DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
UNIVERSITY OF MOHAMED BOUDIAF-M'SILA

FACULTY OF TECHNOLOGY
ELECTRONICS DEPARTMENT
N° :……..

DOMAIN: SCIENCE & TECHNOLOGY
FILIERE: TELECOMMUNICATIONS
OPTION: ENGINEERING OF TELECOMMUNICATIONS

Dissertation Submitted in partial fulfilment of the requirements for the Master Professional Degree
By:

BELDJOUDI Bochra
&
GAHAM Sabrina

Entitled

Authorship Attribution of Specific Arabic Texts using Character N-gram and Classification Techniques

Diplôme de Master dans le cadre du décret ministériel 1275

Presented: June 29th, 2024 in front of the jury composed of :

Pr. LADJAL Mohamed	University of Mohamed Boudiaf - M'sila	President
Dr. KHENNOUF Salah	University of Mohamed Boudiaf - M'sila	Supervisor
Pr. BOURAS Mounir	University of Mohamed Boudiaf - M'sila	Co-Supervisor
Dr. CHALABI Izzeddine	University of Mohamed Boudiaf - M'sila	Examiner
Dr. BRIK Youcef	University of Mohamed Boudiaf - M'sila	CATI Representative
Mme BAKHTI Yamena	University of Mohamed Boudiaf - M'sila	INCUBATEUR Representative
Pr. BARKAT Abdelhak	University of Mohamed Boudiaf - M'sila	Socio-Economic Partner

June : 2024

ACKNOWLEDGMENTS

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

قال الله تعالى " وَآخِرُ دَعْوَاهُمْ أَنِ الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ" [سورة يونس]

In the name of ALLAH, the greatest and the most clement and merciful.

Thanks to him for having guided us towards the right path and for having helped us throughout our studies…

Our thanks go to our supervisors, Dr. KHENNOUF Salah and Pr. BOURAS Mounir, for agreeing to direct this work, for their support, and their foresight that have been invaluable to us. We would like to thank, also, the members of the jury who gave us the great honor for evaluating this work.

Thanks to all lecturers who spared no effort to give us their knowledge during our university studies.

Finally, we would like to thank everyone who contributed in any way to the preparation of this thesis and to the success of our university career.

DEDICATION

I dedicate this humble work, which is the fruit of my long studies, and deep gratitude

To My support in life my dear father Tamim.

This work is also dedicated to whom it was said that heaven is under her feet, my dear mother, Berra Yamina, for all the support and sacrifices she made.

To my wonderful sisters Fayrouz, Saada, Loubna, Chahd and Hanine.

To my dear husband Zaki.

May ALLAH bless you all, with my wishes for continued health and wellness.

To everyone who participated in this work from near and far.

BOUCHRA

DEDICATION

I dedicate this humble work, which is the fruit of my long studies, and deep gratitude

To My support in life my dear father Mahmoud.

This work is also dedicated to whom it was said that heaven is under her feet, my dear mother, Djerad Fatima, for all the support and sacrifices she made.

To my wonderful sister Donia, and my dear brothers Mounir, Abdelmaoula and Kossay.

To my dear husband Sami.

May ALLAH bless you all, with my wishes for continued health and wellness.

To everyone who participated in this work from near and far.

THANK YOU ALL

SABRINA

ABBREVIATIONS LIST

AA: Authorship Attribution

AAR : Authorship Attribution Rate

ANLP : Arabic Natural Language Processing

ATC: Automatic Text Categorization

BOW : Bag-of-Words

CD : Centroid Driver

CNN : Convolutional Neural Networks

DF : Document Frequency

JGAAP: Java Graphical Authorship Attribution Program

K-NN: K-Nearest Neighbor

LTM : Long-Term Memory

STM : Short-Term Memory

NLP: Natural Language Processing

OCR: Optical Character Recognition

TF: Term Frequency

TF-IDF : Term Frequency-Inverse Document Frequency

TREC: Text Retrieval Conference

UTF-8: Unicode Transformation Format – 8 bit

FIGURES LIST

TABLES LIST

TABLE OF CONTENTS

CHAPTER-2
DOTLESS TEXTS AND RESEARCH METHODOLOGY

CHAPTER-3
EXPERIMENTAL WORK AND RESULTS DISCUSSIONS

INTRODUCTION

INTRODUCTION

In this section, we will present an overview of all the work carried out during the preparation of this thesis. We will start with an introduction to the context and motivation of our work, followed by a discussion of the problematic and the proposed methods, as well as the objectives and contributions. Lastly, the structure of this document and a brief description of each chapter will be provided to connect the various parts of this document.

Context and motivation

Authorship Attribution (AA) is one of the earliest research fields of computational linguistics and has a long history in identifying disputed or unknown authors. Several researchers were interested in many applications of the AA such as email authorship verification, categorizing harassing emails and anonymous messages in textual conversations and social media forensics, online criminality, etc. In addition, the AA can be used to identify the document sources, disputed authorship plagiarism detection in student dissertations, etc.

The AA consists of studying the author writing style (or stylometry) to respond to the following question: Who is the author of this document? Accordingly, the suitable set of features is extracted and combined with the more reliable classification technique to find the right author. Over the years, numerous stylometric features have been explored and utilized in AA. These include sentence length, vocabulary richness, function words, punctuation marks, and characters/words n-gram.

Motivated by the aim of improving Authorship Attribution rates in Arabic texts, both with and without diacritics, this work focuses on investigating new features and classification methods.

Objective and contribution

The primary objectives of this thesis are as follows:

- To explore the possibility and effectiveness of using character N-Gram models for the AA in Arabic texts.
- To examine the most efficient classification algorithm and its performance in distinguishing between different authors' writing styles in Arabic.
- To offer models for studying ancient Arabic texts, and evaluate the robustness of the proposed methods with diverse Arabic text genres.
- To provide insights into the challenges and opportunities inherent in automated AA for Arabic texts and suggest avenues for future research.

Thesis organization

This document is constituted of an introduction and three chapters as follows: The first chapter covers some generalities on Authorship Attribution as well as the different types of classification systems, the fundamental concepts of text exploration, stylometry, and text attribution, with a focus on the Arabic texts.

The second chapter expose the research methodology adopted in this work, besides the proposed approaches and techniques for AA in Arabic specific text documents.

In the third and last chapter, we will present the series of the AA experiments conducted on the textual database (or Corpus) that we conceived for this purpose, which contains two types of texts; dotted texts and dot-free (or dotless) texts.

CHAPTER-1

GENERAL INFORMATION ON AUTHORSHIP ATTRIBUTION

CHAPTER-1

GENERAL INFORMATION ON AUTHORSHIP ATTRIBUTION

1.1 Introduction

The authorship attribution of an unknown or doubtful text is one of the oldest problems in applied statistics to literature. In this chapter, we introduce an overview of authorship attribution, followed by stylometry and automatic text categorization. Finally, definitions and some types of plagiarism are presented.

1.2 Generalities on Authorship Attribution

Authorship attribution (AA) is the process of identifying the likely Authorof a given document, given a collection of documents of known authorship. Attribution of authorship is becoming an important issue as the range of anonymous information increases with rapid growth in Internet use worldwide. Enforcement of authorship attribution includes detecting plagiarism, inferring authorship from inappropriate communications that sent anonymously or pseudonymously, as well as resolving historical issues of unclear or disputed authorship [1].

Attribution of authorship is the way to determine the Authorof of a text when it is not clear who wrote it. It is useful when two or more people claim to have written something.

1.2.1 Author Attribution History

In today's digital age, where textual content is abundantly available online, including research papers, literary works, and user-generated text, the ease of reproducing and sharing these texts has significantly facilitated plagiarism and literary theft. To combat this, scholars have turned to a method known as "Authorship Attribution" (AA) to identify authors of anonymous or unattributed text and to verify authorship.

This approach traces back to the 19th century, with Mendenhall's pioneering attempt in 1887 to identify authors based on their writing style. Subsequent exploration by Zipf and Yule using basic statistical methods laid the groundwork for further developments. With technological advancements, computational methods for AA have evolved, ranging from machine learning techniques to the more recent deep learning and transformer-based approaches.

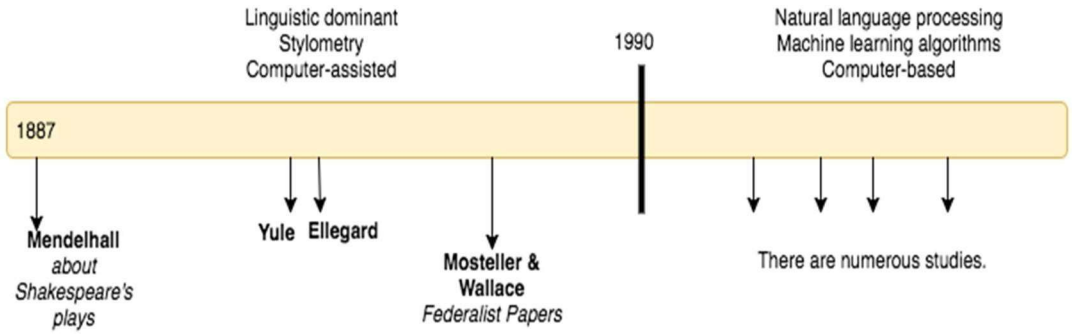


Figure 1.1 : A brief history of Authorship Attribution

In the early 1960s, Mosteller and Wallace's work marked a significant milestone in computer-assisted stylometry, forming the basis for various related fields such as: Authorship Attribution, Authorship Verification, Authorship Discrimination, Plagiarism Detection, and authorship profiling.

The AA serves as a method for identifying the author of an unknown text, distinguishing between different writing styles, and resolving disputes over authorship. Authorship verification aims to determine whether a given text is authored by a specific individual. It is typically framed as a binary classification problem. Authorship discrimination involves assessing whether different texts share the same author [2].

1.2.2 State of the art

Numerous studies have been conducted in the field of authorship attribution in recent years. With the increasing amount of documents on the Internet, and since most writings are anonymous, determining the authorship becomes important. The research focuses on different properties of texts. Two distinct properties of texts are used in classification: the content of the text and the author's style.

Statistical analysis of literary style complements traditional literary scholarship because it provides a means to understand the often-intangible nature of the author's style by quantifying certain characteristics. The majority of stylometric studies use linguistic elements, and most of these elements are lexical-based [3].

1.2.3 Author features

The features employed in Authorship Attribution (AA) can be categorized into various groups:

- Numeric values linked to words (such as word count, characters per word, frequency of character bi-grams/tri-grams within words), also known as lexical traits;
- Values related to sentence syntax (frequency of function words, occurrence of mono/bi/tri-grams of these function words, or sequences of parts of speech);
- Numeric values associated with larger units (such as paragraph count or average paragraph length), essentially structural traits;
- Values associated with thematic content (word bags, keyword n-grams);
- Specificities concerning individual practices (such as spelling or typing errors)

Among these traits, some are specific to types of language and writing. If cutting a text into words is easy in certain cases (by defining a word as a string of characters surrounded by spaces), it is not a trivial task in Chinese or Japanese. Approaches exploiting character n-grams appear to be the simplest for processing any language, as well as the most efficient [4].

The same trait can be attributed to several pairs (text, author), but each text and author do not share a large set of traits. Different sets of traits can be defined to represent texts (and by extension, to represent authors). Considering the existing methods of AA, two main categories of traits can be defined:

- Offline traits: Traits considered a priori relevant for this task with prior knowledge, as widely described by Chaski (2001). These can be defined when the corpus to be processed has not yet been collected.

- Online traits: Traits defined during processing (in the case of supervised methods, based on training and test corpora, such as the character language model described by Peng et al. (2003)). These can only be defined when the corpus to be processed is complete.

Online traits naturally refer to the notion of language independence; no a priori assumptions are made before processing the corpus, and no external linguistic resources are exploited. The method described in this article follows this principle [4].

1.2.4 Author Attribution Steps

A complete author attribution **process** involves several key steps:

- **Collection of texts:** The texts, which are the observations to be classified, are collected. This may include short messages from social networks like Twitter, anonymous letters, contracts, checks, tags, wills, or administrative forms.
- **Feature Extraction:** A feature extraction method is used to calculate the numerical or symbolic information from these observations. This can include techniques such as analyzing word frequency, sentence length, writing style, etc. Convolutional neural networks (CNN) and recurrent short- and long-term memory (LSTM) networks can be used to extract these features.
- **Classification or categorization system:** The classification or categorization system uses the extracted characteristics to classify texts and identify the most likely author. This can be seen as a text classification problem, where supervised learning is used to train the classification models.

These steps make it possible to determine the author of a text by comparing the characteristics of this text with those of other known texts [3].

1.3 Stylometry

Modern technology, particularly computers, enables the analysis of the stylistic framework of a text when the author's identity is disputed. Stylometric analysis, a method that uses computational techniques, allows for determining with a high degree of certainty whether a particular person is the author or not of a given work [3].

1.3.1 Brief history of Stylometry

The first mentions of stylometry to identify authors appeared in 1851. However, given the difficulty of the measurements to be carried out, the first credible studies had to wait for the arrival of modern computers, for their counting precision and their processing at high data speed.

At the beginning of the 1980s, a team of researchers worked to refine and make more efficient the techniques of vitemetry. The work shows that the methods of counting and comparing different texts have been greatly improved.

Stylometry continues to evolve towards ever-greater reliability and sensitivity; it has reached a level, which allows the implementation of a rigorous measurement technique, which gives reliable answers in the analysis of texts of several thousand words. ‘The same author, free flow .

1.3.2 Definition of Stylometry

Stylometry is the quantitative study of literary style using computational distant reading methods. It is based on the observation that each author tends to write in a relatively constant, recognizable, and unique manner. Stylometry combines techniques from linguistics and statistics to identify the style of textual documents.

The style of a text is a characteristic of its broader context of writing, the author, era, and genre. Stylometry aims to show that a text is written in a style different from a collection of other texts. This differentiation allows, to some extent, to determine if an anonymous text was written by a specific author and to determine if a text was not written by a specific author.

Stylometry uses various methods, statistical tests, machine learning, and deep learning, to assess the effect of linguistic features on authorial style. These methods involve analyzing features such as phraseology, punctuation, linguistic diversity, and parts-of-speech to identify the unique characteristics of an author’s writing.

The applications of stylometry are diverse, ranging from identifying the authors of anonymous texts to analyzing literary styles and historical contexts. It is used in various fields, the humanities, literature, and arts, to study and understand literary styles and writing contexts.

1.3.3 Characteristics used in Stylometry

Each individual has their own vocabulary, sometimes rich, sometimes limited. Although an extensive vocabulary is generally associated with quality literature, this is not always the case. Some people write in short sentences, while others prefer complex sentences with multiple clauses. No two authors use semicolons, dashes, and other punctuation marks exactly the same way.

Identifying the author of an anonymous text, however, is one of the most common applications of stylometry. It is sometimes possible to discover the identity of the author of a text by measuring certain characteristics of that text, such as the average length of sentences or the ratio between the number of definite and indefinite articles.

These measurements are then compared with those observed in texts whose authors are known. When we talk about non-contextual words, we refer to words that are often interchangeable or can even be omitted without loss of the general meaning of the text. These words contribute little to contextual information and are often consciously ignored by both the reader and the author.

These words typically constitute 20 to 45% of the total text, which allows for a significant number of statistical choices, and the more statistical measures there are, the more reliable their results are [3].

1.4 Automatic Texts Categorization

1.4.1 Definition of Text Categorization

Text categorization involves seeking a functional relationship between a set of texts and a set of categories (labels, classes). This functional relationship, also known as a prediction model, is estimated through machine learning methods.

For this, it is necessary to have a set of previously labeled texts, called a training set, from which we estimate the parameters of the most accurate prediction model possible, meaning the model that produces the fewest prediction errors [5].

1.4.2 History of Text Categorization

Text categorization is quite an old discipline. In 1627, Gabriel Naudé proposed a classification based on five main themes: theology, jurisprudence, history, sciences and arts, and belles-lettres. The desire to master the universe was evident in the multiplication of encyclopedias. Diderot's encyclopedia (published between 1751 and 1772) was organized alphabetically with associative cross-references, while Panckoucke's encyclopedia (published from 1776 to 1780) followed a methodical organization in a hierarchical order (Fayet & Scribe, 1997).

The thematic classification system, which appeared at the dawn of writing and was institutionalized in Alexandria, led to the creation of a "universal" classification system by Dewey in 1876. This system is a documentary classification of an encyclopedic type. However, the idea of classifying texts by machines dates back to the early 1960s and saw significant progress from the 1990s with the advent of much more efficient algorithms.

Until the early 1980s, building a classifier required substantial human resources. Several experts manually edited rules and refined them during testing. The advent of machine learning resulted in significant time savings, as it was no longer necessary to reconfigure the entire system in case of a change in the hierarchical structure. These technological advancements and advanced algorithms have made categorization a reliable tool today.

In the early 1990s, research mainly came from the Information Retrieval (IR) community. Digitization methods, classification algorithms, and testing methodologies were adapted to text categorization, particularly during the TREC (Text REtrieval Conference) events. The Machine Learning (ML) community also took an interest in this problem about ten years ago, considering it an application domain for pattern recognition algorithms.

Currently, text digitization methods remain largely inspired by IR, while the most efficient classifiers come from ML. Another community, composed mainly of statisticians and linguists, also addresses the problem of text categorization using data analysis methods. The goal here is not to create a system that automatically classifies documents without human intervention but to extract synthetic information from the corpus. Problems addressed here include the study of literary genres or determining the author of a text [6].

1.4.3 Problems of Text Categorization

- Ambiguity of meanings: Texts can have ambiguous meanings, making precise classification challenging.
- Feature selection: It is crucial to choose appropriate features to represent the texts, as this directly affects the performance of the classification model.
- Imbalanced data: Data can be imbalanced, meaning some categories have many more examples than others do. This can lead to biases in classification results.
- Data heterogeneity: Texts can originate from different domains and languages, making it difficult to generalize classification models [7].

1.5 Categorization Systems

The objective of automatic text categorization is to classify text documents into predefined categories automatically, without direct human intervention. This can be achieved using machine learning and natural language processing (NLP) techniques [8].

1.5.1 Applications of Text Categorization

Automatic Text Categorization (CAT) has numerous practical applications in various fields, including:

- **Language Identification:** Text classifiers can detect the language used in a document, which is essential for machine translation tools and linguistic analyses

- **Author Recognition and Multimedia Document Categorization**: CAT techniques can be used to identify the author of a text or to classify multimedia documents such as images or video
- **Document Tagging:** CAT allows for assigning tags or categories to documents to facilitate their search and navigation.
- **Filtering:** Text classifiers can determine whether a document is relevant based on specific criteria, which is useful for filtering emails or online documents.
- **Routing:** CAT enables the assignment of a document to one or multiple categories among n, which is essential for document organization and retrieval.
- **Fraud and Abuse Detection:** CAT techniques can be used to detect fraudulent or abusive messages on social media or in emails.
- **Improving Web Search:** Text classifiers can enhance the relevance of search results by categorizing documents based on their content [9].

1.5.2 Approach for Text Categorization

To carry out the operation of automatic text categorization as we have defined, the common approach is as follows; Text Preprocessing, which is the first phase, involves formalizing the texts so that they are understandable by the machine and usable by learning algorithms. Document categorization, which is the second phase, determines whether learning techniques can produce a good generalization from (Document, Class) pairs. Improving Model Performance that is used to enhance model performance, an evaluation of classifier quality and comparison of results provided by different models is conducted at the end of the cycle. The approach to a standard automatic text classification can be given as follows:

❶ Remove separation characters, punctuation marks, stop words, etc.
❷ The remaining terms are all attributes
❸ A document becomes a vector of <term, frequency>
❹ Train the classification model from (Document, Class) pairs.
❺ Evaluate the classifier's results [6].

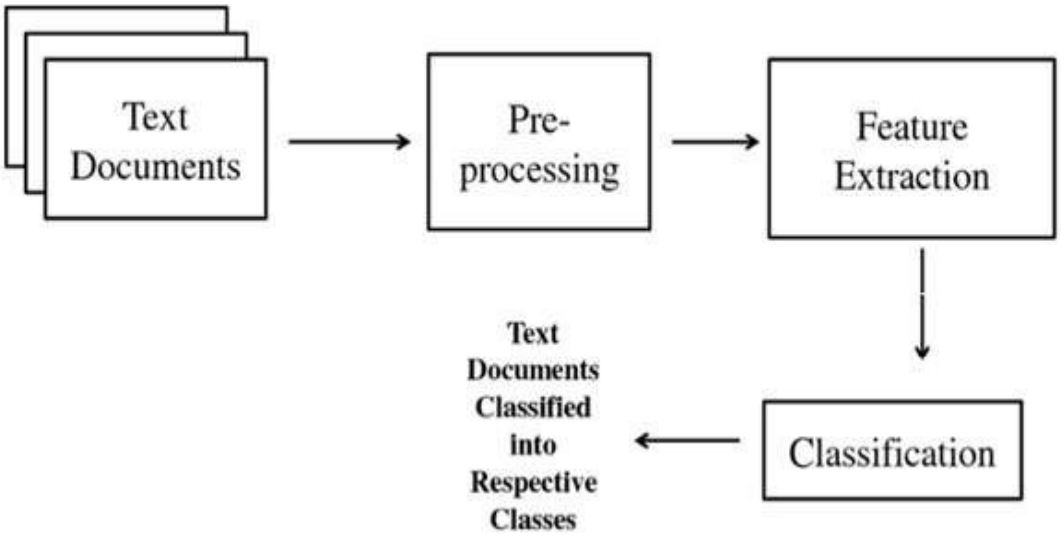


Figure-1.2 : The basic architecture of text categorization

1.6 Plagiarism

1.6.1 Definition of Plagiarism

Plagiarism is a practice that is encountered in all areas of human activity that involve creativity: literature, painting, music, fashion, etc. It is the use of another person's work, ideas, expressions, or intellectual property without proper acknowledgment or permission, and presenting it as one's own original creation.With the aim of manipulating an intriguing location [10].

1.6.2 Types of Plagiarism

We will formulate a general definition of plagiarism, which we will complete by the presentation of twelve types of plagiarism that range on a continuum of severity, ranging from conscious plagiarism to unconscious plagiarism. When necessary, additional distinctions deemed will be made to clarify conduct related to plagiarism without actually being plagiarism [10].

Table-1.1: Different types of Plagiarism

Plagiarism types	Definitions
Direct Plagiarism	Copying someone else's work without giving credit.
Self-Plagiarism	Reusing one's own previous work without permission.
Mosaic Plagiarism (or Patchwriting)	Combining parts of different sources to create a new text without proper citation.
Accidental Plagiarism	Accidentally using someone else's work without right citation.
Global Plagiarism	Passing off an entire text by someone else as one's own work.
Verbatim Plagiarism	Directly copying someone else's words without giving credit.
Paraphrasing Plagiarism	Rephrasing someone else's ideas without giving credit.
Patchwork Plagiarism	Stitching together parts of different sources to create a new text without proper citation.
Hired Plagiarism	Paying someone to write a piece of work and passing it off as one's own.
Idea Plagiarism	Using someone else's ideas without giving credit.
Inaccurate Authorship Plagiarism	Miscrediting authors in an academic research paper.
Concealing Sources	Not making it obvious where one is drawing on someone else's work

1.7 Conclusion

In this chapter, we presented some generalities on author attribution (AA) brief history we showed traits of an author and the stages of author attribution. Subsequently, definitions of vitemetry as well as a brief history of the latter are presented. On the other hand, we cited the automatic categorization of texts and we discussed the definition and some types of plagiarism.

In the next chapter, we will present the dotless texts and the research methodology as well as the techniques used in this research work.

CHAPTER-2

DOTLESS TEXTS AND RESEARCH METHODOLOGY

CHAPTER-2
DOTLESS TEXTS AND RESEARCH METHODOLOGY

2.1 Introduction

In this chapter, we will DISCOVER DOTLESS ARABIC TEXTS (DAT) and present the proposed methods for their analysis. The DAT's, which are texts with diacritics removed, present a unique challenge in Arabic text processing. Diacritics, such as dots above or below letters, are crucial indicators of pronunciation and grammatical structure in Arabic. However, in many cases, the presence of diacritics may not be desired or available, necessitating the development of robust methods for analyzing this type of texts effectively.

We will begin by discussing the significance of diacritics in Arabic text and the challenges met by their absence. Afterward, we will provide a comprehensive review of the proposed methods and approaches for processing dotless texts. These methods will be evaluated based on their effectiveness in understanding and analyzing these texts, considering factors such as precision, efficiency, and applicability in real-world scenarios.

2.2 Generalities on Dotless Arabic Texts

Arabic serves as the official language across 22 nations within the Arab world, wielding a native speaker base exceeding 400 million individuals. Moreover, it functions as a secondary language for numerous non-Arabic adherents within the Muslim community. Recognized as one of the United Nation's six official languages, Arabic language occupies a significant linguistic sphere and has ascended to become the fourth most prevalent language on the internet, experiencing rapid growth in user engagement between 2013 and 2018.

The Arabic language itself exhibits a tripartite classification; Classical Arabic (CA), Modern Standard Arabic (MSA) and Dialectical Arabic (DA). The CA represents the linguistic incarnation within holy texts, strongly interlinked with Islamic cultural and religious heritage, as well as ancient Arabic literature. Whereas, the MSA constitutes the formal register employed in official documentation and news dissemination. Finally, the DA includes a multi Computational Linguistics **[11]**.

2.3 Evolution of Arabic Language Dotting

2.3.1 Original Arabic script

Before the development of the Arabic script, which was originated in the Arabian Peninsula in the 4th century AD, the Arabic language was primarily a spoken language, and various scripts were used for writing in the region, such as Nabatean and Syriac scripts.

The Arabic script is written from right to left and consists of 28 letters. It is a cursive script, meaning that letters within a word are connected to one another. Over time, the Arabic script has been adapted and modified to write various languages, including Persian, Urdu, etc **[12]**.

2.3.2 Adoption of Dots in Arabic language

Originally, Arabic script did not have the familiar dots on certain letters as seen today. However, with the growing number of non-Arab converts to Islam from regions beyond the Arabian Peninsula, difficulties arose in distinguishing between similar letters, leading to misreading. Especially, letters like "ب، ت، ث" were mainly challenging to differentiate, specifically for recent converts to Islam who were not native Arabic speakers.

Thus, the introduction of dots in Arabic script emerged as a solution. The initial document to undergo dotting in Arabic was inscribed on parchment paper, dating back to 22 Hijri (643 AD) **[12]**.

2.3.3 Pioneers of dot introduction in Arabic

The first one to introduce dots in Arabic language was "Abu al-Aswad al-Du'ali," a developer in the field of Arabic grammar. He placed dots as marks of pronunciation aids in the language. A dot above a letter indicated a "fatḥah" (a short vowel sound), while a dot below indicated a "kasrah" (another short vowel sound). If there were two dots, they indicated a "ḍammah" (a third short vowel sound). The method introduced by Abu al-Aswad al-Du'ali became widespread, but primarily in the copies of the Quran, and it was not commonly used in newspapers and books.

This style of writing persisted until "al-Khalil ibn Ahmad al-Farahidi" emerged in the second century of the Islamic calendar. He devised a more precise system than Abu al-Aswad al-Du'ali's, replacing dots with small slanted alifs to indicate the fatḥah, placing small ya below the letter to indicate the kasrah, and adding a small waw above the letter to indicate the ḍammah. In the case of "tanwin" (nasalization), the letter would be repeated twice. He played a significant role in introducing the "hamzah" (glottal stop) and the "shaddah" (gemination) in Arabic language. Additionally, he introduced the "ta' marbutah" (ta with a tied knot) and the "alif maqsura" (short alif) into the language **[12]**.

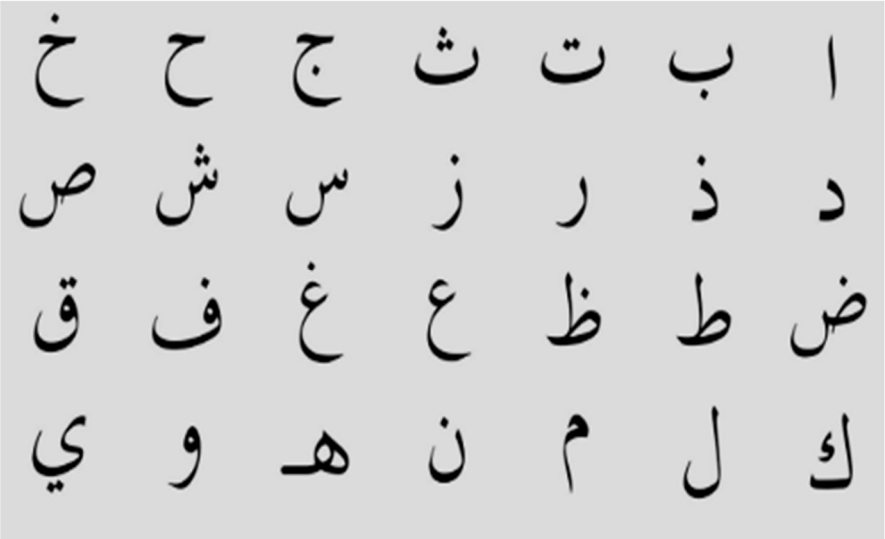


Figure-2.1: Arabic dotted letters

2.3.4 Impact of Diacritical Marks on Arabic Language

Diacritical marks in Arabic, known as 'harakat', include dots and other signs that denote short vowels, doubling of consonants, and other phonetic details. These marks are crucial for precise articulation and disambiguation of words that would otherwise look identical. For example, the word "علم" can mean "science" (ilm) or "flag" (alam) reliant on the diacritics used.

2.3.4.1 Diacritical marks (Tashkeel)

Diacritics are small symbols placed above or below Arabic letters. They indicate vowel sounds, elongation, and other phonetic features. The common used diacritics include **[13]:**

- Fatha : Indicates the short "a" sound.
- Kasra : Represents the short "i" sound.
- Damma : Denotes the short "u" sound.
- Sukun : Indicates a consonant without a vowel.
- Shadda : Indicates consonant doubling.
- Tanween : Represents the nasalized vowels "an," "in," and "un."

2.3.4.2 Function of diacritics:

The main function of the diacritics is to enhance pronunciation accuracy. Diacritics guide readers on how to pronounce words correctly, how to differentiate between letters that look similar (e.g., "ب" vs. "ت") and clarify grammatical structures and verb forms.

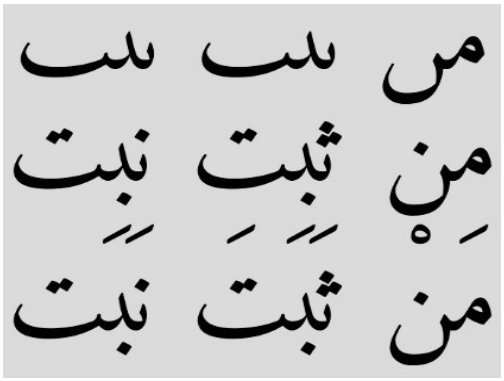


Figure-2.2: Arabic text with and without dots and diacritics

2.4 Outlines of Research Methodology

Our research methodology is based on six steps as follow:

❶ The first step is to convert scanned images into editable texts (word document) using Optical Character Recognition (OCR) system.

❷ The second step is to correct the obtained editable texts in order to prepare them for Author Attribution.

❸ The third step consists of the elimination of the dots and diacritics using the website: ***https://seen-arabic.github.io/Arabic-Services/***, to obtain two types of text: Dotted Texts (DTs) and Dotless Texts (DsT).

❹ In the fourth step, the DTs and DsT are converted to a raw text file using UTF-8 encoding to ensure file compatibility with a range of operating systems and programs.

❺ In the fifth step, we divide (or split) the Dataset into two Datasets : Training Dataset (containing 2/3 of the texts = 60 texts) and Testing Dataset (containing 1/3 of the texts = 30 texts)

❻ In the sixth step, we extract relevant features (n-gram characters) to build a model for each author using.

❼ The seventh and final step was devoted to classification methods using CM and K-NN methods to implement the process of identifying the true author of the text.

2.4.1 Conversion of scanned images to editable texts using OCR

The OCR is a conversion of scanned or printed text images, handwritten text into editable text for further processing. This technology allows machine to recognize the text automatically. It is like combination of eye and mind of human body. An eye can view the text from the images but actually, the brain processes as well as interprets that extracted text read by eye. In development of computerized OCR system, few problems can occur.

First: there is very little visible difference between some letters and digits for computers to understand. For example, it might be difficult for the computer to differentiate between digit “0” and letter “o”. Second: It might be very difficult to extract text, which is embedded in very dark background or printed on other words or graphics.

In 1955, the first commercial system was installed at the reader’s digest, which used OCR to input sales report into a computer and then after OCR method has become very helpful in computerizing the physical office documents **[14]**.

Figure-2.3: Conversion of scanned texts into editable texts using OCR.

2.4.2 Correcting the obtained editable texts

The obtained editable texts using OCR have two types of errors; insignificant or (noise) characters (special characters, numbers, etc.) and incorrect characters. Insignificant characters are characters that do not have a well-defined meaning in the Arabic language and which appear by mistake in texts converted using an OCR system (i.e. ", %, &, £, *, # , $, 0, 1, …, 9, etc.). However, incorrect characters are characters that are incorrectly converted or converted by mistake into characters other than the real characters. The latest are left in order to test the robustness of our proposed method. Consequently, the correction applied in this phase consists of :

- Removing insignificant characters,
- Removing Arabic diacritics,
- Removing multiple word spaces,
- Removing signs,

2.4.3 Elimination of dots and diacritics

Eliminating dots and diacritics from Arabic text involves using an online tools that can process and manipulate text to remove specific characters or modify its format. This service helps to change texts' format for educational or research purposes, and is used in text analysis or linguistic experiments. The site relies on natural language processing techniques to ensure accurate and easy conversion. The site used in our study is : ***https://seen-arabic.github.io/Arabic-Services/***.

Here are general steps followed to achieve this:

- Copy and Paste the editable Text into a text box provided on the site.
- Select Processing Options presented by the website to remove dots (part of some Arabic letters) and diacritics (harakat, such as fatha, kasra, damma, etc.).
- Click on the button to process the text and generate a version of your text with dots and/or diacritics removed.
- Copy the resulting text from the website and use it as needed.

2.4.4 Conversion of word document to UTF-8

After the correction of the different errors, we proceed to the conversion of these word documents to the Unicode Transformation Format – 8 bit (UTF-8) format. It should be noted that UTF-8 encoding was used to encode all the texts in the corpus, because the latter covers a very large number of characters, and is implicitly capable of encoding the majority of languages since it is encoded on 4 bytes. On the other hand, the use of this format costs in terms of calculation time and memory. It is designed to be backward compatible with ASCII and to avoid the complications of endianness and byte-order marks.

2.4.5 Dataset splitting

Dataset splitting refers to the practice of dividing a dataset into distinct subsets for the purpose of training and evaluating machine learning models. This process is crucial in ensuring that the model is capable of generalizing well to unseen data, instead of just memorizing the training data.

There are different strategies for Dataset splitting such as : Train-Test, Train-Validation-Test, Cross-Validation, Leave-One-Out Cross Validation. In this study, we have chosen to use the Train-Test strategy.

2.4.6 Extracting relevant features

Extracting relevant features refers to the process of identifying and selecting specific characteristics from a text that can be used to determine its authorship. These features help differentiate between the writing styles of different authors and are crucial for building accurate models for authorship attribution. Some common types of features used in the AA are :

- Lexical Features (i.e. Word frequency, function word, …)
- Syntactic Features (i.e. Sentence length, use of punctuation, …)
- Stylistic Features (i.e. Use of passive vs. active voice, formal vs. informal, …)
- Character-level Features (i.e. Character n-grams, letter frequency, …)

In the present study, we have chosen to use the Character n-grams as relevant feature.

2.4.7 Classification methods

The classification method refers to the phase where a trained model is used to assign a text to its most likely author based on the extracted relevant features from the text. This phase is crucial for determining the authorship of anonymous texts as it determines the final output of the AA process and controls the learned patterns and relationships to make informed predictions about the new, unknown texts.

The accuracy and reliability of this phase depend on the quality of the feature extraction, the robustness of the model, and the similarity of the test texts to the training data. In our case, we have used the k-Nearest Neighbors (k-NN) which classify based on the similarity of the test text to the training texts and the Centroid Method (CM) which involves texts as vectors in a high-dimensional space and using the concept of centroids, or average vectors, to classify authorship.

2.5 Motivation for Studying Dotless Arabic Texts

The main motivation for studying Dotless Arabic Texts (DAT) is to simplify the learning of Arabic for non-native speakers. This effort aims to prevent difficulties that foreigners may face in understanding and reading Arabic texts due to the presence of diacritics that are not always present in written Arabic texts.

The study of DAT can also contribute to the development of tools for various NLP tasks, such as : Sentiment Analysis and Language Modeling, which are not well-explored in the Arabic texts without diacritics.

Moreover, using DAT can be useful in OCR for ancient manuscripts that were written without diacritics. It can also enhance processing tools for social media platforms, where Arabic content is becoming more common. [15].

2.5.1 Reduction of the number of Arabic characters

When Arabic text is rendered without dots (diacritics), the number of distinct characters is reduced to 15. This is because the following 11 characters are distinguished only by their dots, and without dots, they appear the same as other characters. (See table-2.1 below)

Table-2.1: Reduction of Arabic characters by removing dots and diacritics

Letter	Mapping	Letter	Mapping	Letter	Mapping	Letter	Mapping
ا	ا	س	س	ف	ف	ح	ح
م	م	ش		ق		ج	
هـ	هـ	ص	ص	ك	لـ	خ	
و	و	ض		لـ		ب	ـب
د	د	ط	ط	ي	ى	ت	
ذ		ظ		ى		ث	
ر	ر	ع	ع			ن	
ز		غ					

2.5.2 Importance of dotless texts

Using DAT presents a promising avenue to address various challenges encountered in Arabic (ANLP). The inherent density of Arabic morphology often results in a considerably expansive language vocabulary. However, dotless text adoption could aid in mitigating this issue by consolidating many dotted words into a singular dotless homographic word. This consolidation contributes to reducing the overall vocabulary size, offering a potential solution within ANLP

Moreover, this line of inquiry holds significant relevance in the realm of automatic recognition of ancient parchments where the script used was predominantly dotless. By leveraging dotless text methodologies, researchers and practitioners can enhance their capabilities in deciphering and analyzing historical documents etched in dotless Arabic script.

Our primary objective is to assess the efficacy of context-aware deep learning models specifically when applied to dotless text, contrasting their performance with the dotted variant. To explore this, we propose leveraging language modeling as a fundamental task within NLP [16].

2.6 Proposed method

2.6.1 Feature extraction

Feature extraction is a fundamental step in stylometry, which deals with the analysis of writing style. It involves identifying and extracting quantifiable characteristics from texts that reflect an author's unique way of writing. These characteristics are used to train a machine-learning model to classify text into different categories. The choice of features will depend on the specific text classification task and the type of machine learning model being used.

In this study, we have chosen the N-gram character as features. The N-gram character capture sequences of N character that appear together in a text document. Common n-gram features include: Unigrams (N=1), Bigrams (N=2), Trigrams (N=3), etc…

N-gram models can be trained on large corpora of text data, and the probabilities of character sequences are calculated based on their frequency of occurrence. This allows for the prediction of the next word in a sequence, which is useful in applications such as auto-completion systems. The size of the N-gram can be adjusted to capture more or less context, and larger N-grams can provide predictions that are more accurate but require larger amounts of training data [17].

2.6.2 Technique of classifications

2.6.2.1 k-NN method

The k-Nearest Neighbors (k-NN) method is a simple and powerful algorithm used for both classification and regression tasks in machine learning. It is one of the techniques considered to be among the top 10 for data mining. It is based on the idea that similar data points are close to each other in a feature space.

The k-NN algorithm relies on a distance metric (Euclidean, Manhattan or Minkowski) to measure how similar two instances are. It tries to classify an unknown sample based on the known classification of its neighbors.

Given an unknown sample and a training set, all the distances between the unknown sample and all the samples in the training set can be computed. The distance with the smallest value corresponds to the sample in the training set closest to the unknown sample. Therefore, the unknown sample may be classified based on the classification of this nearest neighbor [18].

The Advantages of k-NN are :

- The k-NN algorithm is one of the simplest machine learning algorithms, making it easy to understand and implement.
- The k-NN does not make any assumptions about the underlying data distribution, making it a flexible algorithm that can be applied to a wide range of problems.
- KNN can be used for both classification and regression tasks.

The disadvantages of KNN are :

- KNN can be sensitive to outliers in the training data. Outliers can have a significant impact on the algorithm's predictions, especially when the value of K is small.
- The computational cost of KNN can be high, especially for large datasets. This is because the algorithm needs to calculate the distance between the new data point and each data point in the training set.
- KNN can suffer from the curse of dimensionality, which means that its performance can degrade as the number of features in the data increases.

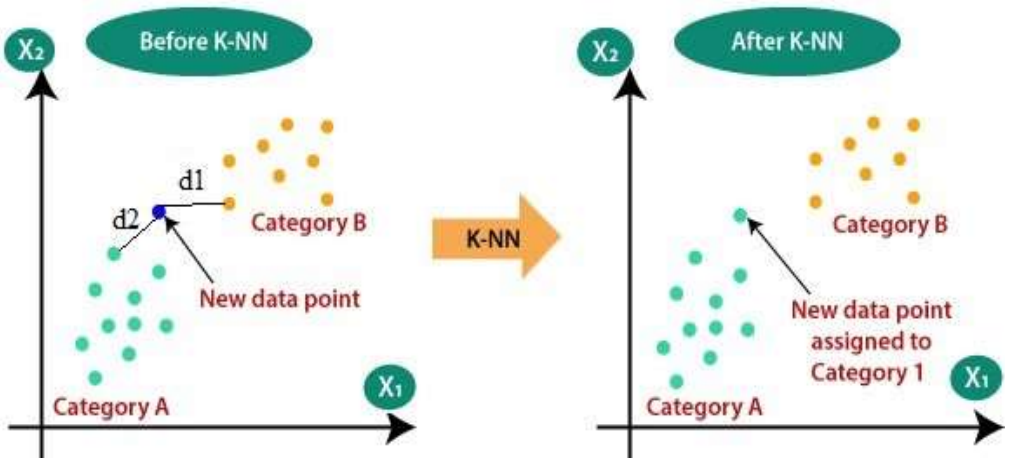


Figure-2.4 : The k-NN method

2.6.2.2 Centroid Method

The Centroid Method (or Centroid Driver), also known as the centroid algorithm or center of gravity algorithm, is a variant of the well-known Rocchio algorithm, used for classifying textual documents using geometric distance measures.

During the learning phase, centroids are computed to represent each class. The centroid vector for a category is defined as the average of the vectors of the texts in that category, which acts as the barycenter of the different texts. So, this method calculates the distance between the geometric centers of gravity of the different categories.

To classify a new document, the similarity between the centroid vector of each category and the vector of the new document is calculated. The new document is then assigned to the category whose centroid is closest, in the case of disjoint multi-class classifications. Otherwise, scores are computed for each class, and a thresholding technique is used to determine the classes to assign [19].

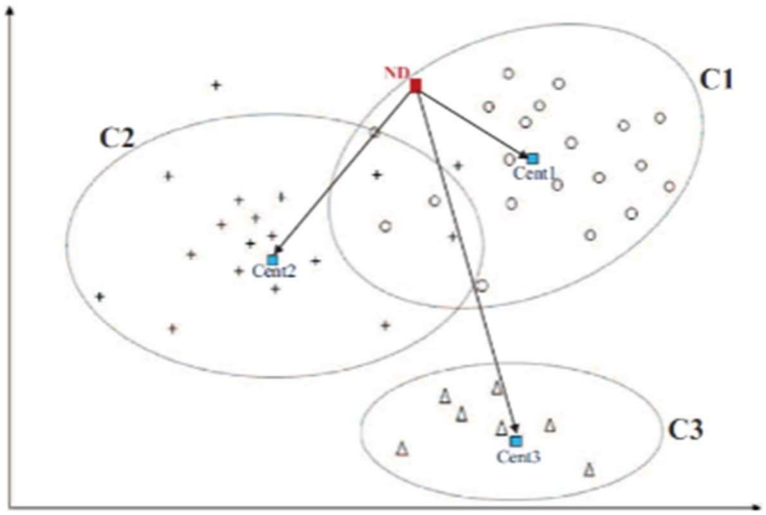


Figure-2.5 : The Centroid Method

The Advantages of the CM algorithm are :

- It is very effective when each class is well defined by a prototype.
- Since comparisons are not only made with centroid vectors, the CM is more efficient than k-NN for real-time categorization.
- It is widely used in signal and image processing, particularly for computing the center of gravity corresponding to spectra or pixels in various treatments, such as segmentation [19].

The disadvantages of the CM algorithm are :

- The CM is influenced by extreme values or outliers in the data.
- If the data distribution is non-spherical or complex, CM may not accurately represent the class.
- The centroid method is known for its efficiency and robustness, but optimization can be challenging, especially with a large number of classes.

2.7 Conclusion

In this chapter, we presented the dotless Arabic texts, the research methodology followed in order to achieve the objectives of this work. Furthermore, we presented the empirical evaluation of the proposed approach for the Author Attribution using the corpus that we designed for this purpose.

In the next chapter, we will present the different experiments carried out and illustrate the results obtained as well as the discussions and conclusions reached.

CHAPTER-3
EXPERIMENTAL WORK AND RESULTS DISCUSSIONS

CHAPTER-3
EXPERIMENTAL WORK AND RESULTS DISCUSSIONS

3.1 Introduction

In this chapter, we will present a series of Authorship Attribution experiments conducted on our corpus. The texts of the corpus underwent a series of experiments to assess how the presence or absence of diacritical marks affects the Authorship Attribution Rate (AAR). Subsequently, we will examine the obtained results and attempt to provide objective interpretations and conclusions.

3.2 Evaluating corpus

Experimental evaluation plays an important role in text classification. Using test corpora, we can observe the impact of diverse parameters on the AAR. Nevertheless, previous studies on texts without diacritical marks have a relatively limited number of corpora. Moreover, the number of possible authors remains limited, as it is challenging to find a significant number of potential candidates who meet multiple constraints (same period, same language, similar cultures, similar themes, etc.).

For this reason, we made the decision to create our own corpus, that we called Dotless Arabic Texts corpus (DAT'24).

3.2.1 Corpus description

The DAT'24 consists of 15 authors (8 men and 7 women), each having written 6 texts with an average length of 2000 words which are numbered from 1 to 6. We have removed the dots and diacritics from them through a special site to finally obtain 6 texts without dots.

Hence, we have 12 texts for each writer: 6 with dots and 6 without dots and diacritics. These texts, which are obtained by using Optical Character Recognition (OCR), are classified into two groups: Dotted Texts (DTs) and Dotless Texts (DsT) based on the presence or absence of diacritical marks.

3.2.2 Constituents of the Corpus

The DAT'24 corpus, that we have conceived, includes 15 contemporary Arab writers (7 female and 8male), namely : Ghada Samman, Huda Barakat, Colette Sohail, May Ziadah, Ahlam mosteghanemi, Nazik malaika, Assia Jabbar, Mahmoud Al-Akad, Haidar Haidar, Abdul Kader Mazini, Taha Hussein, Tawfiq Hakim, Hana Mina, Nadjib Mahfoud, Lotif Al-Manfalouti , and Nadjib Mahmoud.

For each author, we randomly extract a specific number of pages containing the chosen word count. Each author has six texts, averaging a length of 2000 words, which are then subjected to the process of removing dots and diacritics to obtain dotless texts. The considered texts were extracted from the novels of the authors mentioned above. Detailed information about the writers and their texts used in our corpus is provided in the following tables.

Table-3.1: Summary of the Corpus DAT'24 (Female Writers)

Writers	Native Country	Period	Number of books	Used Language	Texts Names	Nbre de Word / text	Utilisation
Assia Djebar	Algérie	1936-2015	26 -livres	AR/FR	Asia_1	2130	Training
					Asia_2	1956	Training
					Asia_3	2408	Training
					Asia_4	2361	Training
					Asia_5	2161	Test
					Asia_6	2510	Test
Ghada Saman	Syrie	1942 – to this day	46-livres	AR	Ghada-1	3088	Training
					Ghada-2	1825	Training
					Ghada-3	2960	Training
					Ghada-4	1697	Training
					Ghada-5	3273	Test
					Ghada-6	3151	Test
Kolite Sohil	Syrie	1931 – to this day	29-livres	AR	Kolite-1	2329	Training
					Kolite-2	1608	Training
					Kolite-3	1995	Training
					Kolite-4	1465	Training
					Kolite-5	1664	Test
					Kolite-6	1969	Test
May Ziada	Palestine	1886 – 1941	19-livres	FR-EN-ES-IT-	May-1	1536	Training
					May-2	1617	Training
					May-3	1580	Training
					May-4	1612	Training
					May-5	1638	Test
					May-6	1606	Test
Ahlam mosteghanemi	tunisia	1952 – to this day	34-livres	AR	Ahlam -1	1880	Training
					Ahlam -2	1844	Training
					Ahlam -3	1878	Training
					Ahlam -4	2225	Training
					Ahlam -5	1896	Test
					Ahlam -6	1868	Test
Houda barakat	liban	1952–to this day	12–livres	AR	Houda-1	2098	Training
					Houda-2	1984	Training
					Houda-3	2073	Training
					Houda-4	1697	Training
					Houda-5	3273	Test
					Houda-6	1929	Test
Nazik malaika	irak	1923–2007	25–livres	AR	Nazik-1	3211	Training
					Nazik-2	1682	Training
					Nazik-3	1035	Training
					Nazik-4	886	Training
					Nazik-5	1015	Test
					Nazik-6	840	Test

Table-3.2 : Summary of the Corpus (Male Writers)

Writers	Native Country	Période	Number of books	Used Language	Texts Names	Nbre de Word / text	Utilisation
Najib Mahfoud	Egypte	1911-2006	49 livres	AR	Najib -1	2749	Training
					Najib -2	3115	Training
					Najib -3	2491	Training
					Najib -4	2426	Training
					Najib -5	2638	Test
					Najib -6	2042	Test
Abdelkader Mazini	Egypte	1889 - 1949	19 livres	AR-EN	Mazini-1	2075	Training
					Mazini-2	2024	Training
					Mazini-3	2088	Training
					Mazini-4	1979	Training
					Mazini-5	2078	Test
					Mazini-6	1974	Test
Haidar Haidar	Syrie	1936- to this day	40 livres	AR	Haider-1	1911	Training
					Haider-2	2282	Training
					Haider-3	2018	Training
					Haider-4	2196	Training
					Haider-5	2710	Test
					Haider-6	2105	Test
Hana Mina	Syrie	1924-2018	21 livres	AR	Hana-1	2026	Training
					Hana-2	2315	Training
					Hana-3	1984	Training
					Hana-4	1960	Training
					Hana-5	2363	Test
					Hana-6	2069	Test
Mahmoud Akad	Egypte	1889 - 1964	689 livres	AR	Akad-1	2033	Training
					Akad-2	2011	Training
					Akad-3	2023	Training
					Akad-4	2036	Training
					Akad-5	2093	Test
					Akad-6	2072	Test
Taha Hocin	Egypte	1889-1973	381 livres	AR-FR-LAT	Taha-1	2028	Training
					Taha-2	2034	Training
					Taha-3	2101	Training
					Taha-4	2082	Training
					Taha-5	2069	Test
					Taha-6	2035	Test
Toufik Hakim	Egypte	1898-1987	233 livres	AR	Toufik-1	2010	Training
					Toufik-2	1973	Training
					Toufik-3	2012	Training
					Toufik-4	1999	Training
					Toufik-5	2024	Test
					Toufik-6	1982	Test
Lotfi Manfalouti	Egypte	1876-1924	107 livres	FR	Lotfi-1	2691	Training
					Lotfi-2	2384	Training
					Lotfi-3	2779	Training
					Lotfi-4	2809	Training
					Lotfi-5	1929	Test
					Lotfi-6	2672	Test

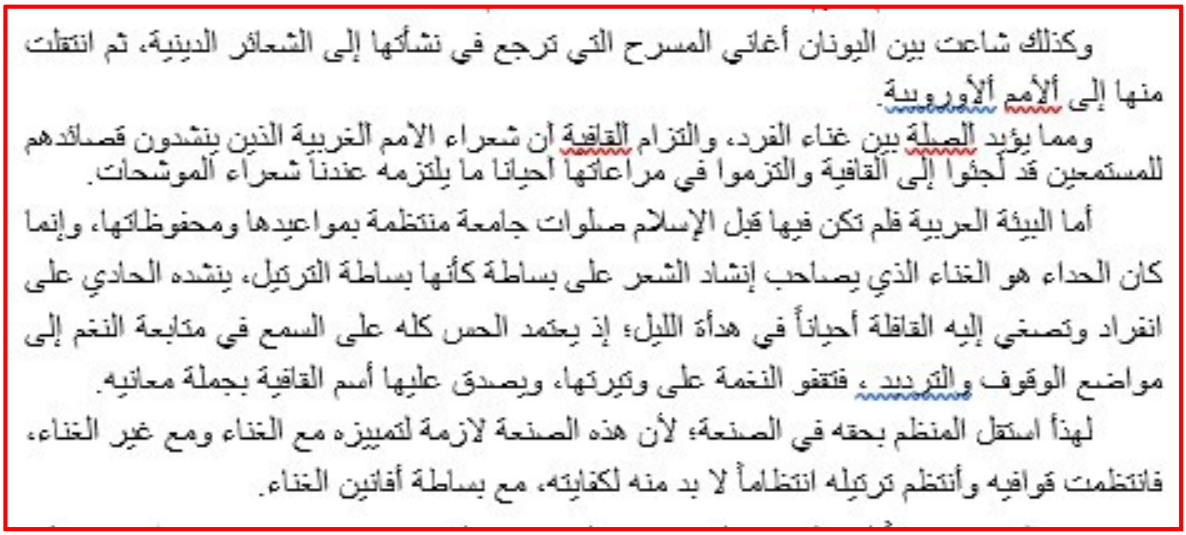
3.3 Preparation of corpus documents

The collected documents must be prepared before they can be used for Author Attribution. The preparation phase can be summarized in the following operations:

- Scan of the selected pages and save them in (.jpeg/png) format,
- Convert the obtained image to Word file using OCR,
- Carry out the pretreatment operations mentioned in section (2.6.2),
- The resulting text documents are saved in UTF-8 format,
- The Dataset is divided into two subsets (training and testing) according to the rule (2/3 for training and 1/3 for testing);

3.3.1 Examples of texts obtained after an OCR operation

After the document-scanning process using OCR, we obtain a document that can be treated as editable documents (in Word format) to correct errors, add, or remove any extraneous content. Below, we review examples of texts that we obtained after the OCR operation in order to use them in upcoming experiments (each color representing a type of error)


وكذلك شاعت بين اليونان أغاني المسرح التي ترجع في نشأتها إلى الشعائر الدينية، ثم انتقلت منها إلى الأمم الأوروبية.

ومما يؤيد الصلة بين غناء الفرد، والتزام القافية أن شعراء الأمم الغربية الذين ينشدون قصائدهم للمستمعين قد لجئوا إلى القافية والتزموا في مراعاتها أحيانا ما يلتزمه عندنا شعراء الموشحات.

أما البيئة العربية فلم تكن فيها قبل الإسلام صلوات جامعة منتظمة بمواعيدها ومحفوظاتها، وإنما كان الحداء هو الغناء الذي يصاحب إنشاد الشعر على بساطة كأنها بساطة الترتيل، ينشده الحادي على انفراد وتصغي إليه القافلة أحياناً في هدأة الليل، إذ يعتمد الحس كله على السمع في متابعة النغم إلى مواضع الوقوف والترديد، فتقفو النغمة على وتيرتها، ويصدق عليها أسم القافية بجملة معانيه.

لهذا استقل المنظم بحقه في الصنعة، لأن هذه الصنعة لازمة لتمييزه مع الغناء ومع غير الغناء، فانتظمت قوافيه وأنتظم ترتيله انتظاماً لا بد منه لكفايته، مع بساطة أفانين الغناء.

Figure-3.1: Example of uncorrected text.

إنما الوزن المقسم بالأسباب والأوتاد والتفاعيل والبحور خاصة عربية نادرة المثال في لغات العالم، وكذلك القافية التي تصاحب هذه الأوزان. ومرجع ذلك إلى أسباب خاصة لم تتكرر في غير البيئة العربية الأولى، أهمها سببان، هما الغناء المنفرد، وبناء اللغة نفسها على الأوزان. فالأمم التي ينفرد فيها الشاعر بالإنشاد تظهر القافية في شعرها؛ لأن السامعين يحتاجون إلى الشعور بموضع الوقوف والترديد، ولكن الجماعة إذا اشتركت في الغناء لم تكن بها حاجة إلى هذا التنبيه؛ لأن المغنين جميعاً يحفظون الغناء بفواصله ولوازمه ومواضع النبر والترديد في كلماته وفقراته، فينساقون مع الإيقاع بغير حاجة إلى القوافي عند نهاية السطور، وإنما تنشأ الحاجة إلى القافية، أو وقفة تشبه القافية عند تفاوت السطور وانقسام القوم إلى منشدين ومستمعين. يقول العلامة جلبرت موري، وهو من ثقات البحث في الأوزان والأعاريض: إن إحدى نتائج هذا الاختلاف زيادة الاعتماد على القافية في اللغات الحديثة، ففي اللغتين اللاتينية واليونانية ينظمون بغير قافية؛ لأن الأوزان فيهما واضحة، وإنما تدعو الحاجة إلى القافية لتقرير نهاية السطر، وتزويد الأذن بعلامة ثابتة للوقوف... وبغير هذه العلامة تثقل الأوزان وتغمض، ولا يستبين للسامع مواضع الانتقال والانفصال، بل لا يستبين له هل هو مستمع لكلام منظوم أو كلام منثور، وقد اختلف الطابعون عند طبع الكتب هذا الاختلاف في بعض المناظر المرسلة من كلام شكسبير، فحسبها بعضهم من المنثور وحسبها الآخرون من المنظوم، ومما يلاحظ أن اللاتين اعتمدوا على القافية حين فقدوا الانتباه إلى النسبة العددية، وأن الصينيين يحرصون على القافية؛ لأنهم يلتزمون الأوزان، وأن انتشار القافية في أغاني الريف الإنجليزية يقترن بالترخص في أوزان الأعاريض.

Figure-3.2: Example of corrected text.

3.3.2 Examples of texts obtained after removing dots and diacritics

After that, we proceed to the removal of dots and diacritics using the following website: ***https://seen-arabic.github.io/Arabic-Services/***, in order to obtain dotless texts. Below we find a sample of text that we obtained after this procedure.

وإذا التمسنا مدخلاً لفن الحركة الموقعة مع الحداء، فهناك إيقاع واحد نتابعه في خطوات الإبل، وفي خطوات الهرولة التي تصاحبها على القدم، وإلى هذا الإيقاع يرجع وزن الرجز على قصد وعلى غير قصد، ومجيئه على غير قصد أدل على تمكن العادة، وعلى أصالتها في الحياة البدوية. أنا وفي سبيل الله ما لقيت وقد تكون حركة الهرولة في الطواف بالكعبة ملحوظة في كل النبي لا كذبْ أنا ابن عبد المطلبْ هل أنتِ إلا أصبعٌ دميت دعاء مروري كيفما اختلف المختلفون في صحة الرواية، كما قيل عن امرأة أحزم بن العاص حين نذرت ولدها للكعبة، فقالت: إني جعلتُ رب من بنيّه ربيطة بمكة العليّة فباركن لي بها إليّه واجعله لي من صالح البرية فهكذا يفهم الناظم كيف تكون حركة الدعاء مع الهرولة، أيّاً كان صاحب النظم أو من ينسب إليه. هذه المرددات الفردية هي التي ميزت النظم العربي باستقلال فنه، ووضوح قافيته وترتيله، ولو وجدت في الجاهلية العربية صلوات جامعة تنشد فيها الدعوات المحفوظة لوجدت فيها القصائد التي تمثل لنا حياتهم الدينية وحياتهم الاجتماعية، إما من أناشيد الصلاة كما عرفها العبرانيون، أو من أناشيد المسرح كما عرفها اليونان، ولكننا نعرف العرب من قصائدهم الفردية، كما نعرف الأمم الأخرى من أمثال تلك القصائد، فلا يفوتنا منها غاية ما تدل عليه. هذا سبب من أسباب تلك الظاهرة النادرة التي ظهرت لنا في القصيدة العربية، وكانت نادرة بين الأمم السامية والأمم الآرية على السواء

Figure-3.3: Example of text before removing dots and diacritics

واذا الممسا مدخلا لفى الحركه الموقعه مع الحدا، فهناك انفاع واحد سابعه فى حطواب الابل، وفى حطواب الهروله الى بصاحبها على القدم، والى هذا الانفاع برجع ورن الرجز على قصد وعلى غير قصد، ومحسه على غير قصد ادل على تمكن العاده، وعلى اصالبها فى الحياه البدويه. انا البى لا كذب انا ابن عبد المطلب هل ابت الا اصبع دمبت وفى سبل الله ما لقبت وقد بكون حركه الهروله فى الطواف بالكعبه ملحوظه فى كل دعا مرورى كبفما احلف المحللون فى صحه الروابه، كما قبل عن امراه احرم بن العاص حبن بدرب ولدها للكعبه، فقالت: انى جعلت رب من بنه ربيطه بمكه العلبه فبارك لى بها البه واجعله لى من صالح الربه فهكذا بفهم الناطم كبف بكون حركه الدعا مع الهروله، انا كان صاحب النظم او من بنسب البه. هذه المرددات الفرديه هى التى مبرب النظم العربى باستقلال فبه، ووصوح فافبه وبربطه، ولو وجدت فى الجاهليه العربيه صلوات جامعه بنشد فبها الدعوات المحفوظه لوجدت فبها القصابد التى بمثل لنا جانبهم الدبنه وجانبهم الاجتماعيه، اما من اناشيد الصلاه كما عرفها العرانبون، او من اناشيد المسرح كما عرفها اليونان، ولكنا بعرف العرب من قصابدهم الفرديه، كما بعرف الامم الاخرى من امثال تلك القصابد، فلا بقوبنا منها عابه ما بدل علبه. هذا سبب من اسباب تلك الظاهره النادره التى ظهرت لنا فى القصبده العربيه، وكانت نادره بين الامم السامبه والامم الاربه على السوا

Figure-3.4: Example of text after removing dots and diacritics

3.3.3 Example of texts converted to UTF-8 format

In this phase, we proceed to the conversion of the Dotted and dotless texts to the Unicode Transformation Format – 8 bit (UTF-8) format. This format is designed to maintain compatibility with ASCII while avoiding issues related to endianness and byte-order marks. Below we find the process of converting word texts to UTF-8 and a sample of the word and UTF-8 text.

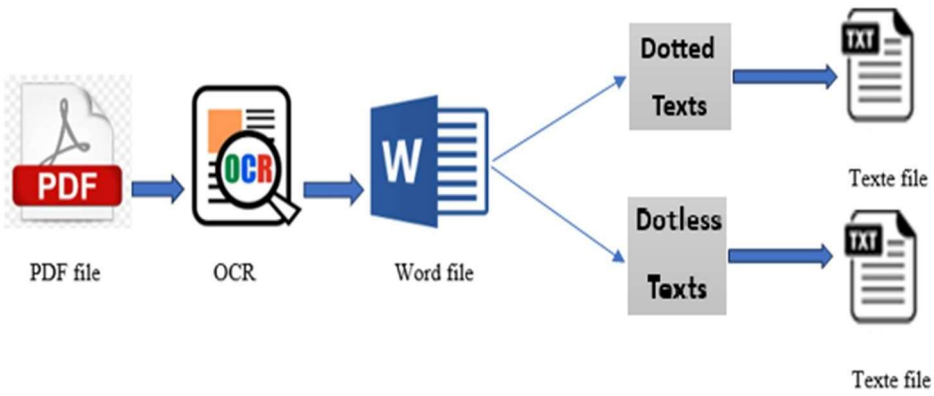


Figure-3.5: Process of converting word texts to UTF-8 texts

Figure-3.6: Example of word converted to UTF-8

3.4 Experimental Work

3.4.1 Experimental protocol

The experimental work was conducted using Dotted Texts (DTs) & Dotless Texts (DsT). Each type of texts underwent 3 series of experiments (Male Vs Male, Female Vs. Female, and Mixed-Gender) and each series contains several cases depending on the N value and the type of classifier. Namely the series of experiments are :

- The 1st serie, Dotted and Dotless Texts written by Male Author are used in the training and testing phases using different values of N and with both classifiers (K-NN and CM).
- The 2nd serie, Dotted and Dotless texts written by Female Author are used in the training and testing phases using different values of N and with both classifiers (K-NN and CM).
- The 3rd serie, Dotted and Dotless texts written by an Author (regardless to his/her gender) are used in the training and testing phases with different values of N and with both classifiers (K-NN and CM).

3.4.2 Series of experiments and obtained results

3.4.2.1 1st Series of experiments

A) Dotted Texts (DTs) written by Male Authors

In this series of experiments, character N-grams were used as features, along with k-NN and CM classifiers, to study the effect of diacritical marks on the AARate for DTs written by Male Author. The obtained results of these experiments are illustrated in the tables and figures bellow:

Table-3.3: AARate for Dotted Texts written by Male Authors (Accuracy; R=5)

Classifier	N=1	N=2	N=3	N=4	N=5	N=6
K-NN	100	100	100	93	87	68
CM	100	100	100	100	100	100

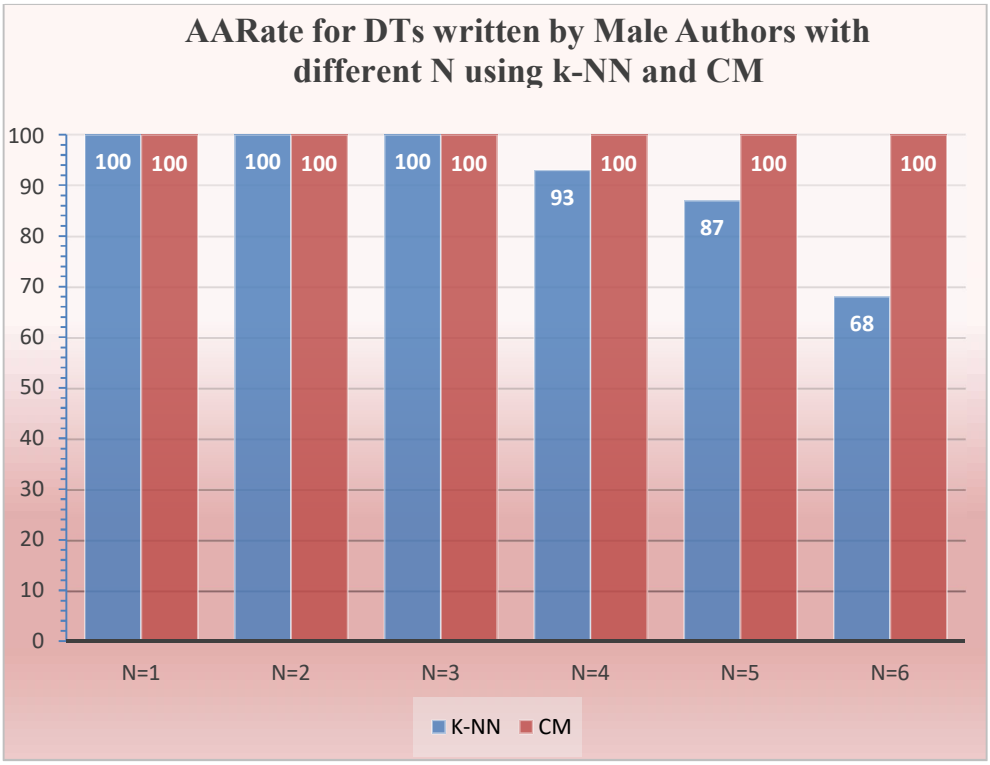


Figure-3.7: AARate for Dotted Texts written by Male Authors (Accuracy; R=5)

From the obtained results, we can clearly see that the recorded AARate for this experiment is between 68-100% for k-NN with an average AARate about (91.33%) and 100% for CM. The lowest AARates are recorded for N=5 and N=6.

B) Dotless Texts (DsT) written by Male Authors

In this series of experiments, character N-grams were used as features, along with k-NN and CM classifiers, to study the effect of diacritical marks on the AARate for DsT written by Male Authors. The obtained results of these experiments are illustrated in the tables and figures bellow:

Table-3.4: AARate for Dotless Texts written by Male Authors (Accuracy; R=5)

Classifier	N=1	N=2	N=3	N=4	N=5	N=6
K-NN	87	100	87	100	100	100
CM	100	100	100	100	100	100

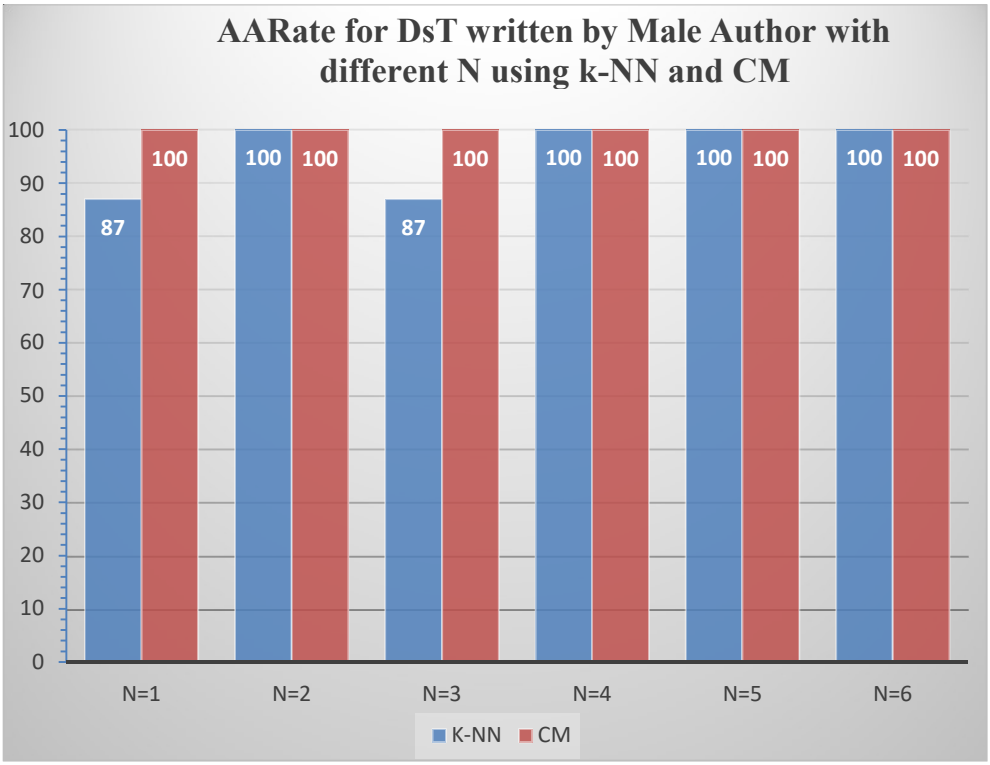


Figure-3.8: AARates for Dotless Texts written by Male Authors (Accuracy; R=5)

The results of this experiment, showed clearly that the AARate for is between 87-100% for k-NN and the lowest AARates (87%) are obtained for N=1 and N=3, with an average AARate about (95.66%). For the CM the AARates are all 100%

3.4.2.2 2nd Series of experiments

A) Dotted Texts (DTs) written by Female Authors

In this series of experiments, character N-grams were used as features, along with k-NN and CM classifiers, to study the effect of diacritical marks on the AARate for DTs written by Female Authors. The obtained results of these experiments are illustrated in the tables and figures bellow:

Table-3.5: AARates for Dotted Texts written by Female Authors (Accuracy; R=5)

Classifier	N=1	N=2	N=3	N=4	N=5	N=6
K-NN	92	85	100	100	100	100
CM	71	71	71	71	85	71

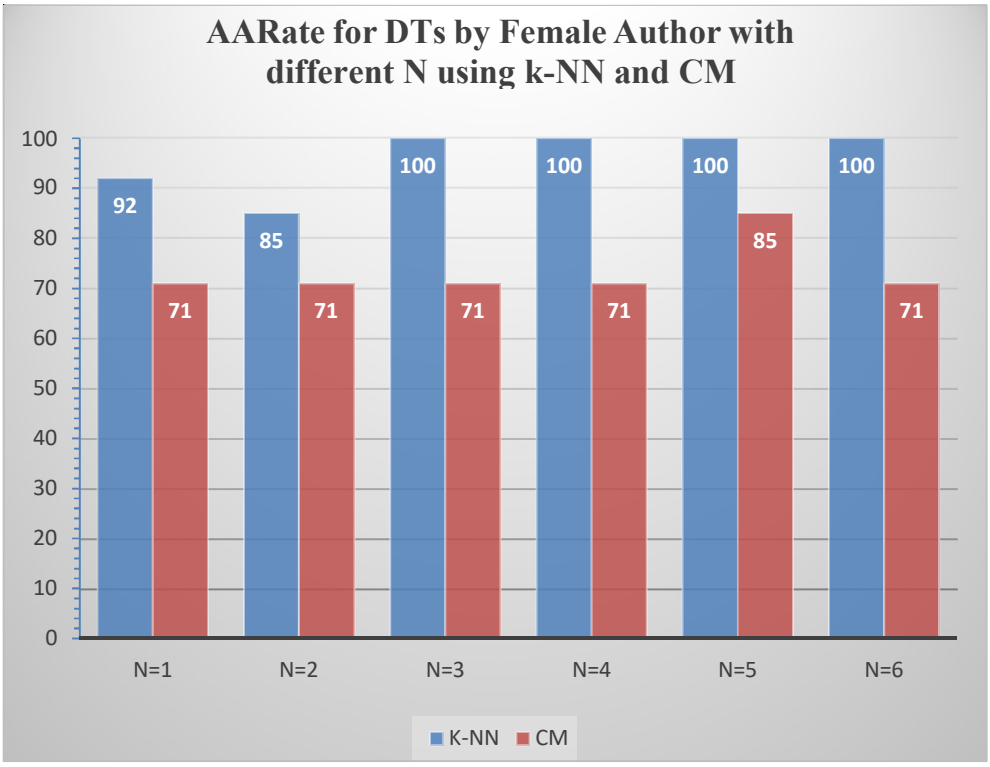


Figure-3.9: AARates for Dotted Texts by Female Authors (Accuracy; R=5)

From the obtained results, we can clearly see that the recorded AARate for this experiment is between 85-100% for k-NN with an average about 96.16%, and between 71-85% for the CM. The lowest AARates are recorded for N=1, 2, 3, 4 et and N=6.

B) Dotless Texts (DsT) written by Female Authors

In this series of experiments, character N-grams were used as features, along with k-NN and CM classifiers, to study the effect of diacritical marks on the AARate for DsT written by Female Authors. The obtained results of these experiments are illustrated in the tables and figures bellow:

Table-3.6: AARates for Dotless Texts by Female Authors (Accuracy; R=5)

Classifier	N=1	N=2	N=3	N=4	N=5	N=6
K-NN	100	100	100	100	100	100
CM	100	100	92	92	100	100

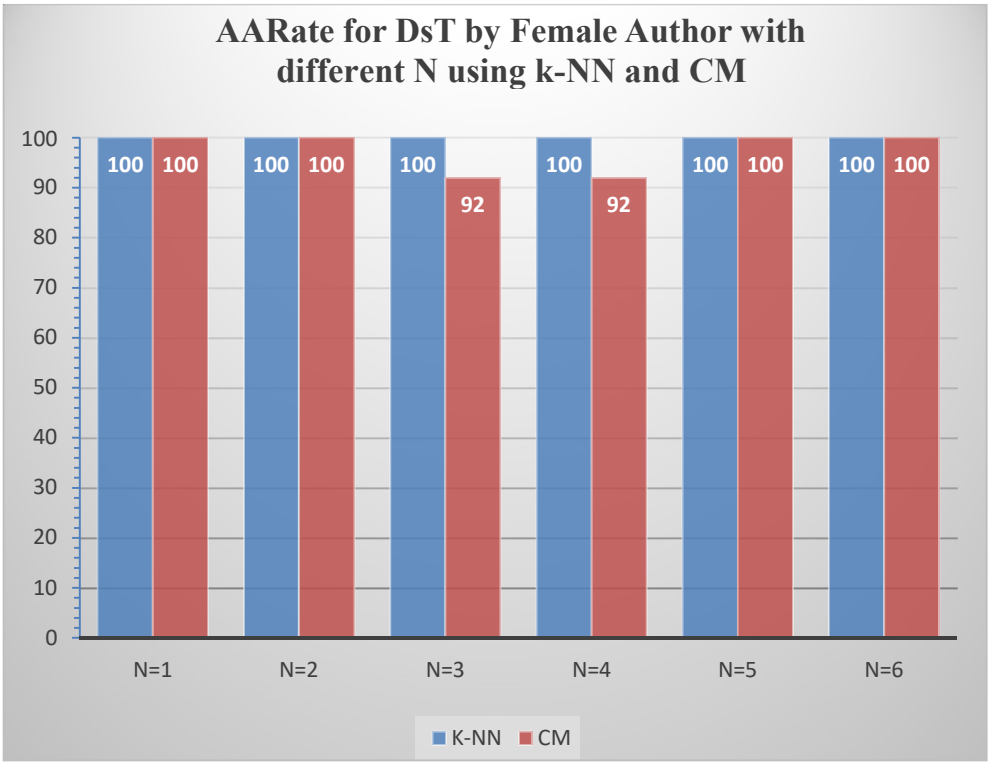


Figure-3.10 : AARate for Dottless Texts by Female Authors (Accuracy; R=5)

From the obtained results, we can clearly see that the recorded AARate for this experiment is between 92-100% for CM with an average AARate about (97.33%) and 100% for k-NN. The lowest AARates are recorded for N=3 and N=4.

3.4.2.3 3rd Series of experiments

A) Dotted Texts (DTs) written by Mixed-Gender Authors

In this series of experiments, character N-grams were used as features, along with k-NN and CM classifiers, to study the effect of diacritical marks on the AARate for DTs written by Mixed-Gender Authors. The obtained results of these experiments are illustrated in the tables and figures bellow:

Table 3.7: AARate for Dotted Texts by Mixed-Gender Authors (Accuracy; R=5)

Classifier	N=1	N=2	N=3	N=4	N=5	N=6
K-NN	96	100	100	100	83	73
CM	86	86	83	96	86	100

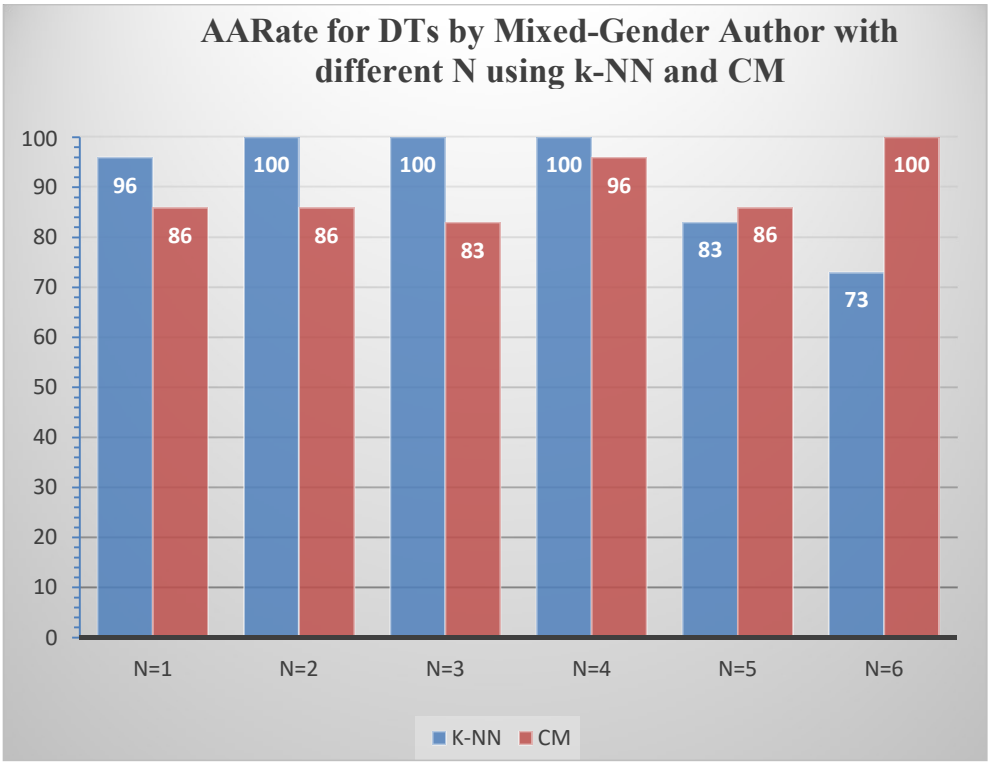


Figure 3-11 : AARate for dotted texts by Mixed-Gender Authors (Accuracy; R=5)

The results of this experiment, showed clearly that the AARate for is between 73-100% for k-NN and the lowest AARates (73%) are obtained for N=6, with an average AARate about (92%). For the CM the AARates is between 83- 100% and the lowest AARates (83%) are obtained for N=3, with an average AARate about (89.5%)

B) Dotless Texts (DsT) written by Mixed-Gender Authors

In this series of experiments, character N-grams were used as features, along with k-NN and CM classifiers, to study the effect of diacritical marks on the AARate for DsT written by Female Authors. The obtained results of these experiments are illustrated in the tables and figures bellow:

Table 3.8 : AAR for Dottless Texts by Mixed-Gender Authors (Accuracy; R=5)

Classifier	N=1	N=2	N=3	N=4	N=5	N=6
K-NN	86	100	100	96	90	83
CM	90	100	90	90	86	90

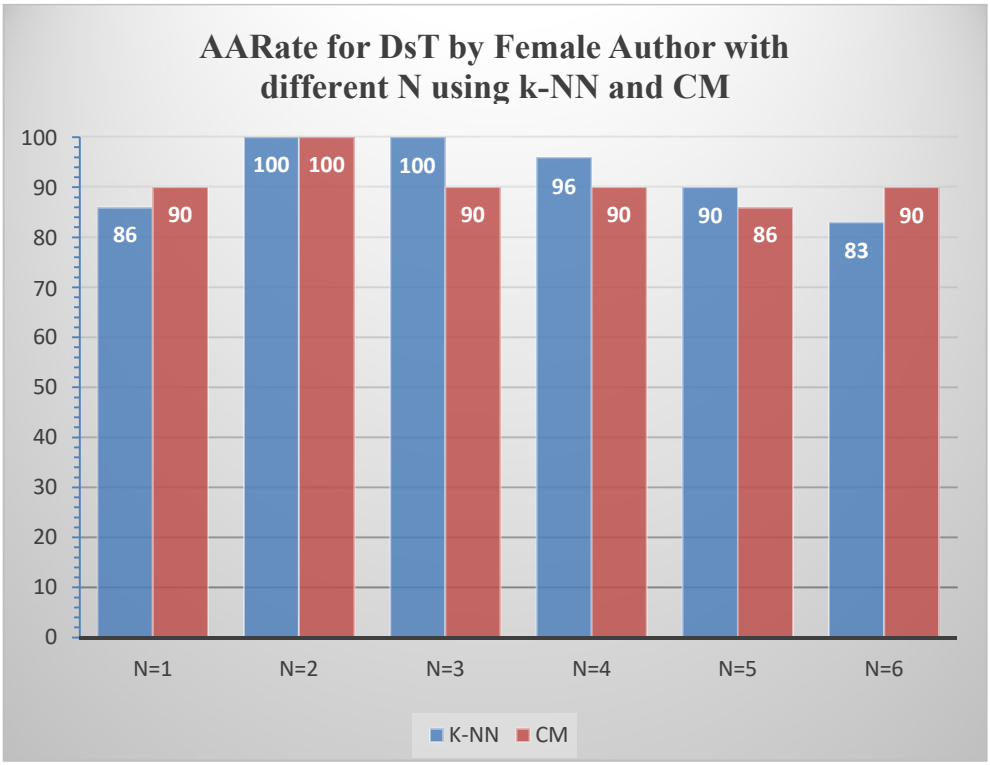


Figure-3.12 : AAR for dottless texts by global authors with accuracy at rank 5

The results of this experiment, showed that the AARate is between 83-100% for k-NN and the lowest AARates (83%) are obtained for N=6, with an average AARate about (92.5%). For the CM the AARate is between 86-100% and the lowest AARates (86%) are obtained for N=5, with an average AARate about (91%).

3.5 Conclusion

In this chapter, we conducted three series of Authorship Attribution experiments using our newly conceived corpus that we called : Dotless Arabic Texts (DAT'24). The experiments were designed to assess the impact of the presence or absence of diacritical marks on the Authorship Attribution Rate.

The obtained results have shown significant variations in AARate depending on the textual features and classification methods used. Notably, the K-NN classifier consistently outperformed CM across various configurations, with dotted texts achieving higher AAR compared to dotless texts.

This underscores the critical role diacritical marks play in the authorship attribution process for Arabic texts.

Conclusion

CONCLUSION

This research focused on developing an authorship attribution system, studying various aspects of data collection and processing, as well as designing and evaluating the proposed used. The previous chapters discussed in detail the methods and techniques used, with particular emphasis on their practical implementation and the results obtained.

The first chapter provided an overview of the field of authorship attribution. We began by introducing the concept and history of authorship attribution, discussing the features used to determine writers' writing styles, and covering the various practical steps involved in authorship attribution. We also reviewed the science of stylistics and its evolution, explained the importance of automatic text classification in improving the accuracy of author identification, and examined the concept of plagiarism and its different types.

The second chapter, reviewed the research methodology we adopted, which included the preparation of a dataset of "Arabic texts without diacritical marks", and the methods of preparation and analysis used in our experiments. We explained how we processed the texts and removed diacritical marks, as well as the techniques employed to test the accuracy of authorship attribution.

The third chapter provided a detailed analysis of the results obtained from the experiments conducted on dotted texts and dotless texts. The results showed that the accuracy of authorship attribution is significantly affected by the presence of diacritical marks, with the classification techniques used (such as K-NN and CM) proving to be more effective on dotted texts compared to dotless texts .

Preliminary experiments identified specific challenges related to the use of Arabic data, leading to the exploration of alternative methods to improve model performance. The final results showed significant improvements in the system's performance, with high precision and recall scores for most categories, although some still require improvements. Confusion AAR provided a detailed evaluation of the model's performance, confirming the effectiveness of our methodology.

Through preliminary experiments, we identified the challenges associated with using the Arabic dataset and explored alternative approaches to enhance the model's performance. Our final experiences have shown significant improvements, as demonstrated by detailed results.

REFERENCES

REFERENCES

REFERENCES

[1] 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).

[2] Holmes, David I. "Authorship attribution." Computers and the Humanities 28 (1994): 87-106.

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

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

[5] Jalam, R. (2003). Apprentissage automatique et catégorisation de textes multilingues. PhD Tesis, Université Lumiere Lyon, 2.

[6] MATALLAH, H. (2011). Classification Automatique de Textes Approche Orientée Agent (Doctoral dissertation).

[7] Mahabal, Abhijit, et al. "Text classification with few examples using controlled generalization." arXiv preprint arXiv:2005.08469 (2020).

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

[9] Mataalah Hocine « classificat ion automatique de textes Orienté Agent » faculté des sciences –algerie2010-2011

[10] Larivée, S. (1995). La notion de plagiat scientifique. Les Cahiers de Propriété Intellectuelle, 8(1), 159-190.

[11] Al-Shaibani, Maged S., and Irfan Ahmad. "Dotless Representation of Arabic Text: Analysis and Modeling." arXiv preprint arXiv:2312.16104 (2023).

[12] ***https://www.cairo24.com/1208795#***

[13] Zitouni, Imed, and Ruhi Sarikaya. "Arabic diacritic restoration approach based on maximum entropy models." Computer Speech & Language 23.3 (2009): 257-276.

[14] Nguyen, Thi Tuyet Hai, et al. "Survey of post-OCR processing approaches." ACM Computing Surveys (CSUR) 54.6 (2021): 1-37.

[15] Al-Shaibani, Maged S., and Irfan Ahmad. "Dotless Representation of Arabic Text: Analysis and Modeling." arXiv preprint arXiv:2312.16104 (2023).

[16] Al-Shaibani, M. S., & Ahmad, I. (2023). Dotless Representation of Arabic Text: Analysis and Modeling. arXiv preprint arXiv:2312.16104.

[17] Nuser, P. (2023). Symbolic Data Representation of Multi-Variate Machine Measurement Data to Identify Quasi-Linguistic Patterns with Machine Learning

[18] Mucherino, Antonio, et al. "K-nearest neighbor classification." Data mining in agriculture (2009): 83-106.

[19] Deka, H., & Sarma, P. (2017). Machine learning approach for text and document mining. Int. J. Comput. Sci. Eng.(IJCSE), 6(5).

ملخص

الكلمة المكتوبة كانت ولا تزال حجر الأساس للحضارات، وهي أساسية للحفاظ على المعرفة ونقلها. يهدف علم الأسلوب، وهو العلم المتواجد عند تقاطع اللغويات والإحصاءات، إلى تحديد أسلوب النص، الذي يميز كاتبه، وكذلك عصره ونوعه. تقوم هذه التقنية بتحليل النصوص المجهولة للتعرف على مؤلفيها وتطبق على النصوص القديمة والحديثة على حد سواء.

يهدف هذا العمل إلى تحديد مؤلف نص عربي محدد واختبار قوة نظام التعرف على المؤلف. تُستخدم ميزات مثل الحروف المتسلسلة (N-Grams) وتقنيات التصنيف.

الكلمات المفتاحية: علم الأسلوب، تحديد المؤلف، الحروف المتسلسلة (N-Grams)، التعلم الآلي.

Abstract

The written word has always been a cornerstone of civilizations, essential for preserving and transmitting knowledge. Stylometry, at the intersection of linguistics and statistics, aims to identify a text's style, inherent to its author, as well as its era and genre. This technique analyzes anonymous texts to recognize their authors and applies to both ancient and modern texts.

This work aims to identify the author of a specific Arabic text and test the robustness of an author recognition system. Features such as character N-Grams and classification techniques are used.

Keywords: Stylometry, Authorship Attribution, N-Grams Character, Machine Learning.

Résumé

La parole écrite a toujours été une pierre angulaire des civilisations, essentielle pour préserver et transmettre le savoir. La stylométrie, à l'intersection de la linguistique et des statistiques, vise à identifier le style d'un texte, propre à son auteur, ainsi que son époque et son genre. Cette technique analyse les textes anonymes pour en reconnaître les auteurs et s'applique aux textes anciens et modernes.

Ce travail vise à identifier l'auteur d'un texte arabe spécifique et à tester la robustesse d'un système de reconnaissance d'auteur. Les caractéristiques comme les character N-Grams et les techniques de classification sont utilisées.

Mots-clés : Stylométrie, Attribution d'auteur, N-Grams Character, Apprentissage.