement liée à la tâche que doit accomplir le réseau de neurones dans l'environnement où il est intégré.

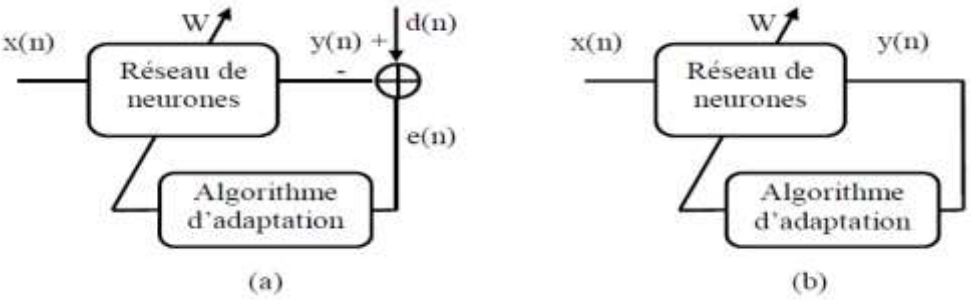


Figure-2.6. Apprentissage supervisé (a) et non supervisé (b) des RNAs

2.3.4.2 Le perceptron multicouche (MLP)

Le perceptron est l'exemple le plus simple des réseaux de neurones linéaires. Il est inventé par Rosenblatt (1957) en s'inspirant du système visuel de l'être humain [……]. Ce type de réseau, dit à propagation avant (ou Feedforward), est utilisé pour résoudre les problèmes de classification binaire et ne contient qu'une seule couche de sortie (monocouche) à laquelle toutes les entrées sont connectées. Il. (Voir figure-2.7).

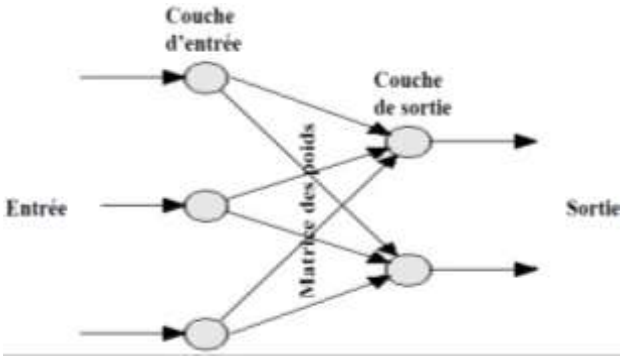


Figure-2.7. Architecture globale du perceptron (a) et du perceptron multicouche (b).

Un perceptron à n entrées (observations) (x_1, x_2,..., x_n) et une sortie (réponse) y défini par la donnée de n poids (coefficients synaptiques) (w_1, w_2,..., w_n) et un biais (seuil) b comme suit :

$$y = sign\left(\sum_{i=1}^{n}(w_i x_i) + b\right) \tag{2.2}$$

La fonction signe, définie ci-dessous, est utilisée comme fonction d'activation.

$$sign(u) = \begin{cases} 1; \text{si } u \succ 0 \\ -1; \text{si } u \leq 0 \end{cases} \tag{2.3}$$

Le type d'apprentissage d'un perceptron est un apprentissage supervisé.

Dans cette étude, nous allons utiliser un réseau de neurones de type perceptron multicouches "MLP" (Multi-layer Perceptron), car ce dernier est le plus utilisés en reconnaissance de formes [KAB, GUE,2005]. Le MLP est un réseau de neurones composé de plusieurs couches successives et chaque couche est connectée à la suivante. Il comporte en général une couche d'entrée, une couche de sortie et une ou plusieurs couches dites cachées. La fonction d'activation des neurones de ce réseau est la fonction sigmoïde. Dans ce type de réseau, le superviseur fournit à l'entrée un ensemble de couples (entrée, sortie désirée). (Voir figure-2.8)

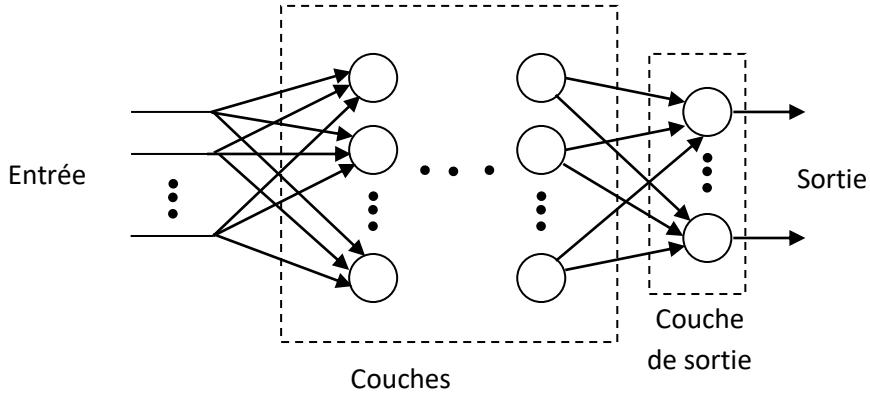


Figure-2.8. Architecture d'un perceptron multicouche MLP.

Les neurones d'entrée sont organisées en une seule couche appelée couche d'entrée sans effectuer aucune opération sur ces signaux. Les neurones qui effectuent les calculs et les traitements intermédiaires sont les neurones des couches cachées. Le théorème des approximations universelles montre que la structure élémentaire à une seule couche cachée est bien suffisante pour prendre en compte les problèmes classiques de modélisation ou d'apprentissage statistique. Il existe plusieurs méthodes pour faire l'apprentissage des réseaux de neurones (MLP), parmi elles on peut citer :

- Algorithme de rétro-propagation du gradient.
- Algorithme de gradient conjugué.
- Méthodes de second ordre.

2.3.4.3 Approche à noyaux

Parmi les méthodes à noyaux, inspirées de la théorie statistique de l'apprentissage de Vladimir Vapnik appelée "théorie de Vapnik-Chervonenkis"introduites en 1995, on trouve les Machines à Vecteurs de Support ou bien en Anglais Support Vector Machines (SVM). Les SVMssont des algorithmes d'apprentissage statistique supervisé destinées à résoudre des problèmes de classification et de régression. Elles sont utilisées pour séparer deux classes différentes ayant comme labels +1et -1 par un hyperplan de séparation [Bur98]. Cette technique de classification discriminative très populaire, est particulièrement adaptée au traitement de données de grandes dimensions, telles que ; la catégorisation automatique de textes, la reconnaissance de locuteurs, etc. [Vap95].

Le principe des SVMs consiste à projeter les données de l'espace d'entrée dans un espace de plus grande dimension appelé espace de caractéristiques, de façon à ce que les données deviennent linéairement séparables. Dans cet espace, on construit un hyperplan optimal séparant les classes tel que :

- Les vecteurs appartenant aux différentes classes se trouvent de différents côtés de l'hyperplan.
- La plus petite distance entre les vecteurs et l'hyperplan (la marge), soit maximale.

A- Notions de base des SVMs

L'objectif principal des SVMs est de trouver un classifieur linéaire qui sépare les données de deux classes d'exemples et maximise la distance entre elles. Ce classifieur est appelé Hyperplan (H). Il est évident qu'il existe une multitude d'hyperplan valide mais la condition des SVM est que cet hyperplan doit être optimal. Cela revient donc, à chercher l'hyperplan qui passe au milieu des points des deux classes d'exemples. Il s'agit donc, de chercher un hyperplan dont la distance minimale aux exemples d'apprentissage est maximale. Cette distance est appelée marge. L'hyperplan séparateur optimal est celui qui maximise la marge.Les points les plus proches de cet Hyperplan, qui seuls sont utilisés pour sa détermination, sont appelés Vecteurs de Support (VS). (Voir figure-2.9)

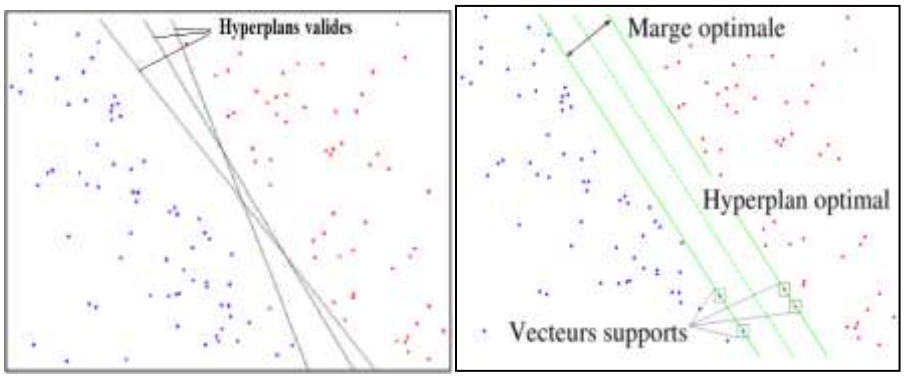
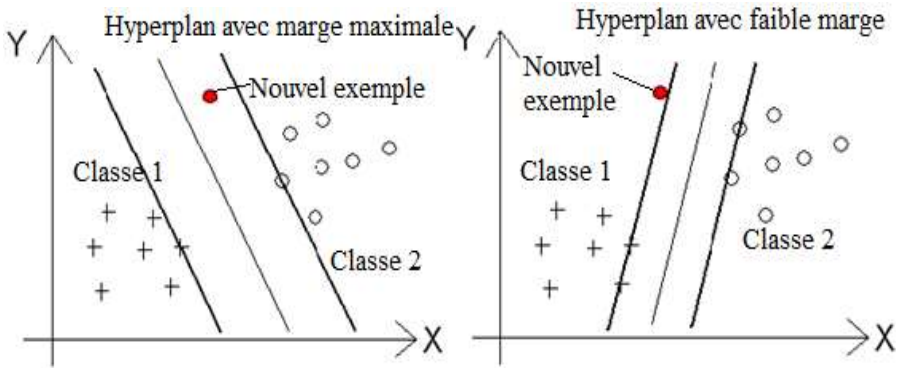


Figure-2.9. Hyperplan optimal, Marge optimale et vecteurs de support

En effet, maximiser la marge assure plus de sécurité lorsque l'on classe un nouvel exemple inconnu. La figure-2.9-a, nous montre qu'avec un hyperplan optimal (marge maximale), un nouvel exemple est bien classé bien qu'il soit dans la marge. Or dans la figure-2.9-b, on constate qu'avec une marge plus petite, l'exemple est mal classé.



(a) Marge maximale (b) Marge Faible

Figure-2.6. Utilité de la maximisation de la marge

B- Théorie de Vapnik-Chervonenkis

Vapnik-Chervonenkis ont défini deux nouvelles notions : la notion de VC-dimension et la notion de VC-entropie, à partir desquelles ils ont établi des conditions nécessaires et suffisantes à la convergence du risque empirique vers le risque réel [ELI] :

- La VC-dimension h d'un modèle d'apprentissage, est la taille maximale d'un ensemble d'exemples qui peut être pulvérisé ou séparé par le modèle ;
- La VC-entropie d'un modèle, est l'espérance de l'algorithme de la diversité de l'ensemble des fonctions que le modèle peut réaliser (du nombre de séparations différentes possibles), sur un échantillon de taille donnée.

Vapnik propose d'appliquer un nouveau principe qu'il nomme principe de Minimisation du Risque Structurel "MRS". Ce principe est basé sur la minimisation conjointe de deux sources d'erreurs : le risque empirique et l'intervalle de confiance Γ(h) (L'intervalle de confiance Γ(h) : Est une fonction croissante de la VC-dimension).

2.4 Conclusion

Dans ce chapitre nous avons présenté les différentes méthodologies suivies afin de réaliser les objectifs visés dans ce travail. Les méthodologies présentées sontcomposées de trois phases principales, la première est représentée par le prétraitement adéquat des textes utilisés dans l'attribution d'auteurs ; la deuxième est représentée par l'extraction des caractéristiques adaptées à ce type de problèmes (attribution d'auteurs) ; et la troisième phase est représentée par les approches, algorithmes etméthodes de classification (identification d'auteurs) choisies.

Dans le chapitre suivant, nous présentons l'évaluation empirique de toutes les approches et techniques proposées pour l'attribution d'auteurs en utilisant lecorpus que nous avons conçu pour cette fin.

Chapitre-3

Expériences et résultats

CHAPITRE-3

EXPERIENCES ET RESULTATS

3.1 Introduction

Dans ce chapitre nous allons exposer les séries d'expériences d'attribution d'auteur effectuées sur notre corpus qui est composé de 11 auteurs dont chacun a écrit 6 textes, d'une longueur moyenne de 2000 mots. Ces textes, numéroté de 1 à 6 et qui ont été obtenus après une opération de Reconnaissance Optique de Caractères (OCR), sont classés en trois classes ; textes corrigés, textes demi-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.

3.2 Corpus d'évaluation

3.2.1 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, demi 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**haracter **R**ecognition of **11 C**ontemporary **A**rab **W**riters "(**OCR11CAW**).

3.2.2 Constituants du Corpus

Le corpus que nous avons conçu contient 11 écrivains arabes contemporains (5 féminins et 6 masculins) qui sont : Ghada Saman, Houda Barakat, Kolite Sohil, May Ziada, Nawel Saadawi, Abbas Mahmoud Akad, Ibrahim Abdelkader Mazini, Mostapha Sadak Rafie, Taha Hocin, Tawfik Hakim et Zaki Najib Mahmoud.

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 textes d'une longueur moyenne de 2040 mots et on les classe en trois catégories ; textes corrigés, textes demi-corrigés et textes non-corrigés.

Les textes utilisés pour l'opération d'apprentissage (qui sont les textes numéros ; 1, 3, 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 2 et 4 pour chaque auteurs) on trouve les trois types de textes (c'est-à-dire on a texte-2-corrigé, texte-2-demi-corrigé et texte-2-non-corrigé et de même pour le texte-4).

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-3.1 : Récapitulatif du Corpus (Ecrivains masculins)

Ecrivains	Pays de Naissance	Période	Nombre de livres	Textes	Nombre de mots / texte	Nombre de mots Moyen	Utilisation
Abbas Mahmoud El Akad	Egypte	1889 - 1964	89 livres	Akad-1	2033	2040	Apprentissage
				Akad-2	2010		Test
				Akad-3	2023		Apprentissage
				Akad-4	2018		Test
				Akad-5	2090		Apprentissage
				Akad-6	2068		Apprentissage
Ibrahim Abdelkader El Mazini	Egypte	1890 - 1949	19 livres (4 traduits)	Mazini-1	2075	2037	Apprentissage
				Mazini-2	2037		Test
				Mazini-3	2086		Apprentissage
				Mazini-4	1979		Test
				Mazini-5	2078		Apprentissage
				Mazini-6	1972		Apprentissage
Mostapha Sadak Rafie	Egypte	1880 - 1937	9 livres	Rafei-1	2056	2008	Apprentissage
				Rafei-2	1943		Test
				Rafei-3	1983		Apprentissage
				Rafei-4	2033		Test
				Rafei-5	2022		Apprentissage
				Rafei-6	2014		Apprentissage
Taha Hocin	Egypte	1889 - 1973	47 livres (6 traduits)	Taha-1	2024	2054	Apprentissage
				Taha-2	2028		Test
				Taha-3	2097		Apprentissage
				Taha-4	2073		Test
				Taha-5	2057		Apprentissage
				Taha-6	2042		Apprentissage
Toufik El Hakim	Egypte	1898 - 1987		Toufik-1	2019	2019	Apprentissage
				Toufik-2	2008		Test
				Toufik-3	2016		Apprentissage
				Toufik-4	2031		Test
				Toufik-5	2045		Apprentissage
				Toufik-6	1993		Apprentissage
Zaki Najib Mahmoud	Egypte	1905 - 1993	21 livres	Nadjib-1	1938	1991	Apprentissage
				Nadjib-2	1975		Test
				Nadjib-3	2015		Apprentissage
				Nadjib-4	2002		Test
				Nadjib-5	2005		Apprentissage
				Nadjib-6	2008		Apprentissage

Tableau-3.2 : Récapitulatif du Corpus (Ecrivains féminins)

Ecrivains	Pays de Naissance	Période	Nombre de livres	Textes	Nombre de mots / texte	Nombre de mots Moyen	Utilisation
Chada Saman	Syrie	1942 – à ce jour	46 livres	Ghada-1	2055	2045	Apprentissage
				Ghada-2	2052		Test
				Ghada-3	2067		Apprentissage
				Ghada-4	1999		Test
				Ghada-5	2037		Apprentissage
				Ghada-6	2061		Apprentissage
Houda Barakat	Liban	1952 – à ce jour	12 livres	Houda-1	2049	2060	Apprentissage
				Houda-2	2087		Test
				Houda-3	2096		Apprentissage
				Houda-4	2038		Test
				Houda-5	2074		Apprentissage
				Houda-6	2018		Apprentissage
Koulite Sohil	Syrie	1937 – à ce jour	29 livres	Koulite-1	2036	2031	Apprentissage
				Koulite-2	1971		Test
				Koulite-3	2061		Apprentissage
				Koulite-4	2025		Test
				Koulite-5	2056		Apprentissage
				Koulite-6	2038		Apprentissage
May Ziada	Syrie	1886 – 1941	19 livres	May-1	1966	2025	Apprentissage
				May-2	2050		Test
				May-3	2024		Apprentissage
				May-4	2025		Test
				May-5	2084		Apprentissage
				May-6	2000		Apprentissage
Nawal Saadawi	Egypte	1931 – à ce jour	34 livres	Nawal-1	2042	2055	Apprentissage
				Nawal-2	2041		Test
				Nawal-3	2068		Apprentissage
				Nawal-4	2040		Test
				Nawal-5	2098		Apprentissage
				Nawal-6	2038		Apprentissage

3.2.3 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 la section (2.3.2) du 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.

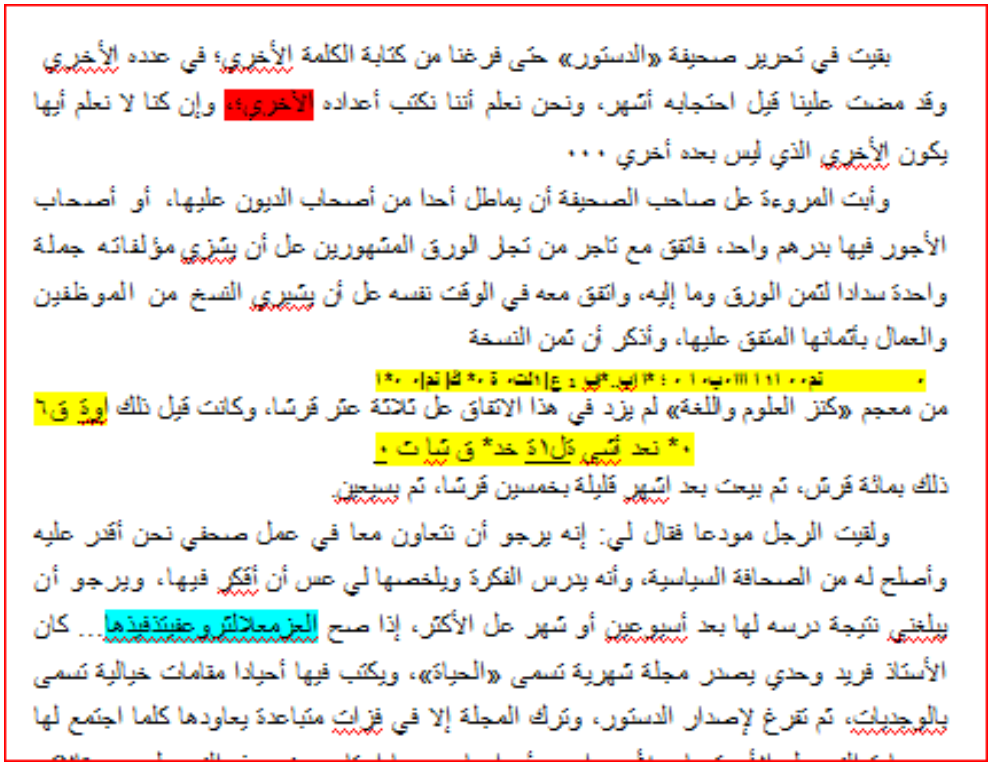
Par la suite, le corpus est divisé en deux sous-ensembles selon la règle (2/3 et 1/3) appliquée dans les bases de données ;

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

Au totale, le corpus contient 110 textes divisés comme suit ; 44 textes corrigés (pour l'apprentissage) et 22 textes corrigés, 22 textes demi-corrigés et 22 textes non-corrigés pour le test. La correction est faite manuellement afin d'obtenir le même texte original (voir section 2.3.2 du chapitre-2).

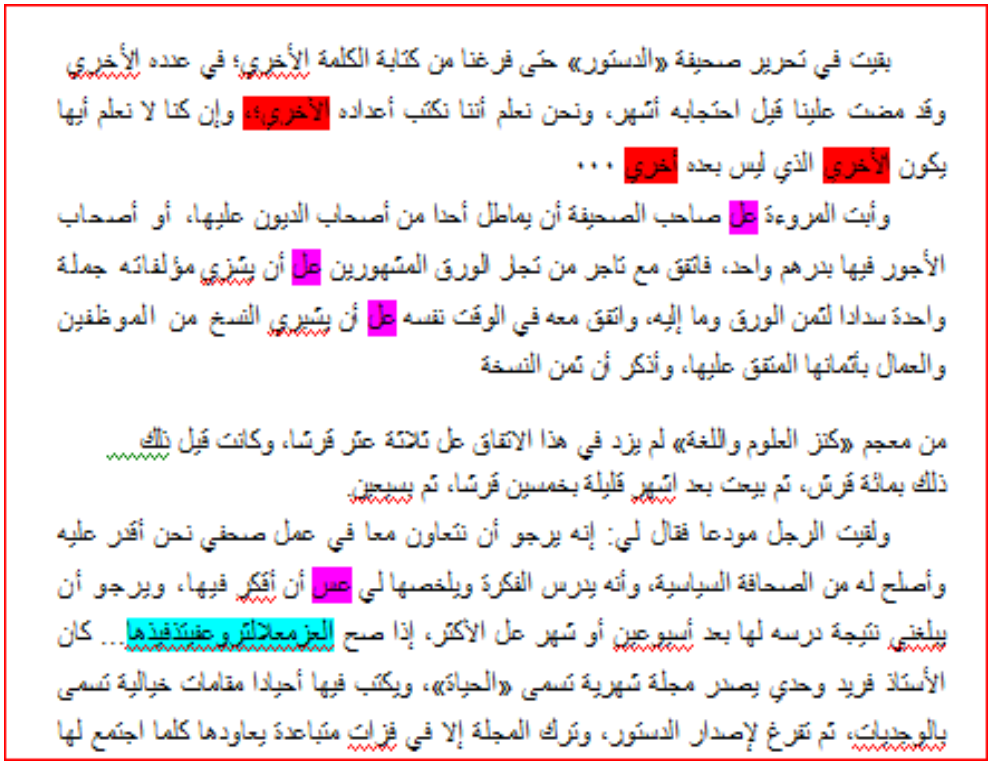
3.2.4 Exemples de textes Word obtenus après une opération OCR

Après le processus de scan des 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 prochaines expériences (chaque couleur exprime un type d'erreurs).



بقيت في تحرير صحيفة «الدستور» حتى فرغنا من كتابة الكلمة الأخيرة؛ في عدده الأخير
وقد مضت علينا قبل احتجابه أشهر، ونحن نعلم أننا نكتب أعداده الأخيرة؛ وإن كنا لا نعلم أيها
يكون الأخير الذي ليس بعده أخير ...
وأبت المروءة على صاحب الصحيفة أن يماطل أحدا من أصحاب الديون عليها، أو أصحاب
الأجور فيها بدرهم واحد، فاتفق مع تاجر من تجار الورق المشهورين على أن يشتري مؤلفاته جملة
واحدة سدادا لثمن الورق وما إليه، واتفق معه في الوقت نفسه على أن يشتري النسخ من الموظفين
والعمال بأثمانها المتفق عليها، وأذكر أن ثمن النسخة
من معجم «كنز العلوم واللغة» لم يزد في هذا الاتفاق على ثلاثة عشر قرشا، وكانت قبل ذلك إردة ق٦
ذلك بمائة قرش، ثم بيعت بعد اشهر قليلة بخمسين قرشا، ثم بسبعين.
ولقيت الرجل مودعا فقال لي: إنه يرجو أن نتعاون معا في عمل صحفي نحن أقدر عليه
وأصلح له من الصحافة السياسية، وأنه يدرس الفكرة ويلخصها لي على أن أفكر فيها، ويرجو أن
يبلغني نتيجة درسه لها بعد أسبوعين أو شهر على الأكثر، إذا صح العزم عملت على تنفيذها... كان
الأستاذ فريد وجدي يصدر مجلة شهرية تسمى «الحياة»، ويكتب فيها أحيانا مقالات خيالية تسمى
بالوجدانيات، ثم تفرغ لإصدار الدستور، وترك المجلة إلا في فترات متباعدة يعاودها كلما اجتمع لها

Figure-3.1 : Exemple de texte non-corrigé.



بقيت في تحرير صحيفة «الدستور» حتى فرغنا من كتابة الكلمة الأخيرة؛ في عدده الأخير
وقد مضت علينا قبل احتجابه أشهر، ونحن نعلم أننا نكتب أعداده الأخيرة؛ وإن كنا لا نعلم أيها
يكون الأخير الذي ليس بعده أخير ...
وأبت المروءة على صاحب الصحيفة أن يماطل أحدا من أصحاب الديون عليها، أو أصحاب
الأجور فيها بدرهم واحد، فاتفق مع تاجر من تجار الورق المشهورين على أن يشتري مؤلفاته جملة
واحدة سدادا لثمن الورق وما إليه، واتفق معه في الوقت نفسه على أن يشتري النسخ من الموظفين
والعمال بأثمانها المتفق عليها، وأذكر أن ثمن النسخة
من معجم «كنز العلوم واللغة» لم يزد في هذا الاتفاق على ثلاثة عشر قرشا، وكانت قبل ذلك
ذلك بمائة قرش، ثم بيعت بعد اشهر قليلة بخمسين قرشا، ثم بسبعين.
ولقيت الرجل مودعا فقال لي: إنه يرجو أن نتعاون معا في عمل صحفي نحن أقدر عليه
وأصلح له من الصحافة السياسية، وأنه يدرس الفكرة ويلخصها لي على أن أفكر فيها، ويرجو أن
يبلغني نتيجة درسه لها بعد أسبوعين أو شهر على الأكثر، إذا صح العزم عملت على تنفيذها... كان
الأستاذ فريد وجدي يصدر مجلة شهرية تسمى «الحياة»، ويكتب فيها أحيانا مقالات خيالية تسمى
بالوجدانيات، ثم تفرغ لإصدار الدستور، وترك المجلة إلا في فترات متباعدة يعاودها كلما اجتمع لها

Figure-3.2 : Exemple de texte demi-corrigé.

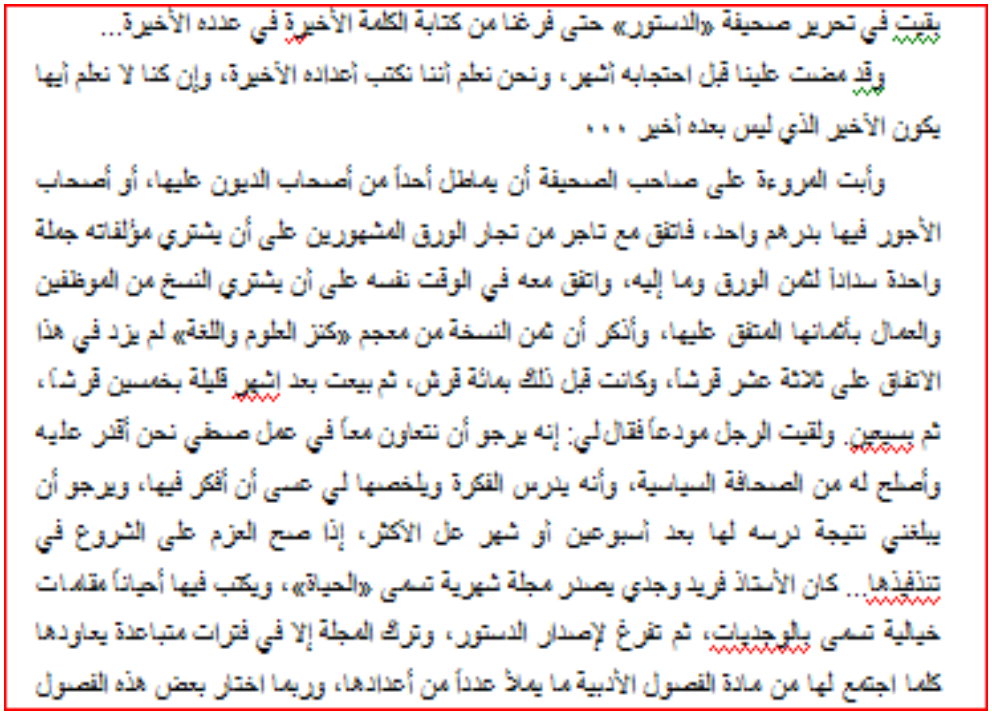

بقيت في تحرير صحيفة «الدستور» حتى فرغنا من كتابة الكلمة الأخيرة في عدده الأخيرة...
وقد مضت علينا قبل احتجابه أشهر، ونحن نعلم أننا نكتب أعداده الأخيرة، وإن كنا لا نعلم أيها يكون الأخير الذي ليس بعده أخير ...
وأبت المروءة على صاحب الصحيفة أن يماطل أحداً من أصحاب الديون عليها، أو أصحاب الأجور فيها بدرهم واحد، فاتفق مع تاجر من تجار الورق المشهورين على أن يشتري مؤلفاته جملة واحدة سداداً لثمن الورق وما إليه، واتفق معه في الوقت نفسه على أن يشتري النسخ من الموظفين والعمال بأثمانها المتفق عليها، وأنكر أن ثمن النسخة من معجم «كنز العلوم واللغة» لم يزد في هذا الاتفاق على ثلاثة عشر قرشاً، وكانت قبل ذلك بمائة قرش، ثم بيعت بعد اشهر قليلة بخمسين قرشاً، ثم بسبعين. ولقيت الرجل مودعاً فقال لي: إنه يرجو أن نتعاون معاً في عمل صحفي نحن أقدر عليه وأصلح له من الصحافة السياسية، وأنه يدرس الفكرة ويلخصها لي عسى أن أفكر فيها، ويرجو أن يبلغني نتيجة درسه لها بعد أسبوعين أو شهر عل الأكثر، إذا صح العزم على الشروع في تنفيذها... كان الأستاذ فريد وجدي يصدر مجلة شهرية تسمى «الحياة»، ويكتب فيها أحياناً مقالات خيالية تسمى بالوجديات، ثم تفرغ لإصدار الدستور، وترك المجلة إلا في فترات متباعدة يعاودها كلما اجتمع لها من مادة الفصول الأدبية ما يملأ عدداً من أعدادها، وربما اختار بعض هذه الفصول

Figure-3.3 : Exemple de texte corrigé.

3.3 Expérimentation et résultats des expériences

3.3.1 Protocole expérimental

Dans ce mémoire, la tâche d'attribution d'auteurs est effectuée en utilisant le N-grams caractère comme caractéristique et deux méthodes de classification ; MLP et 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 :

$$TAA = \frac{nombre\ documents\ correctement\ attribués}{nombre\ document\ testé} \times 100$$

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

- 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 demi-corrigés.

- Dans la troisième 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.

3.3.2 Expériences d'attribution d'auteurs

3.3.2.1 Expérience N°1 : Utilisation des textes corrigés dans le test

Cette série d'expériences vise à déterminer le nombre approprié de caractères (N-gramme) pour avoir le meilleur taux TAA en utilisant les textes corrigés pour les deux phases (apprentissage et test). Après investigation, nous avons choisis d'utiliser la méthode d'analyse des Evénements les Plus Courants (en Anglais Most Commun Events « MCE = 75 »). Les résultats obtenus de cette série d'expérience sont présentés dans les tableaux et les figures qui suivent :

Tableau-3.3 : Taux d'Attribution d'Auteurs pour les textes corrigés

Classifieurs	N=1	N=2	N=3	N=4	N=5	N=6
MLP	100	100	100	95	95	95
SVM	95	100	95	95	90	90

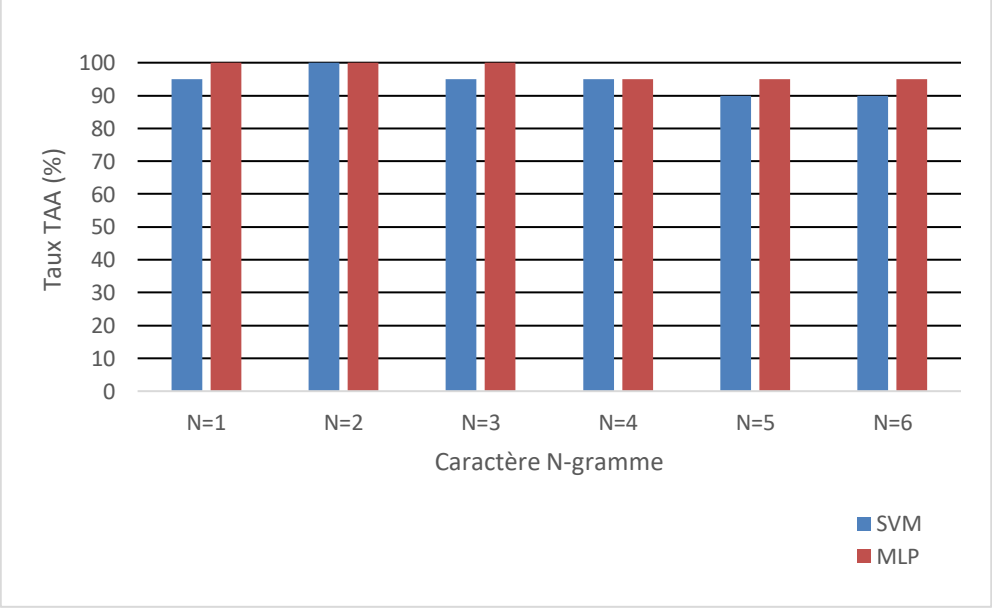


Figure-3.4. Taux d'Attribution d'Auteurs pour les textes corrigés

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 90-100% pour le SVM et 95-100% pour MLP. Ceci peut être expliqué par l'absence des erreurs de conversion entre le texte original et le texte résultant d'une opération OCR.

3.3.2.2 Expérience N°2 : Utilisation des textes demi-corrigés dans le test

Cette série d'expérience vise à déterminer le nombre approprié de caractères (N-gramme) pour avoir le meilleur taux TAA en utilisant les textes demi-corrigés pour la phase de test. Après investigation, nous avons choisis d'utiliser la méthode d'analyse des Evénements les Plus Courants (en Anglais Most Commun Events « MCE = 75 »). Les résultats obtenus de cette série d'expérience sont présentés dans les tableaux et les figures qui suivent :

Tableau-3.4 : Taux d'Attribution d'Auteurs pour les textes demi-corrigés

Classifieurs	**N=1**	**N=2**	**N=3**	**N=4**	**N=5**	**N=6**
MLP	86	95	86	90	77	81
SVM	72	81	77	81	81	86

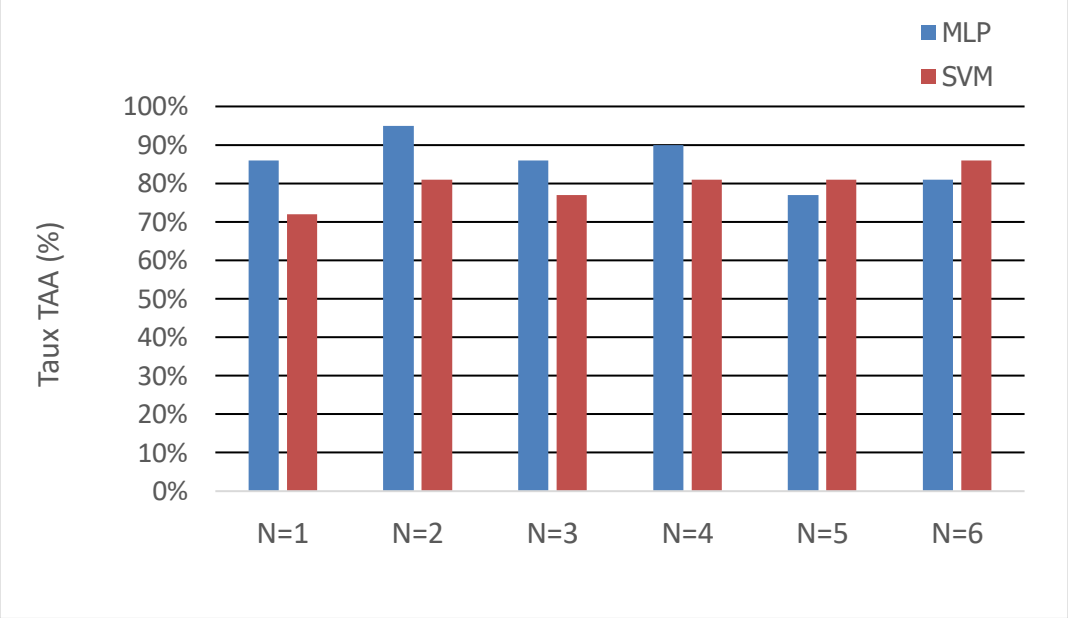


Figure-3.5. Taux d'Attribution d'Auteurs pour les textes demi-corrigés

D'après les résultats obtenus, on constate que le taux TAA de cette expérience est entre 72-86% pour le SVM et 77-95% pour MLP. 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 (demi-corrigé).

3.3.2.3 Expérience N°3 : Utilisation des textes non-corrigés dans le test

Cette série d'expérience vise à déterminer le nombre approprié de caractères (N-gramme) pour avoir le meilleur taux TAA en utilisant les textes no-corrigés pour la phase de test. Après investigation, nous avons choisis d'utiliser la méthode d'analyse des Evénements les Plus Courants (en Anglais Most Commun Events « MCE = 75 »). Les résultats obtenus de cette série d'expérience sont présentés dans les tableaux et les figures qui suivent :

Tableau-3.5 : Taux d'Attribution d'Auteurs pour les textes non-corrigés

Classifieurs	N=1	N=2	N=3	N=4	N=5	N=6
MLP	86	95	95	90	86	86
SVM	75	86	86	86	81	86

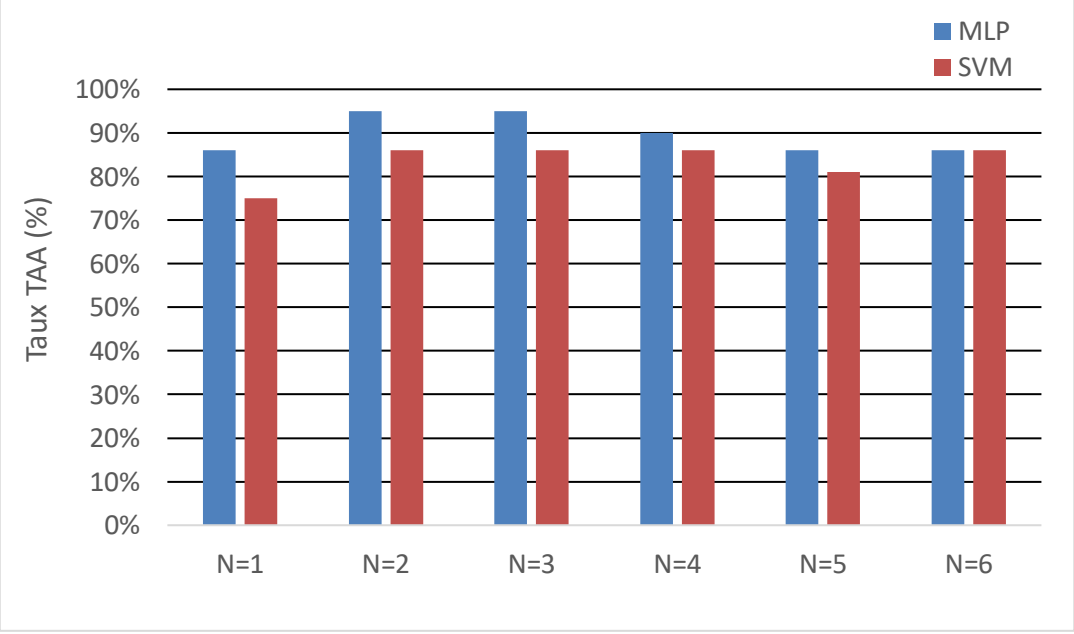


Figure-3.6. Taux d'Attribution d'Auteurs pour les textes non-corrigés

D'après les résultats obtenus, on constate que le taux TAA de cette expérience est entre 75-86% pour le SVM et 86-95% pour MLP. 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 (demi-corrigé).

3.3.3 Expériences de comparaison des performances

Cette série d'expériences vise à comparer les performances des méthodes de classification utilisées (MLP et SVM) en utilisant les différents types de textes (corrigés, demi-corrigés et non-corrigés). Nous avons choisi d'utiliser la méthode d'analyse des Evénements les Plus Courants (en Anglais Most Commun Events « MCE = 75 »).

Les résultats obtenus de cette série d'expérience sont présentés dans les tableaux et les figures suivants :

Tableau-3.6 : Performance du classifieur MLP avec les différents types de textes

Type de texte	**N=1**	**N=2**	**N=3**	**N=4**	**N=5**	**N=6**
Texte corrigé	100	100	100	95	95	95
Texte demi-corrigé	86	95	86	90	77	81
Texte non-corrigé	86	95	95	90	86	86

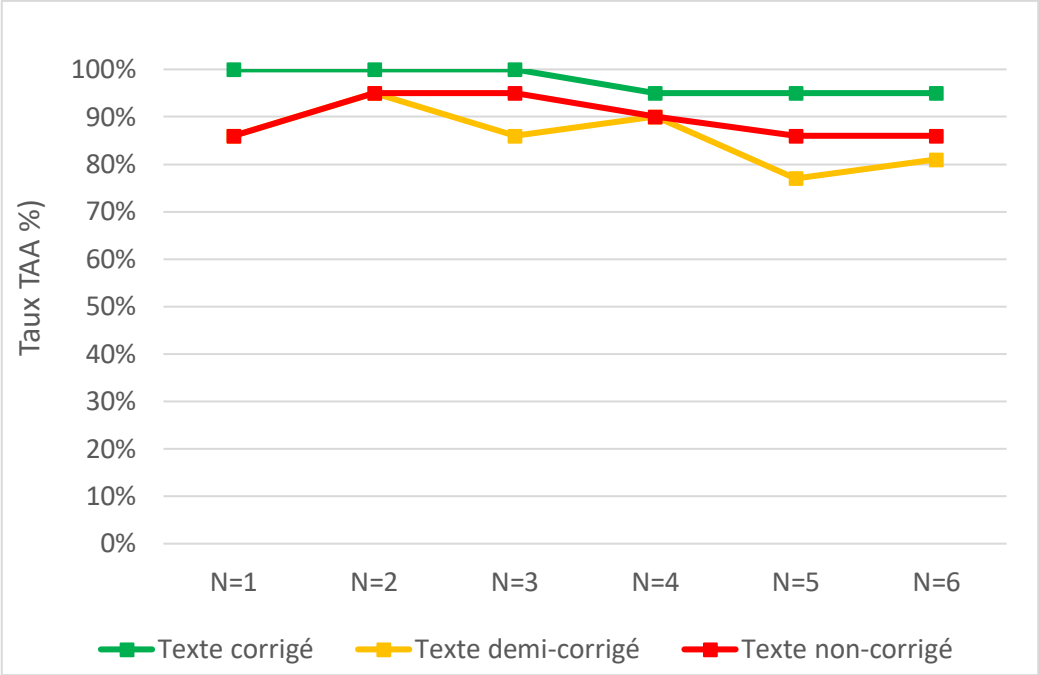


Figure-3.7. Performance du classifieur MLP avec les différents types de textes

D'après les résultats obtenus, on constate que les taux TAA sont excellents (100% avec les textes corrigés) jusqu'à N=3, ensuite on enregistre une baisse importante des taux à partir de N=4 arrivant jusqu'à 77% (pour les textes demi-corrigé). Ceci peut être interprété par le fait que le classifieur MLP est robuste pour les erreurs de conversion dans le texte résultant d'une opération OCR pour N≤3 et sensible pour N>3.

Tableau-3.7 : Performance du classifieur SVM avec les différents types de textes

Type de texte	**N=1**	**N=2**	**N=3**	**N=4**	**N=5**	**N=6**
Texte corrigé	95	100	95	95	90	90
Texte demi-corrigé	72	81	77	81	81	86
Texte non-corrigé	75	86	86	86	81	86

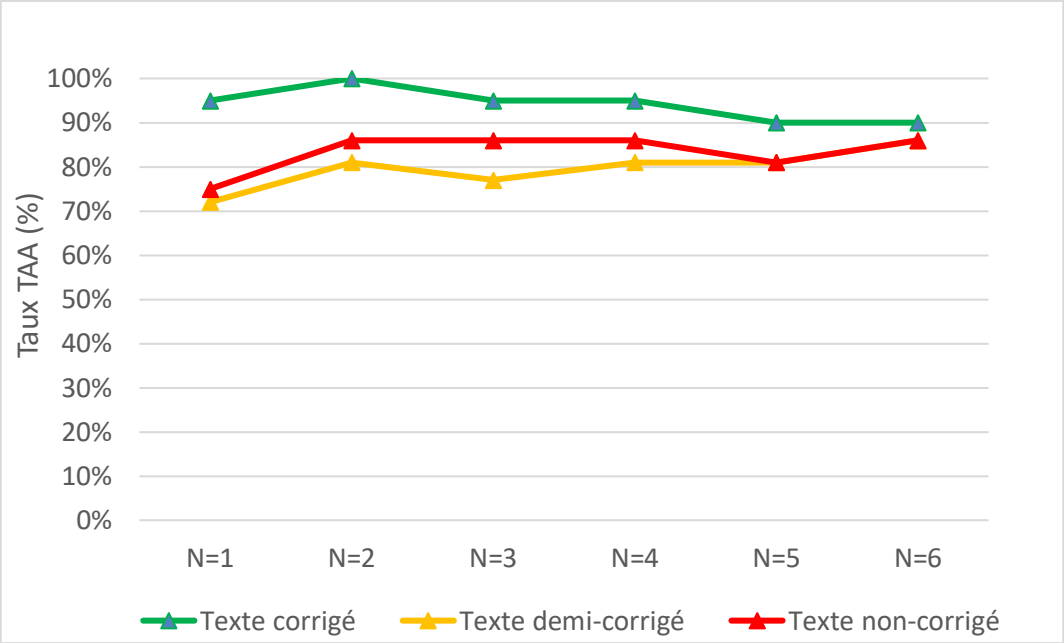


Figure-3.8. Performance du classifieur SVM avec les différents types de textes

D'après les résultats ci-dessus, on aperçoit que les taux d'attribution d'auteurs (TAA) obtenus par le classifieur SVM sont acceptables au début (N=2) avec tous les types de textes, après converge vers une un taux moins important mais stable pour les différents cas (TAA=86%) pour N=6. Ceci peut être interprété par le fait que le classifieur SVM est robuste contre les erreurs de conversion dans le texte résultant d'une opération OCR et robuste aussi pour les valeurs supérieures à N>3.

3.4 Conclusion

Dans ce chapitre nous avons effectué des expériences d'attribution d'auteurs des documents textes arabe obtenus après une opération OCR avec plusieurs degrés d'erreurs. L'évaluation expérimentale a été réalisé en utilisant une base de données qu'on conçu pour cette fin et qu'on a appelé « Optical Character Recognition 11 Arabs Contemporary Writers » (OCR11ACW). Les résultats obtenus ont montré que les erreurs de la conversion OCR influent considérablement sur le taux d'attribution TAA.

Conclusion générale

CONCLUSION GENERALE

Travail réalisé

Les techniques liées au traitement de l'information connaissent actuellement un développement très actif en lien avec l'informatique et ont un potentiel croissant dans le domaine de l'interaction homme-machine. L'homme veut communiquer avec l'ordinateur de la manière la plus simple et la plus naturelle pour faciliter et accélérer l'interaction et l'échange d'informations. Il cherche à rendre ces machines accessibles par la voix, capables de lire, voir, traiter et analyser rapidement les informations reçues.

Dans ce travail, nous avons abordé l'attribution d'auteurs des textes anonymes et bruités. En particulier, nous sommes intéressés par les textes issus d'une opération de reconnaissance optique de caractères. Pour ce type de texte, une application particulière d'identification des textes OCR a été effectuée.

Vu l'indisponibilité de corpus appropriés, pour la tâche adressée, nous avons construit un nouveau corpus que nous avons appelé : « Optical Caractère Recognition de 11 Contemporary Arabs Writers « (i.e. OCR11CAW) contenant des textes bruités issus d'un processus OCR de 11 écrivains arabes, 6 textes pour chaque auteur d'une taille moyenne d'environ 2000 mots. Pour cela, nous avons proposé et implémenté deux approches et algorithmes. Ce mémoire avait pour ambition d'étudier le style des auteurs afin de trouver le véritable auteur, en appliquant des caractéristiques telles que caractère N-grammes et des classifieurs telles que le MLP et SVM.

Dans cette recherche nous avons traité l'attribution d'auteurs des textes arabes extraits d'une image scannée à l'aide d'un logiciel OCR. Ces textes sont classés en trois catégories ; textes corrigés, textes demi-corrigées et textes non corrigées. L'originalité de ce travail réside dans la détermination de l'auteur d'un texte écrits en arabe dans la présence des effets d'acquisition optique à l'aide d'un OCR.

Résultats obtenus

Les approches proposées dans ce travail de recherche, ont été évaluées sur le corpus que nous avons conçus. Les résultats de cette évaluation ont montré que les techniques proposées sont assez intéressantes et encourageantes. Notamment, le classifieur MLP avec caractère n-grammes (n=2) ou nous avons obtenus un taux TAA = 100%. D'autre part, l'évaluation des deux techniques a montré une efficacité particulière du classifieur MLP avec les différents types de textes (Corrigés, demi-corrigés et non-corrigé). Cependant, le classifieur SVM a donné d'excellents résultats lorsque N≤3.

En perspectives

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

- Inclusion de textes poétiques,
- Appliquer cette étude aux textes issus d'un processus OCR et bruités,

REFERENCE

[Agn] La technique de la stylométrie appliquée au Livre de Mormon. Agnès Boltoukhine

[BAL, 2010]. [Baldwin et Lui 2010] 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.

[Bou,Ben, 2005] Boukharouba, A., & Bennia, A. (2005, March). Reconnaissance de caractères imprimés omni-fonte. In 3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, Tunisia

[Boz,Bag,Uya,2007] Bozkurt, I. N., Bağlıoğlu, Ö., & Uyar, E. (2007). Authorship attribution: performance of various features and classification methods. In *22nd International Symposium on Computer and Information Sciences, ISCIS 2007-Proceedings* (pp. 158-162). IEEE

[Cav, 1994] [Cavnar et Trenkle 1994] 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.

[Cla,2013] Clark, J. (2013). Text mining and scholarly publishing. Publishing Research Consortium, 2013

[ELI] Elie, L. Pattern Classification problems in Statistical Arbitrage Colin Umansky.

[GAJ,All] EL GAJOUI, K., & ALLAH, F. A. Vers un système de reconnaissance optique des caractères dans des documents multilingues: Français-Amazighe.

[KAB,GUE,2005] Kabache, M., & Guerti, M. (2005). Application des réseaux de neurones à la reconnaissance des phonèmes spécifiques à l'Arabe standard. SETIT 2005.

[Khe, 2018] S. Kennouf. 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.

[Lam ,Béc ,Ham, Roc ,2010]Laroum, S., Béchet, N., Hamza, H., & Roche, M. (2010). Classification automatique de documents bruités à faible contenu textuel. Revue des Nouvelles Technologies de l'Information, (spécial: Fouille de Données Complexes), 25.

[MED,2010] MEDJILI, F. (2010). Modélisation par Réseaux de Neurones Artificiels (RNA) et commande Prédictive non linéaire d'une station de production d'eau froide (Doctoral dissertation, Université Badji Mokhtar de Annaba).

[Mit,Ind,Div,2013] Mithe, R., Indalkar, S., & Divekar, N. (2013). Optical character recognition. *International journal of recent technology and engineering (IJRTE)*, *2*(1), 72-75

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

[Ouk,2012] Oukacine, N. (2012). Utilisation des réseaux de neurones pour la reconstitution de défauts en évaluation non destructive (Doctoral dissertation, Université Mouloud Mammeri).

[Pew, 2015] http://www.pewresearch.org/

[Sha, 2010]. [Shaalan 2010] K. Shaalan, Rule-based Approach in Arabic Natural Language Processing, International Journal on Information and Communication Technologies, Vol. 3, No. 3, June, 2010, pp. 11-19.

[Sta, 2009] 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.

[Tal,Hna,Aye,Fat,2016] Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text mining: techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, *7*(11), 414-418

[Xia, 2010] [Xia et al. 2010] 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.

[ZEM,2003] Zemouri, R. (2003). Contribution à la surveillance des systèmes de production à l'aide des réseaux de neurones dynamiques: Application à la e-maintenance (Doctoral dissertation, Université de Franche-Comté).

ملخص

أدى التطّور التكنولوجي لمختلف وسائل الاتصال الرقمي إلى زيادة هائلة في مختلف أنواع ملفات الملتيميديا (الصوتية، الفيديو، النصيّة) عبر مختلف الوسائل. تشير الإحصائيات إلى أن الملفات النصية المكتوبة هي الأكثر تداولا مقارنة مع بقية أنواع ملفات الملتيميديا. هذا التطور الضخم والكم الهائل للنصوص المكتوبة نتج عنه صعوبة بالغة في البحث واستخراج المعلومات منها وخاصة التعرف على الكتاب الأصليين لها.

في هذا العمل البحثي قمنا بدارسة أسلوب مجموعة من الأدباء المعاصرين العرب من خلال مؤلفاتهم، بغرض اسناد نصوص أدبية مجهولة لأصحابها، هذه النصوص قد تم الحصول عليها باستعمال برنامج التعرف الضوئي على الحروف (OCR). في هذه الدراسة قمنا بإنشاء قاعدة بيانات جديدة لهذا الغرض كما قمنا باقتراح خوارزميات إحصائية لحل مشكلة التصنيف الأتوماتيكي للمؤلفين والتعرف على الكتاب الأصليين.

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

Résumé

Le développement technologique des divers moyens de communication numérique a conduit à une augmentation considérable des différents types de fichiers multimédias (audio, vidéo, texte) à travers plusieurs moyens. Les statistiques indiquent que les fichiers texte écrits sont les plus couramment utilisés par rapport aux autres types de fichiers multimédias. Ce développement extraordinaire et cette énorme quantité de textes écrits ont entraîné de grandes difficultés pour rechercher et en extraire des informations, en particulier l'identification des auteurs originaux.

Dans ce travail de recherche, nous avons étudié le style d'écriture d'un groupe d'écrivains arabes contemporains à l'aide de leurs livres, dans le but d'attribuer des textes littéraires anonymes à leurs propriétaires, ces textes ont été obtenus à l'aide du programme de reconnaissance optique de caractères (OCR). Dans cette étude, nous avons créé une nouvelle base de données à cet effet, et nous avons proposé des algorithmes statistiques pour résoudre le problème de la classification automatique des auteurs et l'attribution des auteurs originaux.

Mots clés : Attribution d'auteurs, langue Arabe, Reconnaissance Optique des Caractères (OCR).

Abstract

The technological development of various means of digital communication has led to a considerable increase in the different types of multimedia files (audio, video, text) through several means. Statistics indicate that written text files are the most commonly used compared to other types of media files. This extraordinary development and enormous quantity of written texts has created great difficulties in searching for and extracting information, in particular the identification of the original authors.

In this research work, we studied the writing style of a group of contemporary Arab writers using their books, with the aim of attributing anonymous literary texts to their owners, these texts were obtained using the optical character recognition (OCR) program. In this study, we created a new database for this purpose, and we proposed statistical algorithms to solve the problem of automatic classification of authors and attribution of original authors.

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