PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

UNIVERSITY HASSIBA BENBOUALI OF CHLEF

FACULTY OF EXACT SCIENCES AND COMPUTER SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

Doctorate Thesis

Submitted in partial fulfillment of the requirements for Doctorate degree in Computer Science

Option: Information Systems

Presented by **Asmaa MANSOUR KHOUDJA**

Theme

Author Profiling based on Machine Learning Techniques for Modern standard Arabic language

Jury:

President	Rachid BACHAR	Professor	University of Chlef
Director	Mourad LOUKAM	Professor	University of Chlef
Examiner	Freha MEZZOUDJ	Professor	National Polytechnic School of Oran
Examiner	Nassim DENNOUNI	Professor	Higher School of Management of Tlemcen
Examiner	Chakib NAHNOUH	Professor	University of Chlef

June 2025

"Kindness is a language all can understand."

– Salah Addin Yusuf Ibn Ayyub

Dedication

I dedicate this work to my beloved parents, whose love, sacrifices, and unwavering support have been the foundation of my journey. Especially my mother, whose endless strength, prayers, and unconditional love have been my greatest source of motivation—without her, none of this would have been possible.

To my sister, the light that guided me when everything seemed dark. Her unwavering support and belief in me have been a source of strength in moments of doubt. Her kindness and presence made even the hardest days feel a little lighter.

To my husband, my rock, my greatest supporter—his love, patience, and encouragement in every possible way have made this journey easier and more meaningful. Without him, this path would have been much harder.

To my best friend, Selma, whose endless support and companionship have been beyond measure. She has shared every moment of this journey with me, celebrating the highs and standing by me through the lows.

To my family and friends, who have stood by me with love, warmth, and encouragement. Your presence and belief in me have made this journey all the more fulfilling.

To all those who have helped me—directly or indirectly—thank you for your kindness, guidance, and encouragement. This accomplishment is as much yours as it is mine.

With gratitude, I dedicate this work to you all.

Acknowledgements

First and foremost, I am profoundly grateful to Allah, the Most Merciful and Most Gracious, for granting me the strength, patience, and perseverance to complete this PhD journey.

I would like to express my sincere gratitude to the jury members for their time, effort, and invaluable feedback. Their insightful comments and constructive critiques will significantly contribute to refining my research. I would also like to thank Mr. Tahar Abbes Mounir, the head of the Doctoral Training Committee, and Mr. Belalia, director of the LME laboratory, for their roles in supporting and facilitating the doctoral program.

A special and heartfelt thanks to my supervisors, Pr. Loukam Mourad and Pr. Belkredim Fatma Zohra, whose guidance, encouragement, and unwavering support have been instrumental throughout this journey. Pr. Loukam's patience and expertise have shaped not only this research but also my approach to problem-solving and academic growth. I am truly fortunate to have had his mentorship.

I am deeply thankful to my professors, mentors, and educators, whose passion for knowledge has continuously inspired me.

To my family and friends , words cannot express the depth of my appreciation. Their love, sacrifices, and constant encouragement have been my greatest source of strength.

This PhD has been more than an academic pursuit—it has been a journey of resilience, learning, and self-discovery. I am grateful for every challenge that has strengthened me, every person who has inspired me, and every moment that has shaped my path.

To all who have been part of this incredible journey—thank you.

– Asmaa Mansour Khoudja

Abstract

This thesis addresses the challenges of gender profiling and bot detection in Modern Standard Arabic (MSA) using advanced machine learning techniques, including LSTM, ARABERT, and Prompt-Based Learning. The research highlights the scarcity of resources and research in Arabic Natural Language Processing (NLP) compared to high-resource languages like English, aiming to bridge this gap by creating novel datasets and exploring innovative algorithms. Two datasets were curated: one for gender profiling (10,000 MSA texts) sourced from PAN 2018, Arabic Parallel Gender Corpus 2.0, Google Forms, while the other dataset for bot detection (1,100 MSA texts) was sourced from Fake News, and Automatically-Generated Arabic Tweets. Preprocessing steps included tokenization, balancing, and translation of dialectal Arabic to MSA. The experiments evaluated the performance of LSTM, ARABERT, and Prompt-Based Learning, with ARABERT achieving the highest accuracy (92.4% for gender profiling and 88% for bot detection), followed by Prompt-Based Learning (92.3% and 80%) and LSTM (78.5% and 66.8%). The results demonstrate the superiority of transformer-based models and the potential of prompt-based approaches for low-resource languages. Key contributions include the creation of high-quality datasets, the introduction of Prompt-Based Learning to Arabic NLP, and a comprehensive comparison of model performance. Future work include focusing on dataset expansion, optimizing prompt-based approaches, and cross-domain applications such as sentiment analysis and machine translation. This research advances Arabic NLP by providing tailored models and methodologies for author profiling and bot detection, offering valuable insights for addressing similar challenges in low-resource language settings.

Keywords: Gender Profiling, Bot Detection, Modern Standard Arabic, LSTM, ARABERT, Prompt-Based Learning, Natural Language Processing.

Résumé

Cette thèse aborde les défis de la classification du genre des auteurs et de la détection des bots en arabe standard moderne (MSA) en utilisant des techniques avancées d'apprentissage automatique, notamment LSTM, ARABERT et l'apprentissage basé sur les invites (Prompt-Based Learning). La recherche met en évidence la rareté des ressources et des études en traitement automatique du langage naturel (TALN) pour l'arabe, comparé aux langues à fortes ressources comme l'anglais. Elle vise à combler cette lacune en développant de nouveaux ensembles de données et en explorant des algorithmes innovants.

Deux ensembles de données ont été construits : l'un pour la classification du genre (10 000 textes en MSA) provenant de PAN 2018, du corpus parallèle Arabic Parallel Gender Corpus 2.0 et de Google Forms ; et l'autre pour la détection des bots (1 100 textes en MSA) issus de sources de fausses informations et de tweets générés automatiquement. Les étapes de prétraitement comprenaient la tokenisation, l'équilibrage des données et la traduction de l'arabe dialectal en MSA.

Les expériences ont évalué les performances de LSTM, ARABERT et de l'apprentissage basé sur les invites, ARABERT obtenant la meilleure précision (92,4% pour la classification du genre et 88% pour la détection des bots), suivi par l'apprentissage basé sur les invites (92,3% et 80%) et LSTM (78,5% et 66,8%). Les résultats démontrent la supériorité des modèles basés sur les transformeurs et le potentiel des approches basées sur les invites pour les langues à faibles ressources. Les contributions majeures incluent la création d'ensembles de données de haute qualité, l'introduction de l'apprentissage basé sur les invites dans le TALN arabe et une comparaison approfondie des performances des modèles.

Les travaux futurs porteront sur l'expansion des ensembles de données, l'optimisation des approches basées sur les invites et leurs applications interdomaines, notamment l'analyse des sentiments et la traduction automatique. Cette recherche fait progresser le TALN arabe en fournissant des modèles et des méthodologies adaptés au profilage d'auteur et à la détection des bots, offrant ainsi des perspectives précieuses pour relever des défis similaires dans des langues à faibles ressources.

Mots-clés: Profilage de genre, Détection des bots, Arabe standard moderne, LSTM, ARABERT, Apprentissage basé sur les invites, Traitement automatique du langage naturel.

مُلخّص

تتناول هذه الأطروحة تحديات تحديد هوية المؤلف وإكتشاف النصوص المولّدة آليا في اللغة العربية الفصحى الحديثة باستخدام تقنيات متقدمة في التعلم الآلي، بما في ذلك الشبكات العصبية طويلة المدى، ونموذج أرابيرت، والتعلم القائم على التوجيه. تسلط الدراسة الضوء على ندرة الموارد والأبحاث في معالجة اللغة العربية مقارنةً باللغات ذات الموارد العالية مثل الإنجليزية، وتهدف إلى سد هذه الفجوة من خلال إنشاء مجموعات بيانات جديدة واستكشاف خوارزميات مبتكرة.

تم إنشاء مجموعتي بيانات: الأولى خاصة بتحديد هويةالمؤلف (١٠,٠٠٠ نص باللغة العربية الفصحى) مأخوذة من مجموعة بيانات بان ٢٠١٨، والمجموعة الموازية للجنس في اللغة العربية، واستبيانات جوجل فورمز، بينما الثانية مخصصة لاكتشاف النصوص المولّدة آليا (١,١٠٠ نص) مأخوذة من أخبار زائفة وتغريدات عربية منشأة تلقائيًا. تضمنت خطوات المعالجة المسبقة تحليل الكلمات، وتحقيق التوازن بين البيانات، وترجمة النصوص العامية إلى الفصحى.

قيّمت التجارب أداء النماذج المستخدمة، حيث حقق نموذج أرابيرت أعلى دقة (٤٠٩٢% لتحديد هوية المؤلف و٨٨% لإكتشاف النصوص المولّدة آليا)، يليه التعلم القائم على التوجيه (٣٠٩٢% و٨٠%)، ثم الشبكات العصبية طويلة المدى (٥٠٧٨% و٨٠٦٦%). تُظهر النتائج تفوق النماذج المعتمدة على المحولات وإمكانيات التعلم القائم على التوجيه في اللغات منخفضة الموارد. تشمل المساهمات الرئيسية إنشاء مجموعات بيانات عالية الجودة، وإدخال التعلم القائم على التوجيه إلى معالجة اللغة الطبيعية العربية، وإجراء مقارنة شاملة لأداء النماذج.

تشمل الأعمال المستقبلية توسيع مجموعات البيانات، وتحسين استراتيجيات التعلم القائم على التوجيه، وتطبيقاتها في مجالات أخرى مثل تحليل المشاعر والترجمة الآلية. تسهم هذه الدراسة في تطوير معالجة اللغة الطبيعية العربية عبر تقديم نماذج ومنهجيات مخصصة لتحديد هويةالمؤلفين وإكتشاف النصوص المولّدة آليا، مما يوفر رؤى قيمة لمعالجة تحديات مماثلة في اللغات منخفضة الموارد.

الكلمات المفتاحيّة: تحديد هويةالمؤلف، اكتشاف الحسابات الآلية، العربية الفصحى الحديثة، الشبكات العصبية طويلة المدى، أرابيرت، التعلم القائم على التوجيه، معالجة اللغة الطبيعية

List of Publications

Our research findings have been presented at various scientific conferences and published in reputable journals, contributing to advancements in Natural Language Processing (NLP) for the Arabic language.

- Journal Publications

1) Publication in *Ingénierie des Systèmes d'Information* – Our research paper, *Assessment of LSTM, ARABERT, and Prompt-Based Learning for Gender Author Profiling in Modern Standard Arabic Language*, authored by Khoudja, Asmaa Mansour, Mourad Loukam, and Fatma Zohra Belkredim, was published in *Ingénierie des Systèmes d'Information* (Volume 29.6, 2024, Page 2209)[1]. This work presents a comparative evaluation of machine learning models for gender profiling in Arabic text.

- Conference Communications

1) Contribution to the 6th International Congress on Information and Communication Technologies (ICICT 2021)[2] – On February 25, 2021, we contributed to the writing of a research paper presented at this prestigious event held in London, UK, alongside the ICT Excellence Awards. The conference served as a platform to discuss emerging trends in information and communication technologies.

2) Participation in GISDay 2019[3] (Study Day on Geographic Information Systems) – On December 18, 2019, we participated in a study day, GISDay'2019, where discussions focused on advancements in geospatial technologies and their applications.

[1] **DOI: https://doi.org/10.18280/isi.290611**
[2] **DOI: https://doi.org/10.1007/978-981-16-2377-6_69**
[3] **GISDay'2019 Link**

Table of Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

AP:	Author Profiling
ARABERT:	Arabic Bidirectional Encoder Representations from Transformers
BERT:	Bidirectional Encoder Representations from Transformers
BoW:	Bag of Words
CNN:	Convolutional Neural Network
DL:	Deep Learning
LDA:	Latent Dirichlet Allocation
LLM:	Large Language Model
LSTM:	Long Short-Term Memory
ML:	Machine Learning
MSA:	Modern Standard Arabic
MT5:	Multilingual Text-to-Text Transfer Transformer
N-grams:	Sequence of N words or characters
NLP:	Natural Language Processing
RNN:	Recurrent Neural Network
SVM:	Support Vector Machine
TF-IDF:	Term Frequency-Inverse Document Frequency
T5:	Text-to-Text Transfer Transformer
XGBoost:	eXtreme Gradient Boosting

Chapter 1
General Introduction

1 Introduction

In the era of digital communication, the Internet has become a massive repository of textual data, this has surged interest in extracting meaningful information. This surge in online content has given rise to the need for advanced analytical techniques, in areas like author profiling (AP).

Author Profiling seeks to identify and uncover the demographic and personal features of authors based on their written content, such as gender, age, political beliefs, and even linguistic backgrounds of authors from their texts. Through these features we can obtain important insights into the intentions, motives, and objectives of the author. Author Profiling is an effective approach utilized in a variety of applications in areas ranging from security and crime investigations to marketing strategies and consumer behavior analysis.

One significant subset of Author Profiling is gender classification, which has attracted a lot of interest in recent years. However, gender identification task presents a unique set of challenges such as the limited availability of large, diverse, and labeled datasets, especially in under-represented languages like Modern Standard Arabic (MSA). Advances in this field have been also hindered by the linguistic complexity of MSA and the cultural nuances in gender expression.

Beyond traditional author profiling, this thesis expands its scope to Bot Detection, a related yet distinct task. Bot Detection can be viewed as an extension of Author Profiling, where the goal is to classify the author not based on demographics, but on their human or non-human nature. This perspective frames both tasks under the broader objective of inferring who—or what—is behind a given text.

To clarify this connection, the following figures illustrate the two tasks within the broader framework of Author Profiling:

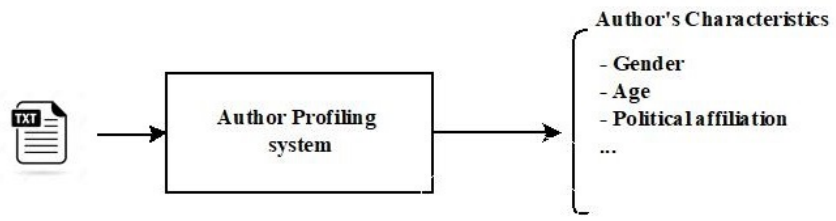

Figure 1: Authors Profiling Aim.

This figure 1 illustrates the author profiling task where the goal is to determine the gender of the author based on textual data.

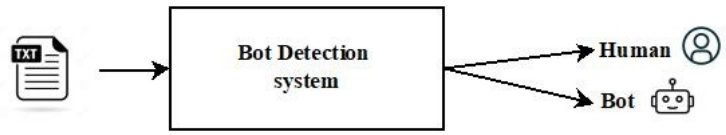

Figure 2: Bot Detection Aim.

This figure 2 illustrates the bot detection task which classifies whether a text is authored by a human or a machine.

This research aims to address the gap in Arabic Natural Language Processing (NLP) by focusing both on gender and bot identification using advanced machine learning techniques, including long short-term memory (LSTM), ARABERT, and the innovative prompt-based learning methods. By leveraging these approaches, we seek to develop and evaluate state-of-the-art models that can accurately profile the gender and nature of Arabic authors online.

2 Problem Identification and Motivation

The tasks of gender and bot identification in author profiling for Modern Standard Arabic remain underexplored due to the lack of linguistic resources and the inherent complexity of the language. Existing research in the field of author profiling has focused predominantly on the English language, leaving a critical gap in the exploration of gender classification for Arabic text. Moreover, the majority of current models are not specifically tailored to handle the linguistic features of Arabic, resulting in suboptimal performance compared to more commonly studied languages.

This research is motivated by the urgent necessity to explore and evaluate novel approaches that can effectively address these challenges and improve the accuracy of these tasks in MSA. By leveraging recent advancements in machine learning and NLP,

we intend to contribute to the advancement of Arabic NLP and offer valuable insights into the use of these techniques for gender and bot identification in this underexplored domain.

3 Objectives and Scope

The primary objective of this research is to develop and evaluate state-of-the-art models for gender profiling in Modern Standard Arabic. Specifically, this study aims to:

1. Assess the effectiveness of three machine learning models—LSTM, ARABERT, and prompt-based learning—in resolving the challenge of gender profiling then of bot detection in MSA.
2. Explore the validity of prompt-based learning as a new paradigm for low-resource languages, such as Arabic.
3. Evaluate the performance of these models on two newly collected Arabic dataset, one of over 10,000 gender labeled text samples and the other of over 1,100 human/bot labeled text samples, thereby contributing to the advancement of Arabic NLP.

By fulfilling these objectives, this research will fill a critical gap in Arabic NLP and provide insightful information into how modern machine learning techniques can be adapted for author profiling in Arabic, particularly gender and bot classification.

4 Research Questions

To guide this research, the following questions will be addressed:

1. How suitable are the three methods—LSTM, ARABERT, and prompt-based learning—for solving the problem of gender profiling in MSA?
2. What are the optimal configurations for each of these methods in the context of gender profiling?
3. Which method outperforms the others in terms of accuracy and overall performance?
4. How does the size of the training dataset impact the performance of each model in identifying gender from MSA texts?

5. To what extent can techniques used in Author Profiling be applied to Bot Detection—when the "author" is, in fact, a machine?

5 Contributions

This thesis provides several significant contributions to the field of Arabic NLP, with a particular focus on gender author profiling:

1. Novel Dataset:
   This research introduces a newly created dataset of more than 10,000 labeled Arabic texts gathered from several sources, offering a valuable resource for future studies in author profiling and gender classification.

2. Employing Prompt-Based Learning:
   This is the first study to introduce the use of prompt-based learning to gender profiling in MSA. The research assesses its effectiveness and adaptability in comparison to more conventional techniques like LSTM and ARABERT.

3. Cross-Model Evaluation:
   Presents a comparative study and analysis of three machine learning paradigms, namely, LSTM, ARABERT, and prompt-based learning, highlighting the advantages and disadvantages of each approach in gender identification tasks.

4. Arabic NLP Advancement:
   This research addresses the lack studies on author profiling in Arabic and provides new methods and results to fill the gap between Arabic and other languages in NLP research.

6 Outline of the Thesis

The thesis is organized into three main chapters:

- **Chapter 1: Introduction**
  This chapter will set the stage by introducing the importance of author profiling and gender identification in Modern Standard Arabic (MSA). It will explain the importance for the study, discuss the challenges, and highlighting the gap in the Arabic NLP domain.

- **Chapter 2: Arabic Language and Author Profiling**
  This chapter focuses on the structure and complexities of the Arabic language,

including MSA and other dialects. Why we conducted our research on MSA. It will also cover definitions and methodologies for author profiling, specifically gender identification, alongside a review of related work. This will provide context for the contribution of this thesis by highlighting the challenges and successes of using author profiling in Arabic texts.

- **Chapter 3: Machine Learning, Natural Language Processing, and Large Language Models**
  Here, we examine the main techniques that are used in gender profiling: an overview of NLP and its uses, comprehensive discussion of machine learning, deep learning architectures and transformer models in author profiling, as well as the progress and importance of LLMs with regards to gender identification and profiling.
- **Chapter 4: Data Collection and Creation**
  This chapter will focus on the dataset utilized for the study. It describes the detailed process of data collection from three different resources and its creation.
- **Chapter 5: Methodologies and Evaluation**
  This chapter will present the three methods used in this research: LSTM, ARABERT, and Prompt-based Learning.
- **Results and Discussion** The findings of the study will be discussed in detail here. It will also address any unexpected outcomes or limitations of the study.

PART ONE: LITERATURE REVIEW

Chapter 2: Arabic Language and Author Profiling: A Linguistic Perspective

Chapter 3: The Evolution of NLP: From Machine Learning to Large Language Models

Chapter 2
Arabic Language and Author Profiling: A Linguistic Perspective

1 Overview

This chapter highlights the importance of the field of Author Profiling (AP) by providing a detailed overview of it in the context of the Arabic language. It begins by delving through Arabic as a language in general and outlining its various forms and complexities.

The first section of part one of this chapter, explores Classical Arabic, the historical and literary foundation and basis of the Arabic language. The second section covers Modern Standard Arabic (MSA), the modern, standardized form used in formal communication throughout the Arab world. In the third section delves into Dialectal Arabic, emphasizing the regional differences and how they affect linguistic analysis. The fourth section provides the reason for focusing on MSA in this research, given the widespread acceptability and standardization of MSA. The last section of part one, covers the challenges of processing Arabic because of its rich morphology and syntactic flexibility.

The second part of the chapter moves to the topic of Author Profiling (AP), which examines writing styles to identify a variety of features about the author. This section begins with an overview of AP and its numerous aspects. Section 2.1 outlines the importance of AP, specifically in fields like marketing, security, and forensics. The sections that follow cover specific aspects of AP, such as approaches for identifying an author's gender, age, political affiliation, and native language as well as psychographic profiling and human vs bot detection. Each of these aspects is addressed with focus on the complexities and difficulties of using these methods on Arabic texts. The chapter concludes with a summary of the discussed topics, with a reminder of how important it is to understand the unique features of Arabic and how AP can be used within this linguistic context. In the following chapters, we will delve deeper into this overview and give a solid foundation for research and real-world applications related to AP for Arabic texts.

2 Arabic Language

Arabic is one of the most ancient and widespread Semitic languages on earth. and with over 12 million words, it stands as one of the most linguistically diverse and versatile languages. Today it is the mother language of about half a billion people, and used as a worship language in Islam by about one and a half billion Muslim people. This has

earned the Arabic language the fifth position in the five most widely used languages on earth. It can be divided into: Classical Arabic, Modern Standard Arabic (MSA), and Dialectical Arabic [1].

2.1 Classic Arabic (CA)

Classical Arabic is the traditional Arabic that is primarily used in the Qur'an, religious texts, and pre-Islamic poetry and literature. It is the foundation of today's standard Arabic language but is more complex in nature. It continues to be used in religious studies and literature till present [1].

2.2 Modern Standard Arabic (MSA)

MSA is the modernized successor of the classic Arabic language that has witnessed some simplification for instance it differs in vocabulary and pronunciation. It is used as the formal language in Arab countries and it is practical for modern communication such as the media, in formal occasions and settings as well as in schools.

2.3 Dialectal Arabic

Unlike MSA dialectal Arabic is the daily spoken Arabic language or as referred to by others: the street language of people in their different regions. It is a challenging language because each dialect in the Arab countries and communities differ by its unique grammar, vocabulary and pronunciation depending on the culture and the historical factors. It is a big part of the identity of people in a specific area or region.

2.4 Why MSA?

We decided to conduct our research on MSA (Modern Standard Arabic) instead of other Arabic dialects due to the following reasons:

- **Wider Coverage and Usage in Formal Contexts**
  MSA is the common language that can link the Arab world speakers together. It is understandable by all of them in various levels, depending on their educational status since it is learned in schools as a first language, also people are exposed to it through the radio, TV, the press, and it is used in formal occasions as well. Therefore, MSA facilitates communication between Arabs living in different geographical locations.

- **Mathematical and Geometric Background**
  In addition to the above reasons, MSA is a very interesting language with a strong mathematical foundation. for instance, all Arabic letters are based on geometric shapes such as triangles and circles, making the script both artistic and logically structured [2]. Unfortunately, not many researchers are focusing on it, which has led to a big amount of underexploited MSA data on the web.

2.5 Difficulties of Arabic Language Processing

Arabic language processing is still lagging behind in the fields of AP and in NLP in general. This is mainly due to the fact that Arabic, like many other languages, is a complicated language in its nature since it has various levels of linguistic representation such as morphology, phonology, etc. which makes word tokenization, lemmatization, and stemming more challenging in comparison to simpler languages. As well as the requirement of sophisticated natural language processing techniques to solve ambiguity problems. And last but not least, the lack of resources for instance there is a significant deficiency of Modern Standard Arabic supervised data on the web.
Fehri [3], Chalabi [4] and Daimi [5] presented the following as the main complexities with the Arabic Language:

- Writing in Arabic is done from right to left.
- Arabic does not use uppercase letters.
- The subject can be removed from a sentence, because the subject can be implied by the verb form, which does not require explicit statement. e.g., ذهب للمنزل Translation: He went home. In this example, the subject "he" is implied by the verb "ذهب"(dhahaba), which is conjugated for the third person singular masculine in the past tense.
- The auxiliary verbs "to be" and "have" are absent from Arabic. e.g., the English translation of 'اسمي أسماء' is: My name (is) Asmaa. Once more, the verb "to be" is not used directly; rather, it is implied.

- In Arabic morphology, the three-letter root is the most common root structure. This system leads to a wide range of meanings, which may result in ambiguity. Figure 3 Derivation of words from a three-letter-root

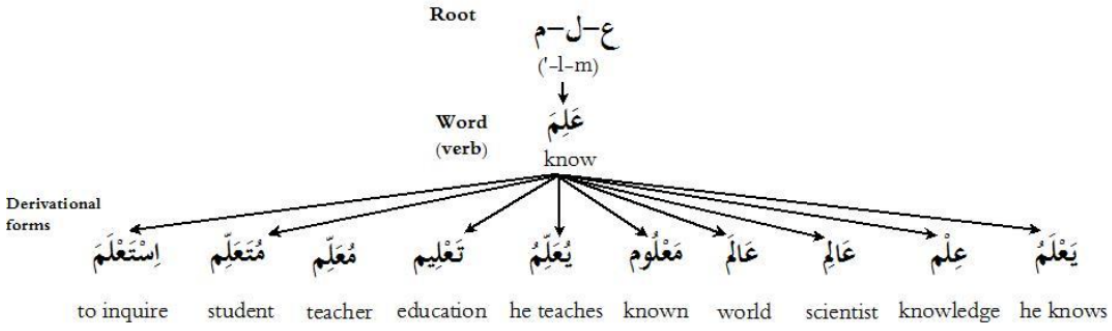

Figure 3: three-letter-root structure

- Vowels and extra consonants are inserted into the root consonantal structure to produce words, forming patterns. A vast range of words are produced when roots and patterns are combined.
- Grammar in Arabic is expressed by inflections, which include tense, gender, number, and case.
- Adjectives, verbs, and nouns all experience inflections, which alter the word's form to reflect different grammatical characteristics.
- In Arabic, feminine nouns are often derived from masculine nouns that act as the base. For example, "عالم" ('alim, male scientist) becomes "عالمة" ('alima, female scientist). Still, some feminine nouns, like "شمس" (shams, sun) and "ليلة" (layla, night), are not derived from a matching masculine noun.
- Also broken plurals Broken plurals are a unique feature of Arabic. The internal structure of the word is modified for pluralization as opposed to just adding a suffix. Let's consider an example that demonstrates the gender of the author through the use of broken plurals:

  Masculine:

  Singular: كاتب (kātib) - Male Author

  Plural: كتّاب (kuttāb)

  Feminine:

Singular: كاتبة (kātiba) - Female Author

Plural: كاتبات (kātibāt)

- In Arabic we can express the same sentence using multiple word orders. These orders divide Arabic sentences into four categories SVO (Subject-Verb-Object), VSO (Verb-Subject-Object), VOS (Verb-Object-Subject), and SOV (Subject-Object-Verb) [6]. We can explain this in table 1.

Table 1: Arabic sentence orders.

Sentence orders	**Arabic**	**Transliteration**	**Translation**
SVO	الطالب حل المسألة	Al-talib halla al-mas'ala	The student solved the problem
VSO	حل الطالب المسألة	Halla al-talib al-mas'ala	Solved the student the problem
VOS	حل المسألة الطالب	Halla al-mas'ala al- talib	Solved the problem the student
SOV	الطالب المسألة حل	Al-talib al-mas'ala halla	The student the problem solved

To conquer these limitations, more studies, researches and collaboration must be initiated in Arabic NLP.

3 Author Profiling

Author profiling (AP) is a discipline in which people's written text is examined and analyzed. AP tries to understand the author behind the text in an attempt to predict the demographic or personal characteristics of the author [7], such as his age group, his gender, his political affiliation, his mother language and psyche traits, etc. Thus, we can obtain important insights into the intentions, motives, and objectives of the author.

3.1 Importance of AP

Author Profiling is an effective approach utilized in a variety of fields.

In order to identify suspects in crimes, AP use forensic techniques to analyze writing collected from documents, signatures, and anonymous letters.

Companies also use AP to determine the age group and gender of their target market in order to develop more effective marketing tactics and products that boost their profits and revenue.

In addition, by examining social media behavior, AP is also used to identify possible violent committers. This helps authorities stop crimes by identifying people who may pose a threat, such as terrorists or sexual predators.

Moreover, in sociolinguistic and psychological studies, AP provides insights into how different genders express themselves, subjects chosen to write about, and language nuances. AP is significant for sociolinguistic and psychological studies.

AP can also be effective in medical studies. Gathering gender-labeled web information aids in understanding symptom differences, illness prevalence across genders, and tailoring treatments based on gender-specific language nuances.

Furthermore, AP protects the originality of academic, literary, and professional works by comparing writing styles to identify fraudulent behavior that lead to plagiarism detection. Author profiling is therefore useful for advancing marketing strategies, protecting integrity in a variety of contexts, and strengthening security.

3.2 Author's Gender

There are several key objectives of gender AP. For instance, such information may be helpful in fields like marketing, where it can assist marketers in targeting specific genders based on the content of their written texts, also in forensic investigations, or psychology or health research, such as using online health forums or medical records, to develop specific interventions to improve health care.

The fact that gender expression varies among cultures is one of the primary obstacles in gender profiling, which can make the task far more difficult to perform. It can also be challenging to extract features from writing styles that accurately identify the gender of the authors because male and female writers have distinct writing styles that are frequently ambiguous and challenging to categorize. For instance, pronouns and first names are ambiguous gender identifiers in the English language. For example, both genders can be addressed by the pronouns "they" and "their" "I" and "my."

3.3 Author's Gender in Arabic

The gender in Arabic is most often determined by morphology, for example, using "ة" (Taa Marbuta) to mark the female gender. E.g., male: "طالب" (talib) "student" and female: "طالبة" (taliba) "student".

However, occasionally we are unable to distinguish between them. For instance, in Arabic, the pronouns "I" and "my" are "أنا" (ana) and "ي" (y) for both genders. For example, when we translate this passage to English "حصلت على شهادة البكالوريا من المدرسة الثانوية الجديدة للبنات" it becomes "I obtained my baccalaureate degree from the new secondary girls school". Thus, based solely on the fact that the author attended a school reserved only for girls, we can assume that she is a female [8]

3.4 Author's Age

With age, an author's writing style either gradually improves or deteriorates. It has been observed that every writer experience style changes as they get older. For instance, as demonstrated by [9], word length expands in proportion to the author's age. It is presumed from the text's content that users between the ages of 13 and 17 write about youth and teens, school-related topics, etc., while users between the ages of 23 and 27 write more about parties, college, travel, and life objectives. Users between the ages of 33 and 47 tend to post more on politics, religion, marriage, and family life [10].

3.5 Author's Age in Arabic

The author's age in Arabic can be determined from the content of the text, for instance "In a few days, I'm going to celebrate my 30th anniversary " its translation in English is 'بعد أيام سأحتفل بعيد ميلادي الثلاثين'. Thus, since the author's age is stated in the text, it may be easily extracted (a content-based information). Additionally, in "منذ عشرين سنة خلت كتبت مقالة في جريدة الأهرام" (I wrote a paper in Al Ahram magazine twenty years ago). Although the author's age is not stated, it is clear that they are middle-aged (perhaps over 40). We can list another example of age extraction from tweets that are

related to examinations, courses, school and college breaks, these tweets may suggest that the individuals were either students at university or high school [11]. Classifying age is highly intricate and challenging because it is often based on textual content.

3.6 Author's Political Affiliation and Ideology

Using AP in multiple languages is a crucial part of fighting online terrorism. This is the reason why a lot of researchers have grown quite interested in the political affiliation prediction. Several contributions have been made to the classification of political texts. For example, in [12], the authors looked up users of a particular US English forum who frequently used insulting words or other abusive language and who promoted the use of violence against a certain group of society. The goal was to identify messages that expressed hatred and anger. In another study, [13] investigated Swedish politicians' speeches and categorized them based on gender, age, and political affiliation. Various feature selections were implemented. After including all forms in the feature sets, the political affiliation categorization showed the best accuracy rate. However, the feature sets that were limited to function words or verbs alone produced the lowest scores for political affiliation classification and the best scores for gender prediction.

3.7 Author's Political Affiliation and Ideology in Arabic

Since there are some extremist organizations in the Arab world, it is crucial that researches should made in the subject of Arabic political affiliation. In [12], the authors worked on extracting linguistic features from online messages in order to establish stylistic features for terrorists' communication patterns. They tried using an already-existing framework, with some modifications, to examine online authors on messages in both Arabic and English that were associated with well-known extremist organizations. In [14], the authors developed a system to categorize Arabic texts according to ideology and organizational affiliation.

3.8 Author's Native Language and Dialects Identification

Native language identification is Identifying the mother tongue of an author who writes in a different language. Furthermore, the ability to distinguish between similar languages is what defines dialect identification or language variety. For instance, Portuguese from Brazil vs Portuguese, Arabic from the Gulf, Egypt, the Maghreb, etc., or English from the United States and Britain (UK). For example, the passage below, "Language Variety Identification analyzes the behavior of..." is clearly

written in American English rather than Brtish, as behavior is spelled with a "u" in British English. However, in the passage "Native Linguage Identification analyzes the behavior of..." we can notice that "Linguage" and "analyzes" may be spelled the way they are in the sentence due to first-language interference from the Italian words "linguaggio" and "analizza" [15]. The authors in [16] attempted to develop a single model for every language provided in PAN17. Word-unigrams and character-grams were employed as features, along with additional features that worsened rather than enhanced performance, such as POS tags, Twitter handle, and geographic entities. For evaluation, a simple single model was created that simultaneously classified each user's gender and language variety. For the same task [17] used a deep learning technique—which has recently shown to be incredibly effective in natural language processing without the need of manual feature extraction. The model that they created was made of bidirectional recurrent neural network combined with a gated recurrent unit (GRU) and an attention mechanism. The model achieved 75,31% gender classification accuracy and 85,22% language variety classification accuracy for each language.

3.9 Author's Native Language and dialects Identification in Arabic

Researchers are particularly interested in native language and dialect identification in Arabic since it is important for the detection of threatening messages, the improvement of online security flaws, and the detection of possible terrorist activities. Nevertheless, the task is more difficult due to the diversity of Arabic dialects, since each dialect requires the availability of an annotated datasets. Because of this, there have only been a few studies done on Language/dialects identification in Arabic. The authors in [11] created a dataset collected from multiple social media apps to build a large manually annotated Arabic dataset that covers 16 Arabic countries and 11 dialectal regions.

3.10 Author's Psychographics (Personality Treats)

Psychographics is the study used to describe characteristics of human personality, values, interests, attitudes and opinions. Psychographics have been investigated by many researchers such as [18] who focused on four important personality characteristics using a corpus of personal weblogs with different sets of n-gram features and they used both SVM and Naïve Bayes (NB) for personality classification for the Binary classification tasks and the Multiple classification tasks. More recently, researchers

started extracting personality characteristics based on people's Facebook profiles. For instance [19] showed that there is a significant relationship between personality characteristics and various features of Facebook profiles. They examined correlations between users' personality and the properties of their Facebook profiles such as the size and density of their friendship network, number uploaded photos, number of events attended, number of group memberships, etc. The best accuracy was achieved for "extraversion" and "neuroticism", the lowest accuracy was obtained for "agreeableness", with "openness" and "conscientiousness" lying in the middle.

In 2015, PAN156, participants were provided with a training data set that consists of twitter tweets in English, Spanish, Italian and Dutch. Their aim was to concentrate on profiling the authors not only into the appropriate gender and age classes, but also in five personality features: openness, conscientiousness, extraversion, agreeableness, and neuroticism. For each feature scores were provided (between -0.5 and 0.5). [20] Tried to solve this task using dimensionality reduction techniques on the top of typical discriminative and descriptive textual features for AP task. Experimental results in AP showed that their idea gave evidence of its usefulness to predict personality profiles. For the same problem [21] proposed a coherent grouping of features combined with appropriate preprocessing steps for each group. They addressed the personality prediction task as a regression problem using Support Vector Machine Regression on documents created by joining each user's tweets.

3.11 Author's Psychographics for Arabic (Personality Treats)

Regarding the Arabic language, Psychographics identification is less investigated comparing to other languages. Few studies had been done in this field. In 2007 [22] applied Text Attribution Tool (TAT) to profiling the authors of Arabic emails. And they showed improvements in psychometric author features predictions.

3.12 Human/Bot Detection for Author Profiling

Several extensive studies have been carried out on the phenomenon of bot spread on social Networks and their impact on society. These studies have highlighted the role of bots in spreading false information that could lead to serious problems such as malicious bots that encourage the spread of hate, particularly political, religious and racial hate. For instance [23]. As well as other studies that covers the effect of bots in influencing and manipulating public opinion in elections. such as [24] and [25] where they collected datasets from popular hashtags and tested on supervised machine learning techniques.

Other bots were designed with the purpose of enlarging social networks or increasing celebrities' followers such as in [26] where they expanded on their previous work for data gathering. They experimented with BotOrNot [27] social bot-detection tool and they used a set of meta-data features.

3.13 Human/Bot Detection in Arabic for Author Profiling

Moreover, while the majority of related work in this area has focused on English, there is a growing need to fill the gap of bot detection problem in the Arabic language. Several recent studies have addressed this gap, such as [28] where they discussed the presence of bots on social media, The paper's main contributions included providing two datasets using different labeling techniques: the Arab Spring in Libya and a dataset of honeypot network of Arabic bots. Furthermore, in 2018 [29] investigated religious hatred spreading Arabic tweets. They collected their dataset that consists of 6,000 Arabic tweets using Arabic terms that represent the most common religious beliefs in the Arab world. They used techniques such as letters normalizing, stemming, diacritics removal, and more in their preprocessing. Unfortunately, the bots' tweets have been permanently deleted, so we can no longer access them based on their respective IDs. Also, [30] utilized an SVM model with multiple features to detect fake accounts that affected politics and manipulated public opinion. In a recent study [31], the GPT2-Small-Arabic2 model was employed to generate a dataset consisting of 3,512 deepfake tweets. the used Arabert model outperformed RNNs in their evaluation demonstrating superior performance.

4 Conclusion

This chapter laid the foundation for understanding the relationship between author profiling and Arabic language. We explored Modern Standard Arabic (MSA), Dialectal Arabic, and Classical Arabic, with a focus on the choice of MSA due to its widespread usage and standardization.

In particular, we addressed the challenges and nuances in Arabic literature. We additionally discussed about the significance of author profiling and covered its aspects such as gender, age, political affiliation, native language, dialects, and personality. This chapter offers a strong foundation for future studies that seek to enhance techniques for author profiling and Arabic language processing.

Chapter 3
The Evolution of NLP: From Machine Learning to Large Language Models

1 Overview

Chapter 3 provides an introduction to the dynamic field of Natural Language Processing and explores the differences and similarities of Deep Learning and Machine Learning in NLP. The chapter starts with a definition of NLP and its different levels of analysis, including both style and content-based features. The chapter also explores the history, evolution and related work of NLP, focusing on significant events and discoveries that have influenced the field. The chapter then, dives into the concepts of Deep Learning and Machine Learning, explaining various machine learning algorithms, and performance evaluation metrics. For each of these algorithms, research studies and related works are provided to point out their usefulness and practical uses. It then proceeds to explore the transition from Machine Learning to Deep Learning, and highlights the remarkable achievements of Deep Learning in NLP applications such as translation and text classification.

2 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a branch of computing sciences which its main point is to analyze, comprehend, produce and manipulate human language, and the input could be both spoken or in a text form. the purpose of NLP is to achieve a human-like language processing for various application in diverse areas [32] such as Information Retrieval, Machine Translation, Question-Answering, etc.
In order to understand the human language, NLP includes multiple levels of analysis. These levels are grouped into two primary types: Style-based and Content-based.

2.1 Style-Based Features

Style-based markers or style features are features of language that reflect the author's style, character and tone [12] there are four main groups of features that we will see: lexical, syntactic, structural and semantic:

2.1.1 Lexical Features

Lexical features are one of the important aspects of text analysis. They treat the choice and frequency of words used in a text as well as the word placement and length, vocabulary richness and the number of letters. These characteristics provide deeper understanding and facilitate analysis of texts [12]. For instance, using text attribution system [22] worked on extracting author profiles from both English and Arabic emails

by computing a feature vector including the count of words and sentences, frequency of punctuation and word length.

2.1.2 Syntactic Features

Syntactic features contain characteristics that specify the grammatical structure and the patterns of words in a sentence. they are used to distinguish between authors [12] for example using the word hence or thus can be used as classification feature to differentiate between authors [33]. For instance [34] in the task of gender identification, discovered that female writers used more pronouns and noun modifiers than male writers.

2.1.3 Structural Features

Structural Features provide information on the organization and layout of the text, such as the number of paragraphs, the spacing, the opening phrases, salutations and signatures of the authors [12]. In [35] the authors used structural features among other features, in their author profiling task. They analyzed the texts according to the number of paragraphs, number of conversations, etc.

2.1.4 Semantic Features

Semantic features adopt another method of linguistic analysis they focus on the meaning of words and the relationships between them. They include features such as synonyms, verbs tenses, nouns count, etc. [36]. In [37] authors described a sarcasm detection model based on semantic features which included usage of frequent vs. rare synonyms.

2.2 Content-Based Features

The content-based approach focuses on classifying the content of the text. One the most commonly used techniques in classifying the text's content is n-gram. Thus, in orders to gain a deeper understanding of the context and topics of texts, the sentences of a given text are split into n-grams [8].

By using the n-grams technique with multiple values of n, calculating the frequencies of different topics mentioned by distinct author groups, could be possible, because of dissimilarities in topics between age groups and gender [38]. e.g., in blog posts, male writers tend to be associated with using topics like football or computers, whereas female writers are associated with using words like shopping, households or spouses.

In their study, [10] used features based both on style and content. Their results revealed that females are more involved in their texts, whereas males focus more on giving information. For the age groups, adults were more concerned about politics and technology while teenagers wrote in their texts more about friendships and emotional states. To conclude [10], found that content-based features worked better in age identification compared to gender identification where, the combination of both style and content features, was more effective.

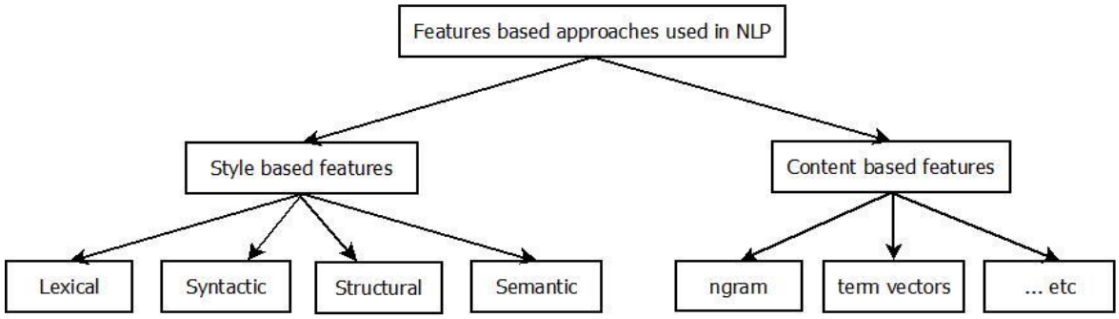

Figure 4: Features-based approaches used in NLP.

2.3 Applications of NLP

NLP is a vast field that continues to evolve every year and researchers today utilize NLP tools in various real-world application. Some of the most common uses of NLP include:

2.3.1 Information Extraction (IE)

IE focuses on recognizing and extracting some important information from raw texts, e.g., entity extraction such as people's names, addresses, important dates and events. Which can be used multiple tasks like question-answering systems.

2.3.2 Sentiment Analysis

Sentiment Analysis is the extraction, identification analysis of expressed sentiments in texts, in order to study the attitude of the author and determine his preferences on a given subject. it is also known as opinion mining [39].

2.3.3 Text Summarization

Text summarization consists of reducing large texts and creating summaries of the original text. it is used in content generation, news summarization [40], etc.

2.3.4 Machine Translation

Machine Translation is one of the first NLP branches. it is the conversion of a text from one language to another without human intervention. it is used in international business, scholar and industrial applications [41].

2.3.5 Question-Answering

Question-Answering provides automated response to user's questions, in the form of a text with an answer rather than just a list of relevant documents. Such as chatbots or virtual assistants [42].

2.3.6 Text Classification

Text Classification is the process of differentiating between predefined categories or labels by capturing contextual information and analyzing features extracted from the texts [43].
NLP has gained significant advancement recently, like the spread of the mind-blowing chatbots, this shows that further advancement of NLP can be very practical in real-world uses.

2.4 Related Work to Author Profiling in NLP

In the field of author profiling NLP, various types of features have been utilized to extract meaningful patterns from texts. Style-based features have been prominently studied, including lexical, syntactic, structural, and semantic aspects. For instance, a text attribution system described in [22] focused on extracting author profiles from both English and Arabic emails by computing a feature vector that included the count of words and sentences, frequency of punctuation, and word length. Syntactic features have also been explored, with [34] demonstrating that female writers tend to use more pronouns and noun modifiers compared to male writers in gender identification tasks. Additionally, structural features were employed by [35], where texts were analyzed based on the number of paragraphs, number of conversations, and other structural elements.

Semantic features have shown their potential in tasks such as sarcasm detection. In [37], a model was described that relied on semantic features, particularly the usage of frequent versus rare synonyms, to detect sarcasm effectively.

Content-based features have also been significant in author profiling, particularly in blog post analyses. It has been observed that male writers tend to be associated

with topics like football or computers, while female writers are more likely to discuss shopping, households, or spouses. This was highlighted in the study by [10], which utilized features based on both style and content. Their findings revealed that female writers are generally more involved in their texts, whereas male writers focus more on providing information. Furthermore, the study differentiated age groups, noting that adults often write about politics and technology, whereas teenagers focus more on friendships and emotional states. Ultimately, the study concluded that content-based features are more effective for age identification, while a combination of style and content features yields better results for gender identification.

As of 2013, in a contest for the text mining community, a group of contestants approached the gender and age classification task. Where, [44] presented an approach that discriminated between authors' age and gender through four steps: the calculation of words' occurrences, the selection of classes, and the creation of ARFF2 files. Although there is potential for improvement, they achieved the highest accuracy in the competition with 36.7%.

In addition to feature extraction, [45] proposed a content-based approach for author's age group and gender detection. They used a different set of features, such as syntactic n-grams, traditional n-grams, and the combination of word n-grams and character n-grams. They used multiple classifiers. The achieved accuracy of 73% signifies that the combination of word n-grams using the SMO classifier can produce good results.

3 Machine Learning

The field of NLP later experienced a new transformation through Machine learning (ML), an interdisciplinary approach that draws from multiple fields such as mathematics, biology, informatics, etc. It enables computers to adapt to the human language and automatically learn from data. Unlike traditional programming, where we manually code rules to create the system, ML includes studying computational techniques that uses prior experiences from large datasets to enhance performance of systems on unseen data [46] It is a fundamental approach in NLP that solves various tasks, like text classification, text generation, and named entity recognition. Thus, ML Allows systems to learn patterns and use manual feature extraction, such as style-based features, for example character-based features that contained properties like the number of the total words, number of words in each sentence, vocabulary

richness, etc.

Some of the common approaches of machine learning that are used in NLP:

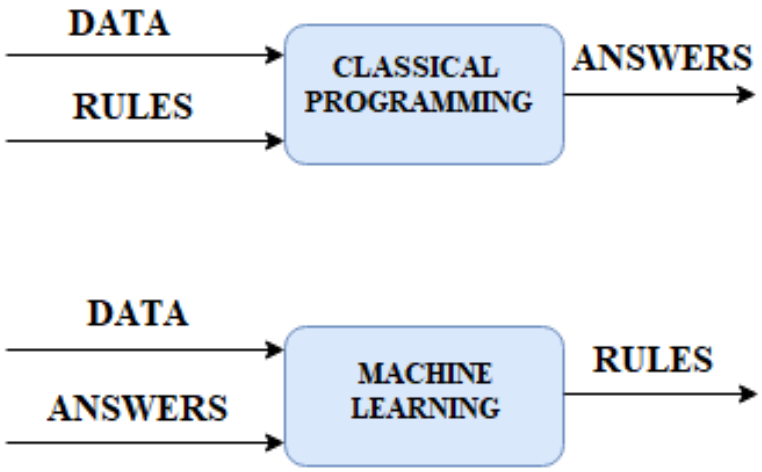

Figure 5: Machine Learning Vs Classical Programming.

3.1 Supervised Learning

In this approach, the model is trained on labeled data, providing inputs (texts) along with their appropriate labels. Thus, the model will be able to predict new data without having seen it, after learning to match the input texts with the appropriate labels [46].

Here are some common examples of supervised learning in NLP:

- Classification problems: gender classification.
- Sentiment analysis: categorizing the expressed preferences of an author on a given subject.
- Machine Translation: the conversion of a text from one language to another without human intervention.

3.2 Unsupervised Learning

This approach is used for unlabeled data. The model here understands the data through visualization and compression. Unsupervised learning is mostly used for:

- Clustering: clustering techniques are used to aggregate similar data points, allowing the identification of similarities within the data.
- Dimensionality reduction: Dimensionality reduction plays a role in limiting the data's dimensionality complexity, allowing to visualize multidimensional data and

uncover concealed patterns [46].

- Topic modeling: Topic modeling technique is used to find hidden themes in a collection of documents. the goal is to discover the basic structure the text without any predefined labels or target outputs. This approach helps identify dominant topics present in the data set.

3.3 Important Terms Used In Machine Learning

There are some common terms used in Machine Learning context. Using ML classifiers depend on understanding them:

Inputs or samples: They are instances of training a set.

Outputs or predictions: They are values or results of an input.

Labels: They are the true output for a specific input.

Classes: Possible categories or labels for the instances of the data. e.g., Positive and Negative are two classes used for a movie review.

Loss or error: An evaluation measure of the model performance and it is the distance between the assigned output and the real output.

Binary classification: It means there are two classes in a specific classification task.

Multi-class classification: It means there are more than two classes in a specific classification task.

Batch: A bunch of input data used during the training process where, for each batch the weights are updated to enhance the model performance.

Epoch: Running the model for one complete pass through all of dataset.

Overfitting: When the model learns from the training set very well, and ends up memorizing its patterns, then when testing with a new test set, it can't generalize and performs poorly.

Underfitting: When a model can't learn from the training set and performs poorly on both the training and test sets.

3.4 Types of Machine Learning

There are almost unlimited number of Machine Learning algorithms, ranging from simple to very complex. Here are some examples of the most used algorithms in the field of NLP:

3.4.1 Decision Trees

Decision trees are one of the supervised ML models that are used for classification instead of regression. They are represented in a tree-like form. In decision trees, decisions are made by running a series of tests, and each internal node in the tree is a test of the attribute value, and the leaf of the node is the class label.

The benefits of decision trees are making feature selection implicit, providing a simple representation of the process of decision-making and reducing uncertainty. In addition to another benefit which is their minimum demand of effort in data preparation. However, like any other technique, decision trees are subject to some limitations, such as instability due to required modification to the overall structure, in case of changes in the data, as well as the problem of overfitting when dealing with continuous values or when the tree becomes very complex. Figure 6 shows the architecture of a decision tree [47].

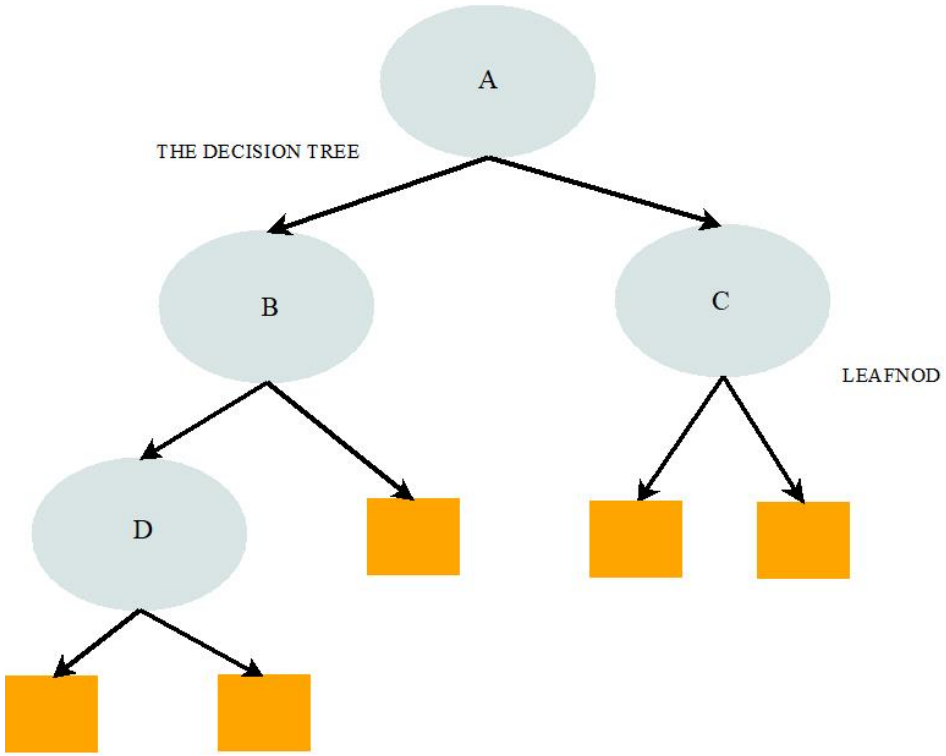

Figure 6: The Architecture of a Decision Tree.

3.4.2 Neural Networks

A neural network is an interconnected neurons that mimic the functioning and the structure of the human brain. The neural network receives inputs with associated weights, and the network generates classified output.

Errors from previous classification are corrected by the network and fed back to it. They are then used in modifying the network's algorithm over multiple iterations to generate the best result. However, its main flaws are the absence of transparency and the learning process is excessively long [48].
Figure 7 represents a Neural Network architecture.

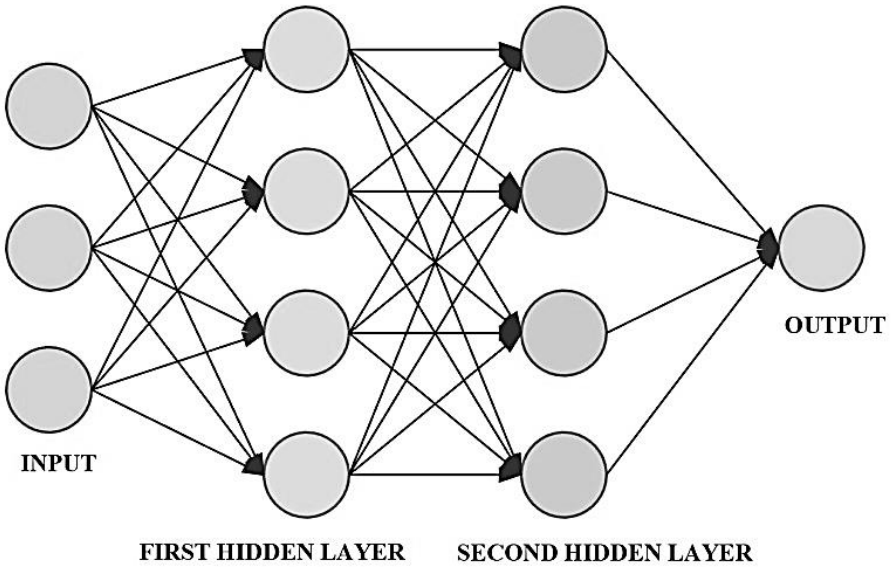

Figure 7: Neural Network Architecture.

3.4.3 Knn (K-Nearest Neighbors)

KNN is a supervised classification algorithm that can be used for classification and regression tasks. and it works very well on multi-class classification problems. KNN is simple to design and it functions in a multidimensional space. Each instance in the training set is represented as a point. To determine the class of a new instance, KNN computes the distances between instances of that dataset. Thus, the algorithm finds the nearest points in the space. The class of the new instance is determined by the majority of its k nearest neighbors. However, some factors should be considered when using KNN such as the high computational cost, for instance, computing the distance between instances becomes very expensive for large datasets also, the problems caused by unbalanced data (Uneven classes) [49].

K. Nearest Neighbor architecture is represented in Figure 8.

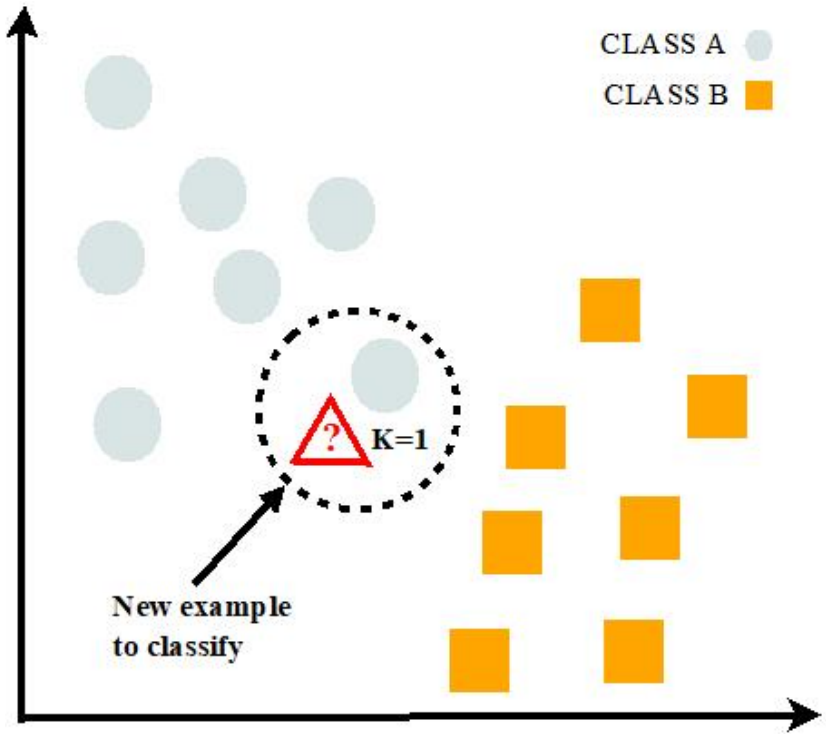

Figure 8: K. Nearest Neighbor Architecture.

3.4.4 Naïve Bayes (NB)

Naive Bayes (NB) is a simple probabilistic classifier. It is based on the Bayes theorem where; it provides the many probabilities of a document belonging to different categories. NB assumes that two features are independent in a document. NB is mostly used for classification and clustering tasks. Classifying and training the data in NB is not excessively time consuming. and it can be easily updated in case of changes in the data. However, the quality of the results could be affected due to the theorem of independent features in case if the features were in fact interrelated [48].

3.4.5 Support Vector Machines (SVM)

SVM is another supervised learning model. Unlike Naive Bayes, SVM is not a probabilistic classifier. It creates a Hyperplane in a multidimensional space to divide data into different classes. It is used for classification and regression tasks. The goal of the margin-maximizing hyperplane is to find the distance between the closest data points of each class and the hyperplane. In general, the model performs better when the margin is large. SVM advantages are: the ability of handling multi-dimensional data and it works well on small datasets. However, it is computationally expensive when using large datasets and the selection of the tuning hyper-parameters require

carefulness [48].

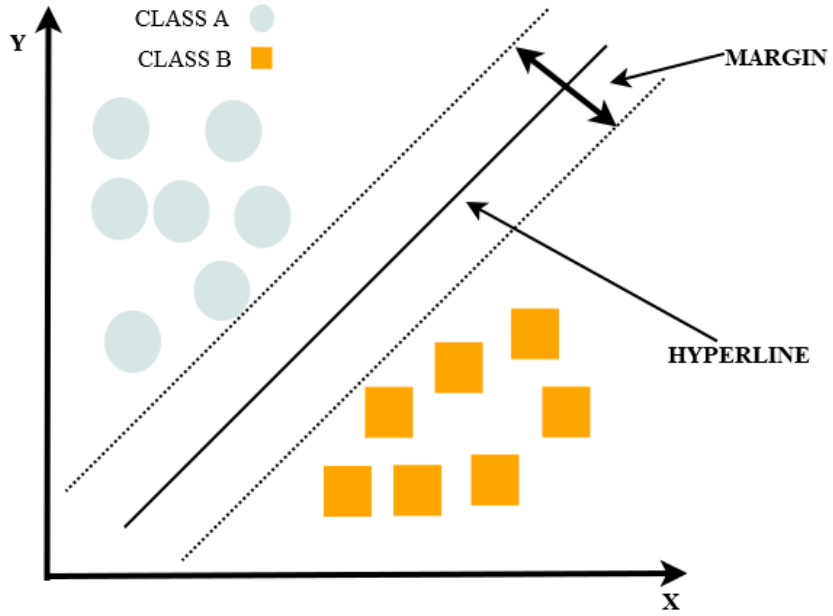

Figure 9: SVM Architecture.

3.5 Measures of Evaluating the Performance of Learning Algorithms

Machine learning uses different ways to evaluate the model's performance. Evaluation measures come from a confusion matrix which gives correct and incorrect predicted classifications for each class. Table 1 shows a binary classification confusion matrix, with TP as the number true positive, FP as false positive, FN as false negative, and TN as true negative [50].

Table 2: A Binary Classification Confusion Matrix.

Class \Predicted	**Positive**	**Negative**
Negative	FP	TN
Positive	TP	FN

The most commonly used performance measures are:

Accuracy

This metric provides an overall assessment of the algorithm by indicating the probability of correctly classified instances from different class labels, considering the prior distribution of classes.

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN} \tag{1}$$

Recall

Recall, also called true positive rate (TPR), it counts the probability of correctly predicting a new positive instance from the test. Often, in medical contexts it is called test sensitivity.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Along with the True Negative Rate, (TNR) that is called the specificity and refer to the probability of correctly predicting a negative new instance from the test [51].

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

Precision

Precision calculates the fraction of correctly predicted as positive out of all the positives in the test. It combines the outcome from the positive and negative samples. in medical contexts, it is referred to as PPV (positive predictive value) [51].

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

F-score

To provide a balanced measurement, F-score takes in account both false positives and false negatives so it combines precision and recall. It is the mean of precision and recall

$$F\text{-}score = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{5}$$

The F-score may be balanced if $\beta = 1$. It leans toward precision if $\beta > 1$ or it leans toward recall if $\beta < 1$ [50].

3.6 Related Work to Author Profiling in NLP

The task of author profiling has been extensively studied using various machine learning techniques, ranging from traditional feature extraction methods to more advanced algorithms, achieving accurate results and demonstrating significant advancements in the field.

[52] Utilized two feature extraction techniques: the Bag-Of-Words (BOW) approach and a sentiment- and emotion-based approach. They employed four classifiers (Naive Bayes, Decision Tree, SVM, and KNN) to investigate whether females write with more

emotions than males. Their results showed no evidence supporting this hypothesis, with Bayesian and SVM classifiers performing the best.

In the PAN 2016 competition, [53] proposed a methodology for cross-genre author profiling. They used a combination of features such as n-grams, average comma count, average dot count, average exclamation count, and average question mark count. Logistic regression with k-fold cross-validation (k = 10) was used for evaluation. Their approach achieved first place for gender detection in English and tied for second place in joint accuracy.

[45] Employed feature extraction for an NLP classification task using syntactic n-grams, traditional n-grams, and combinations of word and character n-grams. Their results indicated that combining word n-grams with multiple classifiers produced good results. Similarly, [54] used a combination of Naïve Bayes, KNN, and SVM, along with content and meta-data features, to detect spam opinion reviews, achieving an impressive 99% F-measure.

In [55] the author focused on author profiling in Arabic tweets and deception detection in Arabic texts. Using classical machine learning models like linear classifiers, SVM, and Multilayer Perceptron classifiers, along with Bag-of-Words and n-grams for feature extraction, Nayel outperformed other teams in deception detection tasks.

In [56] the authors conducted gender analysis on Arabic tweets. They evaluated differences between male and female users in terms of engagement and topics of interest. They proposed a method using feature extraction and the SVM classifier to infer gender by analyzing usernames, tweets, profile pictures, and friend networks. Manual annotation of gender and locations for over 160K Twitter accounts yielded impressive initial results of 82.4%.

These studies highlight the diversity of approaches in machine learning models used in author profiling, demonstrating significant advancements and effective methodologies in the field.

4 From Traditional Machine Learning to Deep Learning

Due to the recent advancements in computing potentials, the huge increase in data and the availability of new frameworks, a more efficient method than the traditional ML was needed. Thus, Deep Learning (DL) has taken the field of machine learning to a new level, revolutionized the field of industry and became an important part of numerous AI tasks. Today, DL with its sophisticated and accurate models can help

solve problems in all domains.

Figure 10 shows the differences between ML and DL for textual data.

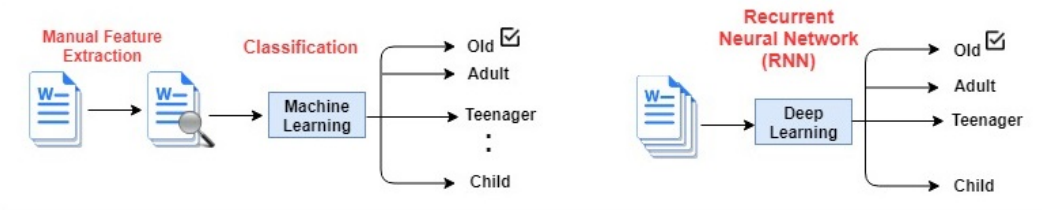

Figure 10: The Differences Between ML And DL For Textual Data.

5 Deep Learning

Deep Learning is an advanced field of Machine Learning. DL expands the basics of ML by using neural networks with multiple layers and trains these networks to learn the hierarchy of concepts straight from data. These networks mimic the functioning and the structure of the human brain. Instead of extracting features manually as in ML, DL uses modern techniques to automate feature engineering.

DL lately has gained significant popularity due to its ability to give accurate results that sometimes has surpassed the human level performance. All thanks to the availability of the huge computing potentials and the large amounts of data [57].

Figure 11 represents Neural Networks, which are organized in layers consisting of a set of interconnected nodes.

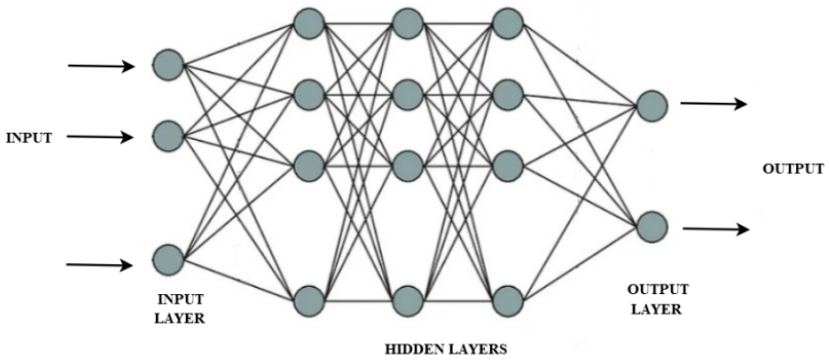

Figure 11: Neural Networks in DL.

5.1 Types of Deep Learning Models

There are various types of DL models, based on different basic architectures. Two very common types of deep neural networks are convolutional neural networks (CNNs) and recursive neural networks (RNNs).

5.1.1 Convolutional Neural Networks (CNNs)

In recent years, CNNs have appeared as impressive models for the computer vision field through convolutional layers. Additionally, CNNs have extended their use beyond visual data, in fields like NLP. However, they aren't very common in NLP tasks. But, with the ongoing research and the recent advancement in the field using CNNs, they are offering promising results in text classification, machine translation, etc.

5.1.1.1 The Basic Architecture Of CNNs

The next section will explain the basic architecture of CNNs

Convolutional Layer:

This layer consists of applying convolutional filters on the input data. Every convolutional filter extracts specific features from the text, such as n-gram, the spacing, or the relationships between. The generated output of each filter is a feature map, representing a feature in the input text.

In contrast to images that uses pixels, textual CNN turns sentences into lists of values (sequences that can be padded to become matrixes) and the filters are applied on them to locate features. The output is considered as a feature map.

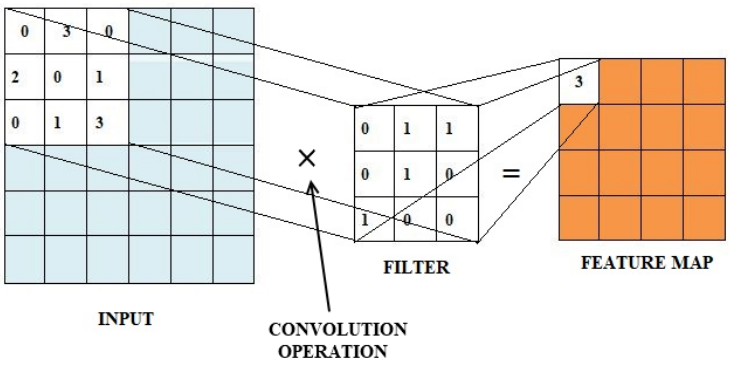

Figure 12: The Features Map of CNN.

Activation Function:

The activation function is applied to each element in the feature maps from the previous convolutional layer. Activation function replaces all negative values to zeros. The most common activation functions are Relu, Tanh and sigmoid.
ReLU:

$$R(z) = (0, z) \tag{6}$$

Pooling Layer:
Similar to visual data, the pooling layer applies operations to textual data to reduce the size of the feature maps and produce a layer with the most important feature (Max-Pooling, Average-Pooling, sum). At the same time, preserving the important information in the text [46].

e.g., when we apply Max Pooling, we take the most important features from feature maps and put them in smaller windows. In Figure 13 below we take the biggest values from the activation function ReLU feature map.

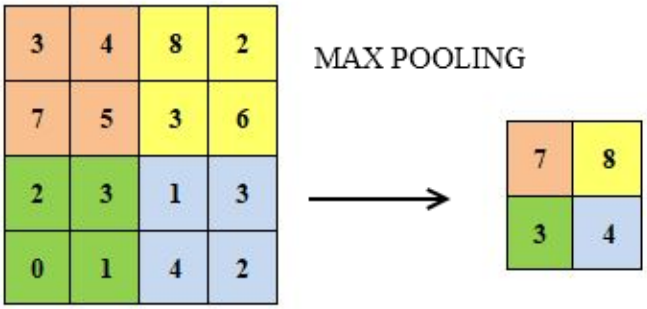

Figure 13: Max Pooling Layer.

- Results of **input1** → we take the maximum of (1, 1, 5, 6) → **6**
- Results of **input2** → we take the maximum of (2, 4, 7, 8) → **8**
- Results of **input3** → we take the maximum of (3, 2, 1, 2) → **3**
- Results of **input4** → we take the maximum of (1, 0, 3, 4) → **4**

Fully-connected layer:
Finally, every neuron in the last layer is connected to every neuron in the following layer which is responsible for classification. The activation function softmax is applied in the output layer to predict which class the input belongs to. Figure 14 explains how the architecture works.

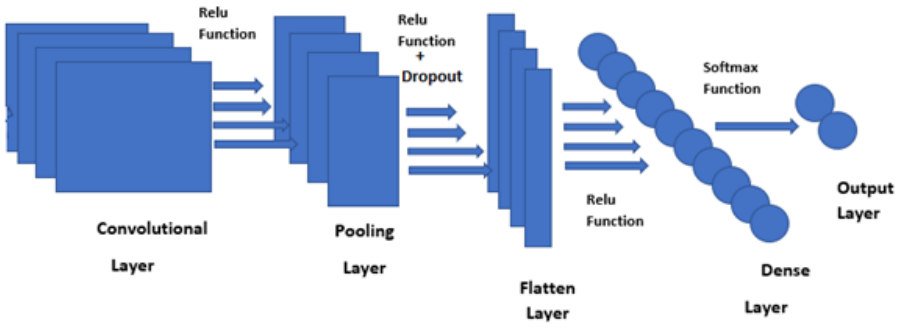

Figure 14: Representation of The CNN Architecture.

5.1.2 Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) are basically neural networks that use information from previous steps because they contain internal memory. In principle, RNN's are based on sequential context. This makes the RNNs suited for problems and tasks where order of inputs is critical, e.g., in NLP, time series, etc.

5.1.2.1 The Basic Architecture Of RNNs

The next section will explain the underlying architecture and key components of RNNs

Embedding Layer:

This layer converts text data into numerical form by mapping words or tokens to dense vectors in a continuous vector space. These embeddings capture semantic relationships between words, enabling the model to understand context and meaning.

Recurrent Layer:

The recurrent layer processes sequential data by maintaining a hidden state (memory) at each time step. This hidden state captures information from previous inputs and passes it forward to influence future predictions.

In some cases, a bidirectional RNN is used to capture information from both past and future instances, enhancing the model's ability to understand context.

Activation Function:

An activation function (e.g., tanh or ReLU) is applied to the output of the recurrent layer. This introduces non-linearity, allowing the network to learn complex patterns in the data.

Output Layer:

For tasks like text classification, a fully connected layer takes the outputs from the

recurrent layer and applies an activation function (e.g., softmax) to produce the final output. This layer is responsible for classification or prediction.

The figure below explains how RNNs work.

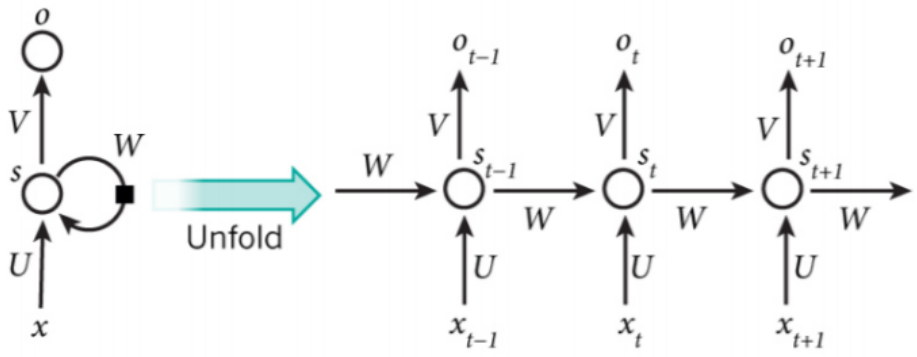

Figure 15: Representation of The RNN Architecture (Source [58].

Mathematical Formulation

In the architecture above, the first layer (embedding layer), at each time step (xt) is the embedded representation of the input text. So, at each time step t, the input (xt) is passed through the embedding layer to obtain its numerical representation.
In the second layer the hidden state (st) is the recurrent layer in the architecture. it represents the memory of the network.

$$st = f\left(Ux_t + Ws_{t-1}\right) \tag{7}$$

Where:

- U and W are weight matrices,
- The function applied to the sum of the input (U_{xt}) and the previous hidden state ($W_{s_{t1}}$) is the activation function (e.g., tanh),
- s_{t-1} is the hidden state from the previous time step.

Finally, the output at each time step (O_t) is the output layer.

$$O_t = softmax\left(v_{s_t}\right) \tag{8}$$

Where, V is the weight matrix for the output layer.

5.1.2.2 Limitations of RNNs

Despite their ability to handle sequential data, RNNs suffer from two major limitations:

- Vanishing Gradient Problem: When processing long sequences, the gradients used to update the network's weights can become extremely small, causing the model to stop learning effectively.
- Difficulty in Capturing Long-Term Dependencies: RNNs struggle to connect inputs that are far apart in the sequence, as the influence of earlier inputs diminishes over time.

5.1.3 Long Short-Term Memory Networks (LSTMs)

To address the limitations of RNNs, Long Short-Term Memory Networks (LSTMs) were introduced [59]. LSTMs are a specialized type of RNN designed to capture long-term dependencies in sequential data. They achieve this through a more sophisticated architecture that includes gates to control the flow of information.

1. **Input Gate:**
   Decides which new information should be stored in the cell state. It uses a sigmoid activation function to decide which values to update and a tanh function to create candidate values for the cell state.
2. **Forget Gate:**
   Determines which information from pervious data to discard or retain. It uses a sigmoid activation function to output values between 0 (completely forget) and 1 (completely retain)
3. **Output Gate:**
   This layer controls which information from the cell state should be output to the next hidden state. It uses a sigmoid activation function to decide which parts of the cell state to output and a tanh function to scale the values.

Mathematical Formulation

The cell state C_t at time step is updated as:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}'_t \tag{9}$$

Where:

- f_t equals the forget gate output,
- i_t equals the input gate output,
- C_t equals the candidate cell state.

The hidden state h_t at time step t is computed as:

$$h_t = o_t \times \tanh \tanh(C^t) \tag{10}$$

Where o_t is the output gate output.

Unlike RNNs, LSTMs handle Long-Term Dependency in NLP which can retain information over long sequences, and their gating mechanism helps mitigate the vanishing gradient problem, enhancing the model to learn more effectively. We can summarize their differences in the table below:

Table 3: Characteristics of RNNs and CNNs

Characteristic	**RNNs**	**LSTMs**
Memory Mechanism	Relies on a simple hidden state to retain information.	Uses a complex gating mechanism (input, forget, and output gates) to control information flow.
Long-Term Dependencies	Struggles to capture relationships between distant inputs in a sequence.	Effectively captures long-term dependencies, even over extended sequences.
Gradient Issues	Problem of vanishing gradients, making it difficult to learn from long sequences.	Solves the vanishing gradients through its gating mechanism, enabling stable training over long sequences.
Suitable Tasks	Suitable for short sequences and simpler tasks.	Ideal for long sequences and complex tasks requiring context over time.

5.2 Related Work to Author Profiling in Deep Learning

The task of author profiling has seen significant progress with the application of deep learning models. In an effort to increase the accuracy of gender prediction, several studies have explored various deep learning architectures and feature representations.

In order to determine the author's gender, [60] examined author profiling and demographic posting behavior on Arabic social media. from the period between 2011 to 2014, they mined metadata from a popular local forum, extracted features such the

top k highest scoring words and stems, normalized using tf-idf. Furthermore, frequent characters were employed as features. Two classifiers were employed, SVM with a linear kernel and 1-NN, using 10-fold cross-validation. With 100 features, the 1-NN classifier achieved an overall accuracy of 93.16%, however, the SVM performed better with bigger feature sets. Although with a small feature vector (≤ 50), both classifiers showed similar performance.

In [17] the authors also addressed gender and language variety classification in the context of Twitter author profiling. They based their approach on a deep learning architecture, and used a bidirectional GRU network with an attention mechanism. Their model demonstrated notable performance, with an average accuracy of 75.31% for gender classification across different languages. The use of GRUs and attention mechanisms showed the effectiveness of deep learning techniques in identifying complex patterns.

In [61] the authors introduced their approach for gender identification that combined text and image. In their Text Image Fusion Neural Network (TIFNN), they used word embeddings and RNNs for text analysis, and for image processing they used ImageNet-based CNN. By combining these methods, they obtained an impressive accuracy of 81.98%, highlighting the potential benefits of multimodal techniques for gender profiling.

In [62] the authors carried out a study of deep neural network approaches for author gender profiling in Turkish. Their study examined different methods, including BoW and Word2vec, fastText, Doc2vec, and they evaluated the performance of various deep learning techniques such as CNN, RNN, LSTM, GRU, etc. Although ML algorithms using BoW showed encouraging results, fastText appeared as a competitive model. This study contributes to the literature by providing a comprehensive evaluation of several instructional methods and models for author gender identification in Turkish.

Another relevant study conducted in 2023 by [63] focused on identifying hate speech on Twitter, particularly targeting sexist, racist comments. They developed a hybrid model based on LSTM and RNN that achieved impressive performance in classifying this kind of hate speech. While the primary goal was hate speech detection, the model's ability to identify sexist comments can be considered a form of gender identification.

In a recent study by [64], the authors conducted a thorough comparison of deep learning methods for author profiling, focusing on age and gender identification. They

evaluated CNN, Bi-LSTM, GRU, CRNN, and ensemble methods on four PAN Author Profiling corpora. Their findings highlighted the effectiveness of ensemble methods for same-genre author profiling and GRU for cross-genre author profiling.

6 From Deep Neural Networks to Transformers and Large Language Models

A huge advancement in the field of NLP have been made, since the evolution from deep learning to transformers appeared. In traditional DL, models struggled to coherently handle long-distance dependencies in sequential data. Thus, transformers produced a new architecture using the Attention Mechanism to address this issue. Transformers have revolutionized NLP and continue to drive advancements in this field.
The chapter starts with an introduction to transformers and their profound impact on NLP, showcasing the popular models such as BERT, GPT, and T5.

6.1 Transformers

The next section will explain the underlying architecture and key components of transformers

6.1.1 The Basic Architecture

A transformer is based on neural networks architecture but it relies on self-attention mechanisms to learn patterns and complex relationships between words or characters in a sequence [65].

The transformer consists of two mechanisms: an encoder, which takes the textual data as input and learns the sequences simultaneously left and right, and a decoder, which generates the output [65].

Transformers are excellent for modeling long-term relationships and context. They have become an important part of modern deep learning models for NLP.

The figure below represents the basic architecture of Transformers[1] where we can see that in the encoder, the self-attention layer processes the input (all words in a sequence simultaneously) and sends it to the feed forward layer where it processes the output and predicts an output. The decoder then, uses another self-attention layer and a feed forward network to generate the output.

[1] https://www.projectpro.io/article/transformers-architecture/840

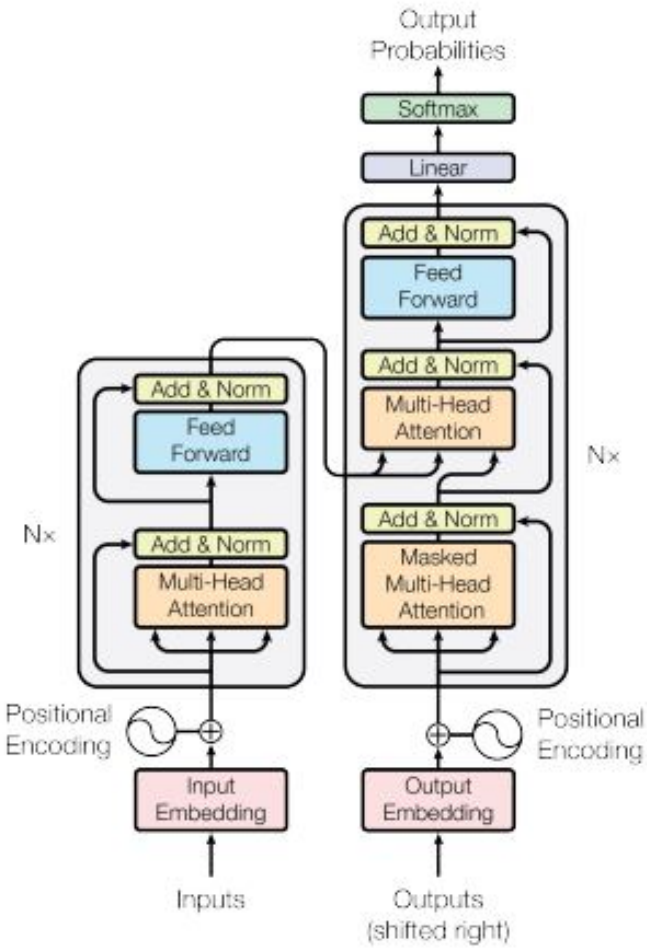

Figure 16: The Basic Architecture of Transformers (Source [65].

6.1.2 Comparison to RNNs

Transformers came solve the problem of traditional DL models. Where, they process the words in a sequence simultaneously. Also, Transformers use the method of Pre-training and Fine-tuning where, they are previously trained on large unsupervised data then fine-tuned on labeled data for a specific task. Unlike traditional DL models such as RNNs that process the words in a sequence separately and need large labeled data for each specific task.

6.2 Large Language Models(LLMs)

Transformers led to significant advances in the field of NLP, and their architecture paved the way for a breakthrough in this field with the advent of Large Language Models (LLMs). LLMs are powerful deep learning models that can efficiently solve various NLP tasks through the use of self-attention mechanisms and pre-train on large

datasets then fine-tune on a specific task. Among these powerful models that have achieved recently impressive results, are BERT, GPT and T5.

6.2.1 Google's BERT

BERT (Bidirectional Encoder Representations from Transformers) is a powerful DL language representation model in NLP. BERT was developed by Google in 2018 [42], its arrival led to some major changes in the field of NLP. BERT uses the pre-train fine-tune technique: It is pre-trained on millions or billions of unsupervised data, so that it can be fine-tuned on smaller sized data for a specific task [42].

Because, previous NLP models could only process textual data from left-to-right or from right-to-left. BERT aimed to overcome this limitation by using multi-layer bidirectional architecture.

During pre-training, BERT uses a masked language model, where it masks a word in the text input and then forces the model to generate the masked word based on the context. This process enables deep understanding of the meaning and context of contextual data. BERT is a transformer-based model with a structure almost identical to that of a transformer. Eg., Transformers have two mechanisms, the encoder and the decoder [65]. However, Unlike Transformers, the purpose of BERT is to generate an LM therefore, it uses solely the encoding part because it can efficiently achieve its objective. The pre-training and fine-tuning procedure illustrated in Figure 17 adapted from the original paper.

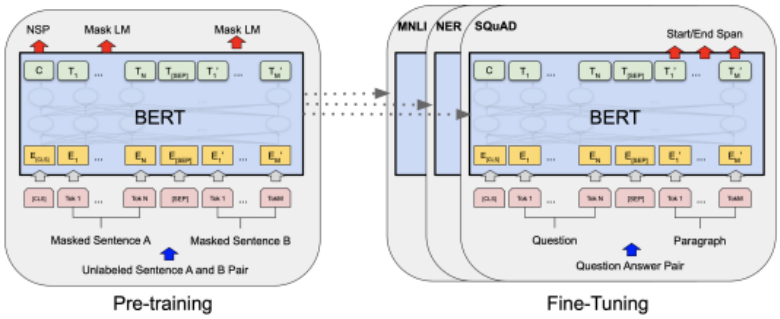

Figure 17: Pre-training and fine-tuning of BERT (Source [66]).

6.2.2 GPT (Generative Pre-trained Transformer)

GPT is another DL language generative pre-trained model. GPT is based on the Transformers structure. It uses the decoder to generate textual data, almost identical to human texts [67]. GPT is previously trained on almost 45 terabytes of unsupervised

data. This data is composed of massive amounts of texts from the web, these texts are mostly in the English language however, they contain other languages too.

The first GPT was introduced by the Open AI team in 2018. Based on the same ground of GPT1, GPT2 was produced a year after it. GPT2 had almost the same model structure of GPT1 with more data for the training [68]. Furthermore, in June 2020, the third successor has appeared as GPT-3, in which parameters and data scaling have been further expanded.

This allowed for deeper understanding of natural language patterns and structures due to training the model on a huge dataset.

GPT3 has been able to be used for a variety of applications including accurate text generation such as the fluent generation of poetry and code generation. GPT3 is also popular by other applications e.g., famous chatbots. With minimal additional training, it can also be adapted to new use cases such as text classification, sentiment analysis, and more (fine tuning for specific tasks).

Despite GPT-3's impressive capabilities, there are ethical, semantic, and mathematical issues due to the fact that it was trained on large textual data that may contain biases in some parts of it that can lead to erroneous or incorrect results [67].

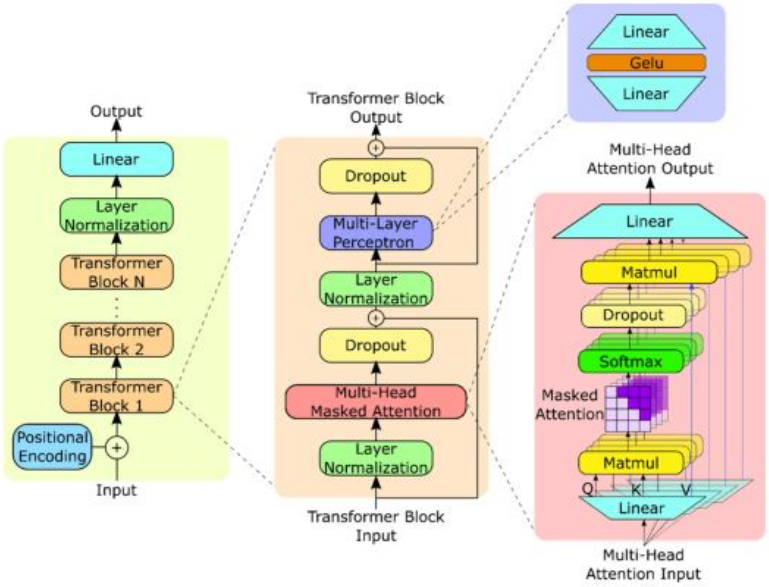

Figure 18: GPT's Basic Architecture (Source [69]).

6.2.3 T5 (Text-to-Text Transfer Transformer)

T5 is another LM developed by Google in 2019. T5 is also a pre-trained model based on transformer architecture. It is flexible and very easy to use and can cover a multiple NLP application. It is inspired by BERT in masking certain percentage of word inputs, the so-called corruption phase, and learns to generate the hidden words based on the relationship between words and context. However, unlike BERT, T5 uses an encoder-decoder architecture. T5 has demonstrated powerful performance in handling tasks like text classification, translation and summarization [70].

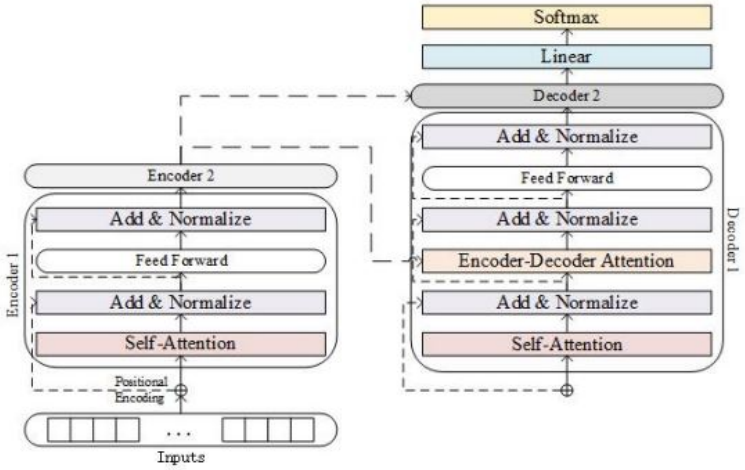

Figure 19: T5's Basic Architecture (Source [71]).

6.2.4 Prompt-Based Learning (Pre-train, Prompt, Predict)

Prompt-Based Learning is one of the latest revolutionizing techniques from transformers in the field of NLP. The idea behind this new technique is that models are self-tuned [72] and try to adapt textual data to the pre-trained model and overcome one of the biggest problems in NLP which is the need for large amounts of labeled data.

Similar to the previous algorithms discussed above, Prompt-based learning appeared in 2021 with the Pre-train and Fine-tune strategy that permits us to apply the same model to multiple problems without re-training on extensive supervised dataset that are often unavailable.

inserting a clear instruction or a template to the textual input is he general idea of Prompt-based learning, to prompt the output by guiding it in other words: prompting the model's response [72].

When taking movie sentiment analysis problem as an example, we may apply a

template to edit the initial input like this: “Input_Text” It was “Mask”.

Then we use a verbalizer to predict the label class of the textual input. The verbalizer often contains a set of label words such as the classes negative or positive for the example of the movie review. These two classes could be defined with words such as: Positive [excellent, good, cool, etc.] and Negative [bad or horrible] [73]. Then we combine the previous steps with a pre-trained language model (PLM) [73].

The PLMs that could be applied in Prompt-based learning are multiple for instance encoder-decoder PLMs like T5 that excel at translation, summarization and classification or Masked encoder only LMs like Bert for tasks like text classification or even decoder only LMs like GPT [72]. After selecting the appropriate PLM, we can move on to the training phase. Selecting the suitable prompt for a specific task in Prompt-based learning is very important for achieving accurate results.

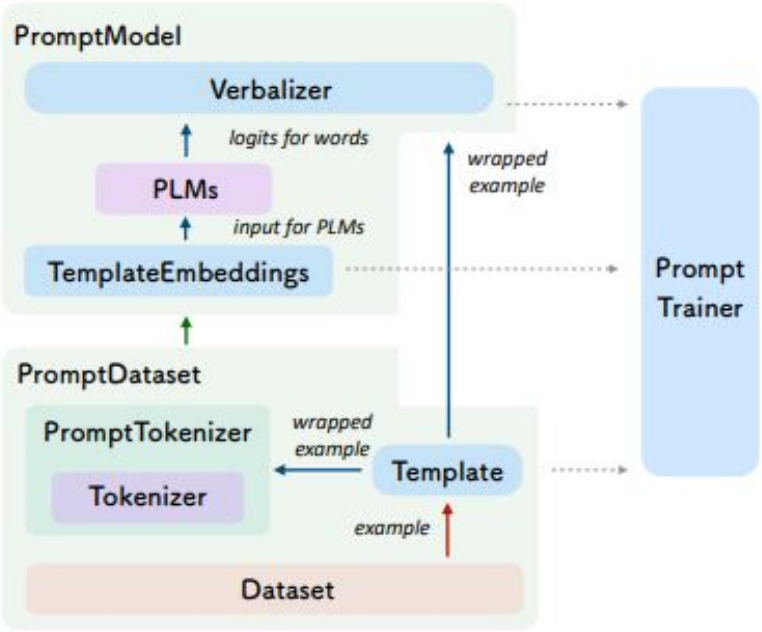

Figure 20: The Basic Architecture of Openprompt (Source [73]).

6.3 Related Work to Author Profiling in Transformers and LLMs

Transformers and large language models (LLMs) have emerged as powerful tools for natural language processing tasks such as author profiling. The following relevant studies have explored their application to gender profiling:

In a recent study by [74], the authors conducted a comprehensive evaluation of language models for author profiling in Spanish. They compared various models,

including LLMs such as BERT, GPT-3 and smaller, distilled versions. Distilled models are trained to imitate the behavior of larger models while being much smaller and more computationally efficient. Additionally, they explored the impact of multilingual models, designed to be proficient in multiple languages. By investigating these factors, the authors aimed to identify the most suitable language models for author profiling tasks in Spanish.

In another recent study by [75], the authors investigated the use of pre-trained transformers and transfer learning for author profiling. They used datasets from PAN15 and PANDORA to explore the effectiveness of these techniques in predicting author characteristics such as gender, age, and Big Five personality features. By utilizing transfer learning with BERT and GloVe, they were able to obtain promising results, indicating the suitability of these approaches for author profiling tasks.

In a recent work by [76], the authors proposed a novel deep learning architecture called Wide & Deep Transformer WD-T for author profiling, specifically focusing on gender, language variety, and personality prediction (PAN competition). Their method combined contextualized word vector representations and handcrafted features using a self-attention mechanism and a novel encoding technique. By using these techniques, they achieved competitive performance on various author profiling tasks, outperformed earlier deep learning models by up to 3.4% on the MBTI dataset.

In [77], the authors suggested a multimodal neural framework for gender profiling using data from Twitter. Their approach combined text and image features using BERT based and EfficientNet. By implementing a direct product-based fusion technique, they outperformed earlier techniques and reached state-of-the-art performance on the PAN-2018 author profiling dataset. They achieved accuracies of 82.05% for pure-image, 86.22% for pure-text, and 89.53% for multimodal setting. Additionally, they performed an extensive analysis to identify gender-specific clues within the data, providing valuable insights into the factors influencing gender prediction.

In [78], the authors explored author profiling in the context of the Arabic author profiling and deception detection shared task (APDA), focusing on age, language variety, and gender prediction. They used pre-trained BERT models, fine-tuning them on shared task data and augmenting it with in-house data to improve performance. Their best models, which were selected by majority vote, achieved competitive results,

with accuracies of 40.97% joint for all three tasks, 81.67% for gender, 93.75% for dialect, and 54.72% for age.

7 Conclusion

In conclusion, Chapter 3 has provided a comprehensive overview of the basic concepts and practical uses of NLP, focusing on the role of Deep Learning and Machine Learning techniques. Where, we have explored the architecture of Convolutional Neural Networks, Recurrent Neural Networks, and the transformer models like BERT, GPT-1, GPT-2, and T5 that have revolutionized the field, by demonstrating the remarkable abilities of these LMs. This chapter also sets the groundwork for further research on prompt-based learning and the promising developments in NLP that lie ahead. By understanding the basics of NLP, this chapter has prepared us to dive deeper into the following chapters.

PART TWO: SCIENTIFIC CONTRIBUTIONS

Chapter 4: Dataset Creation for Gender Profiling and Bot Detection in Arabic NLP

Chapter 5: Experimental Framework and Findings: Gender Profiling and Bot Detection in Arabic

Chapter 4: Dataset Creation for Gender Profiling and Bot Detection in Arabic NLP

1 Introduction

In machine learning and natural language processing (NLP), the quality and size of datasets are crucial to the performance of any model. Datasets constitute the foundation upon which models are built and evaluated. A well selected and sufficiently large dataset is necessary in order to guarantee that the models learn from a diverse and representative range of inputs, which can shape their understanding and capabilities. The quality and size of the dataset is even more crucial when the task at hand involves complicated sociolinguistic aspects, such gender author profiling, since they have a direct impact on the model's ability to identify nuanced patterns in language use between different groups.

The lack of gender-labeled and bot-labeled Modern Standard Arabic (MSA) datasets has significantly limited progress in the fields of gender author profiling and bot detection. Most Arabic studies referenced in the related work section did not provide accessible datasets. For instance, some studies only included tweet IDs without accompanying text, because tweets had been permanently deleted, rendering it unsuitable for our experiments. Other studies unfortunately provided a dataset link that was no longer operational. Also, other existing datasets have proven to be biased and unreliable because of their collection methods, such as the short specific time period, repetitive content, unbalanced distribution and nearly identical texts that follow a predictable pattern.

To overcome this challenge, we have carefully constructed 2 datasets particularly designed for both tasks: gender profiling and bot detection in MSA. Our first dataset consists of 10,000 MSA texts, carefully categorized as male or female, sourced from three distinct sources. And the second dataset for bot detection consists of 1100 MSA texts labeled as Automated or Manual from 2 distinct sources.

For the first task of gender profiling, the first source of the data is the PAN 2018 corpus [79], a widely used resource in the field of author profiling, which contains MSA texts and gender labels. The second source is The Arabic Parallel Gender Corpus 2.0 [80], which was obtained from the Open-Subtitles project, this corpus offers translated texts from English to Arabic. Lastly, we designed a questionnaire via Google Forms, targeting university students to enrich the dataset and introduce more diverse linguistic expressions. This questionnaire collected text samples based on participants' opinions and impressions, in order to capture natural language expressions from various

demographic groups.

By combining these sources, we created a large and representative dataset that covers a variety of linguistic styles and contexts, from formal texts to conversational subtitles. This strong foundation enables us to develop a generalizable gender profiling model capable of accurately identifying the author's gender based on their MSA text, thereby advancing the field of gender studies in Arabic language research.

For the second task of bot-detection, we utilized two primary datasets: Fake News [81] (Dataset 1) and Detecting Automatically-Generated Arabic Tweets [82] (Dataset 2). By combining elements from both datasets and refining the text for MSA, we established a foundational dataset for bot-detection research that is both linguistically consistent and aligned with Modern Standard Arabic.

In this chapter, we will delve into each of these collected resources in detail, explaining how they were selected, the characteristics they bring to the dataset, and how they contribute to the overall task of gender profiling.

2 The First Data Resource for Gender Profiling Dataset

In this section, we present the dataset utilized for gender profiling, which serves as the foundation for evaluating the performance of several machine learning models. The dataset is described in detail below.

2.1 PAN 2018

The PAN 2018 [79] dataset was the first resource we used for data collection. This dataset was created as part of the PAN shared task workshop.

Since 2010, a workshop called PAN (Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection) has organized shared tasks. PAN[1] is now a major hub for digital text forensics since these tasks have attracted a large number of researchers from fields such as information retrieval, NLP, and machine learning. Over the years, PAN has grown to include tasks on gender identification, authorship profiling, and other relevant fields.

In particular, the PAN 2018 shared task worked on author profiling, focusing on gender identification through a multimodal approach that used both text and images. The dataset for this challenge represented three languages: Arabic, English, and

[1]https://pan.webis.de/

Spanish. The dataset contained 100 tweets and ten images for each user, who were labeled according to their gender.

Given that our study focuses on gender profiling in Modern Standard Arabic (MSA), we only extracted the Arabic-language data from this corpus. However, we faced one major challenge, not all the Arabic tweets were written in MSA. A significant portion of the data was written in dialectal Arabic, which is outside the scope of our research. Thus, we manually translated the majority of the tweets into MSA, to ensure that our dataset aligned with the requirements of our study.

2.2 Data Preparation

To retrieve the initial Arabic dataset from PAN 2018, we had to create a program to automate the process which must take as input the absolute path to a dataset, and has to output for each document of the dataset a corresponding XML file. This step resulted in a final dataset with 100,000 tweets, each labeled with the author's gender.

The following algorithm 1 illustrates the process in details:

Algorithm 1 *Dataset XML Converter*

1. **Function** ***generate_xml*** (dataset_path, label_path, output_path):
2. ** Directory Setup **
3. Check if the output directory exists; if not, create it;
4. Open the output file `PAN_ar_dataset.txt` for writing;
5. **For each** XML file in inputDir:
6. Extract the base filename from the XML file;
7. Search the '`label_path`' file to find the corresponding gender;
8. Parse the XML file;
9. **For each** text entry in the XML file:
10. Write the gender label and text content to the output file;
11. End for;
12. End for;
13. End function.

Unfortunately, our research explicitly focuses on using MSA, and the extracted dataset was not entirely in MSA. Therefore, we had to manually translate the majority of the tweets. Eventually, we reached 1000 tweets that were written solely in MSA.

Table 4 illustrates how the original tweets were translated into MSA:

Table 4: Examples of tweets translated from dialect to MSA.

Original Arabic Tweet (Dialect)	**Author's Gender**	**Dialect**	**Revised MSA Translation**	**English Translation**
كيف الحال دكتور، لو فاضي تشيك على المقطع وتعطيني رايك فيه.. اسف على الازعاج	Male	Gulf Arabic	كيف حالك يا دكتور؟ إذا كان لديك وقت فأرجو أن تتفضل بتفقد المقطع وأخبرني برأيك فيه.. اسف على الازعاج	How are you, doctor? If you have time, please take a look at the clip and tell me your opinion on it? Apologies for any inconvenience caused.
وصلني مسج تم استلام طلبك جهاز ريسفر وسيتم التوصيل خلال يومين وانا مو طالب شيء المسج واصله بالخطأ	Male	Gulf Arabic	وصلتني رسالة تم استلام طلبك جهاز ريسفر وسيتم التوصيل خلال يومين وانا لم اطلب شيء الرسالة وصلت بالخطأ	I received a message that your order has been received, a receiver device, and it will be delivered within two days, and I did not order anything. The message arrived by mistake.
حتى انا مافهمتش واش قاعد يشوف السيد هذا ... هدف صحيح وين يشوف في الخطا ؟!	Female	Algerian Arabic	حتى انا لم افهم ما يرى هذا الرجل .... هدف صحيح اين رأى الخطأ؟!	Even I did not understand what this man sees.... a right scored goal and where did he see the foul?!
مليت كل يوم نى انفكر ايش انطيب معاش عندي حتى أفكار جديدة	Female	Libyan Arabic	مللت كل يوم أفكر ماذا اطبخ ليس لدي اي أفكار جديدة	I'm bored every day thinking what to cook I don't have any new ideas.

3 The Second Data Resource for Gender Profiling Dataset

The second data resource used in this study is based on the Arabic Parallel Gender Corpus [80], which is built for gender identification tasks, involving the first and second grammatical persons "I" and "You".

This corpus was taken from the English-Arabic OpenSubtitles 2018 dataset [83] and annotated by [80]. OpenSubtitles is a large, multilingual corpus composed of TV shows and movies' subtitles. The dataset selected for this study is written in Modern Standard Arabic (MSA) with its English translation. The dataset includes a wide variety of conversational texts. The corpus is suitable for studies focused on gender-based language expressions in Arabic because the conversations cover wide range of gender interactions.

The gender mirroring technique used to create this dataset is what makes it unique. In Arabic, gender is grammatically marked in verbs, pronouns, and adjectives, making it possible to transform a text used by a particular gender into their opposite gender. This technique requires changing systematically a statement that was first written using either male or female grammatical markers to reflect the opposite gender. When a female speaker addresses male audience, for instance, it will be rewritten as if a male speaker addresses female audience, with all relevant grammatical elements modified accordingly.

3.1 Gender Labeling in the Dataset

Each text in the dataset is labeled based on the gender of the first and second grammatical person. Both participants' genders are represented by a two-letter code in the labeling method. The first letter indicates the gender of the speaker, and the second letter indicates the gender of the listener. The following are the four potential gender labels:

Table 5: Gender Labeling.

Label	**Description**
M	Male
F	Female
N	Non-existent; the speaker is not addressing anyone in particular
B	Invariant or ambiguous the gender is NOT clear

This labeling results in 16 possible combinations for every sentence:

Table 6: Combinations Possibilities.

Label	Description
BB	Invariant/Ambiguous to Invariant/Ambiguous
FB	Female to Invariant/Ambiguous
MB	Male to Invariant/Ambiguous
BF	Invariant/Ambiguous to Female
BM	Invariant/Ambiguous to Male
BN	Invariant/Ambiguous to Non-existent
NB	Non-existent to Invariant/Ambiguous
NN	Non-existent to Non-existent
FN	Female to Non-existent
MN	Male to Non-existent
NF	Non-existent to Female
NM	Non-existent to Male
MM	Male to Male
FM	Female to Male
MF	Male to Female
FF	Female to Female

This dataset captures a variety of gender-based interactions allowing for a more accurate gender identification.

3.2 Dataset Annotation and Reinflection

Gender reinflection is the process by which annotators rewrite sentences to reflect the opposite gender, thereby enriching the dataset. This reinflection maintains the original word count and guarantees that the original and mirrored sentences are precisely aligned by systematically changing each word in the sentence to match the grammatical rules of the opposite gender. This method makes it possible to directly compare texts written by male and female authors in gender identification tasks.

For instance, if a sentence originally written by a male speaker contained a gender-specific word, such as "مجنون" (crazy [male]), would be reinflected to "مجنونة" (crazy [female]) for its female equivalent. However, Lexical adjustments may be necessary in certain situations where gender-specific words need to be changed out for entirely new ones. For example:

- "أبي" (my father) would change to "أمي" (my mother),
- "ابني" (my son) would change to "ابنتي" (my daughter).

By ensuring that the dataset includes both male and female sentences' versions, these transformations allow researchers to test gender identification algorithms on a dataset that is entirely gender-balanced.

3.3 Tackling Quality Control and Ambiguity

In order to maintain the dataset's neutrality and adaptability for gender identification tasks, all proper names were categorized as gender-ambiguous (B), regardless of how culturally gendered they might appear.

During the annotation process, the annotators were also asked to highlight any poorly translated or distorted sentences. After annotation, the dataset was subjected to a quality control process in which any mistakes—like improperly aligned parallel sentences or malformed Arabic subtitles—were manually corrected. This step was necessary, in order to guarantee the general quality and reliability of the data.

3.4 Our Process for Re-annotation

For our study, we further filtered the data so we could focus on sentences with first-person "I" references. By eliminating any sentences that did not meet this criterion, we intended to streamline the dataset for gender identification tasks. Furthermore, we simplified the gender labeling system to solely male and female. Texts with non-existent or ambiguous gender markers were either eliminated or relabeled.

Texts with the labels FB, MB, FN, MN, MM, FM, MF, or FF, for instance, were kept but re-annotated to conform to our streamlined gender classification. Sentences with labels like BM, BF, BN, NB, NN, BB, NF, and NM were eliminated from the dataset, because they did not fit with our focus on binary gender identification.

We narrowed down the initial collection of 52,000 English-Arabic annotated pairs to a more focused and manageable subset of 8,000 cleaned pairs after this re-annotation

and data cleaning process. This new dataset ensures gender balance, is free of duplicates, and is set up for use in research involving gender identification.

Examples of how the re-annotation and reinflection methods were used on the dataset are shown in the table 7 below. For each English-translated sentence, the corresponding Arabic sentence can be seen with its original label and the re-inflected version with the opposite gender.

Table 7: Example of Re-inflected Texts.

English translated text	**Original Arabic text**	**Label**	**Text Re-inflection**	**Re-inflection Label**
If I told you something strange, would you think I am crazy	إذا اخبرتك بشيء غريب، هل ستعتقدين أني مجنون	Male	إذا اخبرتك بشيء غريب، هل ستعتقدين أني مجنونة	Female
I'm afraid you'll be unlucky there sir	أنا خائف أنك ستكون غير محظوظ هناك سيدي	Male	أنا خائفة أنك ستكون غير محظوظ هناك سيدي	Female
I am a free man and you are a free man	أنني رجل متحرر وأنت رجل متحرر	Male	أنني امرأة متحررة وأنت رجل متحرر	Female
I'm going with you dad	أنا ذاهب معك يا أبي	Male	أنا ذاهبة معك يا أبي	Female

This section provides a thorough description of the second data resource, the re-annotation procedure, and examples of the gender mirroring method, ensuring that all pertinent information is provided.

4 The third Data Resource for Gender Profiling Dataset

4.1 Dataset Purpose and Creation

The third data resource was specifically created to tackle the task of gender profiling within Arabic-speaking populations, With an emphasis on students' perspectives. The process of creating the dataset was unique, gathering the opinion and impressions of university students regarding their personal high school and university experiences. This approach is ideal for gender classification tasks because it allowed the collection of personal, reflective responses which naturally display a variety of linguistic expressions.

We distributed a survey with six specific questions to students using Google Forms. Respondents were encouraged to give concise yet perceptive answers to each question regarding their educational background and personal goals. The responses' diversity

reflects differences in tone, language, and sentence structure that are frequently influenced by the gender of the respondent. This resource offers insightful information about gendered language patterns in a research context, by providing genuine data for machine learning models focused on gender profiling.

4.2 Dataset Collection and Demographics

Students from different academic fields were the target audience for the questionnaire, which was distributed mainly within university settings.

We simplified the labeling process by asking respondents to indicate their gender. despite our goal of obtaining a balanced dataset, the resulting dataset was rather biased, with more responses from female students than male students. Cultural and sociological factors, such as differences in willingness to participate in surveys, particularly when it comes to subjects involving personal expression, may be the cause of this gender imbalance.

In order to ensure consistency in response length and facilitate analysis and annotation, the questions were designed to require answers to be no more than four sentences. We produced a dataset with 1,000 unique texts after data cleaning and verification. This dataset provides a rich resource of information for studying gender language patterns.

4.3 Process of Labeling and Annotation

Each response was annotated based on the gender chosen by the respondent. This labeling helped in creating a reliable ground truth for training machine learning models. The labels used were simple:

- **Male**: Responses provided by male students.

- **Female**: Responses provided by female students.

To make sure that no misclassifications impacted the quality of the dataset, responses were examined for labeling consistency.

4.4 Details of the Questionnaire design

The questions were designed to encourage students to share personal experiences and goals by providing reflective, open-ended answers. The purpose of these questions was to gather information about the students' academic experiences in a variety of areas,

such as their favorite courses, particular experiences and memories, desired careers, and personal projects.

The six questions posed to the students were as follows:

Table 8: The Six Questions Posed To The Students.

Question in Arabic	**Translated Question**	**Explanation**	**Answer in Arabic**	**Translated Answer**	**Label**
ما هي أهم المواد التي كنت تحب دراستها في مرحلة الثانوية؟ لماذا؟	What were the most important subjects that you liked to study in high school? Why?	This question probes students' academic preferences and the motivations behind them, revealing underlying values associated with subject choices.	مادة هندسة الطرائق لأنني كنت جد متعلقة باستاذتي، لهذا كنت ممتازة فيها وأحوز على أفضل النقاط	Process Engineering because I was very attached to my teacher, so I was excellent in it and I got the best points.	Female
			العلوم الطبيعية و الفيزياء: لأنني أحب علوم المادة / الأدب العربي و اللغة الإنجليزية: لأني أحب الكتابة و الشعر.	Natural sciences and physics: because I love exact sciences / Arabic literature and the English language: because I love writing and poetry.	Male

Question in Arabic	Translated Question	Explanation	Answer in Arabic	Translated Answer	Label
لخص أحسن ذكرياتك في الثانوية التي درست بها.	Summarize your best memories of the high school you attended.	This question encourages students to recall positive experiences, providing insight into the aspects of high school life they found fulfilling or memorable.	أحسن ذكرياتي في الثانوية هي أوقات الفراغ لأني كنت سعيدة واستمتعت بوقتي مع زميلاتي وأنسى أنني كنت حزينة	My best memories in high school are break times because I was happy and I enjoyed my time with my classmates and forgot that I was sad.	Female
			أحسن الذكريات كانت في التجمع و الصحبة مع الأصدقاء و الزملاء.	The best memories were gathering and company with friends and colleagues.	Male
عبر عن شعورك وانت تحوز على شهادة البكالوريا.	Express how you felt when you succeeded in the baccalaureate degree.	Here, students describe their emotions upon passing an important milestone, revealing personal sentiments and the weight of academic achievements.	شعرت كأنني فراشة، خفيفة الوزن أكاد أطير، جد سعيدة	I felt like a lightweight butterfly, almost flying, and so happy.	Female
			عدم الرضى نوعاً ما، لعدم حصولي على المعدل الذي أردته.	Somewhat dissatisfied for not getting the grade that I wanted.	Male
هل كان التخصص الذي أرسلت إليه في الجامعة مناسباً لك ؟ لماذا ؟	Was the major you were sent to at the university suitable for you? Why?	This question examines the alignment between students' academic interests and their assigned field of study, offering a perspective on academic satisfaction and choice.	نعم أنا جد راضية على اختياري فمن خلاله اعمل بحرية ولا أنشغل عن أطفالي كأنني ربة بيت	Yes, I am very satisfied with my choice, through which I work freely and do not get busy to take care of my children as if I am a housewife.	Female
			نعم، لأني مهتم بالبرمجة و الحاسوب منذ نعومة أظافري.	Yes, because I have been interested in programming and computers since my childhood.	Male

Question in Arabic	Translated Question	Explanation	Answer in Arabic	Translated Answer	Label
عبر عن طموحاتك المهنية بعد الجامعة.	Express your career aspirations after college.	Students share their future ambitions, indicating both practical and idealistic aspirations, with potential reflections on societal expectations.	من أهم طموحاتي أن أصبح مستقلة ماديا وناجحة في مجال دراستي و إذا حالفني الحظ أصبح أستاذة في يوم ما.	One of my most important ambitions is to become financially independent and successful in my field of study, and if I am lucky, to become a teacher one day.	Female
			العمل كمهندس برمجيات مستقل و إنشاء شركتي الخاصة.	Working as a freelance software engineer and starting my own company.	Male
عبر عن مبادرة تمنيت القيام بها في يوم ما.	Express an initiative that you would like to take one day.	This question explores students' interest in taking initiatives, often touching on themes of altruism, community service, and personal impact.	من أهمها أن أكون إنسانة متطوعة لكل ما فيه الخير لمجتمعي و غيري، أساعد بما أستطيع و أكون مشاركة و مساهمة في كل ما يهدف إلى بناء مجتمع راقي.	One of the most important is to become a volunteer for all that is good for my community and others and to help as I can and to be a participant and a contributor to everything that aims to build a prosperous society.	Female
			تكوين فريق عمل من مبرمجين مختصين في مجالات مختلفة لتغطية مساحة عمل واسعة.	Forming a team of programmers specialized in different fields to cover a wide work area.	Male

4.5 Linguistic Patterns Based on Gender

After analyzing the responses, several gender-based language patterns become apparent:

4.5.1 Emotion and expression

- Female responders frequently use metaphors like "lightweight butterfly" and descriptive words like "attached to my teacher" that express emotions and intimate relationships.
- Male responders are more likely to be brief and real, highlighting accomplishments and practical outcomes, as demonstrated by sentences like "forming a team of programmers".

4.5.2 Concepts and Goals

- Female responders tend to focus on social relationships, community and social engagement, and family-related goals, including "volunteering for my community" or "balancing work and childcare".
- Male responses include technical skills, independence, and career advancement, such as "starting my own company" or "forming a specialized team."

4.5.3 Courses Preferences

- While males concentrate on content and capacities, females tend to show devotion to certain teachers or a supportive environment during the course of their education for example 'I was very attached to my teacher'.

4.6 Applications and Potential Uses of the Dataset

There could be numerous uses for this dataset that extend past gender profiling. A wide range of NLP tasks, including sentiment analysis, language modeling, and emotional tone classification, can be built upon it. it also opens avenues for research on social and cultural influences on language expression through its emphasis on gendered language patterns.

The dataset can be used to:

- Train models for gender classification specific to Arabic.
- Create language models designed to recognize and adapt to gendered expressions in Arabic.
- Contribute to studies on how gender influences language in Arabic-speaking communities, supporting research in gender linguistics.

4.7 Future Enhancements and Expansion

Future work could involve collecting additional responses from male and female students. Furthermore, additional enhancement could involve adding more questions that delve into cultural or societal topics, providing a more comprehensive view of gender in Arabic language use.

This third dataset resource, with its focus on subjective expression and gender profiling, contributes significantly to the field of Arabic NLP. It demonstrates how language reflects personal identity and how gender-specific linguistic features can be harnessed for gender-based studies in author profiling.

5 Final Dataset Preparation and Evaluation Strategy

To ensure a balanced and representative dataset, we merged texts from three distinct resources: PAN, OpenSubtitle, and Google Forms questionnaire. This integration resulted in a corpus of 10,000 samples selected from an initial collection of 13,000 entries across these resources. By creating this merged corpus, we aimed to enhance the diversity and balance of gender representations, which is crucial for robust gender classification in Arabic.

To analyze the impact of dataset size on model performance, we divided the final dataset into progressively larger sub-corpora. This progressive approach allowed for evaluations on datasets of varying sizes, including 2,500, 5,000, 7,500, and the full 10,000 entries. The objective was to observe whether increasing the dataset size would lead to consistent improvements in classification accuracy and generalization across all three methods under study.

For each of these sub-corpora, we ensured a balanced distribution between male and female texts. We developed a custom program that alternates between male and female entries from the different resources. This approach allowed us to maintain an even gender distribution without relying solely on the standard shuffling function. By structuring the data in this way, each sub-corpus maintained a randomized yet balanced selection, which is crucial for reducing potential gender bias in training and evaluation.

The following algorithm 2 explains the process of the gender distribution in dataset:

Algorithm 2 *Alternating Gender Distribution Algorithm*

```
Function AlternateGenderEntries ():
    /* Initialize Sheets and Clear Content */
    Activate Sheet2 and clear contents in columns A and B;
    Set headers in Sheet2: Copy "Text" and "Gender" headers from Sheet1;
    /* Initialize Row Counters */
    i ← 2 (Counter for source sheet rows);
    i_M ← 2 (Counter for male entries in the target sheet);
    While Sheet1.Cells(i, 1) ≠ "":
      If Sheet1.Cells(i, 2) = "M":
      Copy "Text" and "Gender" values to Sheet2.Cells(i_M, 1) and Sheet2.Cells(i_M, 2);
        i_M ← i_M + 2 (Move to the next alternate row for Male entries);
      End if;
      i ← i + 1;
    End while;
    /* Reset Counters for Female Entries */
    i ← 2 (Reset source sheet row counter);
    i_F ← 3 (Row counter for female entries in the target sheet);
    /* Place Female Entries */
    While Sheet1.Cells(i, 1) ≠ "":
      If Sheet1.Cells(i, 2) = "F":
        Copy "Text" and "Gender" values to Sheet2.Cells(i_F, 1) and Sheet2.Cells(i_F, 2);
        i_F ← i_F + 2 (Move to the next alternate row for Female entries);
      End if;
      i ← i + 1;
    End while;
End function.
```

The pre-processing techniques applied were consistent across all models to ensure comparability in results. Each sub-corpus was split into training (80%) and testing (20%) sets, maintaining balance within each subset. This controlled and progressive evaluation framework enabled us to assess the scalability of the methods, observing how performance metrics evolve with increased data availability.

We can present a portion of the final corpus in Table 9 :

Table 9: An Example of The Final Structure Of The Dataset.

Arabic sentence	Label	Translation in English
شكرًا أنا سعيد لأنك تقول هذا	Male	Thanks I'm glad you are saying this
أنا متأكدة أنك مخطئ	Female	I'm sure you're wrong
سأتولى هذا يا رجل أنا رجل	Male	I'll handle this man I'm a man
حسنًا أنت تعلم عندما كنت طفلة	Female	Well you know when I was a kid
آسف لكن عندما تذهب للصيد يلزمك طعم ذكي	Male	Sorry but when you go fishing you need smart bait
الآن إني محبطة بعض الشيء	Female	Now I am a little disappointed
تقول أنني لن أحب نفسي إذا استسلمت الآن	Male	She says I will not love myself if I give up now
آسفة يا سيد هيندرسن	Female	I'm sorry, Mr. Henderson
أنا متأكد أنك مخطئ	Male	I'm sure you're wrong
مرحبًا أنا لست موجودة الآن رجاءا اترك رسالة	Female	Hello, I am not available right now. Please leave a message
حسنًا أنت تعلم عندما كنت طفلا	Male	Well you know when I was a kid
لأنه كما ترى أنني رئيسة هذه اللجنة	Female	Because, as you can see, I am the chair of this committee

6 The First Data Resource for the Bot-Detection Dataset

In this section, we present the first dataset utilized for bot detection. The dataset is described in detail below.

6.1 Fake News Dataset (Bot Dataset 1)

The Fake news dataset [81] was developed by Marc Jones and Wajdi Zaghouani for the Qatar International Fake News Detection and Annotation Contest. The dataset was centered on tweets tagged with "#السعودية" (meaning "Saudi Arabia") and primarily addressed political topics. Unfortunately, this dataset has proven to be biased because of its collection method: the short specific time period, repetitive content (both in the train and test set), unbalanced distribution (13,204 tweets labeled as Automated and

5,180 tweets labeled as Manual) and nearly identical texts that follow a predictable pattern for instance, Automated tweets were characterized by repetitive content and consistent hashtag placement. therefore, it was unreliable to take as whole.

7 The Second Data Resource for the Bot-Detection Dataset

In this section, the second dataset utilized for bot detecion is described in detail below.

7.1 Detecting Automatically-Generated Arabic Tweets (Bot Dataset 2)

This dataset, created by [82], contained a comprehensive collection of tweets. Initially, 11,764 unique user tweets were selected over a four-day period. To capture temporal features, an additional 120 tweets were collected for each user. After preprocessing—such as removing punctuation, stop words, and duplicate content—the dataset included 1,202,815 tweets. Out of these, 3,503 tweets were manually labeled using the CrowdFlower platform. Labels consisted of 55% (1,944) classified as **Automated** tweets and 45% (1,559) classified as **Manual** tweets.

8 Final Dataset Preparation and Evaluation Strategy for Bot-Detection

One major challenge of both datasets was that the text primarily consisted of dialectal Arabic rather than Modern Standard Arabic (MSA), which is the main focus of our research. To address this, we manually curated a balanced subset of 1,100 texts labeled as Automated/Manual (50% from each of the above-mentioned datasets) to create an MSA-compatible dataset. This process involved manually correcting non-Arabic and dialectal texts and translating them into MSA In order to guarantee linguistic consistency and relevance. Furthermore, to ensure the dataset's reliability and reduce biases, URLs were removed from the texts during preprocessing. This decision was made because the majority of automated texts included URLs, making their patterns predictable and easily distinguishable from human texts. However, since bots can strategically delete URLs to imitate human behavior and evade detection, relying solely on URL presence as a feature is insufficient. Thus, it is crucial to incorporate diverse features and techniques to achieve more robust and accurate bot detection.

We therefore, created a foundational dataset for bot-detection research that is linguistically coherent and aligned with Modern Standard Arabic by integrating portions from both datasets and adjusting the text for MSA.

We can present a portion of the final corpus in Table 10 :

Table 10: A Portion of the Final Bot-Detection Corpus.

Arabic sentence	**Label**	**Translation in English**
مراسل قناة الفضائية غرفة القيادة والسيطرة و أرقام البلاغات المسجلة الى هذه الساعة ـ اللهم احفظ الحجيج وسلمهم ـ #السعودية #إيران	Automated	The correspondent of the satellite channel reported from the command and control room and the registered reports up to this hour. O Allah, protect the pilgrims and grant them safety. #SaudiArabia #Iran
RT @faisalbinturki1: يسرني كرئيس لنادي النصر دعوتكم يا جماهير الوفاء لحضور حفل تتويج العالمي بملعب استاد الملك فهد الدولي بالرياض الليلة بكم …	Manual	RT @faisalbinturki1: As president of Al-Nassr Club, I am pleased to invite you, loyal fans, to attend the global coronation ceremony at King Fahd International Stadium in Riyadh tonight with you...
الآن يمكنك الحصول على كل ما تحتاجه بضغطة زر. تسوّق اليوم! 📱	Automated	Now you can get everything you need with a click of a button. Shop today! 📱
لا يطعن العاقل في قلبه مرتين. إن أساء إليك مرة واحدة ف الخطاء منه وإن أساء مرتين ف الخطاء منك وإن أساء ثلاث مرات فعن مثلك رفع القلم.	Manual	A wise person is not hurt in their heart twice. If someone wrongs you once, the fault is theirs. If they wrong you twice, the fault is yours. And if they wrong you three times, accountability has been lifted from someone like you.
أهلا بالعام _الدراسي ايلتس مبتعثين _كندا مبتعثين _أمريكا #الدراسه مبتعث _متميز _في _اللغة #شهادات #الايلتس #معتمدة #موثقة #عروض _ خاصه #اختبار #اختبارات	Automated	#Welcome_Back_to_School IELTS Students_Canada Students_USA #Study Outstanding_Student_in_Language #Certificates #IELTS #Certified #Verified #Special_Offers #Test #Exams
اللهم إن عائلتي هي أجمل هداياك وأغلى ما أملك، فاحفظهم لي وأسعدهم فسعادتهم هي سعادتي .	Manual	O Allah, my family is the most beautiful gift and the most precious thing I own, so protect them for me and make them happy because their happiness is my happiness.
أرسل وزنك وطولك على الخاص وانا سوف اريك كيف تخس وتحصل بوقت قصير للتواصل عبر الخاص	Automated	Send me your weight and height privately, and I will show you how to lose weight and achieve results in a short time. Contact privately.

Arabic sentence	Label	Translation in English
RT @jar_190: إذا لم يتحرك الهلال لبدء التعاقدات والإستقطابات من الأن فمتى يتحرك؟!	Manual	RT @jar_190: If Al-Hilal does not start making contracts and recruiting now, then when will they move?!
قرار السعودية التدريجي حول تأخير إلغاء الاكتتاب لارامكو يعني … البورصة الأمريكية خطر مستقبلا بعد اعترافات المسؤول انتخابات الرئيس #ترامب الأخيرة خطوة ممتازة وهذا يدل على أنها الحدث ليس قرب … منتصف الأحداث … خطوة	Automated	Saudi Arabia's gradual decision to delay canceling Aramco's IPO means... The U.S. stock market poses a future risk after officials' confessions. The last presidential election for #Trump was an excellent step, indicating that the event isn't near...
حرض أبطال الجيش الوطني يخوضون #معارك عنيفه شرق مدينة حرض ويحرزون تقدم متسارع باتجاه المدينه الجهة الشرقيه تحت غطاء كثيف طيران الأباتشي وإسناد مدفعية القوات	Manual	#Haradh National Army heroes are fighting fierce battles east of Haradh city, making rapid progress toward the city. The eastern side is under heavy Apache air cover and artillery support.
جيري ماهر:هناك خلافات نشبت بين النظام السوري حزب الله بسبب منهم يريد السيطرة والهيمنة على بعض المعابر المائية التي تقع على نهر الفرات سوريا.	Automated	Jerry Maher: Disputes have arisen between the Syrian regime and Hezbollah because some of them want control and dominance over certain water crossings on the Euphrates River in Syria.
خطوة جديدة للمرأة السعودية لتحقيق المساواة لأول مرة فتاة تنفذ عملية سرقة …! #فتاة _تسرق _سيارة متوقفة أمام سوبر لأول مرة فتاة :@Hukusfof ماركت بالدمام! تنفذ عملية سرقة …! #فتاة _تسرق _سيارة متوقفة أمام سوبر ماركت بالدمام! سبب تطور السرقات ليصل لهذا الحد؟!! بسبب البطالة والفقر مثلا ؟!	Manual	A new step for Saudi women to achieve equality. For the first time, a girl commits a theft...! #Girl_Steals_A_Car parked in front of a supermarket in Dammam! @Hukusfof: For the first time, a girl commits a theft...! #Girl_Steals_A_Car...

9 Topic Modeling

In this section, we describe the application of topic modeling to the two datasets: gender profiling and bot detection. Using **TF-IDF** and **Non-Negative Matrix Factorization (NMF)**, we extracted latent topics from the text data. This technique is used to uncover recurring patterns and themes and to better understand the content and structure of the dataset.

9.1 Methodology

9.1.1 Preprocessing

The dataset was normalized by removing diacritics, non-Arabic characters, and stop words. Words were also standardized for consistency across forms.

9.1.2 TF-IDF Vectorization

Text data was transformed into a Term Frequency-Inverse Document Frequency matrix to weigh words based on their relevance within the dataset.

9.1.3 NMF Topic Modeling

The NMF algorithm identified latent topics by decomposing the TF-IDF matrix into components representing topics and their associated terms. NMF effectively reveals coherent topics in the dataset.

We then extracted the topics present in the dataset. Each topic was represented by a set of keywords alongside their corresponding weights, which allowed us to interpret and assign meaning to the discovered topics.

Below, we detail the results of topic modeling for the gender profiling dataset. The analysis for the bot detection dataset will be discussed later.

9.2 Topic Modeling For The Gender Profiling Dataset

In the gender author profiling experiment, due to the presence of text mirroring in the original corpus specifically in the **Arabic Parallel Gender Corpus**, we decided to apply the topic modeling on only the first batch that contains 2,500 examples. The results revealed distinct patterns in the male and female classes, emphasizing linguistic features specific to each gender.

9.2.1 Results for the Female Class

The following table presents the results of topic modeling for the female class, highlighting the associated keywords of the most prominent topics.

Table 11: Female Topic Modeling.

Topic	Keywords
Topic 1	بأن، حقا، أعلم ، تكن، هل، سيد، أولا، بشأن، فقط، آسفة
Topic 2	لي، الآن، به، اخترته، هل، متأكدة، حقا، فقط، اسفة، انا
Topic 3	بشدة، صغيرة، لكنني، مهتمة، عندما، رياضيات، احب، لانني، إن، كنت
Topic 4	خائفة، بك، للغاية، إلهي، فى، امرأة، أبي، سيدي، رجل، يا
Topic 5	أما، أنه، الآن، فأنا، أنك، واثقة، لأكون، أعلم ، متأكدة، لست
Topic 6	اختياري، تخصص، الذي، مناسبا، لأنه، لانني، لانه، مناسب، اخترته، نعم
Topic 7	رجل، أريد، أنك، جد، أردت، أكون، سأكون، أكن، لأنك، سعيدة
Topic 8	العلوم، شعور، أحب، لانها، المواد، الله، لي، مادة، كانت، الرياضيات
Topic 9	حلمت، لك، إلهي، إلهتي، نفسي، اليوم، الذي، حسنا، هل، لقد
Topic 10	لأنه، جيدة، كيف، فخورة، مناسب، خائفة، لأنك، جادة، مسرورة، جدا

9.2.2 Female Topic Analysis

The model identified 10 prominent topics within the **female class**, highlighting themes of emotions, personal decisions, and academic preferences.

Key observations:

- **Self-expression and Emotion:** Topics like 4 and 10 include phrases such as "إلهي", "خائفة", and "مسرورة" reflecting emotional expressions.
- **Academic and Career Aspirations:** Topics like 3 and 6 focus on academic subjects and career paths, such as choosing fields like math or expressing ambition to become a lawyer.
- **Politeness and Social Interactions:** Topics 2 and 7 reflect politeness markers like "أسفة" and phrases indicating social expectations.

9.2.3 Results for the Male Class

The following table presents the results of topic modeling for the male class, highlighting the associated keywords of the most prominent topics.

Table 12: Male Topic Modeling.

Topic	Keywords
Topic 1	الرجل، أنه، عني، تكن، فقط، هل، سيد، بشأن، لقد، آسف
Topic 2	مرحبا، فعلا، الآن، قليلا، لي، حقا، متأكد، فقط، اسف، انا
Topic 3	لكنني، أحب، إلهتي، كثيرا، صغيرا، علي، عندما، إن، لقد، كنت
Topic 4	فخور، إلهتي، خائف، بك، للغاية، إلهي، أبي، في، رجل، يا
Topic 5	حقا، ربما، بأن، لكنني، مما، شعرت، أعلم ، لأكون، متأكدا، لست
Topic 6	الحمد، سأكون، الزواج، أعلم ، شخصيا، لأنك، شكرا، أنك، رجل، سعيد
Topic 7	شهادة، خاصة، المعدل، الدراسة، حتى، فرحة، عادي، جميل، رائع، شعور
Topic 8	حلمي، الرياضيات، الآن، اريد، أريد، الان، أصبح، اصبح، الله، أكون
Topic 9	أحب، مناسبا، مهما، حقيقة، رجل، مناسب، لانني، لأنه، التخصص، نعم
Topic 10	حتى، نفسي، قلق، عندما، عادي، انت، لأنك، جاد، مسرور، جدا

9.2.4 Male Topic Analysis

The **male class** also exhibited 10 distinct topics, often reflecting societal roles, ambitions, and emotional restraint.

Key observations:

- **Societal Expectations:** Topics such as 1 and 4 include mentions of "رجل", "فخور", reflecting societal roles, pride, and responsibility. For example, "جعلتني اشعر بالمسؤولية كوني رجل" (Topic 4).
- **Professional and Academic Aspirations:** Topics like 8 and 9 highlight academic fields and career milestones.
- **Emotional Engagement:** Phrases in topics 3 and 5 indicate emotional engagement but often with a sense of rationality.

9.2.5 General Interpretation and Insights

Although the linguistic differences between male and female authors highlight gender-specific themes and patterns, the text mirroring impacted the topic modeling results and led to redundancy because it exceeds 20% of the dataset (500 examples).

9.3 Topic Modeling For The Bot Detection Dataset

The bot detection dataset was divided into two classes: **Automated** and **Manual**. For each class, we applied preprocessing steps including Arabic text normalization, stopwords removal, and TF-IDF vectorization, followed by NMF to extract 10 topics. The table below summarizes the key topics for each class along with representative keywords.

9.3.1 Results for the Automated Class

The following table presents the results of topic modeling for the automated class, highlighting the associated keywords of the most prominent topics.

Table 13: Automated Class Topic Modeling.

Topic	Keywords
Topic 1	الجزيرة، مصر، العالم ، العربية، الحرمين، الشريفين، قطر، المملكة، اليمن، السعودية
Topic 2	معنا، التفاصيل، تطبيق، العروض، موقعك، اسم ، احصل، تسو، اشتر، الآن
Topic 3	الحياة، عمر، كيف، السياسية، تويتر، أكثر، مكان، مسلم ، حزب، الله
Topic 4	الفضائية، احد، مسلم ، داخل، لدينا، مراسل، ايران، قناة، المسلمين، اللهم
Topic 5	الستار، استمتع، التفاصيل، رابط، العالم ، وشارك، غزة، الجديدة، الحلقة، شاهد
Topic 6	سوريا، عبدالعزيز، صورة، كاتب، الإخوان، المسلمين، سلمان، العوين، قطر، محمد
Topic 7	موقع، فقط، أفضل، الجديدة، منتجات، موقعك، منتجاتنا، هل، احصل، اليوم
Topic 8	الحجاج، عدد، أكثر، الحج، مراسلي، أكبر، وزير، الفضائية، قنوات، سعودي
Topic 9	عمان، السعوديه، أبطال، اليمني، قوات، العربي، السعودي، الشعب، التحالف، حتى
Topic 10	صورة، خلال، معنا، الرابط، الخاص، الذي، الرياض، عاجل، عبر، الان

The topics reflect meaningful clusters of words that characterize the different types of communication.

9.3.2 Automated Topic Analysis

In the Automated class texts relates to:

- **Promotional Content:** Topics such as Topic 2 and Topic 7 focus on advertisements, discounts, and products.
- **Religious and Political Themes:** Topics like Topic 4 and Topic 8 address religious figures, events, and political entities.
- **International Regional News and Media:** Several topics, such as Topic 1 and Topic 5 emphasize news coverage and global events.

9.3.3 Results for the Manual Class

The following table presents the results of topic modeling for the manual class, highlighting the associated keywords of the most prominent topics.

Table 14: Manual Class Topic Modeling.

Topic	Keywords
Topic 1	العربية، آخر، سعودي، سلمان، أول، بسبب قطر، العالم، كندا، السعودية
Topic 2	قال، انه، له، فيه، عندما الحج، الوطن، بإذن، يا، الله
Topic 3	تاريخ له، العرب، انت، تغريدة، تكون، الف، الحجاج، لما، فيها
Topic 4	الف، مجلس، وكل، الله، الكويت، الخليج، مبروك، غدا، الدوري، الذي
Topic 5	شركة، معها، مثل، الكثير، غدا، كانت، عبر، هناك، حتى، اليوم
Topic 6	الملك، يتم، اليمني، الشعب، حال، عدم، العام، عاجل، حرب، اليمن
Topic 7	انه، النصر، يارب، الحمدلله، نتمنى، الامارات، المرأة، تحت، شخص، يوم
Topic 8	عدد، العربي، التواصل، الرياض، ريال، عبر، خاص، الحج، فيديو، فقط
Topic 9	جدا، شكرا، رجال، لكل، المملكة، لكم، شخص، الملك، شركة، السعوديه
Topic 10	فلا، له، عبد العزيز، الحياة، يحب، نعمة، ربي، يارب، لي، اللهم

9.3.4 Manual Topic Analysis

The **Manual class** revealed diverse topics, including:

Key observations:

- **Personal Reflections:** Topics like Topic 10 focus on religious invocations and personal gratitude.
- **News, Political Issues, and Commentary:** Topics such as Topic 1, Topic 6, and Topic 8 highlight discussions about Saudi Arabia, political issues, and societal matters.

- **Expressions of Gratitude, Appreciation, and Social Unity:** Topic 9 reveals unity, appreciation, and gratitude.

9.3.5 General Interpretation and Insights

- **Automated Texts:** These are characterized by repetitive patterns associated with media-oriented, promotional campaigns, political, or religious content.
- **Manual Texts:** These show a richer variety of personal and conversational content, with an emphasis on gratitude and engagement with socio-political discussions.

The application of topic modeling illustrates the nature and the significant differences between bot behavior and human communication. The extracted information can be used to understand the audience engagement.

10 Conclusion

In this chapter, we detailed the steps taken to gather, preprocess, and organize a balanced corpus for gender author profiling and for Bot-detection. Starting from three different sources (PAN, OpenSubtitles, and Google Forms) for the gender profiling task, we curated a diverse dataset of 10,000 entries, with an equal representation of male and female authors. Similarly for the bot detection task, we curated a diverse dataset of 1,100 entries. We employed specific data-cleaning techniques to enhance the quality and consistency of the texts, ensuring suitability for our NLP models.

We also introduced a progressive dataset evaluation strategy, where the corpus was divided into subsets of increasing size (2,500, 5,000, 7,500, and 10,000 entries). This approach allows us to assess the impact of corpus size on model performance, adding depth to our analysis. Finally, we developed a custom algorithm to alternate between male and female entries, ensuring balanced training and testing data for our experiments. This methodology aims to minimize bias and enhance the reliability of our gender profiling models.

The next chapter will delve into the experimental setup, model configurations, and evaluation metrics, building on the structured datasets prepared in this chapter.

Chapter 5: Experimental Framework and Findings: Gender Profiling and Bot Detection in Arabic

1 Introduction

This chapter presents the experimental framework, methodologies, and results of the two core tasks of this research: gender profiling and bot detection in Modern Standard Arabic (MSA). The goal of these experiments is to evaluate the performance of state-of-the-art models: LSTM, ARABERT, and Prompt-Based Learning, in addressing these tasks. The chapter is structured to provide detailed insights into the experimental setups, preprocessing steps, model configurations, and evaluation metrics for each task. The findings from these experiments contribute significantly to advancing Arabic NLP by exploring cutting-edge techniques and filling existing gaps in the literature.

2 Gender Profiling Experimentation

The experimentation for gender profiling in Modern Standard Arabic (MSA) was carried out to assess the performance of three advanced models: **LSTM**, **ARABERT**, and **Prompt-Based Learning**. These models were selected due to their demonstrated strengths in natural language processing tasks, particularly in Arabic text classification. The objective was to evaluate their effectiveness in predicting the gender of an author based solely on their written text. This section provides an in-depth description of the experimental setup, data preprocessing, model configurations, and the metrics used for evaluation.

2.1 Experimental Setup

The experimentation was carried out on a newly constructed dataset consisting of **10,000 MSA texts**, evenly distributed between male and female authors. The dataset was constructed from three primary sources, as described in Chapter 4:

- **PAN 2018 Corpus [79]**: This corpus is widely recognized in author profiling studies and was manually translated into MSA. It provided 1,000 texts after translation and filtering.
- **The Arabic Parallel Gender Corpus 2.0 [80]**: This dataset includes translated texts sourced from the Open-Subtitles project. It contains conversational texts in MSA, re-annotated for gender profiling, contributing 8,000 texts.
- **Google Forms Questionnaire**: A Google Forms questionnaire targeted at

university students, labeled by gender. This dataset added 1,000 texts to the overall corpus, ensuring the inclusion of diverse writing styles.

The dataset was divided into four progressive subsets containing 2,500, 5,000, 7,500, and 10,000 texts, respectively, to evaluate the impact of dataset size on model performance. Each subset was further split into 80% training and 20% testing, ensuring a balanced distribution of male and female texts. This progressive approach allowed us to observe the model's scalability with increasing data size.

2.2 Data Preprocessing

The dataset underwent the following preprocessing steps to ensure compatibility with the models:

2.2.1 Tokenization

- For ARABERT and Prompt-Based Learning, subword tokenization was applied using the SentencePiece tokenizer provided by ARABERT.
- For LSTM, word-level tokenization was used to represent the text in sequential format.

2.2.2 Padding and Truncation

- Input sequences were padded to a fixed length to ensure uniformity across batches.
- Longer texts were truncated to the maximum sequence length to align with model constraints.

2.2.3 Balancing

- Each subset was balanced to ensure an equal representation of male and female texts. This was done to prevent any bias in the training process.

2.2.4 Shuffling

- The data was shuffled randomly to ensure diversity and reduce the likelihood of overfitting during training.

2.3 Model Configurations

2.3.1 LSTM Model

The LSTM (Long Short-Term Memory) model was specifically designed to capture long-term dependencies in sequential text data. Its configuration included:

Architecture:

1. Pre-trained word embeddings (e.g., Word2Vec) were used to initialize the embedding layer, to convert tokens into dense vectors, ensuring semantic relationships between words.
2. Two LSTM layers with 128 units each.
3. Dropout layers with a 50% dropout rate to prevent overfitting.
4. An output layer for classification.

Training Parameters:

1. Optimizer: Adam, a popular choice for deep learning models because of its adaptive learning rate capabilities.
2. Loss Function: Binary Crossentropy, suitable for binary classification tasks like gender profiling.
3. Epochs: 5 (with early stopping based on validation loss)
4. Validation Split: 20%, used to monitor the model's performance on unseen data during training.

The following algorithm 3 describes the process of training an LSTM model for gender author profiling:

Algorithm 3 *LSTM model training*

Function train_LSTM (dataset):
/ Preprocessing & Tokenization */*
tokenized_text ← **tokenize** (dataset.text);
padded_sequences ← **pad_sequences**(tokenized_text);
embeddings ← **load_pretrained_embeddings**(); */* e.g., Word2Vec */*
/ Model Architecture */*
model ← **Sequential**();
model.add(**Embedding**(input_dim=vocab_size, output_dim=embedding_dim, weights=[embeddings]));
model.add(**LSTM**(units=128, return_sequences=True));
model.add(**Dropout**(rate=0.5));
model.add(**LSTM**(units=128));
model.add(**Dropout**(rate=0.5));
model.add(**Dense**(units=1, activation='sigmoid'));
/ Training */*
history ← model.fit();
return model, history;

Figure 21 illustrates the architecture and training process of the LSTM model for gender author profiling.

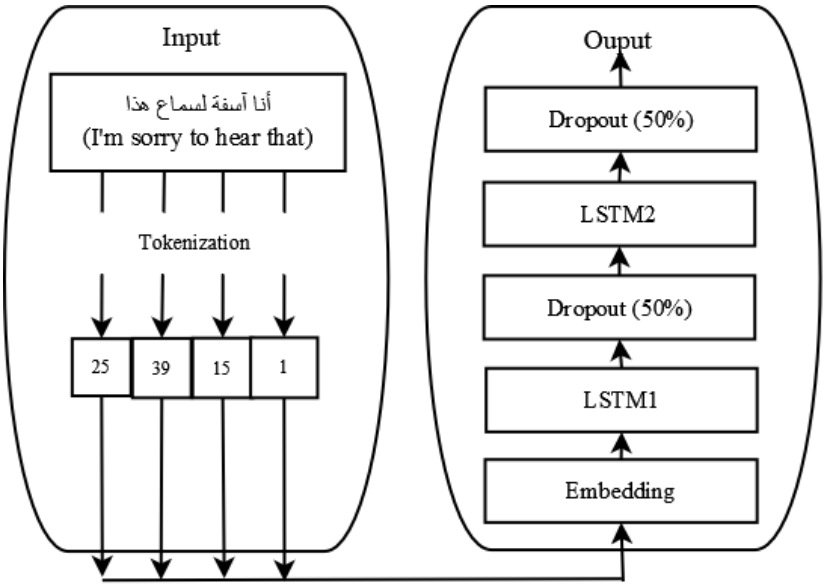

Figure 21: The architecture of the LSTM Model used for gender author profiling (source [33])

2.3.2 ARABERT Model

The ARABERT model is a pre-trained transformer-based model for Arabic NLP tasks. in this work Arabert was fine-tuned for gender profiling, the process involved:
Architecture:

1. ARABERT Utilized the BERT-base architecture, with 12 encoder layers, 12 attention heads, and a maximum sequence length of 512 tokens.
2. The model was pre-trained on a large corpus of Arabic text (70 million MSA and dialectal texts), making it well-suited for MSA tasks. Thus, for the pre-training phase, it uses a masked language model, it hides word in the input sentence and then makes the algorithm, based on the context, predict the hidden word [84].

Fine-Tuning Process:

1. The model was fine-tuned using the CLS token embedding. The CLS token's embedding was used as a summary representation of the input text.
2. The dense layer predicted gender (male/female) based on the CLS embedding.

Training Parameters:

1. Optimizer: Adam

2. Loss Function: CrossEntropyLoss
3. Learning Rate: 2e-5, a common choice for fine-tuning transformer models.
4. Epochs: 5
5. Validation Split: 20%

The following algorithm 4 describes the process of training ARABERT model for gender author profiling:

Algorithm 4 *Arabert model training*

```
Function fine_tune_ARABERT(dataset):
/* Preprocessing & Tokenization */
tokenized_text ← ARABERT_tokenizer(dataset.text);
input_ids ← tokenized_text['input_ids'];
attention_mask ← tokenized_text['attention_mask'];
/* Model Architecture */
model ← ARABERT.from_pretrained('aubmindlab/bert-base-arabertv02');
/* ARABERT uses 12 encoder layers, 12 attention heads,
   and a maximum sequence length of 512 tokens */
/* The CLS token embedding is used as the summary representation of input text */
cls_embedding ← model(input_ids, attention_mask).last_hidden_state[:, 0, :];
/* Add a dense layer for binary classification */
output_layer ← Dense(units=1, activation='sigmoid')(cls_embedding);
model ← Model(inputs=model.input, outputs=output_layer);
/* Fine-Tuning Process */
model.compile(optimizer=optimizer, loss=loss_function, metrics=metrics);
/* Training */
history ← model.fit();
return model, history;
```

Figure 22 illustrates the architecture and training process of the ARABERT model for gender author profiling.

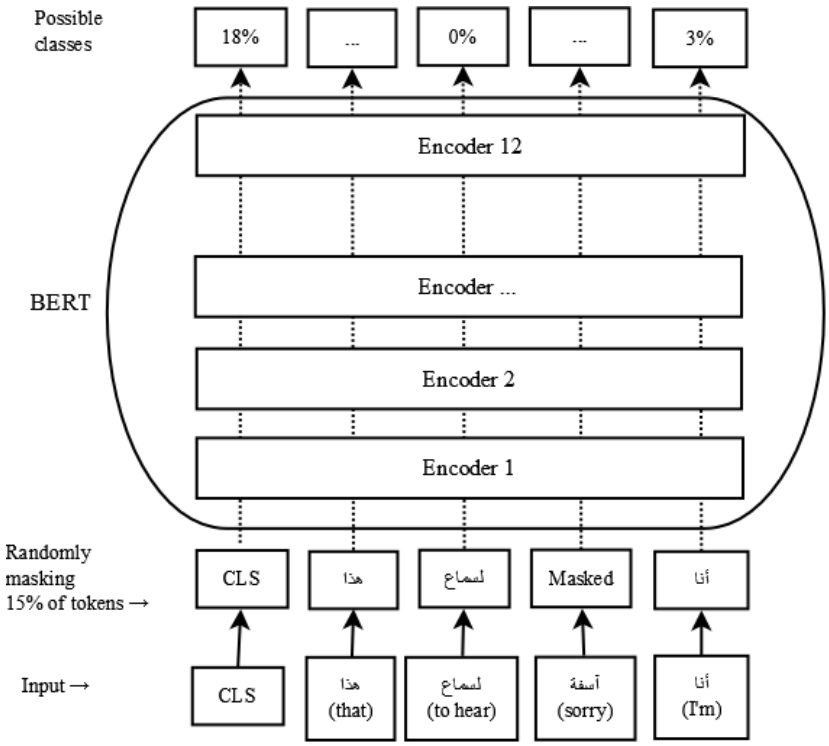

Figure 22: The architecture of the ARABERT Model used for gender author profiling (source [33])

2.3.3 Prompt-Based Learning

The Prompt-Based Learning approach was implemented using the OpenPrompt framework with ARABERT as the pre-trained language model (PLM):

Prompt Design:

Prompt Design: A textual template was designed to modify the input text into the format: "Input_Text this is [MASK]" The model was tasked with predicting the gender based on the masked token as either "Male" or "Female.".

Verbalizer:

The verbalizer mapped the model's predictions to the to the corresponding labels "Male" or "Female". This step translated the model's output into interpretable class labels.

Training Parameters:

1. PLM: ARABERT
2. Template: "Input_Text" this is "Mask"
3. Verbalizer: [Male, Female]

4. Optimizer: Adam
5. Loss Function: CrossEntropyLoss
6. Learning Rate: 1e-5 (PLM), 1e-4
7. Epochs: 5
8. Validation Split: 20%

The following algorithm 5 describes the process of training Prompt-based model for gender author profiling.

Algorithm 5 *Prompt-based model training*

Function predict_gender(text):
/ Preprocessing & Tokenization */*
prompt ← **combine_tokens**(tokenized_text, tokenized_template);
/ e.g., "Input_Text" this is "Mask" */*
model_output ← **ARABERT**(prompt);
/ The choice of PLM */*
/ **Verbalizer** (using a predefined threshold) */*
threshold ← 0.5; */* This value can be adjusted */*
if model_output[0] > threshold **then**
    predicted_gender ← "Male";
else
    predicted_gender ← "Female";
end if
return predicted_gender;

Figure 23 illustrates the architecture and training process of the ARABERT model for gender author profiling.

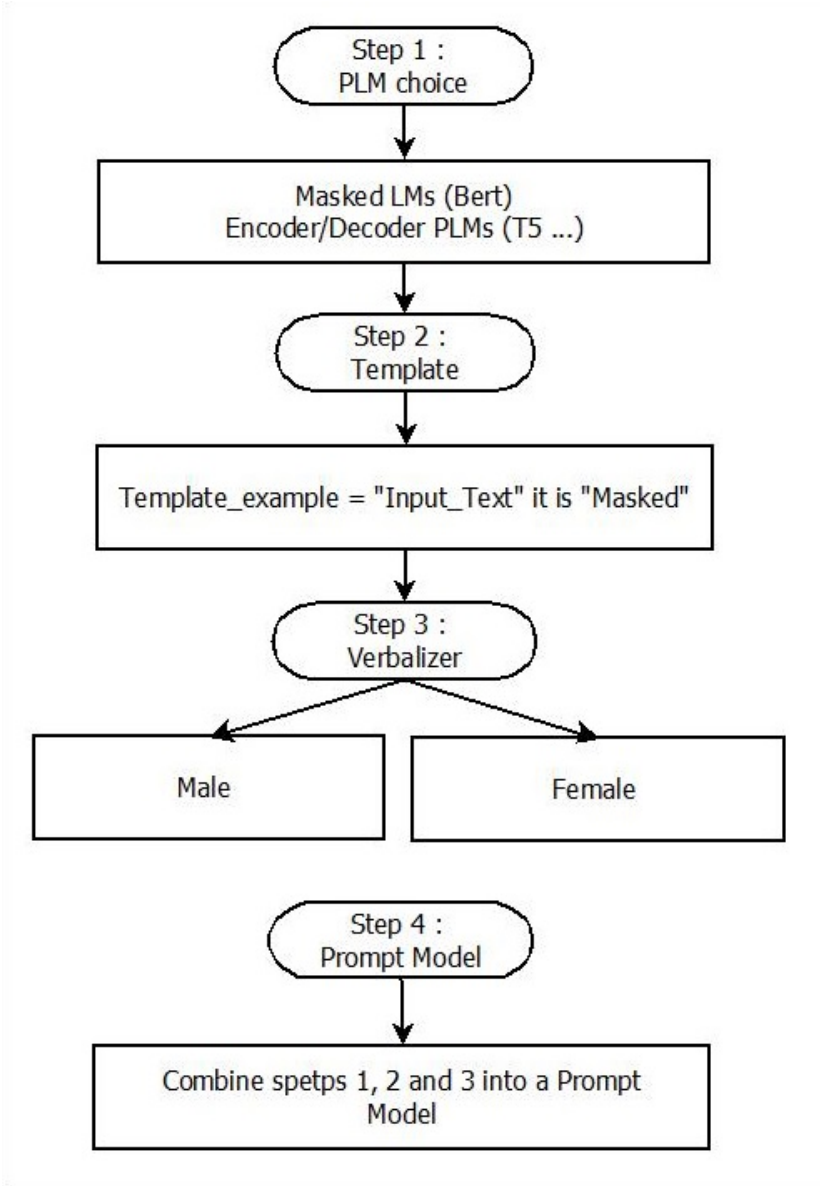

Figure 23: The architecture of the Prompt-based Model used for gender author profiling.

3 Bot Detection Experimentation

For bot detection, the experimentation was conducted on a dataset of 1,100 MSA texts, labeled as either Automated or Manual. The dataset was sourced from two primary resources:

1. Fake News Dataset: A collection of texts labeled as bot or not [81]
2. Detecting Automatically-Generated Arabic Tweets: A dataset of tweets labeled as automated or manual [82].

3.1 Dataset Preparation

The dataset was manually curated to ensure linguistic consistency with Modern Standard Arabic (MSA). Texts containing dialectal Arabic or non-Arabic content were translated into MSA.

URLs were removed from the texts to reduce bias, as automated texts often contained predictable patterns related to URL usage.

The final dataset was balanced, with 550 Automated and 550 Manual texts.

3.2 Experimental Setup

The dataset was split into 80% training and 20% testing sets, ensuring a balanced distribution of Automated and Manual texts.

The same preprocessing steps (tokenization, padding, and shuffling) and evaluation metrics (accuracy) used in the gender profiling experiments were applied here.

3.3 Model Configurations

The same three models—LSTM, ARABERT, and Prompt-Based Learning—were used for bot detection. The configurations and training parameters for these models were identical to those used in the gender profiling experiments (see above section 2.3).

4 Results and Discussion for Gender Profiling

The gender profiling experiments were conducted on datasets of varying corpus sizes, and the performance of the three methods (LSTM, Prompt-Based Learning, and ARABERT) was evaluated using accuracy as the primary metric.

After 5 epochs, we compared the performance of the three methods on four batches of datasets with the corpus sizes using accuracy as the primary performance measure. The overall measurement results are summarized in Table 15 and illustrated graphically in Figure 24

Table 15 presents the performance of the models in terms of accuracy

Table 15: Performance of the Models in Terms of Accuracy.

Corpus Size	**2,500**	**5,000**	**7,500**	**10,000**
LSTM Accuracy	55.7%	68.0%	72.3%	78.5%
Prompt-Based Accuracy	84.0%	89.3%	93.0%	92.3%
ARABERT Accuracy	84.6%	91.8%	92.6%	92.4%

4.1 Discussion

ARABERT consistently outperformed the other models with the highest accuracy, ranging from 84.6% to 92.4%. Its performance improved consistently as the dataset size increased, demonstrating its robustness and ability to handle larger and more diverse datasets and in handling complex linguistic patterns in MSA. Its pre-trained nature and ability to capture contextual information make it particularly suitable for gender profiling tasks. The model's ability to generalize well, even with limited training data, underscores the advantages of transformer-based architectures.

Prompt-Based Learning performed competitively and showed promising results, especially with carefully designed prompts and verbalizers. Its performance was nearly similar to ARABERT with accuracy increasing from 84% to 92.3%, suggesting that prompt-based approaches can be highly effective for low-resource languages like MSA and for gender profiling tasks. The results also indicate that prompt-based learning can achieve competitive accuracy without requiring extensive fine-tuning, making it a practical solution for tasks with limited labeled data.

LSTM, while showing consistency and improvement with larger datasets, struggled to match the performance of ARABERT and Prompt-Based Learning, reaching a maximum accuracy of 78.5%. This highlights the limitations of traditional deep learning models in tasks requiring nuanced language understanding. LSTM's inability to capture long-range dependencies and complex linguistic patterns likely contributed to its lower accuracy.

These results highlight the advantage of pre-trained language models like ARABERT and Prompt-Based Learning over traditional deep learning models such as LSTM. PLMs, having been trained on vast amounts of semi-supervised texts, require less labeled data to generalize effectively to downstream tasks such as gender classification.

The accuracy trend is also illustrated in Figure 24:

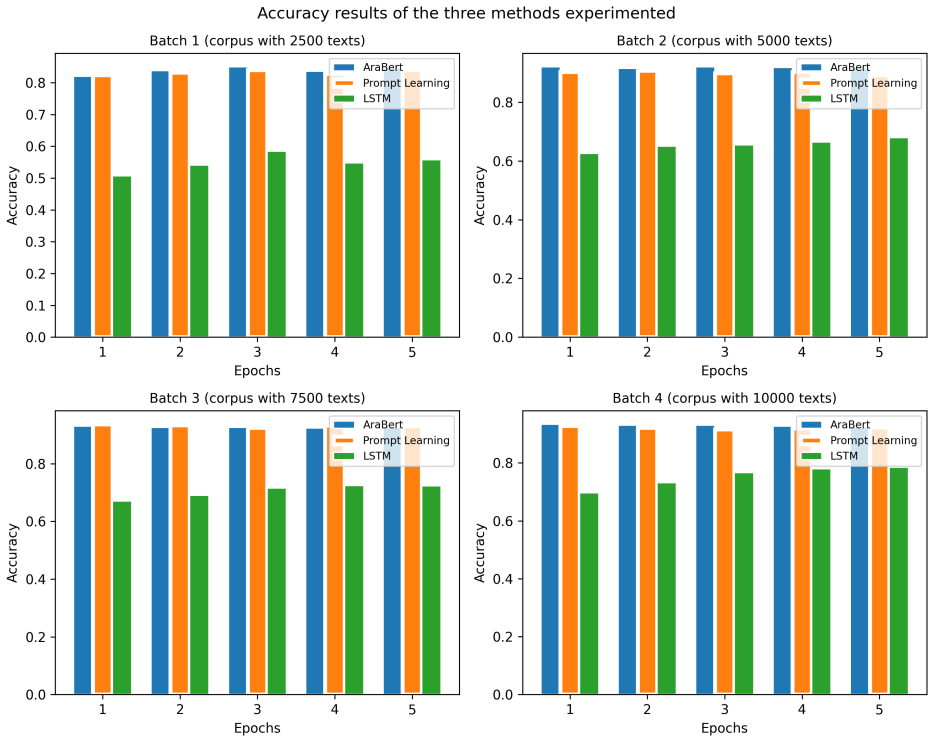

Figure 24: Accuracy Results of the Three Methods (source [33])

From Figure 24, we can observe that:

ARABERT outperforms the Prompt-Based method slightly throughout most of the batches, and shows better results as the dataset size increases.

The minimal performance gap between **ARABERT** and **Prompt-Based Learning** suggests that with further adjustments, such as more carefully designed templates and verbalizers, the prompt-based approach could enhance its performance.

Thus, **Prompt-Based Learning** can achieve competitive results with minimal fine-tuning and offers a promising alternative, especially for low-resource languages.

LSTM improves as the corpus size increases however, it still lags behind compared to the other two models. This limitation highlights the challenges of using traditional deep learning models for tasks requiring nuanced language understanding. While LSTM can be effective for simpler tasks, it may struggle with capturing complex linguistic structures in Arabic.

5 Results and Discussion for Bot Detection

The bot detection experiments were conducted on datasets of 1100 texts, and the performance of the three methods (LSTM, Prompt-Based Learning, and ARABERT) was evaluated using accuracy as the primary metric.

After 5 epochs, we compared the performance of the three methods. The overall results are summarized in Table 16.

Table 16: The Performance of the 3 Models for Bot-detection

Model	**Test Accuracy**
LSTM	66.8%
ARABERT	88.0%
Prompt-Based Learning	80.0%

5.1 Key Findings

ARABERT achieved the highest accuracy of 88.0%, demonstrating its strength in capturing subtle linguistic and contextual patterns and in distinguishing between automated and manual texts.

Prompt-Based Learning performed well, achieving an accuracy of 80.0%. Indicating that prompt-based approaches can solve effectively bot detection tasks.

LSTM lagged behind the other two methods, with an accuracy of 66.8%. This highlights the challenges of using traditional deep learning models for tasks requiring complex language understanding.

5.2 Discussion

ARABERT demonstrated highest performance in distinguishing between automated and manual texts, due to its pre-trained nature and transformer-based architecture make it well-suited for tasks requiring complex language understanding.

Prompt-Based Learning also performed well, suggesting that well-structured prompts can enhance the bot detection task without requiring extensive fine-tuning, making it a practical solution for tasks with limited labeled data.

LSTM struggled to achieve high accuracy. This highlights the limitations of traditional deep learning models for tasks requiring complex language understanding.

The results and discussion highlight the effectiveness of ARABERT and Prompt-Based Learning for bot detection tasks in MSA. However, the inherent biases in the datasets, such as repetitive content and predictable patterns, may limit the models' generalizability.

For instance, the Automatically-Generated Arabic Tweets dataset contains bot-generated texts that often have repetitive topic patterns. Similarly, the second dataset includes a high concentration of tweets featuring repetitive phrasing and a consistent use of the hashtag "#السعودية" in nearly identical positions.

While efforts were made to reduce bias by selecting the most diverse texts available, the presence of these predictable patterns may limit the models' ability to generalize when exposed to different datasets or real-world bot detection scenarios. Future work should consider expanding the dataset collection period, balancing class distributions, and incorporating more varied sources of bot-generated text to ensure robust and adaptable models.

6 Conclusion

This chapter presented a comprehensive evaluation of three state-of-the-art models—**LSTM**, **ARABERT**, and **Prompt-Based Learning**—for gender profiling and bot detection in Modern Standard Arabic (MSA). The results demonstrate the superiority of ARABERT and Prompt-Based Learning over traditional deep learning models like **LSTM**. For gender profiling, ARABERT achieved the highest accuracy of **92.4%**, closely followed by Prompt-Based Learning at **92.3%**, while LSTM reached a maximum accuracy of **78.5%**. In bot detection, ARABERT again outperformed the other models with a test accuracy of 88%, while Prompt-Based Learning achieved **80%** and LSTM lagged behind at **66.8%**.

The findings highlight the potential of pre-trained language models and prompt-based approaches for NLP tasks in low-resource languages like MSA. ARABERT's consistent performance across both tasks underscores its ability to capture

complex linguistic patterns and contextual nuances, while Prompt-Based Learning's competitive results suggest that carefully designed prompts can effectively leverage pre-trained models for specific tasks. The limitations of LSTM, particularly in handling long-range dependencies and subtle language features, further emphasize the advantages of transformer-based architectures.

Although ARABERT and Prompt-Based Learning achieved high accuracy, their effectiveness in real-world applications should be further tested on more diverse and dynamic datasets to validate their generalizability.

Chapter 6
General Conclusion

General Conclusion

This thesis has embarked on a comprehensive exploration of gender profiling and bot detection in Modern Standard Arabic (MSA), addressing critical gaps in Arabic Natural Language Processing (NLP) research. By implementing innovative machine learning techniques—LSTM, ARABERT, and Prompt-Based Learning—this work has advanced the understanding of author profiling and automated content detection in a low-resource language like MSA. The research has not only demonstrated the effectiveness of these models but also contributed novel datasets and methodologies that pave the way for future advancements in Arabic NLP. Below, we synthesize the key findings, contributions, and implications of this work.

The primary objective of this research was to evaluate the performance of advanced machine learning models in addressing the challenges of gender profiling and bot detection in MSA. The results revealed that ARABERT, a pre-trained transformer-based model, consistently outperformed the other models, achieving an accuracy of 92.4% for gender profiling and 88% for bot-detection. Its ability to capture complex linguistic patterns and contextual nuances makes it particularly well-suited for tasks requiring deep language understanding. Prompt-Based Learning, an innovative approach that uses pre-trained language models with carefully designed prompts, also demonstrated competitive performance, achieving 92.3% accuracy for gender profiling and 80% for bot-detection. This highlights the potential of prompt-based techniques as a

flexible and efficient alternative to traditional fine-tuning methods, especially for low-resource languages like Arabic. In contrast, LSTM, while showing improvement with larger datasets, struggled to match the performance of ARABERT and Prompt-Based Learning, achieving a maximum accuracy of 78.5% for gender profiling and 66.8% for bot-detection. This demonstrates the limitations of traditional deep learning models in handling tasks that require nuanced language understanding and long-range dependencies.

A significant contribution of this research is the creation of two novel datasets: one for gender profiling, consisting of 10,000 MSA texts, and another for bot-detection, comprising 1,100 MSA texts. These datasets, carefully prepared from diverse sources such as the PAN 2018 Corpus, the Arabic Parallel Gender Corpus 2.0, and custom questionnaires, address the lack of high-quality, labeled datasets for Arabic NLP tasks. By ensuring balanced representation and linguistic diversity, these datasets provide a valuable resource for future research in Arabic author profiling and bot-detection. Additionally, the progressive evaluation strategy employed in this study—dividing the dataset into subsets of increasing size—allowed for a detailed analysis of how model performance scales with dataset size, offering insights into the data requirements for effective model training.

The exploration of Prompt-Based Learning represents another key contribution of this research. As one of the first studies to apply this approach to gender profiling and bot detection in MSA, this work has demonstrated its potential as a viable alternative to traditional fine-tuning methods. The competitive performance of Prompt-Based Learning, particularly in low-resource settings, suggests that it can be effectively adapted to other Arabic NLP tasks, such as sentiment analysis, dialect identification, and machine translation. This

opens up new avenues for research and application in Arabic NLP, particularly when there is a lack of labeled data.

The findings of this research have important implications for both academic and practical applications. In the realm of gender profiling, the high accuracy of ARABERT and Prompt-Based Learning suggests that these models can be effectively deployed in real-world applications such as security, marketing, and social media analysis. For instance, gender profiling can be used to tailor marketing strategies to specific demographic groups or to identify potential suspects in forensic investigations. Similarly, in bot-detection, the ability to accurately distinguish between automated and manual texts can enhance the detection of fake news, spam, and malicious bots on social media platforms, contributing to a safer and more reliable digital environment. However, the effectiveness of these models in real-world applications should be further validated on more diverse and dynamic datasets to ensure their generalizability.

Despite the promising results, this research also highlights several limitations and areas for future work. The datasets used in the second task of this study, while carefully created, exhibit certain biases due to their collection methods. For example, the Automatically-Generated Arabic Tweets dataset contains bot-generated texts that often rely on repetitive topic patterns, while the Fake News Dataset includes a high concentration of tweets featuring repetitive phrasing and consistent use of a similar hashtag. These biases may limit the generalizability of the models to other contexts and datasets. Future work should focus on creating larger and more diverse datasets to improve the robustness and generalizability of the models. Additionally, further exploration of prompt design and verbalizers could enhance the performance of Prompt-Based Learning, making it an even more powerful tool for Arabic NLP tasks.

In conclusion, this thesis has made significant contributions to the field of Arabic NLP by addressing the challenges of gender profiling and bot-detection in MSA. The development of novel datasets, the evaluation of state-of-the-art models, and the exploration of innovative techniques like Prompt-Based Learning have advanced the understanding of author profiling and automated content detection in Arabic. The findings of this research not only demonstrate the effectiveness of ARABERT and Prompt-Based Learning but also provide a foundation for future research in Arabic NLP. As the digital landscape continues to evolve, the insights and methodologies presented here will play a crucial role in revealing the potential of Arabic language processing, filling the gap between Arabic and other widely studied languages, and enabling new applications in both academic and practical contexts.

References

[1] Kees Versteegh. *Arabic language*. Edinburgh University Press, 2014.

[2] Adnan Nouh, Abobakr Sultan, and Roshdy Tolba. An approach for arabic characters recognition. *J. Eng. Sci., Univ. Riyadh*, 6(2):185–191, 1980.

[3] A Fassi Fehri. *Issues in the structure of Arabic clauses and words*, volume 29. Springer Science & Business Media, 2013.

[4] Achraf Chalabi. Mt-based transparent arabization of the internet tarjim. com. In *Conference of the Association for Machine Translation in the Americas*, pages 189–191. Springer, 2000.

[5] Kevin Daimi. Identifying syntactic ambiguities in single-parse arabic sentence. *Computers and the Humanities*, 35:333–349, 2001.

[6] Arwa Alqudsi, Nazlia Omar, and Khalid Shaker. Arabic machine translation: a survey. *Artificial Intelligence Review*, 42:549–572, 2014.

[7] Michael P Oakes. Author profiling and related applications. 2014.

[8] Asmaa Mansour Khoudja, Mourad Loukam, and Fatma Zohra Belkredim. Towards author profiling from modern standard arabic texts: A review. In *Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London, Volume 1*, pages 745–753. Springer, 2021.

[9] Fazli Can and Jon M Patton. Change of writing style with time. *Computers and the Humanities*, 38:61–82, 2004.

[10] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.

[11] Wajdi Zaghouani and Anis Charfi. Guidelines and annotation framework for arabic author profiling. *arXiv preprint arXiv:1808.07678*, 2018.

[12] Ahmed Abbasi and Hsinchun Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005.

[13] Mats Dahllöf. Automatic prediction of gender, political affiliation, and age in swedish

politicians from the wording of their speeches—a comparative study of classifiability. *Literary and linguistic computing*, 27(2):139–153, 2012.

[14] Moshe Koppel, Navot Akiva, Eli Alshech, and Kfir Bar. Automatically classifying documents by ideological and organizational affiliation. In *2009 IEEE international conference on intelligence and security informatics*, pages 176–178. IEEE, 2009.

[15] Marc Franco-Salvador, Greg Kondrak, and Paolo Rosso. Bridging the native language and language variety identification tasks. *Procedia computer science*, 112:1554–1561, 2017.

[16] Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. N-gram: New groningen author-profiling model. *arXiv preprint arXiv:1707.03764*, 2017.

[17] Don Kodiyan, Florin Hardegger, Stephan Neuhaus, and Mark Cieliebak. Author profiling with bidirectional rnns using attention with grus: notebook for pan at clef 2017. In *CLEF 2017 Conference and Labs of the Evaluation Forum, Dublin, Ireland, 11-14 September 2017*, volume 1866. RWTH Aachen, 2017.

[18] Jon Oberlander and Scott Nowson. Whose thumb is it anyway? classifying author personality from weblog text. In *Proceedings of the COLING/ACL 2006 main conference poster sessions*, pages 627–634, 2006.

[19] Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. Personality and patterns of facebook usage. In *Proceedings of the 4th annual ACM web science conference*, pages 24–32, 2012.

[20] Miguel Ángel Álvarez-Carmona, Adrián Pastor López-Monroy, Manuel Montes-y Gómez, Luis Villaseñor Pineda, and Hugo Jair Escalante. Inaoe's participation at pan'15: Author profiling task. In *CLEF (Working Notes)*, 2015.

[21] Andreas Grivas, Anastasia Krithara, and George Giannakopoulos. Author profiling using stylometric and structural feature groupings. In *CLEF 2015: Conference and Labs of the Evaluation Forum Experimental IR meets Multilinguality, Multimodality and Interaction*. CEUR-WS, 2015.

[22] Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. Tat: an author profiling tool with application to arabic emails. In *Proceedings of the Australasian Language Technology Workshop 2007*, pages 21–30, 2007.

[23] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 436–439, 2019.

[24] Sippo Rossi, Matti Rossi, Bikesh Raj Upreti, and Yong Liu. Detecting political bots on twitter during the 2019 finnish parliamentary election. In *Annual Hawaii International Conference on System Sciences*, pages 2430–2439. Hawaii International Conference on System Sciences, 2020.

[25] Alexander Shevtsov, Christos Tzagkarakis, Despoina Antonakaki, and Sotiris Ioannidis. Identification of twitter bots based on an explainable machine learning framework: the us 2020 elections case study. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pages 956–967, 2022.

[26] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, Liang Wang, and Jon Crowcroft. An in-depth characterisation of bots and humans on twitter. *arXiv preprint arXiv:1704.01508*, 2017.

[27] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*, pages 273–274, 2016.

[28] Fred Morstatter, Liang Wu, Tahora H Nazer, Kathleen M Carley, and Huan Liu. A new approach to bot detection: striking the balance between precision and recall. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 533–540. IEEE, 2016.

[29] Nuha Albadi, Maram Kurdi, and Shivakant Mishra. Hateful people or hateful bots? detection and characterization of bots spreading religious hatred in arabic social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, 2019.

[30] Kareem Darwish, Dimitar Alexandrov, Preslav Nakov, and Yelena Mejova. Seminar users in the arabic twitter sphere. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I 9*, pages 91–108. Springer, 2017.

[31] Fouzi Harrag, Maria Debbah, Kareem Darwish, and Ahmed Abdelali. Bert transformer model for detecting arabic gpt2 auto-generated tweets. *arXiv preprint arXiv:2101.09345*, 2021.

[32] Elizabeth D Liddy. Natural language processing. 2001.

[33] Asmaa Mansour Khoudja, Mourad Loukam, and Fatma Zohra Belkredim. Assessment of lstm, arabert and prompt-based learning for gender author profiling in modern standard arabic language. *Ingenierie des Systemes d'Information*, 29(6):2209, 2024.

[34] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Text & talk*, 23(3):321–346, 2003.

[35] M Czoków1 M Meina1 K BrodziŤska, B Celmer, M Patera, J Pezacki, and M Wilk. Ensemble-based classification for author profiling using various features. 2013.

[36] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

[37] Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. Modelling sarcasm in twitter, a novel approach. In *proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 50–58, 2014.

[38] Eirik Holbæk. Using author profiling to determine the age group of an author. Master's thesis, NTNU, 2019.

[39] Jiawei Yao. Automated sentiment analysis of text data with nltk. In *Journal of physics: conference series*, volume 1187, page 052020. IOP Publishing, 2019.

[40] Hunar Batra, Akansha Jain, Gargi Bisht, Khushi Srivastava, Meenakshi Bharadwaj, Deepali Bajaj, and Urmil Bharti. Covshorts: news summarization application based on deep nlp transformers for sars-cov-2. In *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pages 1–6. IEEE, 2021.

[41] Eduard Hovy, Margaret King, and Andrei Popescu-Belis. Principles of context-based machine translation evaluation. *Machine Translation*, 17:43–75, 2002.

[42] Shreya Acharya, K Sornalakshmi, Bidisha Paul, and Anshul Singh. Question answering system using nlp and bert. In *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*, pages 925–929. IEEE, 2022.

[43] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[44] Seifeddine Mechti, Maher Jaoua, Lamia Hadrich Belguith, and Rim Faiz. Author profiling using style-based features. *Notebook Papers of CLEF2*, 2013.

[45] Iqra Ameer, Grigori Sidorov, and Rao Muhammad Adeel Nawab. Author profiling for age and gender using combinations of features of various types. *Journal of Intelligent & Fuzzy Systems*, 36(5):4833–4843, 2019.

[46] Vishnu Subramanian. Deep learning with pytorch_ a practical approach to building neural network models using pytorch-packt publishing, 2018.

[47] Varun Dogra, Sahil Verma, Kavita, Pushpita Chatterjee, Jana Shafi, Jaeyoung Choi, and Muhammad Fazal Ijaz. A complete process of text classification system using state-of-the-art nlp models. *Computational Intelligence and Neuroscience*, 2022(1):1883698, 2022.

[48] Nihar Ranjan, Abhishek Gupta, Ishwari Dhumale, Payal Gogawale, and Rugved Gramopadhye. Full length review article. 2015.

[49] Klaus Hechenbichler and Klaus Schliep. Weighted k-nearest-neighbor techniques and ordinal classification. 2004.

[50] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer, 2006.

[51] AP Bradley, RPW Duin, P Paclik, and TCW Landgrebe. Precision-recall operating

characteristic (p-roc) curves in imprecise environments. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 123–127. IEEE, 2006.

[52] Kholoud Alsmearat, Mohammed Shehab, Mahmoud Al-Ayyoub, Riyad Al-Shalabi, and Ghassan Kanaan. Emotion analysis of arabic articles and its impact on identifying the author's gender. In *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, pages 1–6. IEEE, 2015.

[53] Pashutan Modaresi, Matthias Liebeck, and Stefan Conrad. Exploring the effects of cross-genre machine learning for author profiling in pan 2016. *Components of an Automatic Single Document Summarization System in the News Domain*, pages 100–107, 2017.

[54] A. Abuhammad and A. El-Halees. An approach for detecting spam in arabic opinion reviews. *International Arab Journal of Information Technology*, 12:9–16, January 2015. doi: 10.34028/iajit.

[55] Hamada A Nayel. A new approach for author profiling and identification of deception in texts. *Journal of Computational Linguistics*, 11(2):73–79, 2020.

[56] Hamdy Mubarak, Shammur Absar Chowdhury, and Firoj Alam. Arabgend: Gender analysis and inference on arabic twitter. *arXiv preprint arXiv:2203.00271*, 2022.

[57] Nikhil Ketkar and Eder Santana. *Deep learning with Python*, volume 1. Springer, 2017.

[58] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[59] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[60] Abdulrahman I Al-Ghadir and Aqil M Azmi. A study of arabic social media users—posting behavior and author's gender prediction. *Cognitive Computation*, 11:71–86, 2019.

[61] Takumi Takahashi, Takuji Tahara, Koki Nagatani, Yasuhide Miura, Tomoki Taniguchi, and Tomoko Ohkuma. Text and image synergy with feature cross technique for gender identification. *Working Notes Papers of the CLEF*, 10, 2018.

[62] Wenqi Wang, Run Wang, Lina Wang, Zhibo Wang, and Aoshuang Ye. Towards a robust deep neural network in texts: A survey. *arXiv preprint arXiv:1902.07285*, 2019.

[63] Sule Kaya and Bilal Alatas. A new hybrid lstm-rnn deep learning based racism, xenomy, and genderism detection model in online social network. *International Journal of Advanced Networking and Applications*, 14(2):5318–5328, 2022.

[64] Muhammad Adnan Ashraf, Rao Muhammad Adeel Nawab, and Feiping Nie. A study of deep learning methods for same-genre and cross-genre author profiling. *Journal of Intelligent & Fuzzy Systems*, 39(2):2353–2363, 2020.

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[66] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota, 2019.

[67] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

[68] Min Zhang and Juntao Li. A commentary of gpt-3 in mit technology review 2021. *Fundamental Research*, 1(6):831–833, 2021.

[69] Steve D Yang, Zulfikhar A Ali, and Bryan M Wong. Fluid-gpt (fast learning to understand and investigate dynamics with a generative pre-trained transformer): Efficient predictions of particle trajectories and erosion. *Industrial & Engineering Chemistry Research*, 62(37):15278–15289, 2023.

[70] C. Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[71] Mingye Wang, Pan Xie, Yao Du, and Xiaohui Hu. T5-based model for abstractive summarization: A semi-supervised learning approach with consistency loss functions. *Applied Sciences*, 13(12):7111, 2023.

[72] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35, 2023.

[73] Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. Openprompt: An open-source framework for prompt-learning, 2021. URL https://arxiv.org/abs/2111.01998.

[74] José Antonio García-Díaz, Ghassan Beydoun, and Rafel Valencia-García. Evaluating transformers and linguistic features integration for author profiling tasks in spanish. *Data & Knowledge Engineering*, 151:102307, 2024.

[75] Viktor From. Transfer learning for automatic author profiling with bert transformers and glove embeddings, 2022.

[76] Roberto López-Santillán, Luis C González, Manuel Montes-y Gómez, and A Pastor López-Monroy. When attention is not enough to unveil a text's author profile: Enhancing a transformer with a wide branch. *Neural Computing and Applications*, 35(13):9607–9626, 2023.

[77] Chanchal Suman, Anugunj Naman, Sriparna Saha, and Pushpak Bhattacharyya. A multimodal

author profiling system for tweets. *IEEE Transactions on Computational Social Systems*, 8(6): 1407–1416, 2021.

[78] Chiyu Zhang and Muhammad Abdul-Mageed. Bert-based arabic social media author profiling, 2019. URL https://arxiv.org/abs/1909.04181.

[79] Francisco Rangel, Paolo Rosso, Manuel Montes-y Gómez, Martin Potthast, and Benno Stein. Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working notes papers of the CLEF*, 192, 2018.

[80] Bashar Alhafni, Nizar Habash, and Houda Bouamor. The arabic parallel gender corpus 2.0: Extensions and analyses. *arXiv preprint arXiv:2110.09216*, 2021.

[81] Marc Owen Jones. The gulf information war| propaganda, fake news, and fake trends: The weaponization of twitter bots in the gulf crisis. *International journal of communication*, 13:27, 2019.

[82] Hind Almerekhi and Tamer Elsayed. Detecting automatically-generated arabic tweets. In *Information Retrieval Technology: 11th Asia Information Retrieval Societies Conference, AIRS 2015, Brisbane, QLD, Australia, December 2-4, 2015. Proceedings 11*, pages 123–134. Springer, 2015.

[83] Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. 2016.

[84] Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*, 2020.

Appendices

Word Embedding Visualization

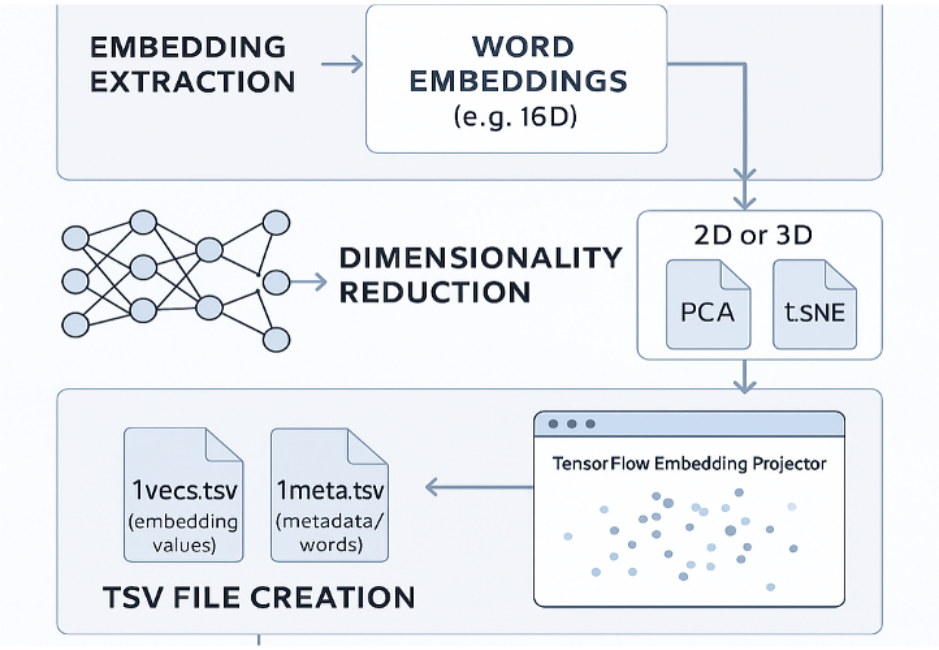

Figure 25: Word Embedding Visualization

To understand what our model learned, we **visualized word embeddings**, which are numerical representations of words capturing their semantic relationships.

We extracted 16-dimensional word vectors from our model's embedding layer (shape: `vocab_size` $\times$ 16).

Dimensionality Reduction: Since 16 dimensions are hard to visualize, we used techniques like t-SNE, PCA, and UMAP to reduce them to 2 or 3 dimensions.

Preparation: We created two TSV files:

- `vecs.tsv` – containing the reduced embedding values
- `meta.tsv` – containing the corresponding words

Using the **TensorFlow Embedding Projector**, we interactively explored these embeddings.

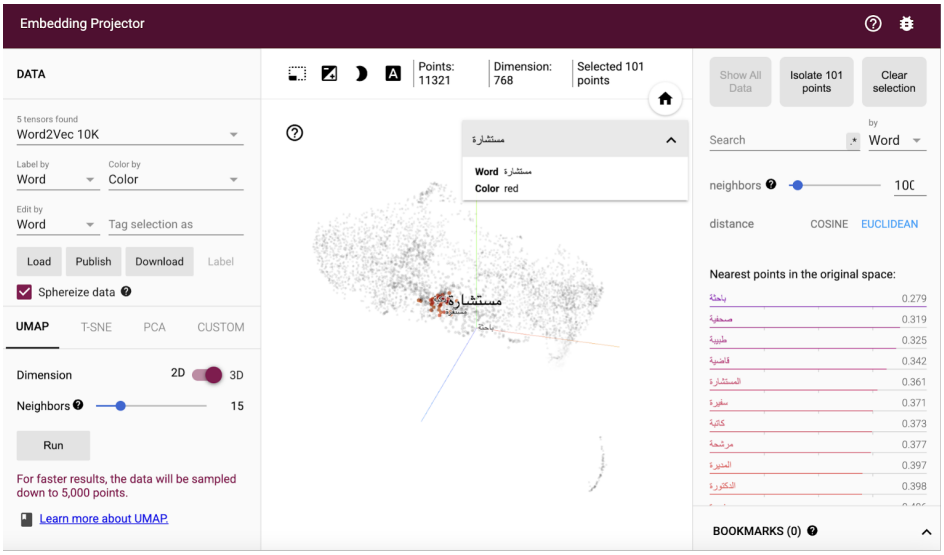

Figure 26: 2D UMAP projection of word embeddings from our trained ArabERT model

The next figure 26 displays a 2D UMAP projection of word embeddings from our trained ArabERT model. Each point represents a word, showcasing the model's ability to learn rich semantic relationships.

Semantic Proximity: Observe how "nearest neighbors" (e.g., مستشارة – female advisor) are grouped with semantically related terms, demonstrating the model's understanding of contextual meanings and associations.

Learned Distinctions: The visualization confirms the model's capacity to distinguish between words and place semantically similar words close together, proving it has learned meaningful representations from our dataset.