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°: … 14 ... /STLC/ 2022

DOMAINE: SCIENCES ET TECHNOLOGIE
FILIERE: TELECOMMUNICATIONS
OPTION: SYSTEMES DE TELECOMMUNICATIONS

Mémoire présenté pour l'obtention
du diplôme de Master Académique

Par : CHABBI Imane et REBANI Samia

Intitulé

Système d'Authentification d'Auteurs des Textes Arabes basé sur la SMO-SVM et la distance de Manhattan

Soutenu publiquement le : 22 /06/ 2022 devant le jury composé de:

Dr. BENNACER Hamza	Université M'sila	Président
Dr. KHENNOUF Salah	Université M'sila	Encadreur
Dr. KENANE El Hadi	Université M'sila	Examinateur

Année universitaire : 2021 /2022

بسم الله الرحمن الرحيم

Remerciements

On remercie tout d'abord Allah qui nous a donné la force pour que nous puissions terminer ce travail.

Nous adressons nos remerciements les plus sincères, à notre encadreur Dr. KHENNOUF Salah, à qui nous devons beaucoup, pour son attention, sa disponibilité, ses conseils et sa sympathie.

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.

Enfin, nous tenons à remercier toutes les personnes qui nous ont conseillé lors de la rédaction de ce travail.

I. CHABBI et S. REBANI

Dédicace

Je dédie ce travail à:

Mes chers parents notamment mon père qui m'a beaucoup soutenu et ma mère; Source de vie

mes frères et mes sœurs.

Mes amies et un merci spécial à mon amie Bouthaina et mon fiancé pour toujours être avec moi merci à tous.

Tout ce qui porte le nom de REBANI.

Samia

Dédicace

Du fond du cœur, je dédie ce travail à ma tante et à mes parents. Aucune dévotion ne peut exprimer mon respect, mon amour éternel et ma gratitude pour les sacrifices qu'ils ont consentis pour mon éducation et ma santé…

Je vous remercie pour tout le soutien et l'amour que vous m'avez donné depuis mon enfance, et j'espère que votre bénédiction sera toujours avec moi.

Que cet humble travail soit l'accomplissement de vos désirs bien formulés, le fruit de vos innombrables sacrifices…

Qu'ALLAH le tout puissant vous accorde santé, bonheur et longue vie …

Je dédie ce travail aussi à tous mes amies, mes sœurs et mes frères…

Imane

Liste des Abréviations

A

AA : d'Attribution d'Auteurs

ACP: l'analyse en composant principale

AHP: hiérarchie analytique Procès

L

LDA: l'analyse discriminant linéaire

LR: régression linéaire

M

ML: Machine Learning.

MLP: Perceptron multisource linear.

S

SMO: Sequential Minimal Optimization

SVM: support vector machine

Liste des Tableaux

N° Tableau	*Titre du tableau*	*Page*
Tableau 3.1	Récapitulatif du Corpus (Ecrivains féminins) ……...…...	35
Tableau 3.2	Récapitulatif du Corpus (Ecrivains masculins) ……..……..	36
Tableau-3.3	Taux d'attribution d'auteurs pour les textes apprentissage et test corrigé avec classifieur liner SVM ……...	39
Tableau-3.4	Taux d'attribution d'auteurs pour les textes apprentissage et test corrigé avec classifier liner SVM caractères (n-gramme) ……	40
Tableau-3.5	Taux d'attribution d'auteurs pour les textes apprentissage et test corrigé avec classifier WEKA SMO caractères (n-gramme) ……...	41
Tableau-3.6	Taux d'attribution d'auteurs pour les textes apprentissage et test corrigé avec classifier WEKA SMO WORD (n-gramme) ……...	42

Liste des figures

N° figure	*Titre de figure*	*Page*
Figure 1.1	Une architecture typique pour la tâche d'attribution de la paternité. (a) approches basées sur les instances, tandis que (b) approches basées sur les profils……..	9
Figure 1.2	Processus typique d'attribution de la paternité basé sur l'apprentissage automatique. La phase de réduction entourée de pointillés est une étape facultative qui dépend de la complexité des dimensions de l'espace…… ……..……..	14
Figure 2.1	Les deux multiplicateurs…….…….	23
Figure 2.2	Apprentissage supervisé d'une machine…….……	24
Figure 2.3	Hyperplan qui sépare les deux ensembles de point…….……	25
Figure 2.4	L'hyperplan avec vecteurs de support ……..……...……	25
Figure 2.5	SVM- L'hyperplan séparateur optimal et la marge ……..……...	26
Figure 2.6	les modèles des SVM ……..	27
Figure 2.7	Exemple de changement de l'espace de données…….....	28
Figure 2.8	Hyperplan de séparation linéaire pour des données linéairement séparables (Les vecteurs de support sont encerclés) ……..……	28
Figure 2.9	Hyperplan séparateur entre 2 classes……..…….	29
Figure 2.10	Distance Manhattan ……	30
Figure 3.1	Exemple de texte corrigé(femme) …….….	37
Figure 3.2	Exemple de texte corrigé(homme) …….	38
Figure 3.3	Taux d'attribution d'auteurs pour classifieur liner SVM …….....	39
Figure 3.4	Taux d'attribution d'auteurs pour classifieur liner SVM ……...….	40
Figure 3.5	Taux d'attribution d'auteurs pour classificateur weka SMO …….…….	41
Figure 3.6	Taux d'attribution d'auteurs pour classifieur weka SMO…….....	42

Table des matières

Remerciements…… i

Dédicace……...……..……. ii

Liste des abréviations…….…… iv

Liste des tableaux……...……. v

Liste des figures……..…… vi

Table de matières…… vii

Introduction générale……..……. 1

CHAPITRE -1

GENERALITES SUR L'ATTRIBUTION D'AUTEUR

1.1 INTRODUCTION ……..…… 6

1.2 HISTORIQUE D'ATTRIBUTION D'AUTEUR ET STYLOMETRIE …...…….…… 6

1.3 ATTRIBUTION D'AUTEUR (AA) ……..…… 7

1.3.1 Etapes d'attribution de l'autour……..…… 9

1.4 ÉTAT DE L'ART…….…….. 10

1.4.1 Définitions des traits ……..…..… 10

1.4.2 Représentation des textes et des auteurs fondée sur les traits …… 10

1.4.3 Catégorisation de textes fondée sur les traits ……..…. 11

1.5 ATTRIBUTION DE L'AUTEUR EN ARABE ……...… 11

1.5.1 Caractéristiques arabe ……..… 12

1.5.2 Défis en contexte arabe …….…….. 12

1.5.3 Les types de la langue arabe ……..…..…… 13

1.6 L'APPRENTISSAGE AUTOMATIQUE …….……..…. 13

1.7 LES CARACTERISTIQUE LEXICALES …….…..... 14

1.8 CLASSIFICATION DES DONNES ……...…… 15

1.8.1 But de la classification …….…... 15

1.9 SYSTEME D'AUTHENTIFICATION …….……..…. 16

1.9.1 Définition et enjeux de l'authentification …….……..…. 16

1.9.2 Les enjeux de l'authentification …….……..... 17

1.9.3 Autre domain d'utilisation d'analyse d'authentification d'auteur…… 17

1.9.3.1 L'enquête policière …… 17

1.10 DEFINITION DE PLAGIAT…… 18

1.10.1 Les différentes formes de plagiat en recherche…… 18

1.11 CONCLUSION…… 19

CHAPITRE-2

APPROCHES PROPOSEES

2.1 INTRODUCTION …… 21

2.2 APPROCHES PROPOSEES POUR L'ATTRIBUTION D'AUTEURS …… 21

2.2.1 L'optimisation minimale séquentielle (SMO)…… 22

2.2.1.1 Algorithme …… 22

2.2.2 Machine à vecteurs de support (SVM) …… 24

2.2.2.1 SVM principe de fonctionnement général …… 24

a- Notions de base : Hyperplan, marge et support vecteur …… 24

b-Linéarité et non-linéarité …… 27

2.2.2.2. SVM linéaire …… 28

2.2.3. Manhattan distance …… 30

2.2.3.1 Bases de la distance de Manhattan …… 30

2.3 CONCLUSION …… 31

CHAPITRE- 3

EXPERIENCES ET RESULTATS OBTENUS

3.1 INTRODUCTION …… 33

3.2 CORPUS D'EVALUATION …… 33

3.2.1 Conception du corpus d'évaluation …… 33

3.2.2 Préparation des documents du corpus d'évaluation…… 37

3.3 LES RESULTATS DES EXPERIENCES …… 38

3.4 EXPERIENCES D'ATTRIBUTION D'AUTEURS ……..…… 38

3.4.1 Attribution d'auteurs par la méthode liner SVM ……...……. 38

3.4.2 Attribution d'auteurs par la méthode WEKA-SMO …….……..…. 40

3.5 CONCLUSION…….…..…….. 43

INTRODUCTION GÉNRÉALE

Introduction générale

L'attribution d'auteur est destinée à identifier l'auteur original d'un texte non vu. L'idée est fondamentalement énoncée de cette façon : chaque auteur a un ensemble de caractéristiques qui distinguent son style d'écriture de tous les autres. Bien que le style d'écriture d'un auteur puisse varier selon le sujet, certains styles et habitudes d'écriture durables et indisciplinés continueront de s'appliquer au fil du temps. Les auteurs de textes anonymes peuvent être identifiés en faisant correspondre les styles d'écriture observés à un candidat du groupe d'auteurs. Plusieurs approches ont été proposées pour résoudre ce problème depuis le 19ème siècle.

Les premières méthodes ont un arrière-plan statistique où la longueur et la fréquence des mots, des caractéristiques et des phrases sont utilisées pour décrire le style d'écriture. Ces méthodes sont souvent basées sur l'expérience humaine, et les applications couvrent également les textes littéraires, religieux et juridiques. Au cours des années 1960 et 1990, les méthodes et les applications ont changé pour couvrir de nouveaux problèmes complexes tels que l'attribution du code source, la détection du spam et l'usurpation d'identité. L'approche à l'époque consistait à définir le style d'écriture en extrayant certains traits du texte. Bien que les méthodes statistiques soient bonnes pour déterminer l'auteur de longs documents, elles souffrent lorsque la longueur du texte étudié est courte.

Plus récemment, les recherches actuelles auxquelles les auteurs attribuent profitent de l'explosion dans le domaine de l'apprentissage automatique, où la tâche peut être considérée comme un problème de classification à une seule étiquette et à plusieurs classes. Essentiellement, l'approche d'apprentissage automatique résout le problème en attribuant des catégories aux étiquettes des échantillons de texte. En consultant la littérature, nous avons trouvé un grand nombre de méthodes et de méthodes pour résoudre le problème, telles que les machines à vecteurs de support (SVM) [NNDT Nearest Neighbor Decision Trees] Bien que les méthodes de clustering aient montré de bonnes performances dans l'amélioration des résultats d'apprentissage automatique, peu de telles études utilisent dans le domaine de l'attribution des auteurs.

L'arabe est la langue maternelle de plus de 250 millions de personnes, vivant principalement sur deux continents différents. Cependant, il y a moins d'œuvres signées en arabe qu'en anglais et dans d'autres langues. Par conséquent, cette étude vise à enregistrer les auteurs de textes arabes basés sur SMO_SVM en utilisant la distance de Manhattan et à fournir quelques suggestions pour choisir les meilleurs paramètres pour la taille de l'ensemble de données afin d'améliorer la précision de l'ensemble de données.

❖ ***Nos objectifs***

Dans ce travail de recherche, nous visons à mener une étude et une analyse sur la performance des techniques d'identification de l'auteur de documents écrits. Textes arabes corrigés Connaître la capacité du programme à identifier les textes volés Pour cela, plusieurs descripteurs seront utilisés pour modéliser le style de chaque auteur et Classificateur SMO-SVM, liner SVM et connaissance des meilleures technologies.

Le travail présenté dans ce mémoire entre dans le cadre générale d'attribution de l'auteurs des textes arabes. Dans cette étude on s'est fixé les objectifs principaux suivants :

- ✓ Démontrer l'identification de l'auteur du texte court arabe et pour examiner la performance d'algorithmes d'apprentissage de la Classificateur SMO-SVM et liner SVM avec distance Manhattan sur le jeu de données de langue arabe.
- ✓ Nous montrons l'effet bénéfique de la N-gramme redondance sur les méthodes utilisant des représentations-vecteur de texte.
- ✓ Concevoir une base de données textuelle pour valider les techniques que nous utilisons est venu avec.

❖ ***Structure de la thèse***

Ce mémoire est organisé comme suit : Le premier chapitre c'est expliquer pour la généralité sur le système d'identification d'auteurs et ainsi que les différents types de systèmes la stylométrie et l'attribution de textes, la langue arabe.

Le deuxième chapitre c'est expliquer pour les différents approches et techniques proposées pour l'attribution d'auteurs (SMO-SVM, liner SVM, avec distance Manhattan).

Le troisième et dernier chapitre nous exposerons une série d'expériences dans lesquelles l'attribution d'auteur sur une base de données textuelle (ou corpus) que nous avons conçue à cet effet.

Enfin, ce mémoire est achevé par une conclusion générale contenant un résumé du travail réalisé dans ce mémoire, les discussions ainsi que les explications possibles des résultats obtenus et enfin quelques suggestions.

CHAPITRE-1

GENERALITES SUR

L'ATTRIBUTION D'AUTEUR

CHAPITRE -1

GENERALITES SUR L'ATTRIBUTION D'AUTEUR

1.1 INTRODUCTION

L'attribution de l'auteur est un processus d'identifier l'auteur d'un texte anonyme compte tenu d'ensemble prédéfini d'auteurs candidats et des échantillons correspondants de leurs textes. Tâche d'identification de l'auteur analyse le style d'écriture de chaque auteur en extrayant les traits stylo métriques du texte et les caractéristiques représenteront le style d'écriture de chaque auteur. Dans un chapitre nous présentons des historiques et généralités sur l'attribution d'auteur, ensuite les état d'art, puis les textes arabes. Enfin, on a présente des définitions et quelques types du plagiat.

1.2 HISTORIQUE D'ATTRIBUTION D'AUTEUR ET STYLOMETRIE

Un excellent survol de l'histoire de la stylométrie est celui de la première publication statistique identifiée par Holmes est que de qui « ont proposé que la longueur du mot puisse être un élément caractéristique des écrivains. D'autres chercheurs retracent les origines d'une lettre par de Morgan à un membre du clergé au sujet de la paternité de l'Evangile, suggérant que l'ecclésiastique « essaie d'équilibrer dans ton propre esprit question si ce dernier [texte] ne traite pas de mots plus longs que l'ancien [texte]. il a toujours couru dans ma tête qu'une petite dépense d'argent réglerait les questions de paternité de cette façon. Certains d'entre eux jours les écritures parasites seront détectées par ce test. **[1]**

Cette idée est au moins superficiellement plausible, en ce que les auteurs avec de grands vocabulaires peuvent généralement utiliser des mots plus longs. Malheureusement, des études, y compris celles dont on a montré que la longueur moyenne des mots n'est ni stable un seul auteur, ni ne fait de distinction entre les auteurs. Chez Smith mots (cités par Holmes), "la méthode de Mendenhall semble maintenant être si peu fiable que tout étudiant sérieux de la paternité devrait le jeter.

Depuis, de nombreuses autres statistiques ont été proposées et largement rejeté, y compris la longueur moyenne de la phrase le mot moyen longueur nombre moyen de syllabes par mot distribution des parties du discours], rapports type/token, ou d'autres mesures de "richesse du vocabulaire" comme l'indice D de Simpson ou la "richesse caractéristique" de Yule K" Aucune de ces méthodes n'a été démontrée être suffisamment distinctif ou suffisamment précis pour être fiable (pour une discussion spécifique et une évaluation fortement négative).

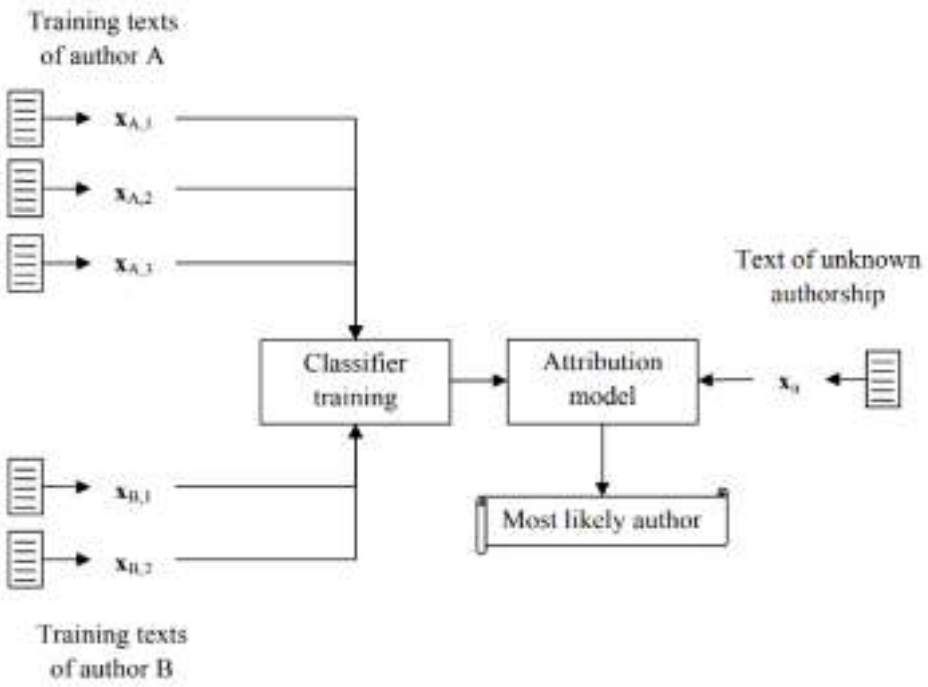
Bien qu'elle n'ait pris une réelle importance en histoire de l'art qu'à partir du xixe s. l'attribution avait déjà existé de temps en temps, au cours des siècles antérieurs, spécialement dans le milieu italien. Il suffira à cet égard de citer le " livre " de Vasari, ce recueil de dessins pour l'encadrement desquels il dessinait des éléments décoratifs, considérés comme caractéristiques du style de l'artiste auquel il attribuait le dessin. À les comparer entre elles, les attributions faites par des historiens comme Vasari, Baldinucci, Lanzi ou d'Agincourt permettent de connaitre l'idée que l'on se faisait jadis de maîtres comme Cimabue, Giotto ou Masaccio : à ce titre, elles nous intéressent surtout du point de vue de l'histoire du gout**. [2]**

1.3 ATTRIBUTION D'AUTEUR(AA)

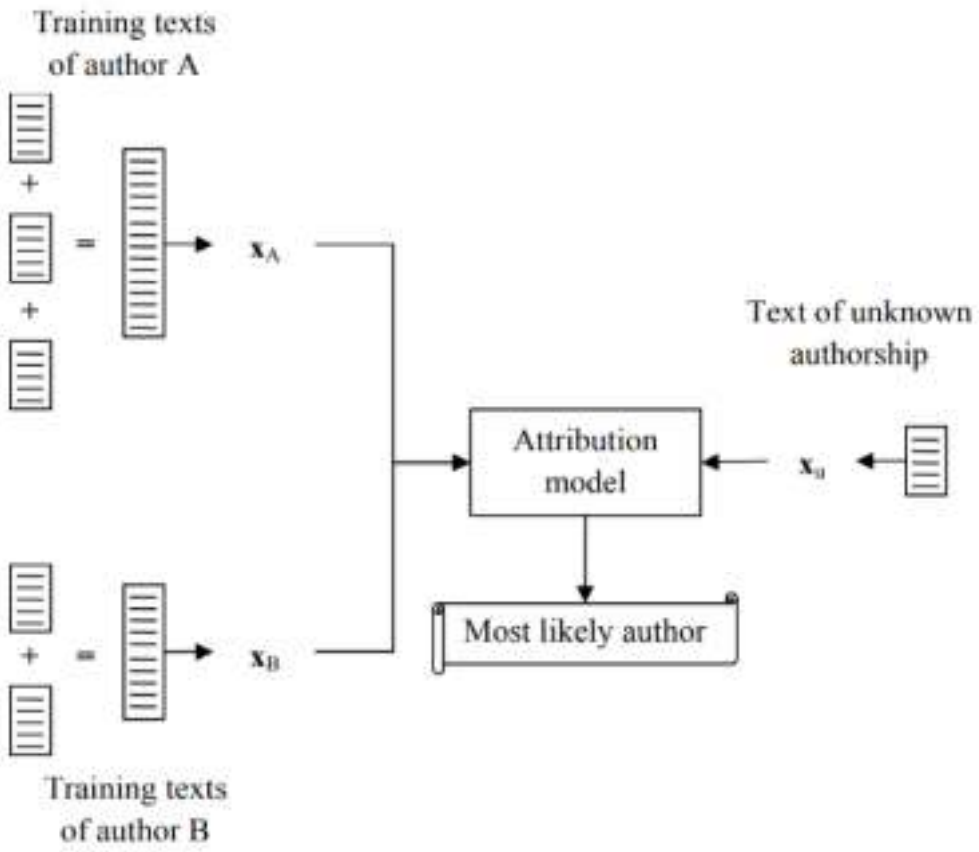
En peut définir "l'attribution de l'auteur" au sens large comme toute tentative pour déduire les caractéristiques du créateur d'un élément linguistique les données. Il s'agit d'une définition délibérément large ; par exemple, ce serait comprennent la plupart des recherches effectuées sur la reconnaissance vocale. Sur le d'autre part, de nombreuses techniques d'attribution d'auteur pourraient être appliqué à la reconnaissance vocale, et peut en fait être plus utile dans certaines circonstances, comme lorsque l'auteur du discours est la personne d'intérêt et différent de l'orateur réel. en général, l'essentiel de l'accent est mis sur le texte écrit (ou sur des aspects de l'oral texte qui sont partagés avec le texte écrit, comme le choix lexical ou structure de la phrase) plutôt que sur des aspects uniquement vocaux comme l'accent ou prosodie. **[3]**

Comme indiqué précédemment, l'attribution de la paternité peut être considérée comme un sous-domaine de l'analyse de la paternité. Il s'agit d'identifier le ou les auteurs d'un document texte anonyme en fonction des caractéristiques ou fonctionnalités du document. Dans les littératures, ces caractéristiques ou caractéristiques sont connues sous le nom de style d'écriture de l'auteur ou caractéristiques de stylo Ces caractéristiques sont extraites de manière différente en fonction de la manière dont l'algorithme AA couvre l'ensemble des échantillons. En général, ces méthodes sont classées en deux grands groupes : les approches basées sur les profils et les approches basées sur les instances. Alors que le premier groupe extrait les stylo-caractéristiques en concaténant tous les échantillons, qui appartiennent à un auteur particulier, dans l'ensemble d'apprentissage dans un gros fichier, le second groupe gère chaque échantillon dans le corpus d'apprentissage de chaque auteur.

Extrait séparément et en conséquence les caractéristiques du style d'écriture de chaque document (voir Fig. 1). De plus, le premier groupe d'approches permet de saisir les habitudes les plus persistantes et les plus incontrôlées dans le style d'écriture de l'auteur, tandis que le dernier groupe permet de détecter toute variation dans le style d'écriture. Ainsi, une combinaison des deux manières est un instrument pratique pour améliorer la précision du processus d'attribution.



(a)



(b)

FIGURE (1.1). Une architecture typique pour la tâche d'attribution de la paternité. (a) approches basées sur les instances, tandis que (b) approches basées sur les profils.

1.3.1 Etapes d'attribution de l'autour:

- ✓ Un processus complet d'attribution de l'auteur consiste en rassemblement des textes qui sont les observations à classer.
- ✓ Une 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.
- ✓ Attribuer statistiquement des textes. Des méthodes telles que l'analyse statistique bayésienne], l'analyse en composantes principales (ACP) l'analyse discriminante linéaire (LDA) et les méthodes basées sur la distance sont utilisées pour attribuer les textes. **[4]**

1.4 ÉTAT DE L'ART

L'AA est une tâche de catégorisation multi classe de textes à label unique. Comme détaillé dans Sun et al. (2012), trois les caractéristiques principales doivent être définies : la nature des traits exploités, l'ensemble des traits représentant un texte et la façon de manipuler ces représentations pour relier un texte à un auteur. **[5]**

1.4.1 Définitions des traits

Les traits utilisés en AA peuvent être séparés en différents groupes (Abbasi & Chen, 2008)

- ✓ Des valeurs numériques associées à des mots (nombre de mots dans les textes, nombre de caractères par mot, nombre de bi-grammes/tri-grammes de caractères au sein de ces mots) autrement dit des traits lexicaux ;
- ✓ Des valeurs associées à la syntaxe des phrases (effectifs des mots outils, des mono-grammes /bi-grammes/tri-grammes de ces mots outils ou des séquences de parties du discours).
- ✓ Des valeurs numériques associées à des unités plus grandes (nombre de paragraphes ou encore longueur moyenne des paragraphes), autrement dit des traits structurels ;
- ✓ Des valeurs associées avec le contenu thématique (des sacs de mots, des n-grammes de mots clefs).
- ✓ Des particularités en rapport avec les pratiques individuelles (telles que les fautes d'orthographe ou de frappe). **[5]**

1.4.2 Représentation des textes et des auteurs fondés sur les traits

Un même trait peut être attribué à plusieurs paires (texte, auteur) mais chaque texte et auteur ne partagent pas pour autant un grand ensemble de traits. Différents ensembles de traits peuvent être définis pour représenter des textes (et par extension, pour représenter des auteurs). Considérant les méthodes d'AA existantes, deux catégories principales de traits peuvent être définies :

- ✓ Les traits hors-hors-ligne: traits a priori considérés pertinents pour cette tâche avec une connaissance préalable, comme ceux largement décrit par Chaski (2001). Ils peuvent être définis quand le corpus à traiter n'est pas encore collecté.

- ✓ Les traits en ligne: traits définis pendant le traitement (dans le cas de méthodes supervisées, en fonction des corpus d'entraînement et de test, comme le modèle de langue de caractères décrit par Peng et al. (2003)). Ils ne peuvent être définis que lorsque le corpus à traiter est complet. **[5]**

1.4.3 Catégorisation de textes fondée sur les traits :

Différentes techniques pour exploiter les traits extraits des textes ont été proposées. SVM (Support Vector Machine ou Séparateur à Vaste Marge) et les réseaux de neurones (neural network) sont des approches efficaces pour mener la tâche d'AA suivant le paradigme d'apprentissage automatique supervisé (Kacmarcik & Gamon, 2006; Tweedieet al., 1996).

Quand l'ensemble des auteurs candidats est extrêmement grand ou incomplet, d'autres approches comparent les textes comme des ensembles de traits avec des fonctions spécifiques pour calculer les similarités entre ces ensembles (Kop-pel et al., 2011). D'autres approches utilisent des ensembles de traits individuels via apprentissage automatique pour construire un classifieur par auteur. chaque classifieur agit tel un expert pour traiter un sous-ensemble de l'espace de recherche lors de la classification d'un corpus, chaque classifieur étant spécialisé dans la détection d'un auteur spécifique.

Les expériences décrites dans cet article utilisent un unique classifieur SVM pour l'ensemble des auteurs en gardant les mêmes paramètres pour chaque expérience, en vue d'analyser finement l'influence du choix des traits sur le traitement. cette analyse sur les traits est alors en principe valide, même pour d'autres méthodes se basant sur ces mêmes traits. **[5]**

1.5 ATTRIBUTION DE L'AUTEUR EN ARABE

Le problème d'attribution de la paternité dans des langues telles que l'anglais, l'espagnol et le chinois est assez bien étudié. Cependant, le problème d'attribution de l'auteur sur les contextes des textes arabes a reçu beaucoup moins d'attention .nous présentons quelques problèmes qui ont un impact direct sur les AA dans le contexte de l'arabe. Certains défis qui compliquent les travaux des chercheurs en arabe sont mis en évidence. Ensuite, nous présentons un examen plus approfondi des travaux récents sur l'attribution de la paternité arabe qui couvre la période de 2005 à 2018. **[3]**

1.5.1 Caractéristiques arabe

Du point de vue morphologique, l'arabe est une langue très riche. la nature et la structure des mots arabes font de l'arabe une langue très fortement dérivée et ineffective .de plus, les structures composées des mots arabes ajoutent plus de complexité/défis, en particulier pour les tâches de traduction automatique où les mots doivent syntaxiquement être considérés comme des phrases.

Plutôt que des mots isolés. L'orientation de l'écriture en arabe, comme on l'appelle, est de droite à gauche et les lettres sont reliées les unes aux autres, ce qui fait que l'écriture arabe diffère nettement de toutes les autres langues basées sur le latin comme l'anglais, le français, etc. **[3]**

En arabe, il existe un ensemble assez restreint de préfixes et de suffixes productifs, cependant, le nombre de mots produits possibles est très élevé. Dans de nombreux cas, il suffit de changer la position de la lettre ou son diacritique2 pour produire un nouveau mot. Bien que l'inflexion et les signes diacritiques augmentent le nombre de mots, l'extraction de caractéristiques stylométriques telles que les mesures de richesse du vocabulaire pourraient influencer.

1.5.2 Défis en contexte arabe

L'arabe est une langue très riche et exigeante. Comme indiqué ci-dessus, l'arabe est une langue très dérivée et ineffective De ce fait, plusieurs défis doivent être relevés avant de travailler sur la tâche d'attribution de la paternité : signes diacritiques, caractéristiques morphologiques, structure et orientation de l'écriture, allongement, longueur des mots et sens des mots. **[3]**

✓ Les signes diacritiques sont des marques spéciales placées au-dessus ou au-dessous des mots. Les signes diacritiques jouent un rôle essentiel dans la représentation des voyelles courtes et dans la modification du sens et de la prononciation des mots.

✓ Caractéristiques morphologiques, l'une des caractéristiques distinctives de l'arabe est un certain nombre de mots produits à partir d'une racine commune. Un tel processus est connu sous le nom d'inflexion où le mot est dérivé en ajoutant des affixes (préfixes, infixes et suffixes Les mots arabes, en général, sont regroupés en quatre groupes : mot, morphème, racine et racine.

- ✓ Structure et orientation de l'écriture : En arabe, les phrases s'écrivent de droite à gauche, pas de majuscules, la forme d'une lettre est modifiée en fonction de sa position dans la phrase.
- ✓ Allongement, pour souligner un sentiment ou une signification, des tirets spéciaux sont insérés entre deux lettres. En plus de cela, ces tirets jouent un rôle stylistique.

1.5.3 Les types de la langue arabe

L'arabe a traversé une histoire de plusieurs variantes :

a) Arabe Littéraire Ancien (ALA) : C'est la langue de la poésie préislamique, Trouvé dans un nombre limité de documents aujourd'hui.

b) Arabe littéraire classique (CLA) : Ce genre représente une autre étape de l'évolution Langue. Il est apparu avec la naissance de l'Islam. Cet usage arabe évolué Les règles de base de la langue du Coran, avec l'ajout d'une grammaire considérée comme une norme idéale.

c) Arabe standard moderne (MSA) : une forme légèrement différente de l'arabe Le classique est la langue écrite de tous les pays arabophones. L'ASM est toujours langue du journalisme et de la littérature, tandis que l'arabe classique appartient Le domaine religieux, pratiqué par le clergé. **[6]**

1.6 L'APPRENTISSAGE AUTOMATIQUE:

Est une branche de l'intelligence artificielle qui s'intéresse à l'apprentissage de systèmes informatiques directement à partir d'exemples, de données et d'expériences.

L'objectif de l'application des méthodes d'apprentissage automatique dans la tâche AA est de construire un vecteur de caractéristiques extraites du corpus de texte d'apprentissage, puis de construire un classifieur qui peut attribuer des textes anonymes sur le corpus de test. La figure (1.2) montre un apprentissage automatique typique basé sur un processus d'attribution de la paternité. **[3]**

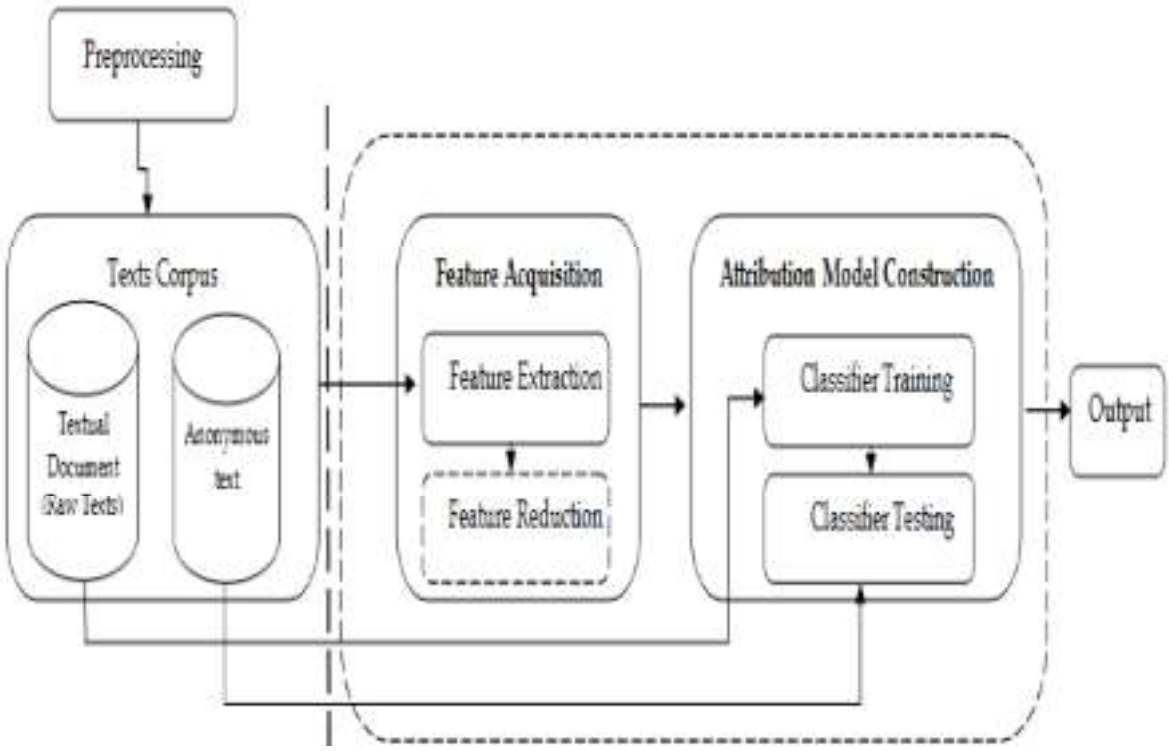


Figure (1.2). Processus typique d'attribution de la paternité basé sur l'apprentissage automatique. La phase de réduction entourée de pointillés est une étape facultative qui dépend de la complexité des dimensions de l'espace.

Le processus d'attribution de la paternité commence par la construction d'un vecteur de caractéristiques extraites du texte considéré. le but de cette étape est d'extraire des traits de "style d'écriture" qui sont des caractéristiques internes du texte. En examinant les études d'attribution de la paternité, ces caractéristiques peuvent être classées en : lexical, caractère, syntaxique, sémantique, spécifique au contenu, structurel et spécifique à la langue.

1.7 LES CARACTERISTIQUE LEXICALES

Les caractéristiques lexicales sont l'une des caractéristiques les plus couramment utilisées pour attribuer la paternité .Ces caractéristiques peuvent être extraites d'un texte en segmentant le texte en une liste de mots, de phrases, de chiffres et même de signes de ponctuation. En effet, dans le cas de l'application des caractéristiques lexicales, les résultats de AA dépendent de la capacité du tokenizer à détecter les limites des mots et des phrases1. **[3]**

- ✓ Caractère, les traits de caractère peuvent être considérés comme un sous-ensemble de traits lexicaux où le contenu du texte est traité comme une séquence de caractères. Les caractéristiques des caractères dépendent partiellement de la langue, ce qui signifie que les

caractéristiques telles que les caractères majuscules et minuscules ne peuvent pas être prises en compte, par ex. Arabe.

- ✓Syntactique, d'un texte à l'autre, l'auteur peut avoir tendance à utiliser inconsciemment des schémas syntaxiques similaires. Ces modèles peuvent constituer une empreinte auctoriale plus fiable que les caractéristiques lexicales. Cependant, ils nécessitent un analyseur spécifique pour analyser le texte. La mesure syntaxique la plus courante est une partie du discours (POS) .
- ✓Sémantique, contrairement aux fonctionnalités susmentionnées, les fonctionnalités sémantiques sont une tâche de traitement du langage naturel de haut niveau. En examinant les littératures, seules quelques tentatives abordent les caractéristiques sémantiques.
- ✓Spécifiques à l'application, ces fonctionnalités peuvent être structurelles, spécifiques au contenu et spécifiques à la langue. la signature de l'auteur, les couleurs de police et la taille de la police sont des caractéristiques structurelles évidentes utilisées pour attribuer l'auteur

1.8 CLASSIFICATION DES DONNES

La classification des données doit utiliser une approche pour rechercher des similitudes dans les données afin de pouvoir placer les données dans les bons groupes. Le regroupement des données divisera l'ensemble de données en plusieurs groupes où la similitude dans un groupe est plus grand que d'autres groupes**. [3]**

1.8.1 But de la classification

Le but de la classification est de trouver un modèle capable d'assigner un objet une classe, c'est-à-dire de reconnaître l'objet représenté par un ensemble de caractéristiques. Dans ce cadre, les sorties du modèle, ici le classifieur, ne prennent que des valeurs discrètes. Un exemple typique d'application est la reconnaissance de caractères manuscrits, dans lequel le modèle doit pouvoir donner en sortie le caractère représenté par l'image d'entrée. En reconnaissance de formes, l'invariance de classe est la forme de connaissances a priori le plus souvent rencontrée.

L'invariance de classe signifie que la sortie du classifieur doit rester inchangée si une transformation particulière est appliquée à la forme en entrée. Par exemple, l'invariance aux translations et rotations de l'image est souvent considérée en reconnaissance de caractères manuscrits. **[3]**

En termes d'action, le fait de classer un objet correspond à prendre une décision sur une base d'une ou plusieurs règles. Dès lors une des premières approches pour automatiser le traitement, fut d'extraire la connaissance sous formes de règles. Ainsi, pour chaque catégorie on disposait d'un ensemble de règles permettant de déterminer l'appartenance d'un objet à ladite classe.

L'approche Machine Learning (ML) devient très populaire. En bref, il s'agit d'apprendre automatiquement les règles de décision sur base d'un ensemble d'objets pré-classées. Il s'agit donc d'un processus inductif suivant lequel un classificateur est construit à partir d'exemples. Dans ce qui suit, nous emploierons à la formaliser de classification des données, en définissent sa forme mathématique, en introduisant le contexte statistique requis par l'apprentissage supervisé.

1.9 SYSTEME D'AUTHENTIFICATION

S'il est bien une notion à laquelle l'identité numérique est intrinsèquement attachée, presque sous forme de corollaire, c'est sans aucun doute l'authentification. l'identité numérique ne vise ,en effet , qua formaliser l'individualisation des accès dans les réseau informatique, laquelle se trouve conditionnée à l'existence de moyens de vérification de l'identité numérique des utilisateurs, voire des objet ainsi que ce soit pour une accès a des réseaux locaux ou étendus que ces réseaux soient filaires ou sans-fils, que ces réseaux soient en architecture client-serveur ou repartie ,l'authentification des équipements ,des services et des homme est nécessaire. Tout ce qui concerne l'accès privé, c'est-à-dire le contrôle de la délivrance de l'information et de la fourniture des ressources réservées à certaines entités, passe par l'authentification. **[8]**

1.9.1 Définition et enjeux de l'authentification :

De cette définition, il est important de souligner A la différence de l'Identification qui consiste a simplement déclarer son identité l'authentification exige à fournir la preuve de son identité nous nous tiendrons donc a la définition suivante plus formelle : L'authentification est la fonction de sécurité qui consiste à apporter et à contrôler la preuve de l'identité d'une personne de l'émetteur d'un message, d'un logiciel, d'un serveur logique ou d'un équipement.

La notion de preuve n'est elle-même pas définie. de fait, elle doit être prise dans son sens le plus large, celui d'un élément que seule l'entité qui s'authentifie aurait la possibilité de connaitre ou de posséder physiquement. Un mot de passe utilisateur, en tant qu'élément confidentiel, est donc considéré comme une preuve, quelle que soit sa robustesse.

L'enrôlement est la fonction critique qui permit d'enregistrer pour la première fois dans le système d'informations l'association entre l'identifiant de l'utilisateur et son secret et qui nécessite donc la vérification préalable de l'identité de l'utilisateur .

1.9.2 Les enjeux de l'authentification

La mise en place de systèmes d'authentification vise à garantir qu'une entité souhaitant accéder à des ressources protégées est bien celle qu'elle prétend être. Par conséquent, l'authentification vise à prévenir majoritairement deux types d'attaque, chacune pouvant avoir des conséquences potentiellement désastreuses :

- ✓ L'intrusion frauduleuse dans un système, et l'accès à des données sensibles qui peut en découler.
- ✓ L'usurpation d'identité, qui peut conduire un individu innocent à être considéré comme l'auteur des agissements menés par l'attaque. **[8]**

1.9.3 Autre domaine d'utilisation d'analyse d'authentification d'auteur

1.9.3.1 L'enquête policière

Une certaine expertise de recherche s'est toutefois développée, particulièrement quant aux homicides (par ex. : Beauregard, Lussier et Proulx, 2005; Mokros et Alison, 2002; Salfati 2000), afin d'orienter, d'améliorer et valider les pratiques policières en matière d'identification, de priorisation et d'authentification des suspects. de pair avec le désir, voire le besoin, grandissant des corps policiers d'être de plus en plus proactifs (plutôt que traditionnellement réactifs) et d'avoir des pratiques efficaces basées sur des données probantes (evidence-based policing), on assiste à l'émergence d'études scientifiques qui visent justement à venir en appui aux corps policiers et aux unités d'enquête spécialisées et améliorer l'efficacité de leurs pratiques. C'est ainsi dans l'optique de mettre en lumière les efforts actuels en recherche dans le domaine de l'enquête et des techniques de support à l'enquête, de même qu'avec l'idée de pouvoir discuter des défis qui leur sont associés et favoriser l'amélioration des pratiques policières en matière d'enquête, que ce numéro a été proposé au comité éditorial de la revue Criminologie. Ce numéro spécial avait aussi pour but de mieux faire connaître ce sujet de recherche et de mettre en lumière les chercheurs qui travaillent sur ce thème encore peu exploité. À ma connaissance, aucun ouvrage francophone recueillant des articles empiriques sur ce sujet n'existe pour le moment.

Il est intéressant de spéculer sur les autres domaines qui pourraient venir sous ce genre d'analyse. [8]

1.10 DEFINITION LE PLAGIAT:

Il est défini comme « l'appropriation d'une idée ou d'un contenu (texte, images, tableaux, graphiques, etc.) total ou partiel sans le consentement de son auteur ou sans citer ses sources de manière appropriée » le plagiat en recherche est avant tout une tromperie vis-à-vis des collègues et du public : « le plagiat est une usurpation du rôle de chercheur, il révèle une imposture. il n'est pas falsification, il est confiscation de la substance de l'idée créatrice à celui qui l'a délivrée ; il n'est pas déformation, il est captation de la pensée novatrice de celui qui l'a avancée . **[9]**

1.10.1 Les différentes formes de plagiat en recherche

Le plagiat dans la recherche peut prendre plusieurs formes et présenter des degrés de gravité variables. Rappelons cependant que, sauf exception13, le plagiat implique une volonté de tromperie et est une pratique incompatible. avec les principes d'intégrité scientifique. [**9**]

1. Le plagiat des textes publiés. les limites de sa détection
2. L'appropriation de résultats et d'idées : quand vol et plagiat se confondent
3. L'auto-plagiat et ses multiples facettes .
4. Les ambiguïtés de l'auto-plagiat.
5. Le plagiat de contrefaçon.

1.11 CONCLUSION

Dans ce chapitre nous avons exposé quelques généralités et un bref historique Attribution d'auteur et stylométrie puis L'attribution d'auteur arabe et ces caractéristiques et défis de contexte arabe. On a situé Par la système d'authentification d'autre part, la classification des données comme une tâche de décision. et le plagiat dans le recherche scientifique.

CHAPITRE- 2
APPROCHES PROPOSEES

CHAPITRE-2
APPROCHES PROPOSEES

2.1 INTRODUCTION

L'attribution de la paternité est une approche concernée par l'analyse de textes dans l'exploration de texte, par exemple, des romans et des documents historiques. Dans ce chapitre, nous parlerons les différentes techniques proposées pour identifier ou attribuer des auteurs (SMO-SVM et liner SVM avec distance Manhattan).

2.2 APPROCHES PROPOSEES POUR L'ATTRIBUTION D'AUTEURS

Dans le contexte de l'attribution de la paternité, diverses méthodes d'attribution de textes arabes ont été utilisées. Abassi et Chen ont été les premiers à aborder l'attribution de la paternité dans le contexte arabe. La machine à vecteurs de support (SVM) et les arbres de décision C4.5 ont été appliqués sur les messages du forum Web arabe Pour faire face au défi de l'allongement, ils ont proposé un filtre qui est utilisé pour supprimer l'allongement du texte.

Cependant, le nombre de caractères d'allongement est calculé et il est utilisé plus tard comme fonctionnalité. Abassi et Chen (35) ont répété l'expérience avec les mêmes méthodes d'apprentissage automatique (SVM et C4.5) et ont été appliqués sur les massages du forum Web arabe, mais les racines des mots ont été extraites par de Roeck et l'algorithme d'Al-Fares. [10]

Sayoud a abordé le problème de la discrimination de l'auteur. à cette fin, le Coran et les déclarations du prophète ont été utilisés. Le SMO-SVM, régression linéaire (LR)et Perceptron multicouche (MLP) ont été utilisés. Toutes les classes ont prouvé sa capacité à discriminer l'auteur du texte à l'étude avec une précision de 100 % la meilleure précision obtenue était de 96,7 %. Ils ont également répété l'expérience avec l'application de NB, SVM et SMO le l'ensemble de fonctionnalités comprend les fonctionnalités qui ont été utilisées dans et le mètre de la poésie arabe et suivi la même régulation. La meilleure

précision moyenne qu'ils ont obtenue était de 72,83 % lorsque l'ensemble de toutes les fonctionnalités a été utilisé et que SMO a été appliqué.

Bourib et Khennouf ont abordé le problème d'attribution de la paternité lorsque le genre et le sujet sont assez proches. La taille du texte dans l'ensemble de formation variait de 100 mots à 3000 mots par texte. Le caractère n-gramme et les mots étaient employés et SMO-SVM, MLP et LR ont été utilisés. Les résultats montrent que la performance des classes dépend. [11]

2.2.1 L'optimisation minimale séquentielle (SMO)

Est un algorithme permettant de résoudre le problème de programmation quadratique (QP) qui se pose lors de la formation de machines à vecteurs de support (SVM). Il a été inventé par John Platt en 1998 chez Microsoft Research [12].SMO est largement utilisé pour la formation de machines à vecteurs de support et est implémenté par le populaire outil LIBSVM. la publication de l'algorithme SMO en 1998 a suscité beaucoup d'enthousiasme dans la communauté SVM, car les méthodes précédemment disponibles pour la formation SVM étaient beaucoup plus complexes et nécessitaient des solveurs QP tiers coûteux. [13]

2.2.1.1 Algorithme

SMO est un algorithme itératif pour résoudre le problème d'optimisation décrit ci-dessus. SMO divise ce problème en une série de sous-problèmes les plus petits possibles, qui sont ensuite résolus analytiquement. en raison de la contrainte d'égalité linéaire impliquant les multiplicateurs de Lagrange , le plus petit problème possible implique deux multiplicateurs de ce type. et ce problème réduit peut être résolu analytiquement : il faut trouver un minimum d'une fonction quadratique unidimensionnelle. est le négatif de la somme sur le reste des termes de la contrainte d'égalité, qui est fixée à chaque itération**[14].** La SMO choisit deux coefficients de Lagrange ∝i et ∝j pour les optimiser ensemble, trouver ses valeurs optimales étant donné que toutes les autres sont fixes, et actualiser le vecteur solution **∝.[15**]

$$\text{LD}=\alpha 1+\alpha 2-\frac{1}{2}k(x1.x2)\alpha_1^2-\frac{1}{2}k(x1.x2)\alpha_1^2-2\gamma 1\gamma 2k(x1.x2)\alpha 1\alpha-\gamma 1\alpha 1\alpha v2+cte$$

Avec: $\text{Vi}=\sum_{i=3}^{l} yi\alpha ik(xi.yj)$

La SMO optimise deux coefficients à chaque itération. un des deux doit violer les conditions de KKT pour être choisi dans l'itération courante.

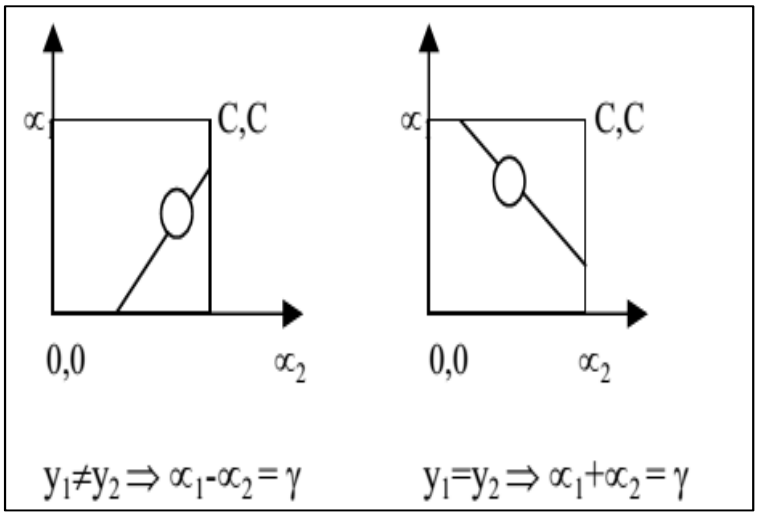


Figure 2.1 : Les deux multiplicateurs.

Algorithme de minimale séquentielle (SMO) pour la formation SVM est simple, plus rapide et plus évolutif. SMO utilise une analyse Étape QP (programmation quadratique), plutôt que QP numérique que les méthodes précédentes de formation SVM utilisent. La SMO dépense plus de temps à évaluer les fonctions de décision plutôt que d'exécuter le QP, il peut donc exploiter des ensembles de données clairsemés efficacité. SMO utilise moins le stockage matriciel, donc très de grands problèmes de formation SVM peuvent tenir dans la mémoire d'un ordinateur personnel, en raison de l'évitement SMO de grands manipulation matricielle. SMO évolue entre linéaire et quadratique dans la taille de l'ensemble d'apprentissage des problèmes de test, ce qui rend il est efficace pour une grande quantité de données d'entraînement de protéines utilisées dans cette recherche. SMO met à jour deux multiplicateurs de Lagrange en tant qu'Étape SMO comme indiqué ci-dessous : **[16]**

$$\alpha_2^{new} = \alpha_2 + \frac{y_2(E_2 - E_1)}{\eta}$$

Avec :

$$\eta = K(\vec{x}_1, \vec{x}_1) + K(\vec{x}_2, \vec{x}_2) - 2K(\vec{x}_1, \vec{x}_2)$$

2.2.2 Machine à vecteurs de support (SVM)

SVM est une méthode de classification binaire par apprentissage supervisé, elle fut introduite par Vapnik en 1995. Cette méthode est donc une alternative récente pour la classification. Cette méthode repose sur l'existence d'un classificateur linéaire dans un espace approprié. Puisque c'est un problème de classification à deux classes, cette méthode fait appel à un jeu de données d'apprentissage pour apprendre les paramètres du modèle.

Elle est basée sur l'utilisation de fonction dites noyau (kernel) qui permettent une séparation optimale des données. **[16]**

La classification de données se fait dans un contexte supervisé où on cherche à estimer une fonction f(x) à partir des exemples x, comme l'indique la figure (2.2).

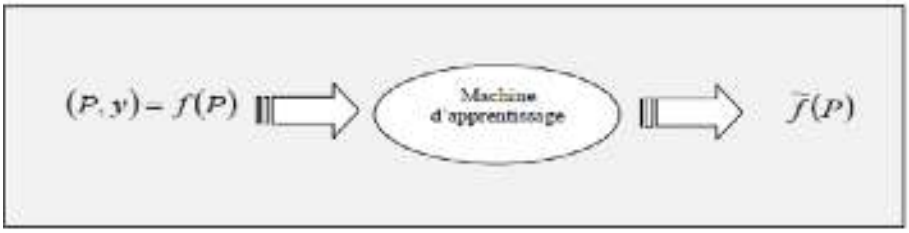


Figure 2.2 : Apprentissage supervisé d'une machine.

2.2.2.1 SVM principe de fonctionnement général

a- 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. **[16]**

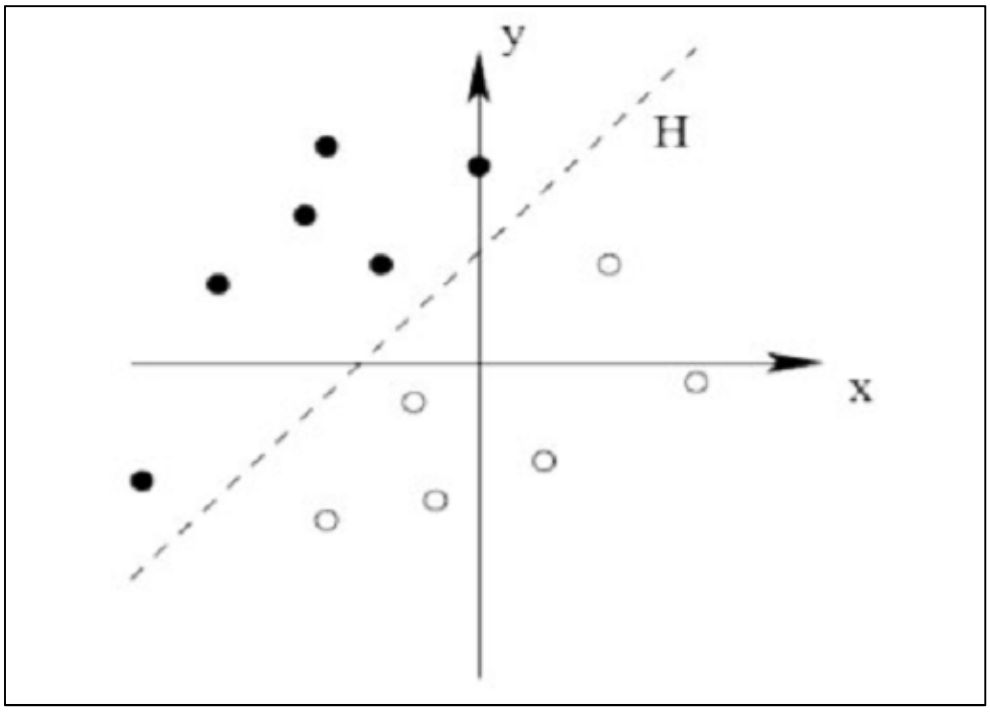


Figure 2.3 : Hyperplan qui sépare les deux ensembles de point.

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

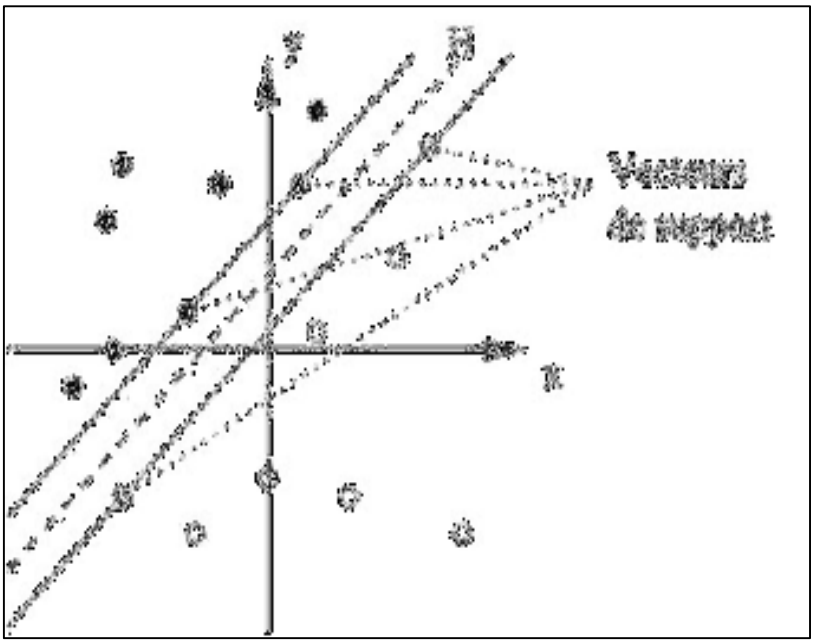


Figure 2.4 : L'hyperplan avec vecteurs de support.

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. **[17].**

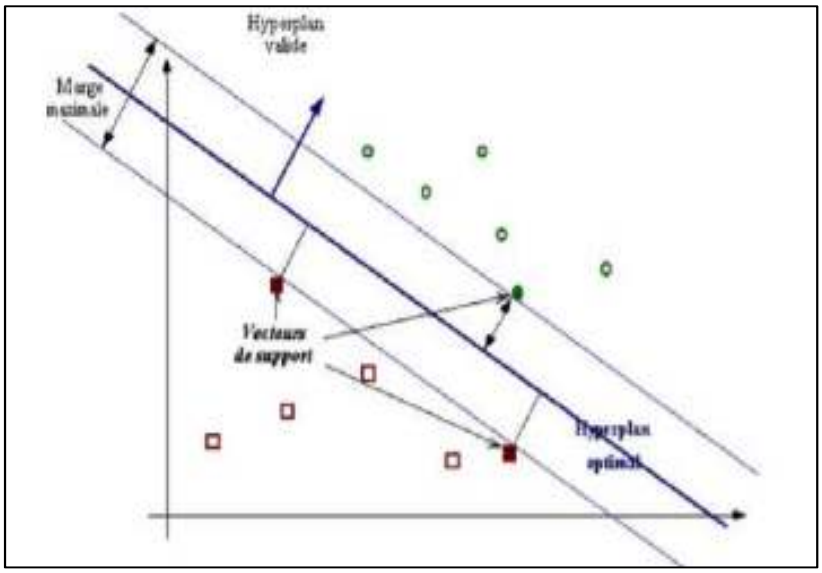


Figure 2.5 : SVM- L'hyperplan séparateur optimal et la marge.

b- 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. **[17]**

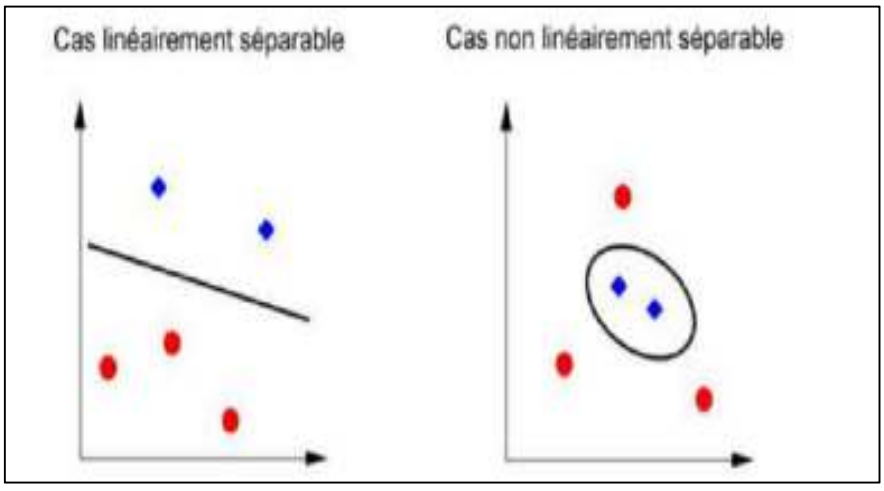


Figure 2.6 : les modèles des SVM.

Pour surmonter les inconvénients des cas non linéairement séparable, l'idée des SVM est de changer l'espace des données. La transformation non linéaire des données peut permettre une séparation linéaire des exemples dans un nouvel espace. On va donc avoir un changement de dimension.

Cette nouvelle dimension est appelé « espace de description ». En effet, intuitivement, plus la dimension de l'espace de description est grande, plus la probabilité de pouvoir trouver un hyperplan séparateur entre les exemples est élevée. Ceci est illustré par le schéma suivant : **[17]**

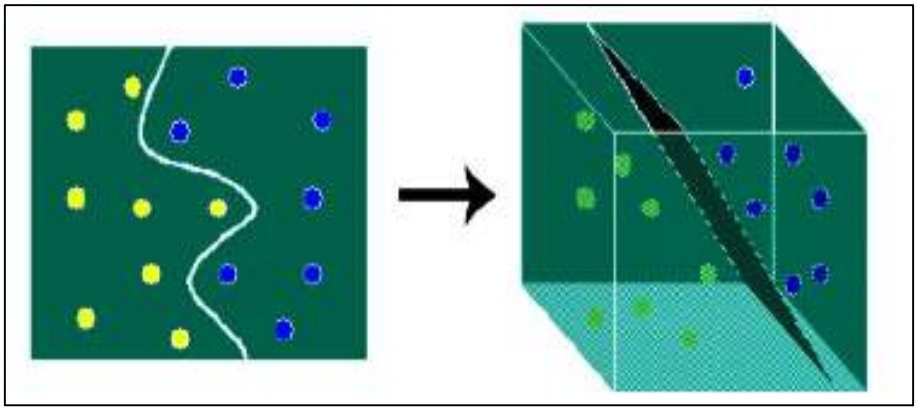


Figure 2.7: Exemple de changement de l'espace de données.

2.2.2.3 SVM linéaire

Soit un ensemble de données d'apprentissage : $\{x_i, y_i\}, \mathrm{i} = 1; \ldots, \mathrm{l}$ avec : $x_i \in R^d$ et $y_i \in \{-1; +1\}$. Supposons qu'on dispose d'un hyperplan séparant les données positives des données négatives. Les x_i qui appartiennent à l'hyperplan vérifient la relation: $x_i.w + b = 0$, Où ww est la normale de l'hyperplan tandis que $|b| / \|w\|$ est la distance perpendiculaire entre l'hyperplan et l'origine (Figure 2.7). **[17].**

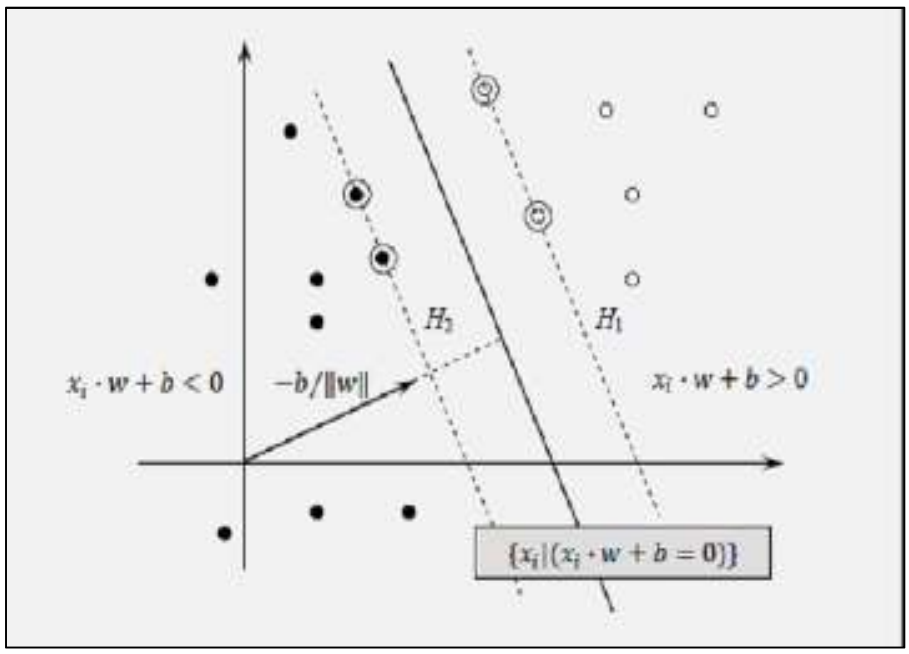


Figure 2.8 : Hyperplan de séparation linéaire pour des données linéairement séparables (Les vecteurs de support sont encerclés)

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

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) = signe (w^txi + b) = yi, i = 1,…, M

Ce qui est équivalent à :

y_i ($w^t x_i$+ b) > 0, i = 1,…, M.

Si un tel hyperplan existe, alors l'ensemble d'apprentissage est linéairement séparable : [19]

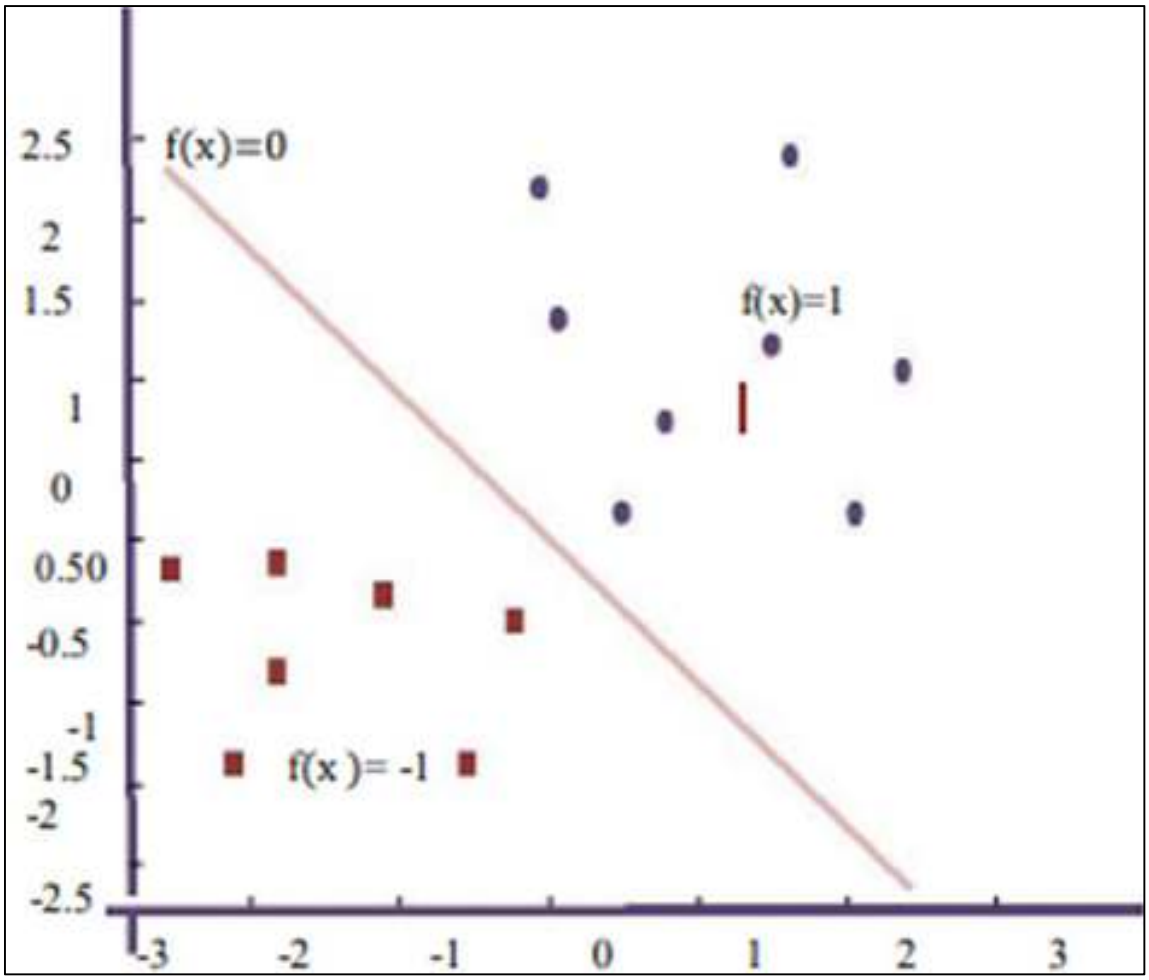


Figure 2.9 : Hyperplan séparateur entre 2 classes.

2.2.3 Manhattan distance

2.2.3.1 Bases de la distance de Manhattan:

La fonction de distance de Manhattan calcule la distance qui serait parcourue pour aller d'un point de données à l'autre si un chemin en forme de grille est suivi. Le Manhattan distance entre deux éléments est la somme des différences de leurs composants correspondants. La formule pour cela distance entre un point X= (X1, X2, etc.) et un point : Y= (Y1, Y2, etc.) est : **[20]**

$$d = \sum_{i=1}^{n} |X_i - Y_i|$$

Où n appartient au nombre de variables, et Xi et Yi sont les valeurs de i variable, aux points X et Y respectivement. **[21]**

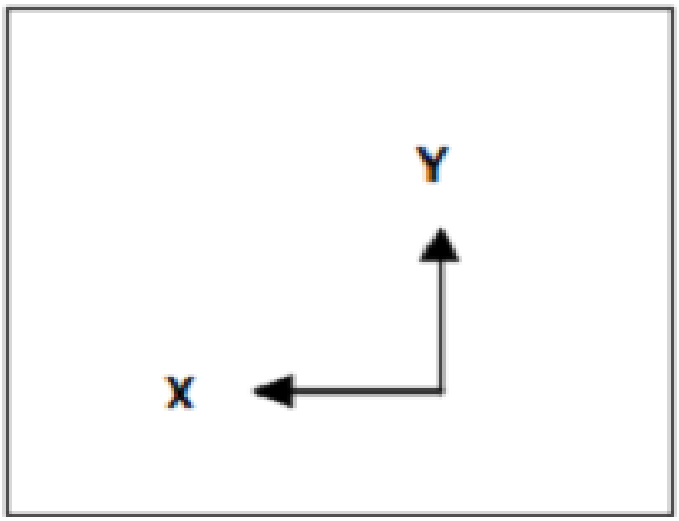


Figure 2.10 : Distance Manhattan

Diagramme de distance de Manhattan Dans les premiers temps de l'infographie, l'écran est composé de pixels, qui sont des entiers, et les coordonnées des points sont généralement des nombres entiers. La raison en est que les opérations en virgule flottante sont coûteuses, lentes et comportent des erreurs. Si vous utilisez directement la distance euclidienne de AB (Distance euclidienne: la distance euclidienne dans un espace bidimensionnel et tridimensionnel est la distance entre deux points), vous devez effectuer des calculs en virgule flottante, si vous utilisez AC et CB, il vous suffit de calculer l'addition et la soustraction. Cependant, cela améliore considérablement la vitesse de calcul, et quel que soit le nombre de fois que le calcul est accumulé, il n'y aura aucune erreur.

2.3 CONCLUSION

Dans ce chapitre, nous avons expliqué les différentes approche d'attribution de la paternité à suivre pour atteindre les objectifs de ce travail, et nous présentons une évaluation empirique dans toutes les méthodes et techniques d'attribution d'auteur proposées, utilisez le corpus que nous avons conçu à cet effet.

CHAPITRE-3

EXPERIENCES ET RESULTATS OBTENUS

CHAPITRE- 3

EXPERIENCES ET RESULTATS OBTENUS

3.1 INTRODUCTION

Dans le chapitre précédent, nous avons présenté les méthodes d'attribution d'auteur proposées ; SMO-SVM, Linear SVM et distance Manhattan. Dans le présent chapitre, nous allons exposer et commenter les résultats obtenus des expériences conduites.

Nous allons exposer les séries d'expériences d'attribution d'auteur effectuées sur notre corpus qui est composé de 14 auteurs, 7 femmes et 7 hommes dont chacun a écrit 6 textes d'une longueur moyenne de 2000 mots. Ces textes ont fait l'objet d'une série d'expérimentations pour comprendre les effets de l'apprentissage Point de vue de l'attribution de l'auteur. Par la suite, les résultats obtenus ont été vérifiés et discute et donne des explications et des conclusions objectives.

3.2 CORPUS D'EVALUATION

L'évaluation expérimentale joue un rôle important dans la classification texte. En utilisant le corpus de test, nous pouvons voir l'effet de l'acquisition de documents texte sur l'attribution de l'auteur. Cependant, la recherche sur la paternité des textes le nombre de corpus obtenus après opération OCR est relativement faible, moins de texte corrigé. Aussi, le nombre d'auteurs possibles est encore limité car il est difficile de trouver un grand nombre de candidats potentiels qui obéissent à de multiples contraintes (même période et même langue, culture proche, sujet et volume d'études similaires important).

3.2.1 Conception du corpus d'évaluation

Le corpus que nous avons conçu contient 14 écrivains arabes contemporains (7 femmes et 7 hommes) sont : Gada Samane, Houda Barakat, Kolite Sohil, May Ziada, Assia Djabar, Nazik Malaika, Latifa Zayat, Djobran Khalil, Abdelkader El Mazini, Saddak Rafie, Taha Hocin, Najib Mahfoud, Moustafa Al Manfalouti et Mahmoud Akad.

Nous choisissons un livre pour chaque auteur, puis extrayons au hasard un livre spécifique nombre de pages contenant le nombre de mots choisis, et pour chaque auteur 6 sont sélectionnés Textes d'une longueur moyenne de 2000 mots Textes corrigés. Les textes utilisés pour l'opération d'apprentissage (qui sont les textes numéros ; 3, 4, 5 et 6 pour chaque auteurs) sont tous du texte corrigé. Cependant, chacun des textes utilisés pour l'opération de test (qui sont les textes numéros 1 et 2 pour chaque auteurs).

Au total, le corpus contient 84 textes ;56 textes l'ensemble d'apprentissage, et 28 textes l'ensemble de test. 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 (1) : Récapitulatif du Corpus (Ecrivains féministes)

Ecrivains	Pays de naissance	Période	Nombre de livre	textes	Nombre de mots/texte	Nombre de mots moyen	utilisation
Assia djabar	Algérie	1936-2015	26 livre	Assia-1	2138	2254	Test
				Assia -2	1957		Test
				Assia-3	2401		Apprentissage
				Assia-4	2361		Apprentissage
				Assia-5	2161		Apprentissage
				Assia-6	2510		Apprentissage
May ziada	syrie	1886-1941	19 livres	May-1	2180	2080	Test
				May-2	1960		Test
				May-3	2149		Apprentissage
				May-4	1956		Apprentissage
				May-5	1978		Apprentissage
				May-6	2258		Apprentissage
Nazik malaika	Bagdad	1923-2007	25 livres	Nazik-1	3211	1443	Test
				Nazik-2	1676		Test
				Nazik-3	1034		Apprentissage
				Nazik-4	896		Apprentissage
				Nazik-5	1007		Apprentissage
				Nazik-6	839		Apprentissage
Ghada saman	syrie	1942- à ce jour	46 livres	Ghada-1	3088	2665	Test
				Ghada-2	1825		Test
				Ghada-3	2960		Apprentissage
				Ghada-4	1696		Apprentissage
				Ghada-5	3273		Apprentissage
				Ghada-6	3151		Apprentissage
Kolit sohil	syrie	1937 à ce jour	29 livres	Kolit-1	2329	1838	Test
				Kolit-2	1609		Test
				Kolit-3	1995		Apprentissage
				Kolit-4	1465		Apprentissage
				Kolit-5	1664		Apprentissage
				Kolit-6	1969		Apprentissage
latifa zayat	Egypte	1923-1996	12 livres	Latifa-1	2209	2005	Test
				Latifa-2	1865		Test
				Latifa-3	2232		Apprentissage
				Latifa-4	1983		Apprentissage
				Latifa-5	1993		Apprentissage
				Latifa-6	1753		Apprentissage
Houda baraka	Liban	1952- à ce jou	12 livres	Houda-1&	2098	2044	Test
				Houda-2	1984		Test
				Houda-3	2073		Apprentissage
				Houda-4	2116		Apprentissage
				Houda-5	2069		Apprentissage
				Houda-6	1929		Apprentissage

Tableau(2) : Récapitulatif du Corpus (Ecrivains hommes)

Ecrivains	Pays de naissance	période	Nombre de livre	texte	Nombre de mots /texte	Nombre de mots moyenne	utilisation
Djobran Khalil	Liban	1883-1931	163 LIVRES	Djobran-1	2306	1890	Test
				Djobran-2	2514		Test
				Djobran-3	2457		Apprentissage
				Djobran-4	2328		Apprentissage
				Djobran-5	1875		Apprentissage
				Djobran-6	2064		Apprentissage
Saddek rafie	Egypte	1880-1973	9 livres	Rafie-1	3260	2423	Test
				Rafie-2	1932		Test
				Rafie-3	2846		Apprentissage
				Rafie-4	2006		Apprentissage
				Rafie-5	1938		Apprentissage
				Rafie-6	2560		Apprentissage
Taha hocin	Egypte	1889-1973	47 livres	Taha-1	2888	2125	Test
				Taha-2	1761		Test
				Taha-3	1983		Apprentissage
				Taha-4	1884		Apprentissage
				Taha-5	2089		Apprentissage
				Taha-6	2145		Apprentissage
Abdelkader Mazini	Egypte	1890-1949	19 livres	Meziani-1	2697	2143	Test
				Meziani-2	2019		Test
				Meziani-3	2055		Apprentissage
				Meziani-4	2077		Apprentissage
				Meziani-5	1970		Apprentissage
				Meziani-6	2041		Apprentissage
Mohamed Akad	Egypte	1989-1964	89 livres	Akad-1	3293	2195	Test
				Akad-2	2514		Test
				Akad-3	2457		Apprentissage
				Akad-4	2328		Apprentissage
				Akad-5	1875		Apprentissage
				Akad-6	708		Apprentissage
Moustafa al manfalouti	Egypte	1876-1924	107 livres	Lotfi-1	1961	2576	Test
				Lotfi-2	2384		Test
				Lotfi-3	2779		Apprentissage
				Lotfi-4	2809		Apprentissage
				Lotfi-5	1921		Apprentissage
				Lotfi-6	2672		Apprentissage
Najib Mahfoud	Egypte	1905-1993	21 livres	Najib-1	2749	2576	Test
				Najib-2	3115		Test
				Najib-3	2491		Apprentissage
				Najib-4	2426		Apprentissage
				Najib-5	2638		Apprentissage
				Najib-6	2041		Apprentissage

3.2.2 Préparation des documents du corpus d'évaluation

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)
- ✓ Correction des erreurs dans les textes
- ✓ Les documents textes obtenus sont enregistrés sous forme UTF-8 (Encodage basé sur l'Unicode qui peut être codé sur 4 octets).

Après le processus de numérisation du document PDF, le résultat obtenu est traité comme un document éditable (format Word) pour corriger les erreurs, ajouter ou supprimer des extras.

ولدت في بغداد في من شهر آب أغسطس سنة وكنت كغيري إخوتي وهم أربع بنات وولدان

وقد تدرجت في دراستي من الابتدائية إلى المتوسطة فالثانوية وتخرجت في الثانوية عام وكنت منذ صغري أحب اللغة العربية والإنجليزية والتاريخ ودروس الموسيقى كما كنت أجد لذة في دراسة العلوم بخاصة علم

الفلك وقوانين الوراثة والكيمياء ، ولكني كنت أمقت الرياضيات مقتاً شديداً

وأحد السنين يوماً يوماً لأصل إلى إنهاء مرحلة الثانوية فأتخصص بدراسة الآداب ثم دخلت دار المعلمين العالية فرع اللغة العربية وخرجت منها بليسانس الآداب من مرتبة الامتياز وهي أعلى مرتبة تمنح وخلال سنوات دراستي فيها تعرفت إلى موضوع الفلسفة وأحببته حباً شديداً

فساعدني على تكوين ذهن منطقي وكانت دراساتي الكثيرة للنحو العربي هي

أصوله القديمة قد هيأتني له تهيئة واضحة وقد بدأت نظم الشعر وحبه منذ طفولتي الأولى والواقع أنني سمعت أبي وجدي يقولون عني أنني شاعرة قبل أن أفهم معنى هذه الكلمة لأنهم لاحظوا على القفية وأذنا

Figure (3.1) : Exemple de texte corrigé (femme)

جاء في تاريخ الواقدي أن المقوقس عظيم القبط في مصر زوج بنته أرمانوسة من المسلمين من هرقل وجهزها بأموالها حشدنا لتسير إليه حتى يبني بها في مدينة قيسارية فخرجت الى بلبيس وأقامت بها وجاء عمرو بن العاص إلى بلبيس فحاصرها

حصارًا شديدًا وقاتل من بها وقتل منهم زهاء ألف فارس والتجأ من بقي إلى المقوقس وأخذت أرمانوسة وجميع مالها وأخذ كلّ ما كان للقبط في بلبيس فأحب عمرو ملاطفة المقوقس فسيّر إليه ابنته مكرّمة في جميع مالها مع قيس بن أبي العاص السهمي فسار بقدومها

هذا ما أثبته الواقدي في روايته ولم يكن معلولًا إلا بأخبار المغازي والفتوح فكان يقتصر عليها في الرواية أما ما أخذته فهو ما تقصّته ألسن

كانت لأرمانوسة وصيفة سوداء تسمّى مارية ذات جمال يونانيّ أتمته مصر ومسحته بسحرها فزاد جمالها على أن يكون مصريًا ونقص الجمال اليوناني أن يكونه فهو أجمل منهما وتمصر طبيعة خاصة في الحسن فهي قد تُهمل شيئًا في جمال نسائها أو تُلحق منه وقد لا توفيه جهد محاسنها الرائعة ولكن متى نشأ فيها جمالٌ يتزع إلى أصل أجنبي أفرجت فيه سحرها إفراطًا وأبت إلا أن تكون الغالبة عليه وجعلته آيتها في المغايرة بينه في طابعه المصري وبين أصله في طبيعة أرضه كالثقة ما كانت تغار على سحرها أن يكون إلا الأعلى

Figure (3.2) : Exemple de texte corrigé (homme)

3.3 LES RESULTATS DES EXPERIENCES :

Dans ce mémoire, la tâche d'attribution de l'auteur est effectuée en utilisant les caractères N-grams et Word N-grams comme caractéristiques et deux méthodes de classification SMO et SVM. Ces techniques sont utilisées pour évaluer la robustesse de notre système d'attribution d'auteur des fichier textes.

Le taux d'Attribution d'Auteurs (TAA) est défini par la relation suivante:

$$TAA = \frac{nombre\ document\ correctemment\ attrbués}{nombre\ documents\ testé} \times 100$$

Ce travail expérimental est organisé en deux séries d'expériences et chaque série comporte plusieurs cas d'applications selon la valeur de N-grams dans deux caractère.

3.4 EXPERIENCES D'ATTRIBUTION D'AUTEURS :

Dans cette expérience, nous étudierons un groupe de textes arabes corrigés de différents auteurs, femmes et hommes, de différentes méthodes et caractéristiques nous l'étudierons comme suit :

3.4.1 Attribution d'auteurs par la méthode Liner SVM

Cette série d'expériences vise à déterminer le nombre approprié de caractères (N-gramme et world N-gramme) pour avoir le meilleur taux TAA en utilisant les textes

corrigés pour les deux phases (apprentissage et test). Les résultats obtenus de cette série d'expérience sont présentés dans les tableaux et les figures qui suivent :

A- Caractères (N-Gramme) :

Nombre (N) / Taux d'attribution (%)	N=2	N=3	N=4	N=5	N=6
Male / Male	71	92	92	71	85
Femelle / Femelle	78	71	71	78	78
Male / Femelle	85	92	89	85	85

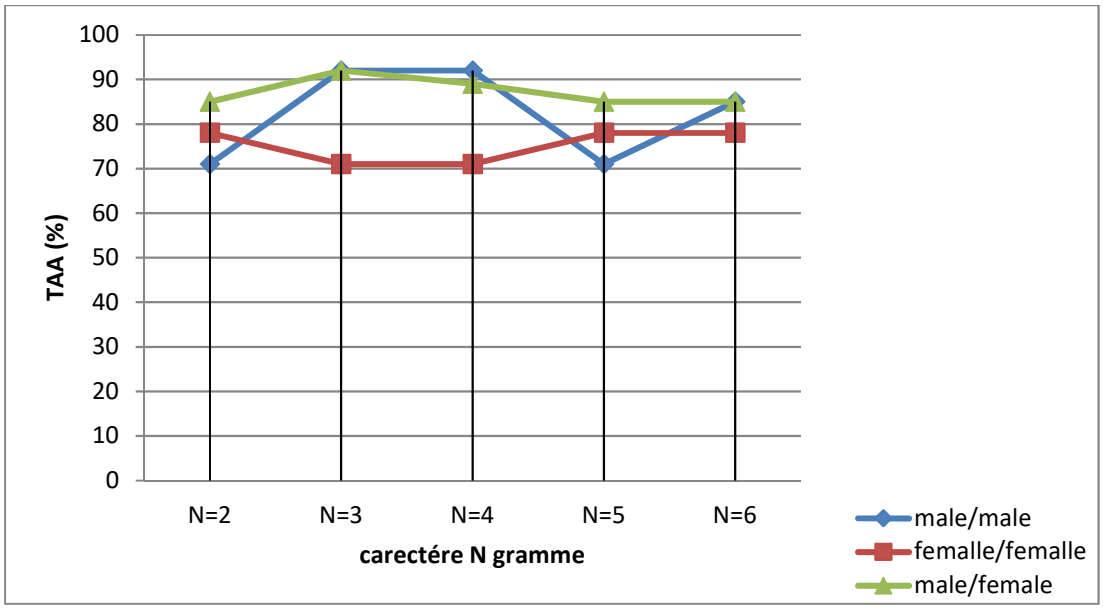


Figure (3.3) : Taux d'attribution d'auteurs pour classifieur liner SVM

On remarque que la courbe représentée dans la figure (3-3), on peut voir clairement que le taux TAA de cette expérience est entre 85-92% pour les auteur male /femelle et le meilleur résultat c'est 92% correspond à N=3. Et pour les auteur male/male on peut avoir le taux TAA entre 71-92% et meilleure résultat c'est 92% correspond à N=3, N=4 femelle/femelle entre 71-78% meilleure résultat c'est 78% correspond à N=3, N=5, N=6.

B- Word (N-Gramme) :

Nombre (N) / Taux d'attribution (%)	N=2	N=3	N=4	N=5	N=6
Male / Male	85	71	57	64	64
Femelle / Femelle	85	71	57	64	71
Male / Femelle	85	85	64	85	85

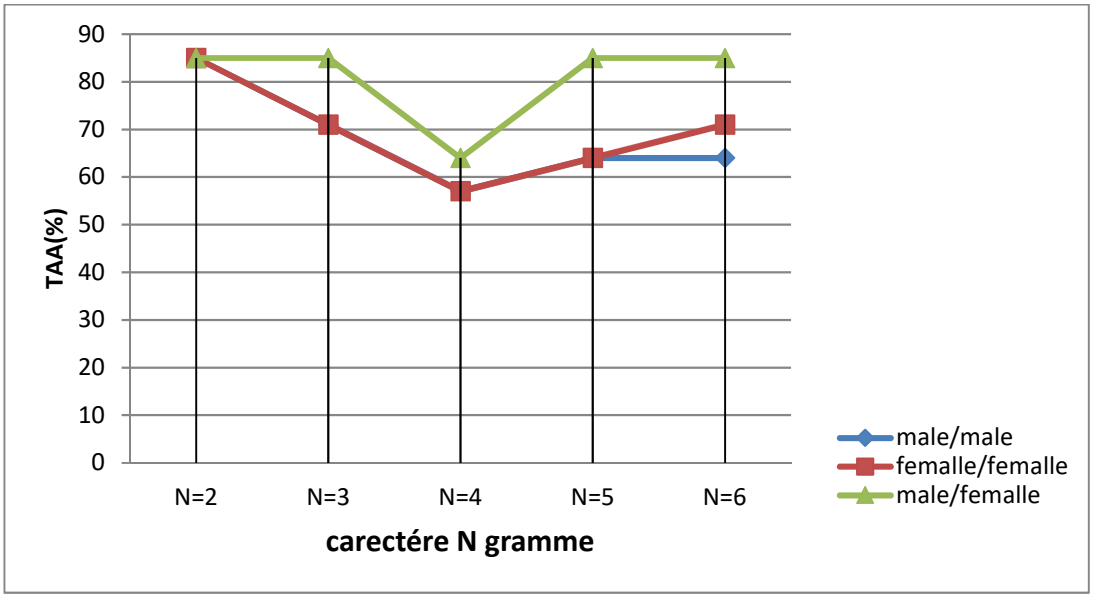


Figure (3.4) : Taux d'attribution d'auteurs pour classifieur liner SVM

On remarque que la courbe représentée dans la figure (3-4) : on peut voir clairement que le taux TAA de cette expérience est entre 64-85% pour les auteur male /femelle et pour les auteur male/male on peut avoir le taux TAA entre 57-85% et femelle/femelle entre 57-85% ,et pour N=4 le taux TAA entre 57-64% , ceci peut être expliqué, aussi ,les auteurs n'utilisent pas abondamment des mots de quatre lettres.

3.4.2 Attribution d'auteurs par la méthode WEKA-SMO

Dans cette expérience on vise à déterminer le nombre optimal de caractères (N) et – le mot gramme et Word gramme) pour avoir le meilleur taux TAA en utilisant les textes corrigés pour les deux phases (apprentissage et test). les résultats obtenus de cette série d'expérience sont présentés dans les tableaux et les figures qui suivent :

A. CARACTÈRES (N-GRAMME)

Nombre (N) / Taux d'attribution (%)	N=2	N=3	N=4	N=5	N=6
Male / Male	92	100	100	71	85
Femelle / Femelle	78	78	78	78	85
Male / Femelle	92	89	96	96	92

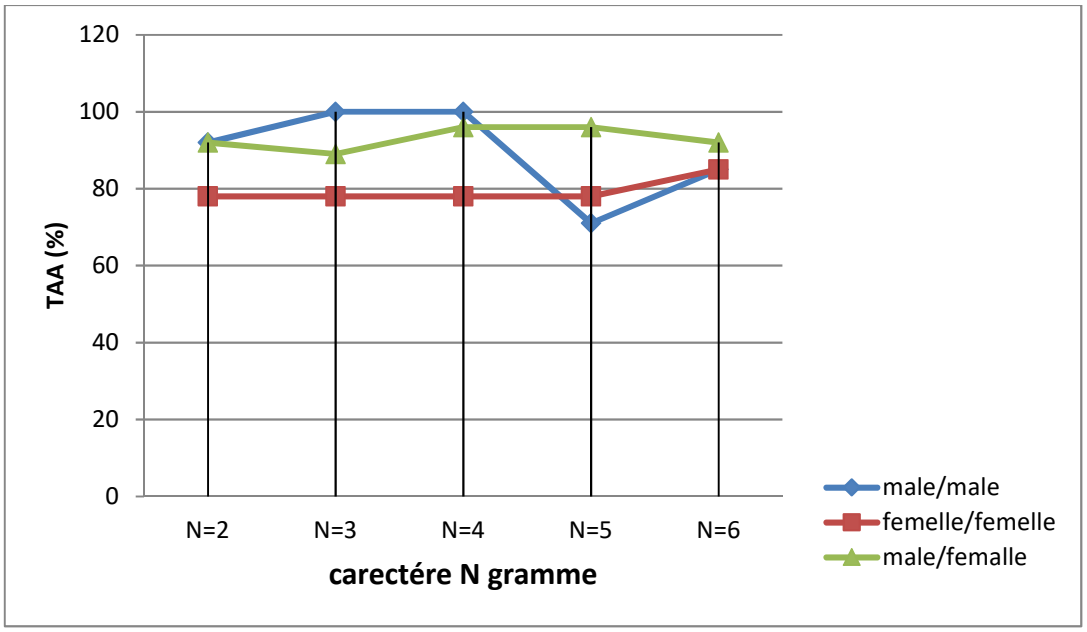


Figure (3.5) : Taux d'attribution d'auteurs pour classificateur Weka SMO

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 71-100% pour les écrivains male/male, 78-85% pour les écrivains femelle/femelle et 89-96% pour les deux (male /femelle). Pour les résultats des auteurs male/male on a la meilleure valeur de TAA =100% correspond à N=3 et N=4 . Pour les résultats des auteurs femelle/femelle on a la meilleure valeur de TAA =85% correspond à N=6. Pour les résultats mâle / femelle on a la meilleure valeur de TAA =96% correspond à N=4 et N=5.

Ceci peut être expliqué Weka SMO est fort contre les erreurs de conversion dans le texte généré par le procédé OCR et également fort pour les valeurs inférieures à 5.

B. WORD (N-GRAMME)

Nombre (N) / Taux d'attribution (%)	N=2	N=3	N=4	N=5	N=6
Male / Male	100	92	64	71	71
Femelle / Femelle	71	71	64	64	71
Male / Femelle	85	85	92	92	85

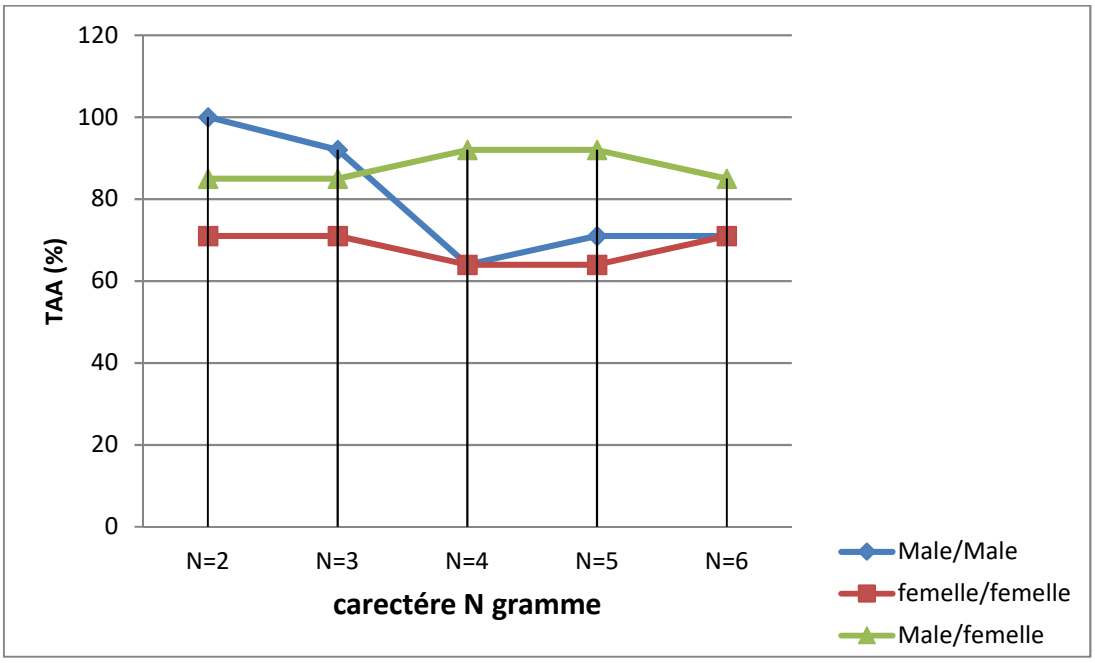


Figure (3.6) : Taux d'attribution d'auteurs pour classifieur Weka SMO.

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 64-100% pour les écrivains male/male, 64-71% pour les écrivains femelle/femelle et 85-92% pour les deux (male /femelle). Pour les résultats des auteurs male/male on a la meilleure valeur de TAA =100% correspond à N=2. Pour les résultats des auteurs femelle/femelle on a la meilleure valeur de TAA =71% correspond à N=2,3,6. Pour les résultats male / femelle on a la meilleure valeur de TAA =92% correspond à N=4 et N=5.

3.5 CONCLUSION

Dans ce chapitre, nous menons des expériences avec l'attribution d'auteur le document texte arabe obtenu après l'opération OCR comporte plusieurs erreurs. Des évaluations expérimentales ont été réalisées à l'aide d'une base de données conçue pour 14 auteur (Arabes Contemporary Writer). Les résultats obtenus montrent la possibilité d'identifier les auteurs par des classificateurs SMO-SVM et liner SVM, en utilisant la distance Manhattan avec des erreurs.

CONCLUSION GÉNÉRALE

CONCLUSION GÉNÉRALE

❖ Travail réalisé

Dans cette mémoire, nous abordons le travail d'évaluation de la force d'attribution En identifiant les auteurs, notre base de données conçue effectue Nos experts sont composés de 14 auteurs arabes, et nous en avons sélectionné 6 pour chaque auteur Texte qui a été converti à l'aide du processus OCR.

Le but est d'étudier le style de l'auteur afin de trouver le véritable auteur et d'appliquer des choses comme N-grammes et classificateurs (linéaire SVM, SMO-SVM avec distance Manhattan). Nous avons classé les auteurs par sexe, femme et homme pour examiner leur style d'écriture, ensuite nous, il récupère au hasard un certain nombre de pages contenant (2000) mots. Pour chaque auteur nous avons sélectionné 6 textes d'une longueur moyenne de 2000, textes corrigés utilisés dans le processus d'apprentissage (ce sont des fichiers Textes numérotés 3, 4, 5 et 6 pour chaque auteur. Cependant, chacun des scripts également utilisés pour le processus de test (qui sont des fichiers Textes numérotés 1 et 2 par auteur).

❖ Résultats obtenus

Avec ce travail, nous montrons que les résultats obtenus sont très encourageants compte tenu des contraintes liées à la taille du texte. Élu (2000 mots seulement) dont on voit l'importance et l'efficacité dans l'action n-grammes pour l'affectation AA.

A partir de ces résultats, nous observons la valeur TAA= 100% pour N=2, N = 3 et N = 4, de classificateurs Weka SMO ont été trouvées bons résultats pour les textes male/mal et male/femelle alors sont plus adaptées aux tâches AA.

.

❖ **Perspectives suggérées**

Point de vue proposé Dans le cadre du développement futur de ce travail, nous proposons Complétez les travaux suivants :

- Utiliser le texte correction après avoir (apprentissage approfondi) Étendre la base de données.
- Généraliser cette étude à d'autres types de textes écrite, tels que : Discours politiques, commentaires sportifs, etc.
- Utilisez un petit texte pour tester l'attribution de l'auteur.

REFERENCES BIBLIOGRAPHIQUES

REFERENCES BIBLIOGRAPHIQUES

[1] P. Juola and G. K.Mikros, “Cross-Linguistic Stylometric Features: A Preliminary Investigation”, in JADT2016: International Conference on Statistical Analysis of Textual Data, France, 2016.

[2]https://www.google.com/url?q=https://www.larousse.fr/encyclopedie/peinture/attribution.

[3] M. Al-Sarem, F. Saeed1, A.Alsaeedi1, W. Boulila1, T. Al-Hadhrami article “Ensemble Methods for Instance-based Arabic Language Authorship Attribution’’1College of Computer Science and Engineering, Taibah University, Medina 344, Saudi Arabia.2017.

[4] Mémoire de master B. BACHA et I. HADLI. (2021). « Attribution d’auteurs des textes arabes traduits en plusieurs langues en utilisant les traducteurs automatiques ».

[5] Brixtel, R., Lecluze, C., & Lejeune, G. (2015, June). Attribution d’Auteur: approche multilingue fondée sur les répétitions maximales. In Actes de la 22e conf\'erence sur le Traitement Automatique des Langues Naturelles. Articles longs (pp. 208-219.

[6] Mémoire de master R. MENASRI et M. YAKOUBI. (2020). « Etude et analyse des effets d'acquisition optique à l'aide d'un OCR des textes arabes sur l’attribution d’auteurs ». 2020.

[7] Technique de la stylométrie appliquée au Livre de Mormon. Agnès Boltoukhine.

[8] H. Saleh, « Système d’authentification de locuteurs base sur les machines à vaste marge (SVM) » Mémoire de master en électronique université de M’sila, 2012.

[9] C. KIENNERT, S. BOUZEFRANE et P. THONIEL. “La gestion des identités numériques” France, 2015.

[10] A. N. de Roeck and W. Al-Fares, ‘‘A morphologically sensitive clustering algorithm for identifying Arabic roots,’’ in Proc. 38th Annu. Meetine Assoc. Comput. Linguistics (ACL), 2000.

[11] S. BOURIB and S. KHENNOUF, ‘‘Author identification using different sizes of documents: A summary,’’ Hidden Data Mining Sci. Knowledge Discovery J.,vol. 1, pp. 9–12, 2015.

[12] A b c d e Platt, John (1998). « Optimisation minimale séquentielle : un algorithme rapide pour les machines à vecteurs de support de formation » (PDF). Cite Seer X 10.1.1.43.4376. Citer le journal nécessite |journal= (aide.

[13] Rifkin, Ryan (2002). Tout ce qui est ancien est nouveau : un nouveau regard sur les approches historiques de l'apprentissage automatique (thèse de doctorat). Massachusetts Institute of Technology. p. 18. hdl : 1721,1/17549 .

[14] https://stringfixer.com/fr/Sequential_minimal_optimization.

[15] A. Amirou. M. Djeddi. Application des SVMs basés sur l'algorithme SMO pour la détection d'anomalies cardiaques. March 25-29, 2007 – TUNISIA.

[16] Mohamadally Hasan Fomani Boris BD Web, ISTY3 Versailles St Quentin, France 16.

[17]. Deep Learning using Linear Support Vector Machines Department of Computer Science, University of Toronto. Toronto, Ontario, Canada. Yichuan Tang.

[18]. BELHADJ Mohammed et BOUREZG Ala Eddine Evaluation de la robustesse d'attribution en reconnaissance d'auteur 2021.

[19]. AN OVERVIEW ON CLUSTERING METHODST. Soni Madhulatha

Associate Professor, Alluri Institute of Management Sciences, Warangal-2012.

[20]. Application of Vector Quantization for Audio Retrieval Volume 88 – No.17, February 2014. Shruti Vaidya-Kamal Shah, PhD.

[21]. https://www.codetd.com/fr/article/11967736.

ملخص

في هذه المذكرة، ندرس باستخدام SMO_SVM نظام مصادقة مؤلف النص العربي باستخدام مسافة مانهاتن. يتم استخدام ميزات مختلفة كمدخلات إلى) SMO-SVM أي آلة متجه تعتمد على الحد الأدنى من التحسين المتسلسل). تُظهر تجارب مصادقة المؤلف، في قاعدة البيانات النصية هذه، نتائج مثيرة للاهتمام بدقة تصنيف تصل إلى 80٪. كشف هذا العمل للبحث في النص باللغة العربية عن عدة نقاط مثيرة للاهتمام.

Résumé

Dans cette mémoire, nous examinons un Système d'Authentification d'Auteurs des textes arabes base sur la SMO_SVM en utilisant la distance de Manhattan. Diverses fonctionnalités sont utilisées comme entrées de SMO-SVM (c'est-à-dire une machine vectorielle basée sur une optimisation séquentielle minimale). Les expériences D'authentification d'Auteurs, dans cette base de données textuelles, montrent des résultats intéressants avec une précision de classification de 80%. Ce travail de recherche prouesse rare de texte en langue arabe, a révélé plusieurs points intéressants.

Abstract

In this thesis, we examine a SMO_SVM based Arabic Text Author Authentication System using Manhattan distance. Various features are used as inputs to SMO-SVM (Support Vector Machine based on Minimal Sequential Optimization). The Author Authentication experiments, in this textual database, show interesting results with a classification accuracy of 80%. This work is a rare feat of research of text in Arabic language, revealed several interesting points.