PEOPLES DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

UNIVERSITY OF DJILALI BOUNAAMA KHEMIS MILIANA

FACULTY OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

THESIS PRESENTED TO OBTAIN

THE MASTER DEGREE ON COMPUTER SCIENCE

Option :

< SOFTWARE ENGINEERING AND DISTRIBUTED SYSTEMS >

TITLE

Prophetic Hadith Knowledge Mining Using Large Language Models

Master student : Chahrazed BELKACEM

In front of a jury composed of :

President	Dr. D. HANICHE	University of Khemis Miliana
Examiner	Dr. D. BOUZIANE	University of Khemis Miliana
Supervisor	Dr. D. BAHLOUL	University of Khemis Miliana

Academic year : 2022/2023.

Acknowledgements

First and foremost, I express my sincere gratitude to Almighty Allah for granting me the strength, knowledge, ability, and patience to undertake this modest work. Without His blessings, this achievement would not have been possible.

I would like to extend my heartfelt thanks to my research promoter, Mr.Djamel BAHLOUL. I consider myself fortunate to have had him as my supervisor, who entrusted me with this research and provided invaluable guidance and support at every step of the process. His experience, constant orientation, invaluable advice, availability, and extreme kindness have been instrumental in the completion of this work.

I would also like to express my sincere appreciation to the members of the jury who have honored me by evaluating this thesis and examining my work. I am grateful for their valuable feedback and suggestions, which have enriched my research.

I am deeply grateful to my parents and sisters for their unwavering support and constant encouragement throughout my journey. Their love, belief in my abilities, and sacrifices have been the driving force behind my success.

Furthermore, I extend my deep thanks to all the teaching, administrative, and educational staff of the Mathematics and Computer Science Department at the University of Djilali BOUNAAMA in Khemis Miliana. Their dedication to providing a nurturing academic environment has played a significant role in shaping my education and research.

Dedication

I dedicate this work to the following individuals, who have played a significant role in shaping my journey and have been a constant source of love, support, and inspiration :

To my beloved father , your unwavering belief in me and your endless encouragement have been the driving force behind my pursuit of knowledge. Your sacrifices and guidance have shaped the person I am today. This achievement is a testament to your love and dedication.

to my mom ,You have been my rock, providing a safe haven where I can share my joys, sorrows, and ambitions. Your wise counsel, encouragement, and gentle words of wisdom have shaped my character and nurtured my passions.

To my sisters Nariman and Douaa your presence in my life has been a source of joy and strength. Your unwavering support, understanding, and sisterly bond have provided me with the motivation to overcome challenges and strive for excellence.

To my friends Amira , Aicha and Nesrine For their help and support in difficult times.

To my late late grandfather Tiab and my late uncle Zoubir

I dedicate this work to all the mentioned individuals, as well as to the countless others who have supported and encouraged me throughout my journey. Your belief in my abilities has fueled my determination and has made this accomplishment possible. Thank you for being a part of my life and for inspiring me to reach for the stars.

عن مُعَاوِيَةَ بن أبي سفيان رضي الله عنهما قال :

سَمِعْتُ النَّبِيَّ صَلَّى اللهُ عَلَيْهِ وَسَلَّمَ يَقُولُ :

(مَنْ يُرِدِ اللَّهُ بِهِ خَيْرًا يُفَقِّهْهُ فِي الدِّينِ).

رواه البخاري ومسلم

ملخص

يشير الحديث النبوي في الإسلام إلى كل ما ورد عن النبي محمد صلى الله عليه وسلم من الأقوال والأفعال والتقريرات و الصفات الخَلْقية و الخُلُقية . يُعتبر الحديث النبوي المصدر الثاني من مصادر التشريع بالنسبة للمسلمين في جميع أنحاء العالم بعد القرآن الكريم. يستخدم العديد من المسلمين البيانات الرقمية والإنترنت كمصادر معرفية رئيسية، بما في ذلك عند البحث عن النصوص الدينية مثل الحديثمع ظهور أساليب ونماذج تعلم عميقة جديدة، تم فتح فرص جديدة أمام الباحثين في مجال المعالجة الآلية للغة العربية الطبيعية. يتم تكريس معظم الأعمال لتحليل محتوى الأخبار ووسائل التواصل الاجتماعي إلا أن الأعمال المتعلقة بمعالجة النصوص الإسلامية وعلى وجه الخصوص الفهم الآلي للحديث لا تزال بعيدة عما تنتظره المجتمعات المسلمة.

يهدف المشروع المقترح إلى دراسة تأثير استخدام المناهج الجديدة في التعلم المعمق في استخراج المعارف من الأحاديث النبوية. يمكن اعتبار استخراج المعارف كفرع من فروع التنقيب الآلي في النصوص، حيث يقوم النظام بالتنقيب في محتوى حديث معين لاستخراج أو إنتاج مجموعة من الأفكار والفوائد. في إطار العمل في هذا المشروع ركزنا على استعمال النماذج اللغوية العربية الجديدة في استتخراج الأبواب الحديثية وفوائد الأحاديث النبوية.

الكلمات الدليلية : فقه الحديث، التنقيب الآلي في النصوص، الفهم الآلي للغة العربية، النماذج اللغوية العربية، التوليد الالي لنصوص

Résumé

Les Hadiths du Prophète dans l'islam font référence à tout ce qui a été rapporté sur le Prophète Muhammad صلى الله عليه وسلم incluantses paroles, ses actions, ses approbations et ses qualités personnelles et morales. . Les hadiths prophétiques sont considérés comme la deuxième source de référence pour les musulmans du monde entier, après le Saint Coran. De nombreux musulmans utilisent les données numériques et Internet comme principales sources de connaissance, notamment lorsqu'ils recherchent des textes religieux tels que les hadiths. Avec l'apparition de nouvelles méthodes et modèles d'apprentissage approfondi, de nouvelles opportunités se sont ouvertes aux chercheurs en traitement du langage naturel. La plupart des travaux sont consacrés à l'analyse de contenus d'actualités et de médias sociaux, cependant, les travaux sur le traitement des textes islamiques et en particulier la compréhension automatique des hadiths, sont encore loin de ce que la communauté musulmane attend.
Le projet proposé dans cette thèse vise à étudier l'impact de l'utilisation de nouvelles approches d'apprentissage approfondi et en particulier les modèles de l'IA générative afin d'extraire des connaissances des hadiths prophétiques. L'extraction de connaissances peut être considérée comme un problème d'exploration de textes, où à partir d'un hadith donné, le système explora le contenu pour générer un ensemble d'idées et de bénéfices. Dans ce travail, nous nous concentrons sur l'utilisation des nouveaux modèles de langage dédiés à l'arabe pour la génération des sections et des bénéfices à partir des Hadiths prophétiques.

Mots clés : Jurisprudence des Hadiths, Fouille de textes, Compréhension automatique de la langue arabe, Modèles de langage arabes, Génération de textes

---

Abstract

The Prophet's Hadith in Islam refers to everything that has been reported about the Prophet Muhammad (صلى الله عليه وسلم), including his sayings, actions, approvals, and personal and moral qualities. Prophetic Hadith It is considered as the second source of reference for Muslim all over the world after the Holy Al-Quran. Plenty of Muslims used digital data and internet as primary knowledge sources including when searching for religious texts such as Hadith. With the appearance of new deep learning methods and models, new opportunities have been opened to NLP researchers. Most of works are dedicated to news and social media content analysis, however, works on Islamic text processing and Hadith understanding in particular are still far from what the Muslim community is waiting for.

The proposed project aims to study the impact of using new deep learning approaches and especially generative AI models in order to extract and generate knowledge from Prophetic Hadith. Knowledge extraction can be considered as a text mining problem and where from an input Hadith the system will mine the content in order to generate a set of ideas and benefits. In this work, we focus on generating Hadith sections and benefits using Arabic large language models (LLMs)

Keywords : Hadith jurisprudence, Text mining, Arabic natural language understanding, Arabic language models, Text generation

---

Table of Contents

Table of Contents 7

List of Figures 2

List of Tables 4

List of Abbreviations 5

Introduction 6

1 Hadith jurisprudence science (Fiqh Al-Hadith) 2

1.1 Introduction . . . 2
1.2 Definition of Hadith . . . 2
1.3 Components of Hadith . . . 3
1.3.1 Sanad . . . 3
1.3.2 Matn . . . 3
1.3.3 Taraf . . . 4
1.3.4 Takhridj . . . 4
1.4 Hadith Science . . . 4
1.4.1 Hadith Science riwayah . . . 4
1.4.2 Hadith Science Dirayah . . . 5
1.5 Classification of hadith . . . 5
1.5.1 According to the reference to a particular authority . . . 5
1.5.2 According to the links of isnad - interrupted or uninterrupted . . . 5
1.5.3 According to the number of reporters involved in each stage of isnad . . . 6
1.5.4 According to the nature of the text and isnad . . . 6
1.5.5 According to the reliability and memory of the reporters . . . 6
1.6 Definition of jurisprudence . . . 7
1.7 Definition of hadith jurisprudence (Fiqh al-Hadith) : . . . 7

1.8 The importance of hadith jurisprudence . . . 8
1.9 Benefits of Hadith Jurisprudence . . . 9
1.10 Conclusion . . . 9

2 Text mining approaches 11
2.1 Introduction . . . 11
2.2 Definition . . . 11
2.3 Text mining Process . . . 12
2.3.1 Text Preprocessing : . . . 12
2.3.2 Text Transformation / Feature Generation : . . . 12
2.3.3 Feature Selection / Attribute Selection : . . . 13
2.3.4 Text mining methods : . . . 13
2.3.5 Interpretation or Evaluation : . . . 13
2.4 Data Mining vs Text Mining . . . 13
2.5 Natural Language Processing . . . 13
2.5.1 Data preprocessing : . . . 14
2.5.2 Feature extraction : . . . 14
2.6 Applications of Text Mining : . . . 14
2.6.1 Sentiment analysis : . . . 15
2.6.2 Customer relationship management : . . . 15
2.6.3 Competitive intelligence : . . . 15
2.6.4 Human resources management : . . . 15
2.6.5 Biomedical research : . . . 15
2.6.6 Fraud detection : . . . 15
2.6.7 Security applications : . . . 15
2.6.8 Company resource planning : . . . 15
2.7 Text mining techniques . . . 17
2.7.1 Text categorization/classification : . . . 17
2.7.2 Clustering : . . . 17
2.7.3 Text summarization : . . . 18
2.7.4 Sentiment analysis : . . . 18
2.7.5 Topic modeling : . . . 18
2.7.6 Named entity recognition : . . . 18
2.7.7 Text mining with deep learning : . . . 18
2.7.8 Information Retrieval (IR) : . . . 18
2.7.9 Information Extraction (IE) : . . . 18
2.7.10 Text generation : . . . 19

2.8 Neural Text Generation . . . 19
2.8.1 Neural Network . . . 19
2.8.2 Recurrent Neural Network (RNN) . . . 20
2.8.2.1 Types of RNN : . . . 20
2.8.2.2 Applications of Recurrent Neural Network : . . . 21
2.8.3 Transformer Models : . . . 21
2.8.3.1 The Transformer Architecture : . . . 21
2.8.4 Large Language Models : . . . 23
2.9 Arabic text mining . . . 24
2.9.1 Arabic Language : . . . 24
2.9.1.1 Arabic Language Peculiarity : . . . 25
2.10 Arabic text mining related works . . . 25
2.10.1 Arabic Opinion Mining : . . . 26
2.10.2 Arabic Web Documents Mining : . . . 27
2.10.3 Holy Quran Mining : . . . 28
2.10.4 Prophetic Hadith Mining : . . . 29
2.11 Conclusion . . . 31

3 Contribution **32**
3.1 Introduction . . . 32
3.2 Datasets Description . . . 32
3.3 Transformer model . . . 34
3.4 Section Abstraction . . . 36
3.4.1 mbert2mbert Model : . . . 36
3.4.2 araT5 Model : . . . 36
3.4.3 Processing steps : . . . 37
3.4.3.1 Preprocessing : . . . 37
3.4.3.2 Data Splitting : . . . 38
3.4.3.3 Tokenization : . . . 38
3.4.3.4 Model Embedding : . . . 39
3.4.3.5 Model fine-tuning and training : . . . 40
3.4.3.6 Section Prediction : . . . 40
3.5 Benefits Extraction : . . . 41
3.5.1 Beam search : . . . 41
3.5.2 Processing Steps : . . . 42
3.5.2.1 Preprocessing : . . . 43
3.5.2.2 Data Splitting : . . . 43

3.5.2.3 Model training : . . . 43
3.5.2.4 Benefits Prediction : . . . 43
3.5.3 Text Completion : . . . 45
3.5.3.1 araGPT2 Model . . . 45
3.5.4 Processing Steps : . . . 45
3.5.4.1 Preprocessing : . . . 45
3.5.4.2 Data splitting : . . . 45
3.5.4.3 Model training : . . . 46
3.5.4.4 Benefits Prediction : . . . 46
3.6 Conclusion : . . . 50

4 Implementation and Evaluation 51
4.1 Introduction . . . 51
4.2 Software Configuration for implementation . . . 51
4.2.1 Kaggle platform . . . 51
4.2.2 google colab pro . . . 51
4.2.3 Language and libraries . . . 52
4.3 Experimental Results . . . 52
4.4 Section Abstraction Results . . . 53
4.4.1 AraT5 Training results . . . 54
4.4.1.1 Training and validation loss . . . 54
4.4.1.2 AraT5 Training Rouge score . . . 55
4.4.2 mbert2mbert Training results . . . 55
4.4.2.1 Training and validation loss : . . . 55
4.4.2.2 mbert2mbert Training ROUGE SCORE . . . 56
4.4.3 Section abstraction Test results . . . 57
4.5 Benefits Extraction Results . . . 58
4.5.1 Beam search . . . 58
4.5.1.1 AraT5 Training results . . . 65
4.5.1.2 mbert2mbert Training results . . . 67
4.5.1.3 Beam search test results . . . 69
4.5.2 Text Completion . . . 76
4.5.2.1 AraGPT2 Training results . . . 76
4.5.2.2 AraGPT2 Test results . . . 77
4.6 Some related works results . . . 78
4.7 Conclusion . . . 80

Conclusion **81**

Bibliography **82**

Appendix **86**

List of Figures

1.1 Components of Hadith [44] . . . 3
1.2 Exemple of Sanad and Matn . . . 4
1.3 classification of hadith . . . 10

2.1 General text mining process flow[46] . . . 12
2.2 Inter-relationship among different text mining techniques and their core functionalities [57] . . . 17
2.3 Recurrent Neural Network Schema[62] . . . 20
2.4 The Transformer - model architecture [60] . . . 22

3.1 Encoder-Decoder structure of Transformer . . . 35
3.2 Representation of Section abstraction using araT5 mbert2mbert model . . . 36
3.3 AraT5 encoder-decoder model and prompt samples from four investigated tasks, namely : title generation, machine translation, question generation, and paraphrasing [30] . . . 37
3.4 Section abstraction Process using Arabic T5 and mbert2mbert Models . . . 38
3.5 Splitting Data for Machine Learning . . . 39
3.6 Representation of Benefits extraction using araT5 and mbert2mbert model . . . 42
3.7 Representation Beam search technique . . . 42
3.8 Representation of Benefits extraction using GPT2 model . . . 45
3.9 Different variants of araGPT2 model . . . 46

4.1 Training and validation loss of Arat5 . . . 54
4.2 Training Rouge score metric of araT5 . . . 55
4.3 training and validation loss of mbert2mbert . . . 56
4.4 Training Rouge Score metric of mbert2mbert . . . 57
4.5 ROUGE scores of araT5 and mbert2mbert model . . . 58
4.6 Training and validation loss of AraT5 . . . 66
4.7 Training rouge score of AraT5 . . . 67
4.8 Training and validation loss of mbert2mbert . . . 68

4.9 Training rouge score of mbert2mbert . . . 69
4.10 ROUGE Scores for Permutation 1 . . . 70
4.11 ROUGE Scores for Permutation 2 . . . 71
4.12 ROUGE Scores for Permutation 3 . . . 72
4.13 ROUGE Scores for Permutation 4 . . . 73
4.14 ROUGE Scores for Permutation 5 . . . 74
4.15 ROUGE Scores for Permutation 6 . . . 75
4.16 ROUGE Scores for concatenation technique . . . 76
4.17 Training and validation loss of AraGPT2 . . . 77
4.18 Training and validation loss of AraGPT2 . . . 78
0.19 Validation Loss . . . 86
0.20 Validation Steps per Second . . . 86
0.21 validation samples per second . . . 86
0.22 Training loss . . . 87
0.23 Training epochs . . . 87
0.24 Training learning rate . . . 87
0.25 Training learning rate (all sweeps) . . . 87
0.26 train global steps . . . 87
0.27 Training Steps per second . . . 87
0.28 GPU power usage (W) . . . 88
0.29 GPU power usage (%) . . . 88
0.30 GPU Time Spent Accessing Memory (%) . . . 88
0.31 GPU Memory Allocated (%) . . . 88
0.32 GPU Utilization (%) . . . 88
0.33 GPU Temp (C) . . . 88

List of Tables

2.1 Research studies conducted on Arabic text mining in terms of sentiment analysis [49] 26
2.2 Research studies conducted on Arabic text mining in terms of web documents [49] 28
2.3 Research studies conducted on Arabic text mining in terms of the Holy Quran[49] 29
2.4 Research studies on Arabic Hadith analysis 31

3.1 Our first Dataset 33
3.2 Basic parameters of model training 40
3.3 Research studies conducted on Arabic text mining in terms of the Holy Quran[49] 41
3.4 Representation of predicted benefits with mbert2mbert and AraT5 model 44
3.5 Basic parameters defined for Aragpt2 model 46
3.6 Representation of benefits prediction with araGPT2 model 47
3.7 Representation of benefits prediction with araGPT2 model ,excluding the inclusion of hadiths 49

4.1 An example of how to calculate ROUGE metric 53
4.2 All possible permutations of predicted benefits 59
4.3 Concatenation of three benefits 64
4.4 Comparison among different techniques for Arabic summarization 79

List of Abbreviations

NLP	*Natural Language Processing*
ANN	*Artificial Neural Networks*
ML	*Machine learning*
ROUGE	*Recall Oriented Understanding for Gisting Evaluation*
RNN	*Recurrent Neural Networks*
AI	*Artificial Intelligence*
NN	*Neural Network*
LSTM	*Long Short-Term Memory*
LLM	*Large Language Model*
TF	*Term Frequency*
IDF	*Inverted Document Frequency*
BoW	*Bag of Words*
GPU	*Graphics Processing Unit*
BLEU	*Bilingual Evaluation Understudy*

Introduction

Research background and motivation

The field of Islamic studies encompasses a wide range of disciplines, including the study of the Quran, Hadith, jurisprudence, and Islamic ethics. Within this rich tradition, Hadith literature holds a special place as a primary source of guidance for Muslims. The Hadith, which consists of the sayings, actions, and approvals of Prophet Muhammad (peace be upon him), provides valuable insights into the practical implementation of Islamic teachings. However, the sheer volume and complexity of Hadith texts pose significant challenges in extracting relevant information and deriving meaningful insights. This has motivated researchers and scholars to explore computational methods and natural language processing techniques to facilitate the processing and understanding of Hadith. .

Context of study and problematic

The context of this study revolves around the need to enhance the process of extracting knowledge from Prophetic Hadith. Traditional methods of Hadith analysis and interpretation rely on human expertise and manual examination of texts, which can be time-consuming and prone to subjectivity. With advancements in computational techniques and natural language processing, there is an opportunity to develop automated methods that can assist scholars, researchers, and the broader Muslim community in accessing, analyzing, and applying Hadith teachings more efficiently and accurately. The problematic at hand is to explore how deep learning approaches can be employed to tackle the challenges of Hadith processing and facilitate a deeper understanding of Islam text,and especially the newly generated artificial intelligence models.

Contribution

The primary contribution of this project is to study the impact of utilizing new deep learning approaches in the field of digital Hadith processing. By leveraging advanced computational methods and natural language processing techniques, the project aims to extract knowledge and insights from Prophetic Hadith. The focus will be on improving models that can perform tasks such as section abstraction and benefit extraction

from Hadith texts. The outcomes of this research endeavor have the potential to enhance the quality and efficiency of Hadith analysis, enabling scholars, researchers, and the Muslim community to access and apply the teachings of Prophet Muhammad (peace be upon him) in a more comprehensive and accurate manner. Our contribution can be summarized in the following main points :

— Utilizing Arabic language models such as AraT5, mbert2mbert, and AraGpt2.

— Conducting section abstraction and benefit extraction from Hadith texts.

— Employing the ROUGE metric for evaluation purposes.

— Propose a new Prophetic hadith benefits dataset

Thesis organization

This thesis is structured as follows :

— **Chapter 1 :** This chapter explores the field of Hadith jurisprudence

— **Chapter 2 :** It delves into various text mining approaches and text generation

— **Chapter 3 :** The third chapter describes the proposed approach, including the process of dataset formation, system processing tasks, and the different parameters of the proposed models.

— **Chapter 4 :** This chapter focuses on the implementation of the proposed solution, along with an evaluation section.

Chapitre 1

Hadith jurisprudence science (Fiqh Al-Hadith)

1.1 Introduction

Without a doubt, working with the Prophet's Sunnah and studying its chain of transmission, its texts and jurisprudence is one of the greatest and most important things Muslims do and scholars turn to.
As we all know, hadith jurisprudence is one of the hadith sciences whose subject matter is related to the jurisprudence that studies hadith texts, and it is undoubtedly one of the most important sciences in the service of the Qur'an and hadith.
By this, we can know the rulings of the hadith and the rulings derived from the hadith, because the sunnah is the track of most jurisprudential rulings. In what follows, we will provide an overview of the science of Hadith jurisprudence and its significance.

1.2 Definition of Hadith

The term hadith (the plural is ahadith) is defined as follow : "Everything attributed to Prophet Muhammad ﷺ,whether in the form of speech, deed, judgment or nature [32]
Hadiths work to support and strengthen the Holy Quran and to provide interpretations and details of what is written in the Holy Quran.[13]

1.3 Components of Hadith

Each hadith consists of three parts as shown in the diagram. Figure 2[44].

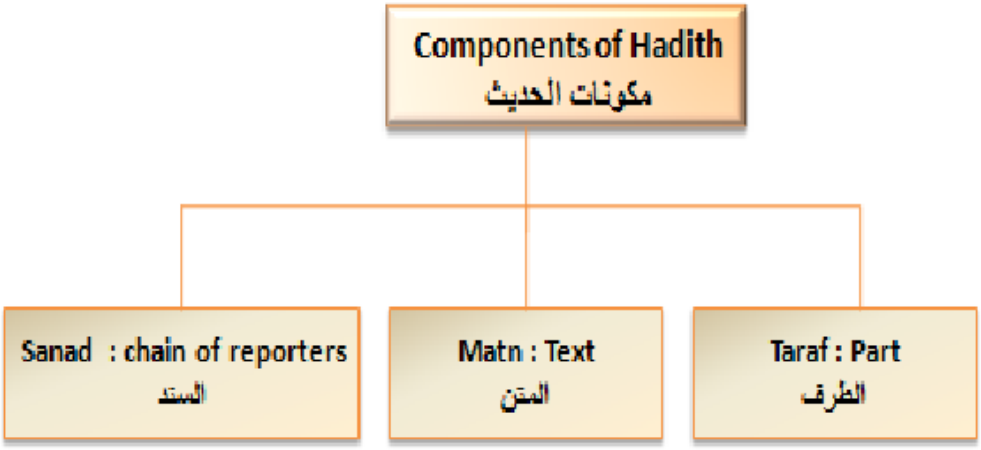

FIGURE 1.1 – Components of Hadith [44]

1.3.1 Sanad

The first component is the Narrator (Reporter) chain.Also known as Silsila Isnad or Sanad السَّنَد ,This chain contains the start narrator or "originator"and the final narrator of the hadith. Between the start and the final narrator, there are any numbers of transmitters (narrators) who have passed on the hadith orally from one to the other[11][44]

1.3.2 Matn

The second essential component is the text known as the 'matn' المَتْنُ and is carried from the originator[11]

FIGURE 1.2 – Exemple of Sanad and Matn

1.3.3 Taraf

The third part is called taraf (الطَّرَفُ) the part, or the beginning sentence, of the text which refers to the sayings, actions or characteristics of the Prophet[11]

1.3.4 Takhridj

Hadith Takhridj (التخريج) refers to attributing the hadith to the respected imams and scholars of hadith, and talking about it after examining its condition and the men of its director. The science of hadith graduation is one of the sciences of the Prophets Sunnah in Islam.

1.4 Hadith Science

The science of hadith (عِلْمُ الحَدِيْث) is a field of study within Islamic scholarship that is dedicated to investigating and verifying the authenticity of existing hadiths. This field includes the sayings, actions, characteristics, stories, and origins of the Prophet Muhammad (peace be upon him) and his companions.
The hadith science is divided into two categories :

1.4.1 Hadith Science riwayah

This category deals with the Matn's meanings and with explaining it. The subject of this science is the words of the Prophet (صلى الله عليه وسلم), his deeds, his approval, or its natural or ethical explanation.

1.4.2 Hadith Science Dirayah

Derayah درايـة is a field of study in Islamic sciences that focuses on the principles and methods of verifying the authenticity and reliability of Hadith (narrations of the Prophet Muhammad). It involves the analysis of both the chain of narrators (Sanad سنـد) and the content of the narration (Matn متـن).
The study of derayah includes the principles for evaluating the reliability of the narrators and their character traits, as well as the procedures for receiving and transmitting Hadith. Scholars in this field also examine the historical context in which the narrations were transmitted, as well as any potential biases or motivations of the narrators.
Through these methods, derayah aims to distinguish authentic Hadith from those that are fabricated or unreliable, providing a basis for the development of Islamic jurisprudence and ethical conduct.

1.5 Classification of hadith

A number of classifications of hadith have been made. Five of these classifications are shown in the figure [below], and are briefly described subsequently.[1]

1.5.1 According to the reference to a particular authority

Four types of hadith can be identified.

— **Qudsi - Divine :** a revelation from Allah ; relayed with the words of the Prophet.

— **Marfu‘ - elevated :** a narration from the Prophet, e.g., I heard the Prophet saying...

— **Mauquf- stopped :** a narration from a companion only, e.g., we were commanded to...

— **Maqtu‘ - severed :** a narration from a successor.

1.5.2 According to the links of isnad - interrupted or uninterrupted

Six categories can be identified :

— **Musnad - supported :** a hadith which is reported by a traditionalist, based on what he learned from his teacher at a time of life suitable for learning ; similarly - in turn - for each teacher until the isnad reaches a well known companion, who in turn, reports from the Prophet

— **Muttasil - continuous :** a hadith with an uninterrupted isnad which goes back only to a companion or successor.

— **Mursal - hurried :** if the link between the successor and the Prophet(P) is missing, e.g., when a successor says "The Prophet said...".

— **Munqati‘ - broken :** is a hadith whose link anywhere before the successor (i.e., closer to the traditionalist recording the hadith) is missing.

— **Mu'adal - perplexing :** is a hadith whose reporter omits two or more consecutive reporters in the isnad.

— **Mu'allaq - hanging :** : is a hadith whose reporter omits the whole isnad and quotes the Prophet directly (i.e., the link is missing at the beginning).

1.5.3 According to the number of reporters involved in each stage of isnad

Five categories of hadith can be identified :

— **Mutawatir - Consecutive :** is a hadith which is reported by such a large number of people that they cannot be expected to agree upon a lie, all of them together.

— **Ahad - isolated :** is a hadith which is narrated by people whose number does not reach that of the mutawatir.
It is further classified into :

— **Mash'hur - famous :** hadith reported by more than two reporters.

— **Aziz - rare, strong :** at any stage in the isnad, only two reporters are found to narrate the hadith.

— **Gharib - strange :** At some stage of the isnad, only one reporter is found relating it.

1.5.4 According to the nature of the text and isnad

— **Munkar - denounced :** is a hadith which is reported by a weak narrator, and whose narration goes against another authentic hadith.

— **Mudraj - interpolated :** an addition by a reporter to the text of the hadith being narrated.

1.5.5 According to the reliability and memory of the reporters

This provides the final verdict on a hadith - four categories can be identifiedSahih - sound. Imam al-Shafi'i states the following requiremetts for a hadith, which is not Mutawatir, to be acceptable "each reporter should be trustworthy in his religion ; he should be known to be truthtul in his narrating, to understand what he narrates, to know how a different expression can alter the meaning, and to report the wording of the hadith verbatim, not only its meaning".

— **Hasan - good :** is the one where its source is known and its reporters are unambiguous.

— **Da'if - weak :** a hadith which fails to reach the status of Hasan. Usually, the weakness is : a) one of discontinuity in the isnad, in which case the hadith could be - according to the nature of the discontinuity - Munqati (broken), Mu'allaq (hanging), Mu'adal (perplexing), or Mursal (hurried), or b) one of the reporters having a disparaged character, such as due to his telling lies, excessive mistakes, opposition to the narration of more reliable sources, involvement in innovation, or ambiguity surrounding his person.

— **Maudu‘ - fabricated or forged :** is a hadith whose text goes against the established norms of the Prophet’s sayings, or its reporters include a liar. Fabricated hadith are also recognized by external evidence related to a discrepancy found in the dates or times of a particular incident.

1.6 Definition of jurisprudence

Jurisprudence, or Fiqh, is a term that has both linguistic and technical meanings. Linguistically, it refers to the knowledge, understanding, and perception of a particular subject matter [9].
Ibn Faris said :

الفاء والقاف والهاء أصل واحد صحيح يدل على إدراك الشيء والعلم به، تقول : فقهتُ الحديث أفقهه، وكل علم بشيء فهو فقه...

"Faa, Qaf and distraction are one true origin, indicating the awareness of the thing and knowledge of it, it says : the jurisprudence of hadith is its horizon, and every knowledge of something is jurisprudence"[19]
Ibn Manzur said :

ألفقه : العلم بالشيء والفهم له، وغلب على علم الدين؛ لسيادته وشرفه وفضله على سائر أنواع العلم.

"Fiqh : knowledge of something and understanding of it, and prevailed over the science of religion, because of its sovereignty, honor and virtue over other types of knowledge" [21].
Technically, "Fiqh" means "knowledge of practical Islamic rulings based on their detailed evidence."[20].

Therefore, jurisprudence is the study and application of Islamic law, which involves understanding the sources of Islamic law, interpreting them, and applying them to practical situations in daily life.

1.7 Definition of hadith jurisprudence (Fiqh al-Hadith) :

It is the extraction of meanings, judgments and rulings from the hadiths of the Prophet [10].

Judge Ayyyad said [17] :

ثم التفقه فيه، [وهو] استخراج الحكم والأحكام من نصوصه ومعانيه، وجلاء مشكل ألفاظه على أحسن تأويلها، ووِفق مختلفها على الوجوه المفصلة تنزيلها

Tibi said :[35]

هو ما تضمنه متن الحديث من الأحكام والآداب المستنبطة

1.8 The importance of hadith jurisprudence

Imam Ali bin Al-Madani said :

التفقه في معاني الحديث نصف العلم ومعرفة الرجال نصف العلم

The Emir of the Muhaddiths and the master of jurists, Imam al-Bukhari (may Allah have mercy on him), said : "... You have to follow the jurisprudence that you can learn while you are in your home stationary, you do not need to travel and step home, and ride the seas, which is with the fruit of the hadith. The reward of the jurist is not without the reward of the modernizer in the hereafter, nor his glory is less than the glory of the modernist"[16].

... فعليك بالفقه الذي يُمكنك تعلمه وأنت في بيتك قارٌّ ساكن، لا تحتاج إلى بعد الأسفار ووطء الديار، وركوب البحار، وهو مع ذا ثمرة الحديث. وليس ثواب الفقيه بدون ثواب المحدث في الآخرة، ولا عزّه بأقل من عِزّ المحدّث

Al-Hakim Al-Nisaburi warned that the jurisprudence of hadith is one of the sciences branching from the science of hadith, he said : "From the science of hadith knowledge of the jurisprudence of hadith, as it is the fruit of these sciences, and it has the strength of the Sharia, as for the jurists of Islam, the owners of measurement, opinion, deduction, controversy and consideration, they are known in every era and the people of each country, and we remember the will of God in this place the jurisprudence of hadith, about its people to indicate that the people of this industry who sail in it are not ignorant of the jurisprudence of hadith, It is a kind of this science."[58]

مِنْ علم الْحَدِيْث مَعْرِفَة فقه الْحَدِيْث، إِذْ هُوَ ثمرة هَذِهِ العلوم، وبه قوام الشريعة، فأما فقهاء الإسلام أصحاب القياس والرأي والاستنباط والجدل والنظر فمعروفون في كُلّ عصر وأهل كُلّ بلد، ونحن ذاكرون بمشيئة الله في هَذَا ا لموضع فقه الْحَدِيْث، عَنْ أهله ليستدل بِذَلِكَ عَلَى أن أهل هَذِهِ الصنعة من تبحر فِيْهَا لا يجهل فقه الْحَدِيْث، إِذْ هُوَ نوع من أنواع هَذَا العلم

Al-Khattabi said in the introduction to his book "Milestones of the Sunan" : "After knowing the authenticity of the hadith, we must work to understand it, as it is the fruit of this knowledge. The foundation without the building is a ruined house."

بعد معرفة صحة الحديث؛ يجب الاشتغال بفهمه، إذ هو ثمرة هذا العلم. فإن الأساس بدون البناء بيت خرب

Ibn Hajar said : "The truth is that both of them (meaning knowledge of the jurisprudence of hadith and knowledge of its validity and weakness) in the science of hadith is important and does not preponder one over the other, yes; if he said : working with the first art (meaning the jurisprudence of hadith) is more important was a Muslim with what is in it, and there is no doubt that whoever collected it won the mug.[35]

فالحق أن كلاً منهما (يعني معرفة فقه الحديث ومعرفة صحته وضعفه) في علم الحديث مهم لا رجحان لأحدهما على

الآخر، نعم؛ لو قال : الاشتغال بالفن الأول (يعني فقه الحديث) أهم كان مسلماً مع ما فيه، ولا شك أن من جمعها حاز القدح المعلى، ومن أخل بهما فلا حظ له في اسم المحدث

1.9 Benefits of Hadith Jurisprudence

Hadith Jurisprudence provides numerous benefits for those studying and practicing Islamic law. It allows one to understand the formulation and maintenance of hadith narrations, the differences between narrators in the wording of hadiths, and the control of the novel's words. This knowledge also enables one to grasp the reasons for the content of hadiths and the corresponding jurisprudential rulings. Additionally, it allows for the effective control and issuance of fatwas and rulings, the ability to distinguish between commands and prohibitions, and to deduce and consider the correct inferences. Hadith Jurisprudence also encompasses the use of scientific terminology employed by Sharia scholars and the ability to balance legal rulings in cases of apparent conflict. Finally, it enhances one's abilities to persuade, communicate with eloquence, and make effective statements in legal and religious contexts.

1.10 Conclusion

In conclusion, the science of Hadith jurisprudence plays a crucial role in the understanding and application of Islamic law. Its importance lies in the fact that it allows us to derive rulings from the hadith and understand the jurisprudence of the texts. With the help of hadith jurisprudence, we can understand the practical application of the Prophet's Sunnah and follow his guidance in our daily lives. Therefore, it is imperative for scholars and students of Islamic studies to have a strong understanding of hadith jurisprudence in order to derive accurate and authentic rulings from the hadith.In what follows, we will focus on Text mining and text generation

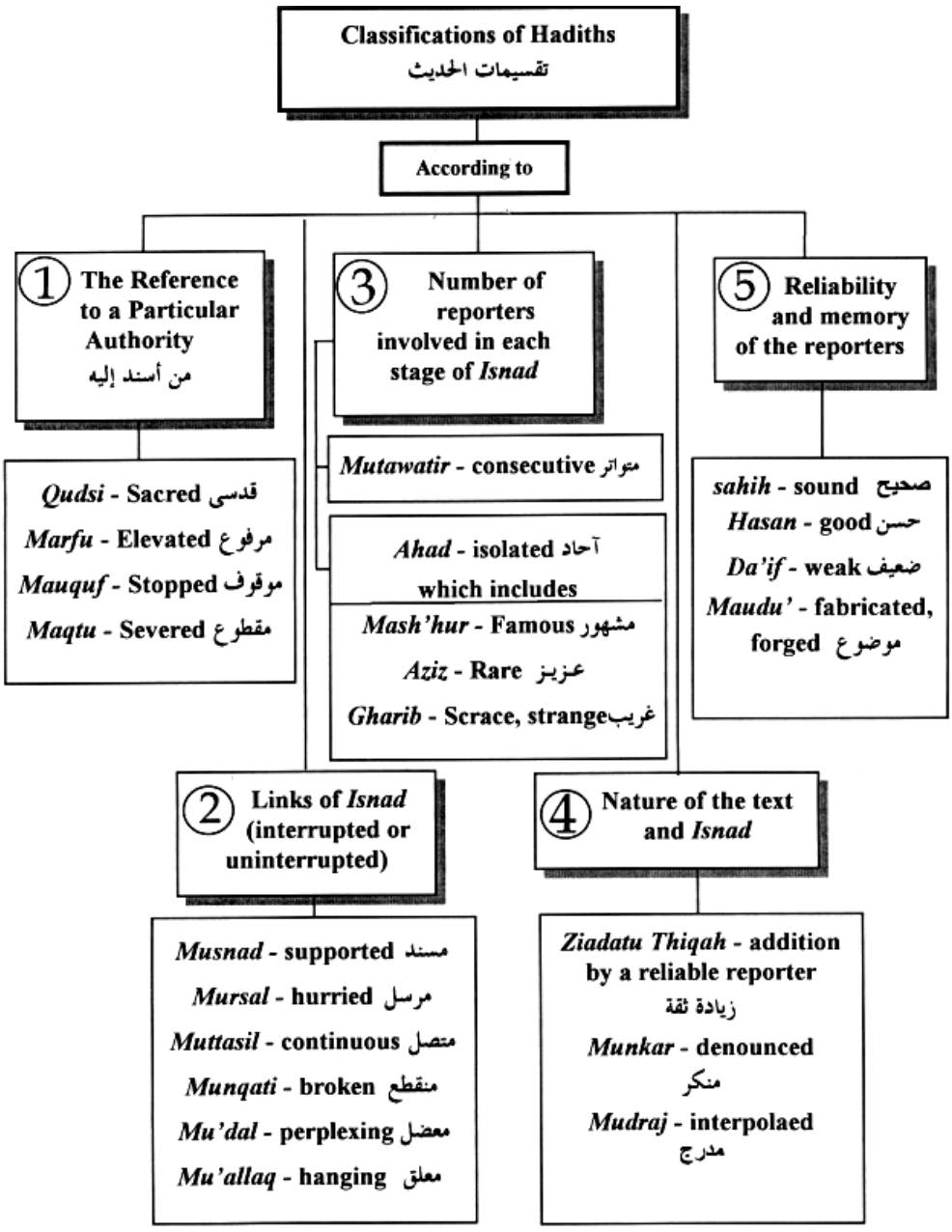

FIGURE 1.3 – classification of hadith

Chapitre 2

Text mining approaches

2.1 Introduction

The field of text mining has gained significant attention in recent years due to the abundance of digital data in the form of text. Text mining aims to extract valuable and relevant information from unstructured text data, which can then be used to make informed decisions. This chapter provides an overview of text mining, covering its definition, process, techniques, and applications. Furthermore, the chapter delves into the methods and models employed in text mining, exploring the distinctions between data mining and text mining, as well as the relationship between NLP and text mining. The chapter also discusses text generation, Recurrent Neural Networks, Transformers, and Pre-trained Language Models (PLMs). Finally, the chapter concludes with an examination of the complexities and challenges of Arabic text mining.

2.2 Definition

Text mining is a field that attempts to glean meaningful information from natural language texts. It is used to discover new, previously unknown information "not found in any individual document" by automatically extracting information from different written resources [3]. It consists of the collection, organization, and analysis of large volumes of documents in order to provide a targeted type of information to analysts and decision makers and to discover relationships between relevant facts across a wide range of research disciplines.

Text mining can also be defined as the discovery or creation of new knowledge from a collection of documents. The new knowledge may be the statistical discovery of new patterns in available data (standard text mining). It may also incorporate AI abilities to interpret patterns and provide more advanced capabilities such as hypothesis suggestion (intelligent text mining). Artificial intelligence, and especially natural language processing, can be used to simulate the human capabilities needed for intelligent text mining [3].

The key concepts emphasized in the various definitions of text mining include the extraction of valuable insights and knowledge from unstructured text data, the use of computational and statistical techniques, and the application of natural language processing methods. Additionally, text mining can involve transforming unstructured data into structured data for easy analysis and interpretation. Another important aspect of text mining is the incorporation of artificial intelligence and machine learning algorithms to enable intelligent text mining, which goes beyond statistical discovery to suggest hypotheses and provide more advanced capabilities[46]. Overall, text mining is a powerful tool for discovering hidden insights and patterns in large volumes of unstructured text data, with the potential to create new knowledge and generate actionable insights.

2.3 Text mining Process

Mining a text should go through certain stages in order to extract valuable insights and knowledge. The steps involved in the overall process of text mining are depicted in Figure 2.1 [46].

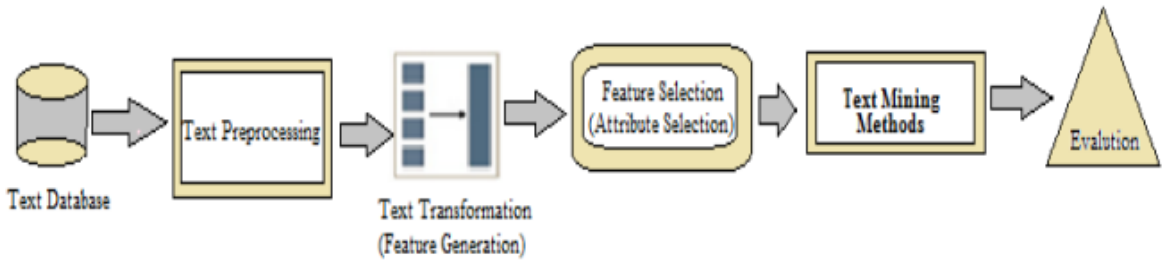

FIGURE 2.1 – General text mining process flow[46]

2.3.1 Text Preprocessing :

Text preprocessing is a critical step in text mining, involving several sub-steps.

Tokenization is the first step in which the text document is divided into individual words by removing spaces, commas, and other delimiters. The next step involves the removal of stop words like :["in" "على" , "on" "في"],and other commonly used words that do not add significant meaning to the text. HTML and XML tags are also removed from web pages in this step. Finally, stemming techniques are used to find the root or stem of a word. A stemmer is applied to convert words to their stems,

2.3.2 Text Transformation / Feature Generation :

Text transformation means in most of cases the conversion of the text document into bag of words or Vector space document model notation, which can be used for further effective analysis task.

2.3.3 Feature Selection / Attribute Selection :

This phase mainly performs removing features that are considered irrelevant for mining purpose. This procedure gives advantage of smaller dataset size, less computations and minimum search space required.

2.3.4 Text mining methods :

There are different text mining methods as in Data mining had been proposed such as : Clustering, Classification, Information retrieval, Topic discovery, Summarization, Topic extraction.

2.3.5 Interpretation or Evaluation :

This phase includes Evaluation and Interpretation of results in terms of calculating Precision and Recall, Accuracy, F measure etc .

2.4 Data Mining vs Text Mining

Text mining and data mining are both about finding patterns, but they differ in the type of data they analyze and the nature of the patterns they seek to uncover.
Data mining deals with implicit information that is hidden within data, while text mining deals with explicit information that is stated in text but not in a form that is easily processed by machines.
Text mining aims to extract this information and present it in a way that can be easily understood and used by computers.

2.5 Natural Language Processing

Text mining and natural language processing (NLP) are closely related, but they are not the same thing. Text mining is a broader term that includes NLP, as well as other techniques for analyzing unstructured text data.
NLP is a subset of text mining that focuses specifically on using computational techniques to understand human language. NLP involves using algorithms and statistical models to analyze text data, with the goal of understanding the meaning and structure of human language.
This can involve tasks such as part-of-speech tagging, named entity recognition, and sentiment analysis. NLP techniques often involve deep linguistic analysis and modeling of grammar, syntax, and semantics. In contrast, text mining typically focuses on extracting insights and patterns from large volumes of unstructured text data, often with the goal of discovering new knowledge or trends.
Text mining techniques may include NLP, but they may also involve other techniques such as machine learning, data visualization, and network analysis.

2.5.1 Data preprocessing :

Before a model processes text for a specific task, the text often needs to be preprocessed to improve model performance or to turn words and characters into a format the model can understand.
Various techniques may be used in this data preprocessing :

- **Stemming and lemmatization :** Stemming and lemmatization are techniques used in text preprocessing to reduce words to their base or root form. The goal is to normalize the text and reduce variations of words that have the same meaning.
- **Sentence segmentation :** Sentence segmentation involves splitting a paragraph or a larger text into individual sentences.
- **Stop word removal :** Stop words are common words that do not carry significant meaning and are often removed during text preprocessing.
- **Tokenization :** Tokenization is the process of breaking down text into smaller units called tokens. These tokens are typically words, but they can also be subwords or characters, depending on the requirements of the task. Tokenization is a crucial step in text processing as it forms the basis for subsequent analysis. Tokenization can be as simple as splitting text on whitespace or more complex, involving techniques such as using regular expressions or language-specific rules , for example :

  - Original sentence : "أنا أقرأ كتابًا جيدًا في المكتبة"
  - After tokenization : ["أنا","أقرأ","كتابًا","جيدًا","في","المكتبة"]
  - After removal of stop words : ["أقرأ","كتابًا","جيدًا","المكتبة"]
  - After stemming : ["قرأ","كتاب","جيد","مكتب"]

2.5.2 Feature extraction :

Most conventional machine-learning techniques work on the features generally numbers that describe a document in relation to the corpus that contains it created by either Bag-of-Words, TF-IDF, or generic feature engineering such as document length, word polarity, and metadata (for instance, if the text has associated tags or scores). More recent techniques include Word2Vec, GLoVE, Word embedding and learning the features during the training process of a neural network.

2.6 Applications of Text Mining :

Text mining has many applications across various industries and fields. Some of the most common applications of text mining are :

2.6.1 Sentiment analysis :

Text mining can be used to analyze the sentiment of social media posts, reviews, and customer feedback to understand the public's perception of a product or service.

2.6.2 Customer relationship management :

Text mining can be used to analyze customer interactions and feedback to improve customer service and satisfaction.

2.6.3 Competitive intelligence :

Text mining can be used to collect and analyze data on competitors, market trends, and industry developments to help companies make informed decisions.

2.6.4 Human resources management :

Text mining can be used to analyze employee feedback, resumes, and performance reviews to improve recruitment, training, and retention.

2.6.5 Biomedical research :

Text mining can be used to analyze and extract information from scientific literature and medical records to support biomedical research and drug discovery.

2.6.6 Fraud detection :

Text mining can be used to identify patterns and anomalies in financial transactions, insurance claims, and other types of data to detect fraudulent activity.

2.6.7 Security applications :

Text mining can be used for monitoring and analyzing online plain text sources such as Internet news, blogs, etc. for national security purposes.

2.6.8 Company resource planning :

Text mining can be used to analyze company reports and correspondences to identify problems and plan future actions.

In addition, Text mining can be used to analyze religious texts, such as the holy quran and prophetic Hadith, to extract knowledge and insights. This can help scholars and researchers better understand the context, meaning, and interpretation of these texts, as well as identify patterns and themes within them.

Religious text knowledge mining can also aid in the development of new Islamic scholarship, educational resources, and religious guidance for the community.

2.7 Text mining techniques

There are various techniques used in text mining, some of which are represented in the figure bellow 2.2 :

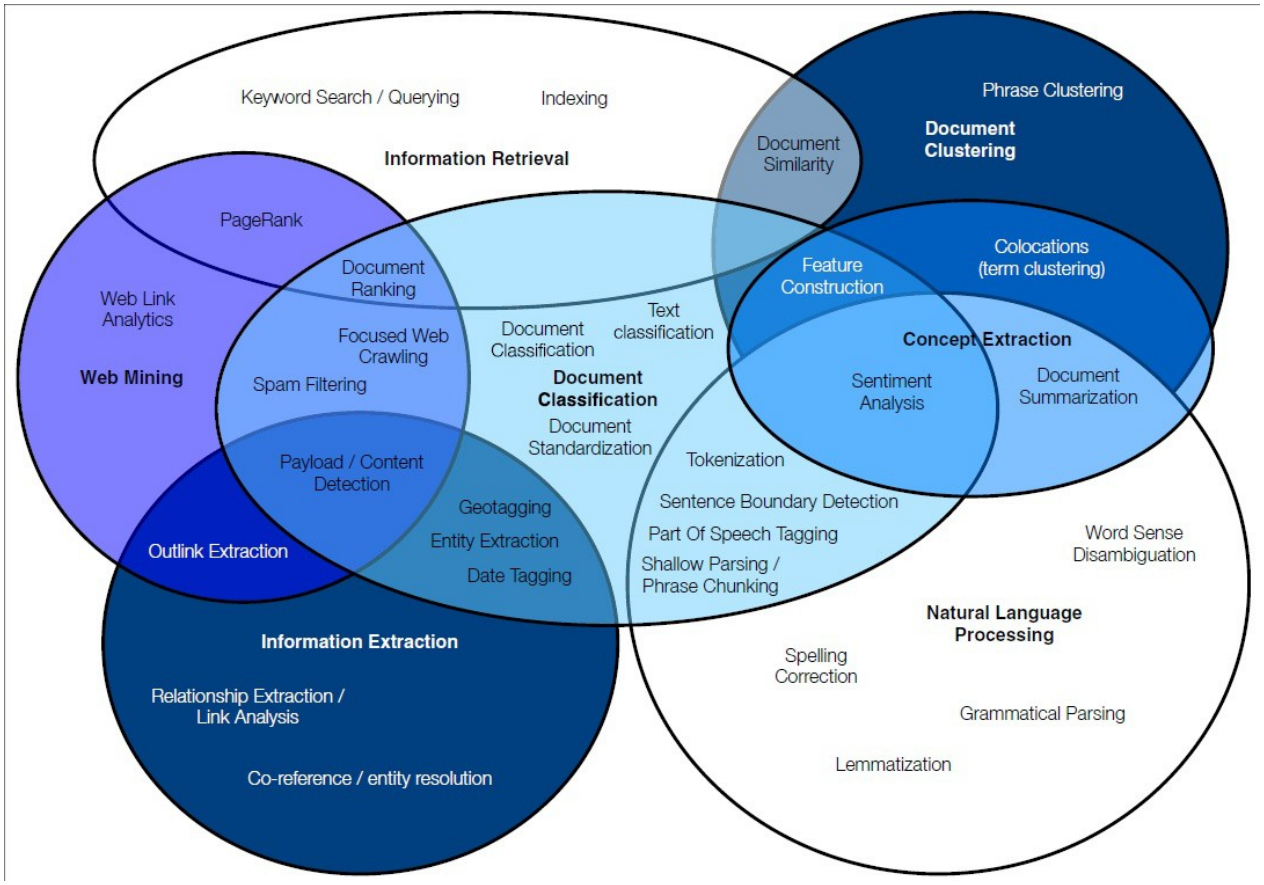

FIGURE 2.2 – Inter-relationship among different text mining techniques and their core functionalities [57]

2.7.1 Text categorization/classification :

This technique involves assigning pre-defined categories to a given text document based on its content. It is a supervised learning method that uses machine learning algorithms to classify text documents into predefined categories.

2.7.2 Clustering :

This is an unsupervised learning technique that involves grouping similar documents together based on their content, without any pre-defined categories.

2.7.3 Text summarization :

This involves creating a shorter version of a long text document that retains the most important information.
Summarization can be done either with extraction or abstraction. Abstractive methods involve in most of cases text generation processes.

2.7.4 Sentiment analysis :

This technique involves analyzing the sentiment or emotion expressed in a given text document, which can be useful in applications such as social media monitoring and customer feedback analysis.

2.7.5 Topic modeling :

This technique involves identifying the topics or themes present in a collection of documents, using unsupervised learning algorithms such as Latent Dirichlet Allocation (LDA).

2.7.6 Named entity recognition :

This technique involves identifying and extracting entities such as people, organizations, and locations mentioned in a text document.

2.7.7 Text mining with deep learning :

Deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can be used to perform tasks such as text classification, sentiment analysis, and text generation.

2.7.8 Information Retrieval (IR) :

This involves retrieving relevant documents or information from a large collection of text data based on a user query.
IR techniques use indexing and ranking algorithms to retrieve the most relevant documents based on the user's query.

2.7.9 Information Extraction (IE) :

This technique involves automatically extracting structured information from unstructured text data. IE techniques use natural language processing and machine learning algorithms to extract entities, relationships, and events from text data.
IE is useful in applications such as resume parsing, customer feedback analysis, and news article summarization.

2.7.10 Text generation :

It is the task of generating text with the goal of appearing indistinguishable to human-written text.
This task if more formally known as "natural language generation" in the literature. Currently this task is managed using neural networks.

2.8 Neural Text Generation

Text Generation is a fundamental part of various NLP tasks, such as generating open-ended responses, summarizing long documents, translating text from one language to another, and more.
Moreover, it is also utilized in mixed-modality applications where text is used as an output, such as speech-to-text and vision-to-text.
To achieve text generation, numerous models have been developed over time, including Recurrent Neural Networks and Transformer-based language models such as GPT2, XLNet, OpenAI GPT, CTRL, TransformerXL, XLM, Bart, T5, GIT, Whisper, among others.
These models are based on deep learning techniques, such as neural networks and transformer architectures, that can understand the context and generate coherent and natural-sounding text. They are trained on vast amounts of textual data, which allows them to learn the intricacies of language and produce high-quality text with accuracy and efficiency.

2.8.1 Neural Network

A Neural Network consists of different layers connected to each other, working on the structure and function of a human brain. It learns from huge volumes of data and uses complex algorithms to train a neural net[55]. For example, a neural network can be trained to identify the breed of a dog based on its features. The input layer receives the image pixels of the dog, which are then processed in the hidden layers for feature extraction. The output layer produces a result that identifies whether the dog is a German Shepherd or a Labrador.

We will now examine several neural networks that can be utilized to solve different business problems.
Feed-Forward Neural Network is used for general Regression and Classification problems, while a Convolutional Neural Network is used for object detection and image classification. A Deep Belief Network is used in healthcare sectors for cancer detection, and an RNN is used for speech recognition, voice recognition, time series prediction, and natural language processing.

2.8.2 Recurrent Neural Network (RNN)

A Recurrent Neural Network (RNN)[62] is a type of neural network in which the output of a previous step is fed as input to the current step. In a traditional neural network, all inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words.
Thus RNN came into existence, which solved this issue with the help of a Hidden Layer .The main and most important feature of RNNs is the Hidden state, which remembers some information about the sequence.

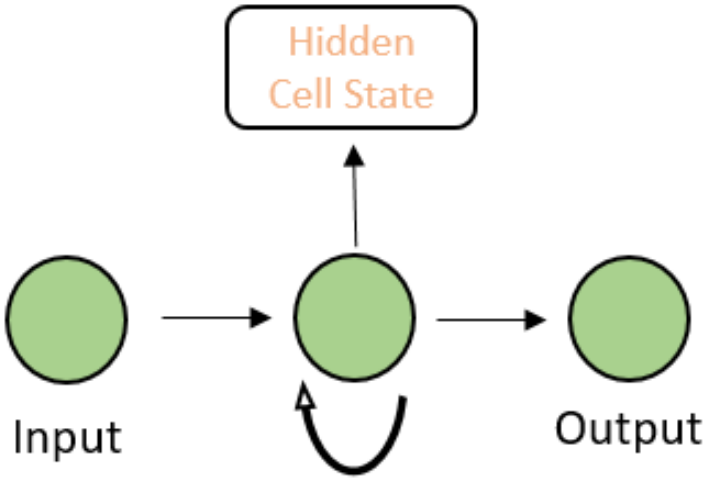

FIGURE 2.3 – Recurrent Neural Network Schema[62]

An RNN has a "memory" that stores all information about what has been computed. It performs the same task and produces the output for all inputs or hidden layers, so it uses the same parameters for each input. In contrast to other neural networks, this reduces the parameter complexity.

2.8.2.1 Types of RNN :

There are four types of RNNs based on the input and output configurations :
one-to-one, one-to-many, many-to-one, and many-to-many [62].
The one-to-one RNN takes a single input and produces a single output, making it similar to a traditional feedforward neural network. The one-to-many RNN takes a single input and generates multiple outputs, such as image captioning, where a single image is used to generate a sequence of words describing it.
The many-to-one RNN takes multiple inputs and produces a single output, making it useful for sentiment analysis, where a sequence of words is used to predict the sentiment of a sentence. The many-to-many RNN takes multiple inputs and produces multiple outputs, such as machine translation, where a sequence of words in one language is used to generate a sequence of words in another language.
The ability of RNNs to capture temporal dependencies and their diverse input-output configurations make

them a powerful tool for a wide range of sequential data processing tasks.

2.8.2.2 Applications of Recurrent Neural Network :

Recurrent Neural Networks (RNNs) are used in a wide range of applications that involve sequential data. One of the most popular applications of RNNs is language modelling and text generation[33]. RNNs are also used in speech recognition systems to convert speech to text, and in machine translation systems to translate text from one language to another. In image recognition, RNNs can be used for tasks such as face detection and recognition. Additionally, RNNs are widely used in time-series forecasting applications, such as predicting stock prices or weather conditions. The ability of RNNs to capture temporal dependencies in sequential data makes them a versatile tool for a variety of applications that involve time-varying data.
One of the significant advantages of RNNs is their ability to remember all information over time, including long-term memory, making them useful for time series forecasting. RNNs are also used in convolutional layers to extend valid pixel neighborhoods[33].
However, One of the primary issues is the gradient vanishing and exploding problems, which can hinder their performance. Additionally, training an RNN is a challenging task. Furthermore, if using tanh or relu as an activation function, RNNs cannot process very long sequences effectively. Despite these disadvantages, RNNs remain a popular choice for various applications.

2.8.3 Transformer Models :

The Transformer is a deep learning model that uses self-attention mechanism to weight the importance of each part of the input data[60]. It is commonly used in natural language processing and computer vision applications. Unlike recurrent neural networks, transformers process the entire input at once using parallel processing. This enables faster training and development of large pre-trained systems like BERT and GPT. Transformers are especially useful for tasks like text summarization and translation where sequential input data is involved.

2.8.3.1 The Transformer Architecture :

The Transformer model follows an encoder-decoder architecture that doesn't rely on recurrence or convolution operations for generating an output.
The encoder component transforms the input sequence into a set of continuous representations, while the decoder component takes the output from the encoder and the decoder's previous output to generate the final output sequence.
The architecture is illustrated in a diagram below3.1 :

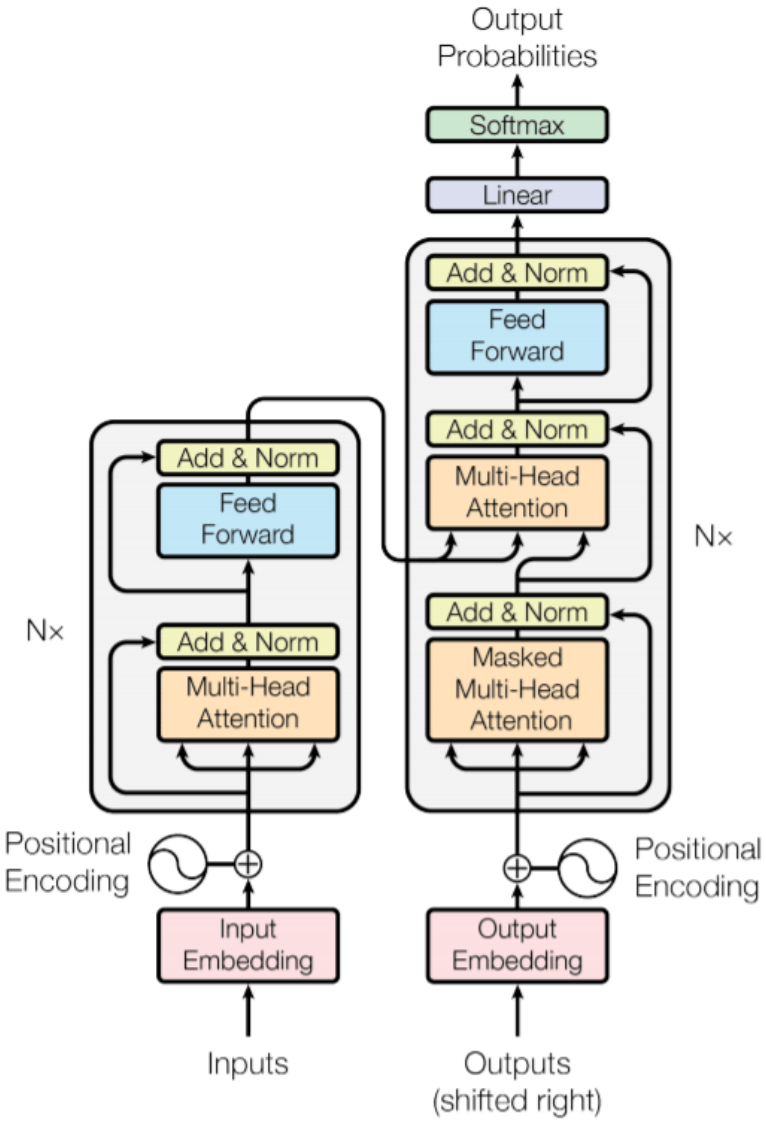

FIGURE 2.4 – The Transformer - model architecture [60]

1.The Encoder :

The encoder is comprised of six identical layers, each consisting of two sublayers. The first sublayer is a multi-head self-attention mechanism, while the second sublayer is a feed forward network. These sublayers are coupled positionally, and the output of each sublayer is normalized using LayerNorm(x+Sublayer(x)).

The multi-head self-attention mechanism utilizes h heads, where each head receives a linearly projected version of queries, keys, and values. This allows for the generation of h outputs simultaneously, which are then combined to obtain the final result.

The feed-forward network in each layer involves two linear transformations with ReLU activation in between

$$F(x) = \text{ReLu}(W_1 x + b_1)W_2 + b_2 \tag{2.15}$$

It applies weight parameters (W1, W2) and bias parameters (b1, b2) to the input.

2.The Decoder :

The decoder consists of six identical layers, each containing two sublayers from the encoder, as well as an additional sublayer called Masked multi-head attention.
The Masked multi-head attention sublayer takes the previous output of the decoder stack and applies a masking technique. This masking ensures that the self-attention mechanism focuses only on the tokens in the input sequence that come before them, rather than considering the entire sequence as a whole.
The multi-head self-attention mechanism in the decoder accepts the keys and values from the encoder's output, along with the queries from the preceding sublayer in the decoder. This allows the decoder to attend to every word in the input sequence, enabling comprehensive analysis.
Similar to the second sublayer in the encoder, the decoder's third layer includes a fully connected feed-forward network. This network consists of two linear transformations with a Rectified Linear Unit (ReLU) activation function applied in between.

2.8.4 Large Language Models :

Pretrained language models are becoming increasingly important in the field of natural language processing (NLP), and several models have emerged as leaders in this space. Here is a brief explanation of some of the most important ones :

- **GPT-2 [47] :** Generative Pre-trained Transformer 2. GPT-2 is a large-scale transformer-based language model developed by OpenAI. It is capable of generating high-quality text that is often difficult to distinguish from text written by humans.
- **T5 [48] :** Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. T5 is a text-to-text transformer-based model developed by Google. It is capable of performing a wide range of NLP tasks, including language translation, summarization, and question answering, using a single unified architecture.
- **BERT[26] :** Bidirectional Encoder Representations from Transformers. BERT is a transformer-based model that uses unsupervised learning to pretrain deep bidirectional representations for a wide range

of NLP tasks. It was developed by Google AI and has achieved state-of-the-art results on many NLP benchmarks.

— **XLNet[63] :** Generalized Autoregressive Pretraining for Language Understanding. XLNet is another transformer-based language model that was developed by researchers at Carnegie Mellon University, Google Brain, and Stanford University. It uses an autoregressive formulation that allows it to capture dependencies between all possible permutations of words in a sentence.

— **RoBERTa[40] :** A Robustly Optimized BERT Pretraining Approach. RoBERTa is a modification of BERT that was developed by Facebook AI. It uses additional training data and improved training techniques to achieve better results than the original BERT model on many NLP benchmarks.

— **ALBER[39]T :** A Lite BERT for Self-supervised Learning of Language Representations. ALBERT is a more lightweight version of BERT that was developed by researchers at Google. It achieves similar or better results than BERT while using fewer parameters, making it more computationally efficient.

— **GPT-3[23] :**Generative Pre-trained Transformer 3. GPT-3 is the successor to GPT-2 and is currently the largest language model available, with 175 billion parameters. It is capable of performing a wide range of NLP tasks, including few-shot learning and even generating code.

— **ELECTRA[25] :** Pre-training Text Encoders as Discriminators Rather Than Generators. ELECTRA is a transformer-based model developed by Google that uses a discriminator to pretrain the model instead of a generator. This results in faster training times and better performance on some NLP tasks.

— **DeBERTa[37] :** Decoding-enhanced BERT with Disentangled Attention. DeBERTa is a transformer-based model developed by researchers at Microsoft that uses disentangled attention to improve the model's ability to capture long-term dependencies in text.

— **PaLM[53] :**Path-augmented Language Model. PaLM is a transformer-based model developed by researchers at Google that uses a novel "pathway" architecture to improve the model's ability to handle long sequences of text.

2.9 Arabic text mining

2.9.1 Arabic Language :

Arabic is one of the most widely spoken languages in the world, with over 672 million people speaking it as a first or second language. It has a unique alphabet of 28 letters plus the Hamza, which are written from right to left.

Arabic has a complex morphology, with most words having a tri-letter root, and others having quad-letter, penta-letter, or hexa-letter roots. Nouns in Arabic have three grammatical cases, and words are classified into three main parts of speech : nouns (including adjectives and adverbs), verbs, and particles.

Due to its unique features, text mining of Arabic documents requires specialized techniques and tools that

are tailored to the language's specific characteristics.

2.9.1.1 Arabic Language Peculiarity :

Arabic is a very complex language, with a number of unique features that make it challenging to learn and process. These features include :

1.Orthography :

The Arabic writing system is based on a cursive script and is written from right to left. It also includes diacritical marks that can change the meaning of words or indicate the presence of short vowels, which are not always written in standard Arabic texts.

2.Morphology :

Arabic is a highly inflected language with a complex system of prefixes, suffixes, and infixes. The majority of Arabic words have a three-letter root, with additional letters added to indicate tense, aspect, voice, and other grammatical features.

3.Synonyms :

Arabic has a rich vocabulary with many synonyms for common words, making it difficult to determine the precise meaning of a word in context.

4.Polysemy :

Many Arabic words have multiple meanings depending on the context in which they are used.

5.Lack of corpora :

Compared to other languages, there is a scarcity of freely accessible Arabic corpora, making it difficult for researchers to train and test natural language processing algorithms.

2.10 Arabic text mining related works

Arabic text mining has gained increasing attention in recent years, with a growing number of research papers published on the topic. This field encompasses various techniques that allow for the extraction of meaningful insights and knowledge from Arabic language content, such as sentiment analysis [49] , text mining of web documents, Text Mining of the Holy Quran and Prophetic Hadith.
In this section, we will briefly discuss each of these works and their potential applications.

2.10.1 Arabic Opinion Mining :

According to a study by [4], sentiment analysis is a research field that is connected to various areas of study such as NLP, text mining, and computational linguistics.

It involves extracting specific data from individuals through textual data and is sometimes referred to as opinion mining, subjective opinion analysis, or emotion extraction [45].

Each term has a specific objective : opinion mining involves classifying negative and positive opinions presented in text, while emotion extraction focuses on differentiating between different emotions (e.g., happy, angry, and sad). In this section, different studies related to sentiment analysis are presented, which differ in their preprocessing methods, analysis techniques, and review designs.

Some studies have utilized supervised learning, while others have used unsupervised learning. [4] Proposed a multi-stage approach based on semantic orientation, which included a lexical-based classifier to handle unlabeled tweets and a machine learning SVM classifier to recognize the polarity of Arabic tweets. However, the main challenge of this hybrid approach was to handle the practice of tweeting in dialectical Arabic.

Another approach was proposed by [24], which utilized sentiment classification and Support Vector Machines to handle the syntactic and grammatical complexity of the Arabic language and analyze Arabic reviews and comments. The study utilized a dataset consisting of 625 Arabic reviews and opinions obtained from the official website of Trip Advisor.

[27] Put forth an approach that recommended preprocessing reviews as a means of improving the performance of classifiers, including Support Vector Machines, Naive Bayes, and K-nearest neighbor classifiers.

[56] Addressed the significant challenge of mining Arabic slang language, for which they utilized efficient techniques such as Slang Sentimental Words and Idioms Lexicon (SSWIL) and the Gaussian kernel SVM classifier.

Table 2.1 summarized the work described earlier including the results they obtained.

TABLE 2.1 – Research studies conducted on Arabic text mining in terms of sentiment analysis [49]

Study	Problem	Methodology	Results
[4]	Recognize the polarity of Arabic tweets and the practice of dialectical Arabic tweets	Hybrid classifiers, lexical classifiers, feature extraction and support vector machine classifiers	The general quality of the obtained results in this study from the hybrid classifier, quantified by F-measure, is 84% and the accuracy is 84.01%

Continued on the next page

Table 2.1 :

Study	Problem	Methodology	Results
[24]	Evaluate the impact of Arabic grammatical richness on opinion-generating accuracy, develop new and precise statistical approaches to support the complexity of Arabic syntax and grammar, and review and commentary on Arabic in more detail	Sentiment classification, Support Vector Machines, a dataset composed of 625 Arabic reviews and opinions of the public obtained from the official website of Trip Advisor	Results obtained were rooted in Support Vector Machines depicted that this method greatly affects the identification of opinions
[27]	The effects of stemming feature correlation and n-gram models for Arabic text on sentiment analysis	Support Vector Machines, Naive Bayes, and K-nearest neighbor classifiers	Experimental results suggest that choosing a method to preprocess the texts improves the performance of the classifier
[56]	The Arabic slang language suffers not only from unstructured forms but also from new expressive (opinion) words and idioms. Arabic slang mining requires efficient techniques to extract young people's opinions on various topics such as : Example : News website	A Slang Sentimental Words and Idioms Lexicon (SSWIL) of opinion words, Gaussian kernel SVM classifier	The results indicated that the accuracy is 86.86% ; precision and recall are 88.63% and 78% respectively

2.10.2 Arabic Web Documents Mining :

According to [15], an innovative method for building a comprehensive Arabic dictionary involved using linguistic methods to extract relevant single and compound Frequently Association (FA) terms from corpora related to specific fields, with the help of Arabic POS tagging.

In [50], a proposed approach for identifying the language of Arabic web pages involved using a decision tree-ARTMAP approach (DTA) to detect the overall characteristics of the web document, which would ultimately enhance the language identification performance.

The authors of [22] investigated several text stemming techniques for Arabic topic modeling. They found that using lemma-based stemming allowed users to identify interesting topics discussed in press articles published between 2007 and 2009.

Additionally, they used the Dormant Dirichlet Allocation model to extract latent topics from three different Arabic corpora. The following Table 2.2 summarizes the studies described above.

TABLE 2.2 – Research studies conducted on Arabic text mining in terms of web documents [49]

Study	Problem	Methodology	Results
[15]	Lacking of an effective method to automatically extract relevant FA terms to build a comprehensive dictionary	Text classification methodology using Arabic FA terms, Na-ve Bayes (NB) and KNN classifiers	The new approach achieved a precision of 80.65% followed by NB (72.79%) and KNN (36.15%)
[50]	Optimizing the performance of language identification based on Arabic script web documents by proposing a DT-ARTMAP (DTA) approach	Decision tree and ARTMAP approaches	The results revealed that the proposed approach has outperformed both the decision tree and the default ARTMAP approaches
[22]	In the tasks related to semantic analysis,it is preferable to directly deal with texts in their original anguage.Studies on topic models which provide a good way to automatically deal with semantic embedded in texts are not complete enough to assess the effectiveness of the approach on Arabic texts	Support vector machine (SVM),The Latent Dirichlet Allocation model	The results indicated a high effectiveness for the approach. The BBW lemma-based stemmer reduces significantly vocabulary dimension and under- and over-stemming errors. In addition,classification performance is improved slightly compared to classification of raw and light stemmed texts

2.10.3 Holy Quran Mining :

A research study by [5], the authors applied various techniques such as analyzing the most frequent words, creating word clouds, constructing document-term matrices, and using clustering techniques to evaluate Arabic text and provide statistical and factual information about the Holy Quran. In [42], the authors used various methods to enable the analysis of unstructured information within the Quran text. These methods included the vector space model and pronoun tagging.

Finally,a thesis research that was conducted by [51] to discover and extract the interesting, non-trivial knowledge out of the Holy Quran uses methods like information extraction, text categorization, concept linkage and discovery of associations[49].

Table 4.4 Research studies conducted on Arabic text mining in terms of web documents

TABLE 2.3 – Research studies conducted on Arabic text mining in terms of the Holy Quran[49]

Study	Problem	Methodology	Results
[5]	Evaluation of Arabic texts to present and provide statistical and factual details	Most Frequent Words, Word Cloud, Document-Term Matrix, and Clustering	The obtained results from this study carry different features of the Holy Quran like some significant words, its word cloud and different chapters with high term frequencies
[42]	Creating a platform that enables the analysis of unstructured information in searching of interesting knowledge embedded within the Qur'an text	Vector space model, The pronoun tagging, the part-of-speech tags (POS), the verse relatedness dataset, and concept clouds	Results revealed that world knowledge is required to reach the human upper bound in certain computational tasks such as detecting text relatedness, question answering, and textual entailment
[51]	Discovering and extracting interesting, non-trivial knowledge out of the Holy Qur'an by applying text-mining methods	Information extraction, text categorization, concept linkage, and discovery of associations	POS tagged corpus improved the results over the raw text and the intention of this project is to improve the results further through added layers of annotation

2.10.4 Prophetic Hadith Mining :

In the past 12 years, researchers have employed various methods to analyze and study hadiths, focusing on improving efficiency and accuracy. These approaches include data mining, fuzzy systems, machine learning techniques, natural language processing (NLP) tools, neural networks, and context-free grammar. Given that the authenticity of hadiths relies heavily on the reliability of the sources involved in their transmission, researchers have primarily focused on analyzing the Isnad (chain of narrators) of hadiths, considering multiple perspectives.

They have addressed issues such as validating hadiths, identifying trustworthy chains of narration, retrieving hadiths based on queries, generating and extracting Isnads, cross-validating information, and employing various other techniques to process Isnads. Due to the significance of hadiths as fundamental scriptural sources in Islamic beliefs, there has been considerable research on them from a computer science standpoint.[36]

In 2008, researchers of Iran took initiative and presented a paper [34] on determining the rate of validity of a hadith, which came out 94%, through a fuzzy system through extracting essential rules and parameters based on narrators background, religion, number of hadiths they narrated, the continuity of Isnads and also

graded the narrators from their biographies.

In 2010, [18] presented a system called iTree which stands for Isnad Tree that automatically generates the transmission chain of a hadith and graphically displays the chain as a tree. Their solution involves parsing and annotating the hadith text upon domain specific context-free grammar of hadiths and identifying the narrators names. Their system got a success rate of 87%. This works with Arabic language, which is quite challenging, but has nothing to do with hadith verification.

In 2012, [64] considered information extraction from sets or related documents where one set complements and facilitates the information extraction task in the other, which is called cross-document technique of NLP. Their results show a significant increase in accuracy from 41% to 93% after applying this cross-document NLP methodology to hadith and biography documents. To recognize identical narrator names, but their application is limited to graph extraction and not trying to verify/validate hadiths.

In [54] Moath M. Najeeb proposed that there is a big opportunity to build an automatic information system to classify hadith as Sahih (accepted) or Daif (poor) using Associative Classification technique. This incorporates NLPs classification technique with Data Minings Association Rule Mining (ARM).

[7] Used an NLP technique called N-gram modeling to build a system that segments a hadith into its two main components, Isnad and Matn. Their hadith segmentation system is built and tested with hadiths from six major hadith books. More recently, [8] contributed in creating an Arabic- English parallel Hadith corpus to support further researches in this field, along with an improved hadith segmentation tool of their own using ML that successfully segments and annotates Isnad and Matn with 92% accuracy.

They scraped pages from Sunnah.com and using their HTML tags extracted hadiths along with their location information to create bilingual corpus named Leeds University and King Saud University (LK) Hadith corpus.

TABLE 2.4 – Research studies on Arabic Hadith analysis

Study	Problem	Methodology	Results
[34]	Validity of hadiths	Fuzzy system with essential rules and parameters	Validity rate of 94% obtained by analyzing narrators' background
[18]	Generating transmission chain of hadiths	iTree system with parsing and annotation	Success rate of 87% achieved in generating and displaying hadith chains
[64]	Information extraction from related docs	Cross-document technique of NLP	Accuracy increased from 41% to 93% using cross-document NLP methodology
[54]	Classification of hadiths as Sahih or Da'if	Associative Classification technique	Proposal for an automatic system using NLP and Data Mining techniques
[7]	Segmentation of hadiths into Isnad and Matn	N-gram modeling	Hadith segmentation system tested with major hadith books
[8]	Creation of Arabic-English Hadith corpus	ML-based hadith segmentation tool	92% accuracy achieved in segmenting and annotating Isnad and Matn

2.11 Conclusion

In summary, text mining has become an indispensable instrument for obtaining valuable insights and extracting valuable information from vast collections of documents. By employing a variety of techniques and models, text mining has made it possible for researchers to uncover previously unknown patterns, trends, and associations within textual data.

Despite facing significant obstacles, such as dealing with intricate languages like Arabic, text mining has fundamentally transformed the way we analyze and interpret textual data. Therefore, it is anticipated to remain an essential discipline in the age of big data and artificial intelligence, powering advancements in text generation and other natural language processing tasks.

Chapitre 3

Contribution

3.1 Introduction

In this chapter, we present our Hadith knowledge mining proposal, which consists of two main parts : Section Abstraction (باب الحديث) and Benefits Extraction (فوائد الحديث). In the first part, Section Abstraction, we utilize the two transformer models T5 and mbert2mbert in the abstraction process, known for their language understanding and generation capabilities, to enhance the efficiency and accuracy of the abstraction process. These models have demonstrated success in various natural language processing tasks, making them ideal for our proposition.

Moving on to the second part, Benefits Extraction, we introduce two methods to extract benefits from a given hadith. The first method employs beam search, a technique commonly used in natural language processing to explore multiple potential solutions. By utilizing beam search with the T5 model and mbert2mbert, we can effectively identify and extract the most relevant benefits from the hadith.In addition to the beam search method, we also incorporate Text Completion using the GPT2 model for Benefits Extraction.

To evaluate the performance and effectiveness of our proposal, we employ two datasets. One dataset is specifically used for Section Abstraction, while the second dataset for Benefits Extraction.

3.2 Datasets Description

The first dataset we utilized is LK-Hadith-Corpus [52], a bilingual parallel corpus of English-Arabic Islamic Hadith. It was extracted from the six canonical Hadith books (Bukhari, Muslim, Tirmizi, Ibn Majah, Abu Daoud) and consists of 39,038 annotated Hadiths, comprising more than 10 million tokens. Each component of the Hadith is extracted and allocated to a specific column in the dataset.

The dataset includes :

Table 3.1 – Our first Dataset

Chapter_Number	96.0
Chapter_English	Holding Fast to the Qur'an and Sunnah
Chapter_Arabic	كتاب الاعتصام بالكتاب والسنة
Section_Number	7.0
Section_English	Judging made on the basis of opinion or Qiyas
Section_Arabic	مَا يُذْكَرُ مِنْ ذَمِ الرَأيِ وَتَكَلُف الْقِيَاسِ
Hadith_number	7307
English_Hadith	Narrated 'Abdullah bin 'Amr : I heard the Prophet (صلى الله عليه وسلم) saying, "Allah will not deprive you of knowledge after he has given it to you, butit will be taken away through the death of the religious learned men with their knowledge. Then therewill remain ignorant people who, when consulted, will give verdicts according to their opinionswhereby they will mislead others and go astray."
English_Isnad	Narrated 'Abdullah bin 'Amr :
English_Matn	I heard the Prophet (صلى الله عليه وسلم) saying, "Allah will not deprive you of knowledge after he has given it to you, butit will be taken away through the death of the religious learned men with their knowledge. Then therewill remain ignorant people who, when consulted, will give verdicts according to their opinionswhereby they will mislead others and go astray."
Arabic_Hadith	حَدَّثَنَا سَعِيدُ بْنُ تَلِيدٍ، حَدَّثَنِي ابْنُ وَهْبٍ، حَدَّثَنِي عَبْدُ الرَّحْمَنِ بْنُ شُرَيْحٍ، وَغَيْرُهُ، عَنْ أَبِي الأَسْوَدِ، عَنْ عُرْوَةَ، قَالَ حَجَّ عَلَيْنَا عَبْدُ اللَّهِ بْنُ عَمْرٍ فَسَمِعْتُهُ يَقُولُ سَمِعْتُ النَّبِيَّ صلى الله عليه وسلم يَقُولُ إِنَّ اللَّهَ لاَ يَنْزِعُ الْعِلْمَ بَعْدَ أَنْ أَعْطَاهُمُوهُ انْتِزَاعًا، وَلَكِنْ يَنْتَزِعُهُ مِنْهُمْ مَعَ قَبْضِ الْعُلَمَاءِ بِعِلْمِهِمْ، فَيَبْقَى نَاسٌ جُهَّالٌ يُسْتَفْتَوْنَ فَيُفْتُونَ بِرَأْيِهِمْ، فَيُضِلُّونَ وَيَضِلُّونَ

Continued on the next page

Table 3.1 : Data Dictionary

Arabic_Isnad	حَدَّثَنَا سَعِيدُ بْنُ تَلِيدٍ، حَدَّثَنِي ابْنُ وَهْبٍ، حَدَّثَنِي عَبْدُ الرَّحْمَنِ بْنُ شُرَيْحٍ، وَغَيْرُهُ، عَنْ أَبِي الأَسْوَدِ، عَنْ عُرْوَةَ، قَالَ حَجَّ عَلَيْنَا عَبْدُ اللَّهِ بْنُ عَمْرٍ فَسَمِعْتُهُ يَقُولُ
Arabic_Matn	سَمِعْتُ النَّبِيَّ صلى الله عليه وسلم يَقُول إِنَّ اللَّهَ لاَ يَنْزِعُ الْعِلْمَ بَعْدَ أَنْ أَعْطَاهُمُوهُ انْتِزَاعًا، وَلَكِنْ يَنْتَزِعُهُ مِنْهُمْ مَعَ قَبْضِ الْعُلَمَاءِ بِعِلْمِهِمْ، فَيَبْقَى نَاسٌ جُهَّالٌ يُسْتَفْتَوْنَ فَيُفْتُونَ بِرَأْيِهِمْ، فَيُضِلُّونَ وَيَضِلُّونَ
Arabic_Comment	nan
English_Grade	Sahih-Authentic
Arabic_Grade	صحيح

For our purposes, we selected two specific columns from the dataset : Section_Arabic, and Arabic_Matn. These columns contain the original Arabic texts of the section, and the main body of the Hadith, respectively. In the second dataset, we collected the necessary data manually to feed the learning models for benefits extraction. Since there was a lack of existing similar work or readily available collections, we took the initiative to curate our own dataset.

The data was collected from the following sources [38] [2] [14] :

- الفوائد المستنبطة من الأربعين النووية وتتمتها الرجبية
- الجامع لفوائد بلوغ المرام
- الفوائد المنتخبة من أحاديث الصيام من صحيح البخاري

For each Hadith we do associate three benefits. The combined data from these sources contains a total of 13,167 lines. This data serves as the foundation for training and evaluating the benefits extraction models. By manually collecting and curating this data, we ensure that it aligns with the specific requirements and objectives of our proposition.

3.3 Transformer model

In this part, we utilize the Transformer model for section abstraction and benefits extraction. Transformers have the advantage of not needing to process sequential input in a strict order. Unlike traditional models, they can process different parts of a sentence simultaneously, thanks to their parallel processing capabilities. The transformer architecture consists of an encoder and a decoder. The encoder employs multiple cipher layers to recursively process the input, while the decoder applies a similar process to generate

the output based on the encoder's representation.

The fundamental units of a transformer are note units, which produce weighted connections between symbols. These weights measure the relationships between units during the machine learning process, where symbols represent letters or elements from the languages the machine is learning. The observation module generates contextual embeddings for each symbol, which not only contain information about the symbol itself but also incorporate a weighted combination of other related symbols based on the note weights.
Transformer models consist of an encoder and a decoder :

— **Encoder :** It takes an input and constructs its representation or features. The model is optimized to gain understanding from the input data.

— **Decoder :**It utilizes the representation or features from the encoder, along with other inputs, to generate a target sequence.The model is optimized for generating desired outputs.

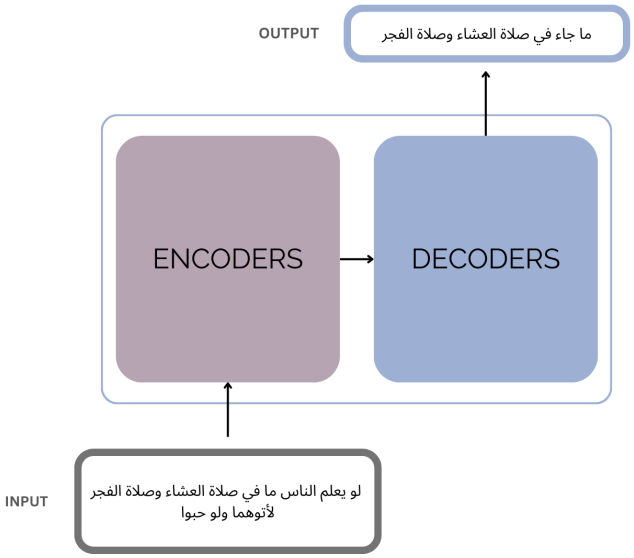

FIGURE 3.1 – Encoder-Decoder structure of Transformer

By employing the encoder-decoder architecture of transformers, we can leverage their powerful capabilities for analyzing and generating content.

3.4 Section Abstraction

In the Section Abstraction step, our goal is to utilize transformer models to predict the section or topic of a given Hadith 3.1. We employ two transformer models to achieve this : malmarjeh/mbert2mbert-arabic-text-summarization [41] and UBC-NLP/AraT5-base-title-generation[59] .

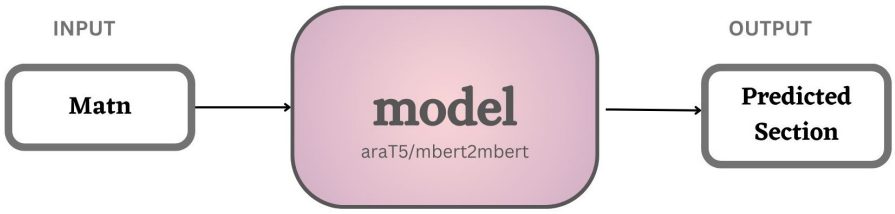

FIGURE 3.2 – Representation of Section abstraction using araT5 mbert2mbert model

3.4.1 mbert2mbert Model :

The first model is a BERT2BERT-based model. It initializes its parameters with mBERT weights and has been fine-tuned on a dataset consisting of 84,764 paragraph-summary pairs.This model is known for its language understanding and generation capabilities, making it suitable for our task of section abstraction in the context of Hadith analysis.
In our case, we utilize the malmarjeh/mbert2mbert-arabic-text-summarization model to predict the section or topic of a given Hadith. We provide the Hadith text as input to the model, and it generates an output that represents the corresponding section of the Hadith. This allows us to automatically categorize and organize Hadiths based on their content .
By leveraging the mbert2mbert model, we aim to enhance the efficiency and accuracy of the section abstraction process in our proposed framework.

3.4.2 araT5 Model :

The second model we utilize is UBC-NLP/AraT5-base-title-generation.

This model is part of the AraT5 repository, which introduces three powerful Arabic-specific text-to-text Transformer-based models : AraT5MSA, AraT5Tweet, and AraT5. AraT5-base-title-generation is specifically designed for title generation in Arabic text. Given a Hadith, this model generates a concise and meaningful

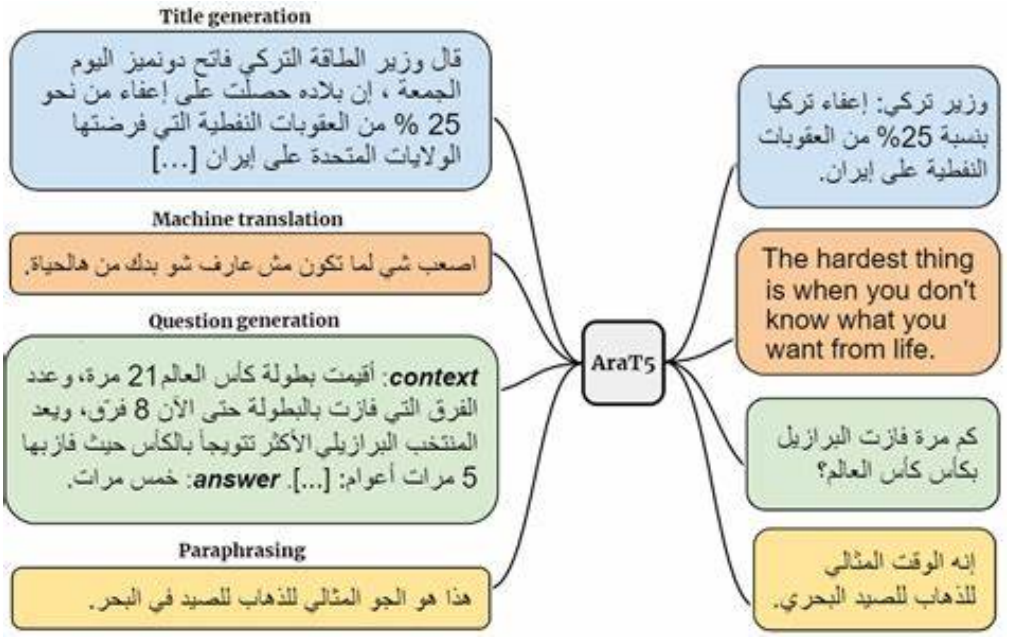

FIGURE 3.3 – AraT5 encoder-decoder model and prompt samples from four investigated tasks, namely : title generation, machine translation, question generation, and paraphrasing [30]

title that represents the section or topic of the Hadith.

These transformer models have been trained on large-scale datasets and fine-tuned for their respective tasks. By utilizing their language understanding and generation capabilities, we can enhance the efficiency and accuracy of the abstraction process for Islamic Hadiths.

Themalmarjeh/mbert2mbert-arabic-text-summarization model and UBC-NLP/AraT5-base-title-generation model work by leveraging the power of transformer-based architectures, which have been highly successful in natural language processing tasks.

3.4.3 Processing steps :

The process of Model transformers involves several processing stages, as shown in Figure . The first stage is preprocessing, where the text is cleaned and prepared. Following that, the data undergoes tokenization to prepare it for the model input. The final stage is prediction . The detailed descriptions of these processing stages are provided in the following sections.

3.4.3.1 Preprocessing :

Data preprocessing is an essential step in Machine Learning that plays a vital role in improving the quality of data and enabling the extraction of valuable insights. It involves various techniques aimed at cleaning and organizing raw data to make it suitable for building and training Machine Learning models. In essence, data

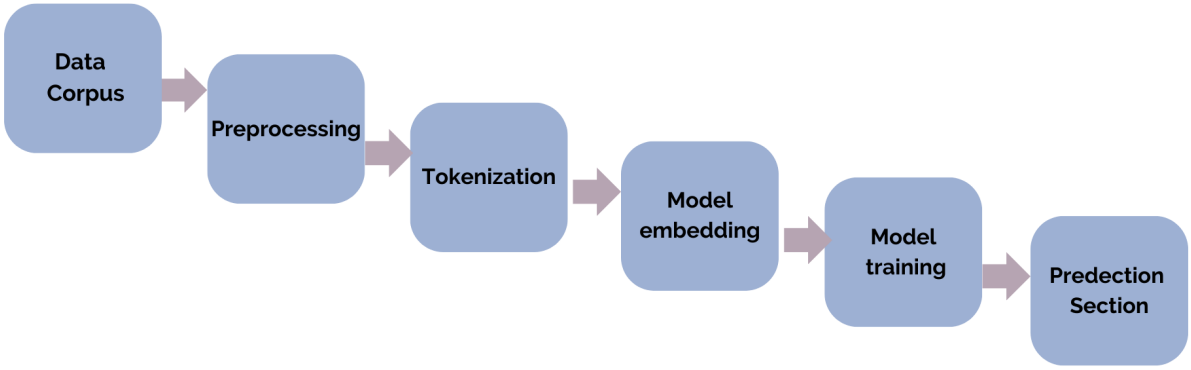

FIGURE 3.4 – Section abstraction Process using Arabic T5 and mbert2mbert Models

preprocessing can be seen as a data mining approach that converts raw data into a format that is easily comprehensible and usable for analysis.

3.4.3.2 Data Splitting :

Before training the models, it is common practice to split the dataset into training and testing sets. This splitting allows us to evaluate the performance of the trained models on unseen data. In our case, we follow the standard practice of splitting the dataset as follows :

— **Training Set :**This subset of the dataset is used to train the models. We use 70% of the available data for training. During the training process, the models learn the patterns and relationships within the training set to generate accurate predictions. .

— **Validation Set :**The validation set is separate from the training set and is used during the training process to tune hyperparameters and assess the model's performance. It typically constitutes around 10% of the dataset. The validation set helps in preventing overfitting by providing an unbiased evaluation of the model's performance on data that it hasn't seen during training.

— **Testing Set :** The testing set, also referred to as the holdout set, contains the remaining portion of the dataset, typically around 20%. This set is used to evaluate the final performance of the trained models. Since the testing set is completely unseen during the training and validation stages, it provides an objective measure of how well the models generalize to new, unseen data.

By splitting the data into training and testing sets, we ensure that the models are trained on a representative portion of the dataset and evaluated on unseen examples, enabling us to gauge their performance accurately.

3.4.3.3 Tokenization :

Tokenization is a crucial step in natural language processing that involves breaking down sentences or text into individual words or tokens. By splitting the text into tokens, we can analyze and process the language

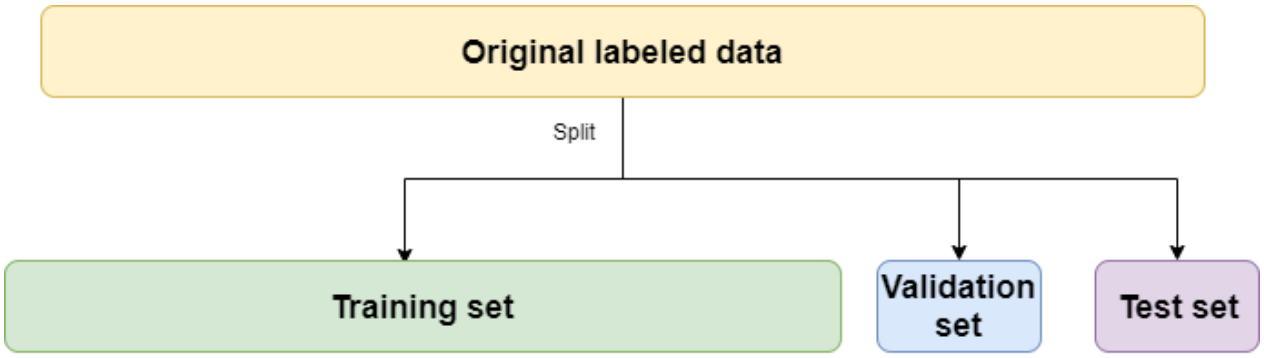

FIGURE 3.5 – Splitting Data for Machine Learning

more effectively.

Here is an example of tokenization for the matn (main body) of a Hadith :

— **Original Text :** من صام رمضان إيمانًا واحتسابًا غفر له ما تقدم من ذنبه

— **Tokenized Text :** من، صام، رمضان، إيمانًا، واحتسابًا، غفر، له، ما، تقدم، من، ذنب

In this example, the original text is tokenized into individual words or subwords. Each word is represented as a token, and punctuation marks are treated as separate tokens as well. The tokenized text consists of a list of tokens that can be further processed and analyzed.

3.4.3.4 Model Embedding :

After tokenization, the next step in natural language processing is often embedding. Embeddings are numerical representations of words or tokens that capture the semantic and syntactic relationships between them. Embedding techniques aim to transform the textual data into a numerical format that can be easily processed by machine learning models.

There are various embedding techniques available, such as word embeddings and contextual embeddings. Here, we'll focus on word embeddings.

Word embeddings represent words or tokens as dense vectors in a high-dimensional space, where similar words are closer to each other in the vector space. These embeddings are learned from large corpora using unsupervised learning techniques like Word2Vec, GloVe, or FastText.

In the context of our Section Abstraction task for Hadiths, we can use pre-trained word embeddings specific to the Arabic language. These embeddings are trained on a large Arabic corpus and can capture the semantic meaning of Arabic words effectively.

During the training process, the embedding model learns to assign similar vector representations to words that have similar meanings or are used in similar contexts. This allows the model to capture the semantic relationships between words, which can be beneficial for downstream tasks.

Once the words in the Hadiths are embedded, the models can learn from these representations to predict the corresponding sections or topics effectively.

3.4.3.5 Model fine-tuning and training :

The training process involves utilizing the datasets and splitting them into appropriately sized portions. Each model is then trained over multiple epochs. We fine-tune the selected models for our target task (Section abstraction) by involving the following strategies and including the parameters mentioned in Table 1 :

Epoch and Batch size :

An epoch means training the neural network with all the training data for one cycle. In an epoch, we use all of the data exactly once. A forward pass and a backward pass together are counted as one pass.due to epochs large size, they are divided into smaller chunks known as batches.due to epochs large size, they are divided into smaller chunks known as batches.

The following table summarizes the basic set of parameters defined for the model :

TABLE 3.2 – Basic parameters of model training

Learning rate	2e-5
Input size	100
Output size	30
weight_decay	0.01
batch_size	5
num_train_epochs	10
fp16	True

By dividing the data into batches, the training process becomes more manageable. The model is trained on each batch sequentially until all the batches within an epoch have been processed. This approach allows for efficient training on large datasets, as the model gradually learns from different subsets of data within each epoch.

3.4.3.6 Section Prediction :

Once the training process is completed, we will evaluate our model using the specified parameters, including the "repetition_penalty" set to 3.0. Subsequently, we can utilize the trained model to generate forecasts and construct a new section, as depicted in the accompanying Table 2.

TABLE 3.3 – Research studies conducted on Arabic text mining in terms of the Holy Quran[49]

Matn	Original Section	Mbert Predicted Section	AraT5 Predicted Section
إن الله عز وجل يبسط يده بالليل ليتوب مسيء النهار ويبسط يده بالنهار ليتوب مسيء الليل حتى تطلع الشمس من مغربها	قبول التوبة من الذنوب وإن تكررت الذنوب والتوبة	ما جاء في التوبة	من فضائل الله عز وجل
من بنى لله مسجدا بنى الله له مثله في الجنة	من بنى لله مسجدا	في بناء المساجد	ما جاء في بناء المساجد
سمعت النبي صلى الله عليه وسلم يقول إن في الليل لساعة لا يوافقها رجل مسلم يسأل الله خيرا من أمر الدنيا والآخرة إلا أعطاه إياه وذلك كل ليلة	في الليل ساعة مستجاب فيها الدعاء	الدليل على زيادة الإيمان	ما جاء في فضل صلاة الليل

3.5 Benefits Extraction :

The second part consists of Benefits Extraction, we presented two methods : the first one involves using beam search, and the second method utilizes text completion. We begin by discussing the approach using beam search for Benefits Extraction.

3.5.1 Beam search :

Beam search is a common technique used in natural language processing tasks to explore multiple potential solutions and select the most promising ones. In the context of Benefits Extraction from Hadiths, we employ beam search to predict three benefits for a given Hadith as shown in the figure In our case, we utilize two

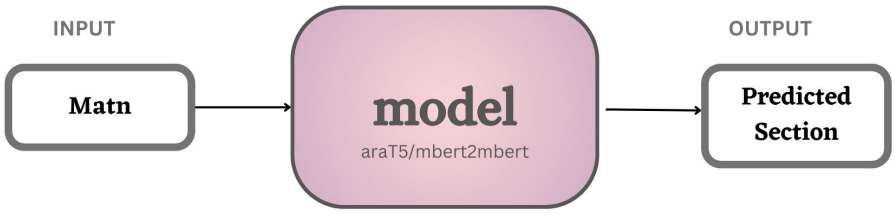

FIGURE 3.6 – Representation of Benefits extraction using araT5 and mbert2mbert model

transformer models : Arabic T5 (arat5) and mbert2mbert. These models have been trained on large-scale Arabic language data and are capable of generating meaningful text.

Here's a simplified schema that illustrates how beam search operates in selecting three benefits for a given Hadith :

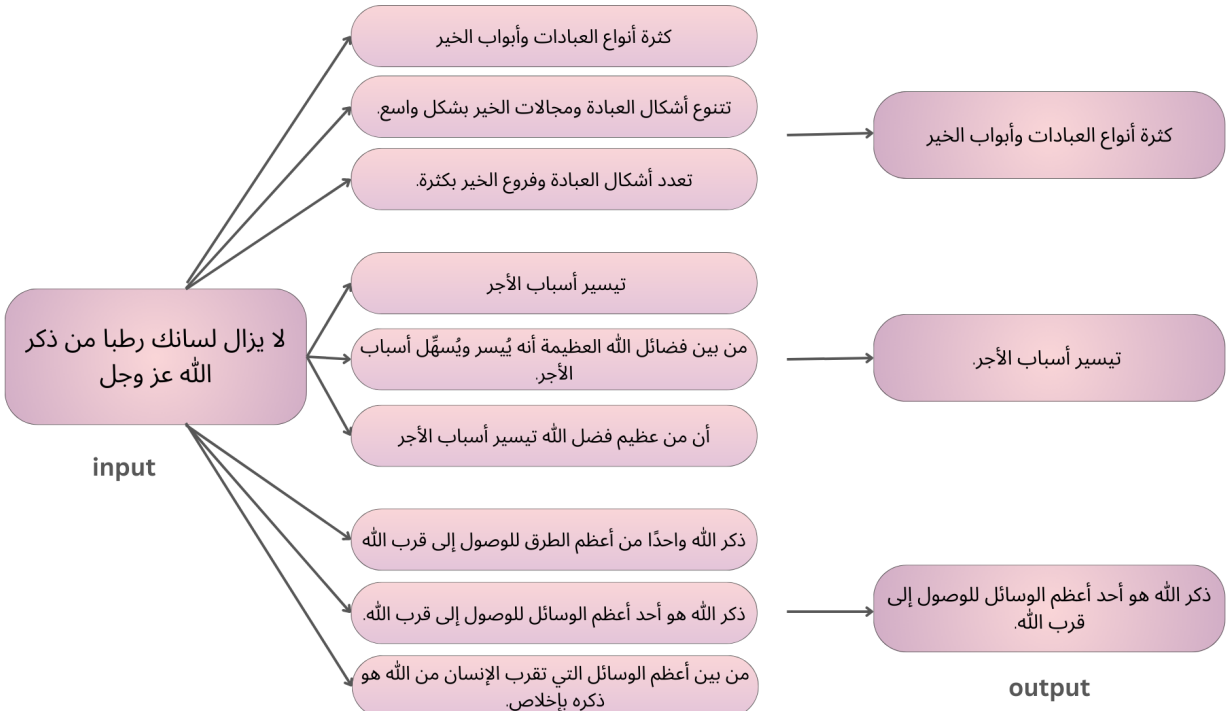

FIGURE 3.7 – Representation Beam search technique

By utilizing beam search with the Arabic T5 (arat5) and mbert2mbert models, we can effectively generate and select three relevant benefits for each Hadith. The beam search process allows for exploration of multiple possibilities, ensuring that diverse and meaningful benefits are extracted.

3.5.2 Processing Steps :

The process of benefits extraction using the Arabic T5 and mbert2mbert models with beam search involves several steps. Firstly, the data is preprocessed to ensure its cleanliness and suitability for further

analysis,followed by model training, and finally leading to the generation of predicted benefits.
The Arabic T5 (arat5) and mbert2mbert models used for beam search benefit extraction follow a transformer architecture. This architecture is composed of multiple transformer layers that incorporate self-attention mechanisms, enabling the models to capture contextual relationships and generate coherent and relevant text. The specific details of the model architecture, such as the number of layers, hidden dimensions, and attention heads, depend on the chosen transformer variant (arat5 or mbert2mbert).

3.5.2.1 Preprocessing :

Preprocessing is the initial step in benefits extraction where the data is prepared and cleaned for further analysis. This involves tasks such as data cleaning, normalization, and any necessary transformations to ensure the data is in a suitable format for training and testing.

3.5.2.2 Data Splitting :

The data will be divided into three parts : training, validation, and testing. The first part, comprising 70% of the corpus, will be used for training the learning model. The second part, which accounts for 20% of the corpus, will be set aside for testing the model's performance. Finally, the remaining 10% will be allocated for validation, enabling us to fine-tune and validate the model during the training process.

3.5.2.3 Model training :

The datasets for training the models in Benefits Extraction from Hadiths are appropriately divided into sizes suitable for each model. These divisions ensure efficient training and better utilization of the available data. The models undergo training over multiple epochs, which involves iterating through the entire dataset multiple times to improve their performance.
To maintain consistency and comparability, the same set of hyperparameters as in the previous models are used. Hyperparameters are adjustable parameters that determine the behavior and performance of the model during training. By keeping the hyperparameters consistent across models, it becomes easier to assess their relative effectiveness and make fair comparisons between them.

3.5.2.4 Benefits Prediction :

Once the models are trained, The beam search algorithm, when applied to generate benefits for new or unseen Hadiths, can be enhanced by setting the repetition penalty to three. This penalty discourages the model from repeating the same benefit multiple times in a sequence. During the beam search process,

candidate benefit sequences are explored and expanded, with the repetition penalty promoting diverse and unique outputs. The top-ranked sequences obtained from this process, considering the repetition penalty, are considered the predicted benefits for the given Hadith. Applying the repetition penalty helps generate more informative and varied results by reducing redundancy.

TABLE 3.4 – Representation of predicted benefits with mbert2mbert and AraT5 model

Matn	Reference Benefits	Mbert Predicted Benefits	AraT5 Predicted Benefits
ـ تسحروا فإن في السحور بركة	ـ سمي سحورا لأنه قرب السحر وكانوا يسمونه الغداء لأنه بدل منه ـ أن في السحور بركة دينة ودنيوية ـ السحور يعطي الطاقة اللازمة للصيام	ـ يذكر هذا الحديث أهمية تناول وجبة السحور قبل صيام الصبح ـ السحور يحصل به على ثواب من الله ـ سمي سحورا لأنه يقام قبل طلوع	ـ السحورفيه الطاقة للصيام ـ استحباب السحور ـ أن السحور ينبغي للمسلم

3.5.3 Text Completion :

In the second method, we utilized text completion, which is a technique that utilizes language models to generate text that is coherent and contextually suitable based on partial input. In this particular case, we employed the ARAGPT2 model for text completion to produce benefits associated with the provided Hadiths, ensuring that the resulting text is paraphrased.

FIGURE 3.8 – Representation of Benefits extraction using GPT2 model

3.5.3.1 araGPT2 Model

The ARAGPT2 model [12] is used for text completion in our project to generate benefits associated with the given Hadiths. ARAGPT2 is a stacked transformer-decoder model trained using the causal language modeling objective. It is trained on a large corpus of Arabic text, totaling 77GB.
ARAGPT2 comes in four different variants : base, medium, large, and xlarge. The base variant has the same size as ARABERT-base, making it accessible for a larger number of researchers. The larger model variants (medium, large, xlarge) offer improved performance but are harder to fine-tune and computationally more expensive.

3.5.4 Processing Steps :

Here are the process steps for the second method, which involves text completion using the ARAGPT2 model, as previously described :

3.5.4.1 Preprocessing :

As in the previous method, we begin by performing preprocessing on the collected data. This includes tasks such as text cleaning, handling special characters, and ensuring the data is in a suitable format for training and testing.

3.5.4.2 Data splitting :

In the next step, we split the collected data into training and test sets. Similar to the previous models, we allocate 70% of the data for training,additionally, we allocate an additional 10% for validation and the

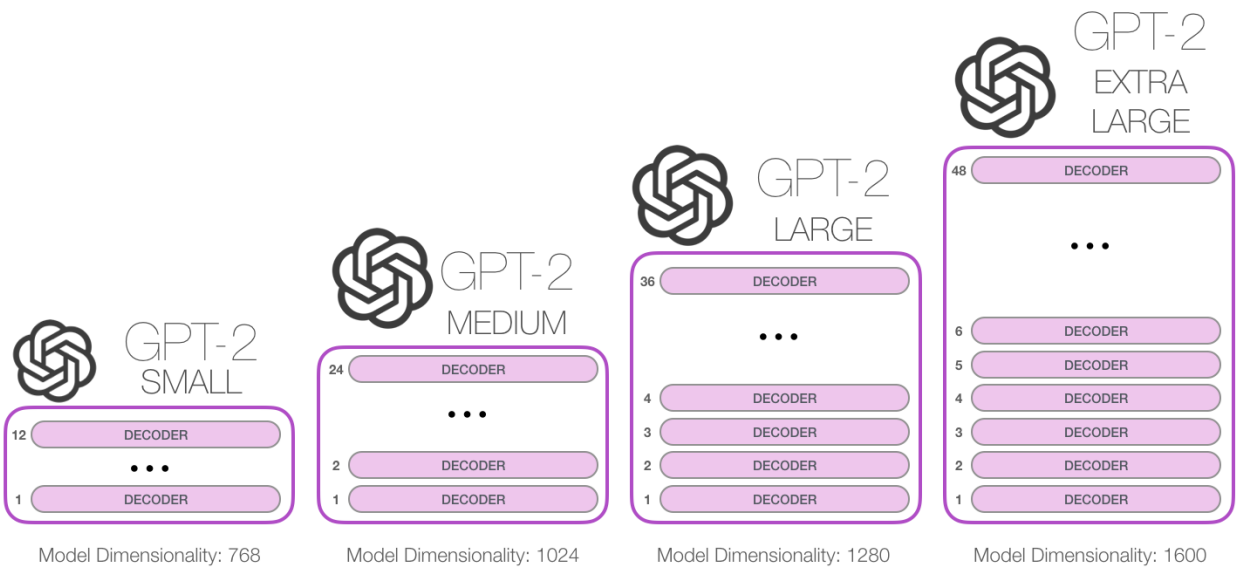

FIGURE 3.9 – Different variants of araGPT2 model

Learning rate	2e-5
weight_decay	0.01
batch_size	5
num_train_epochs	5

TABLE 3.5 – Basic parameters defined for Aragpt2 model

remaining 20% is reserved for testing and evaluating the model's performance.

3.5.4.3 Model training :

We followed a standard training procedure to train the ARAGPT2 model for text completion. The training began by initializing the model with pre-trained weights and then fine-tuning it specifically for generating benefits associated with the provided Hadiths. Our training dataset consisted of 70% of the collected data, and we used suitable batch sizes and adjusted the learning rate to ensure optimal convergence given our available computational resources. To ensure the model's performance and prevent overfitting, we monitored the training progress using evaluation metrics such as perplexity and accuracy on the validation set. These metrics guided us in making necessary adjustments to the training process, such as early stopping or learning rate scheduling. Additionally, due to time constraints and specific resource limitations, we decided to use only 5 instead of 10 epochs during the training process.
The following table summarizes the basic set of parameters defined for the model :

3.5.4.4 Benefits Prediction :

Section benefits : Using the trained ARAGPT2 model, we generated predictions for the benefits associated with the provided Hadiths. Given a partial input, the model leveraged its understanding of the Arabic

language and the contextual information from the training data to produce appropriate completions.
During the prediction phase, we passed the partial input, which consisted of the Hadith, to the ARAGPT2 model. The model then generated completions based on the learned patterns and semantic knowledge it acquired during training. To ensure diverse and non-repetitive output, we applied a repetition_penalty of 3.0, which helped in reducing repetitive patterns in the generated text. This parameter discourages the model from generating the same phrases or words multiple times, resulting in more varied and informative predictions.

TABLE 3.6 – Representation of benefits prediction with araGPT2 model

Hadith	Reference Benefits	Predicted Benefits
عن عائشة رضي الله عنها قالت كان النبي صلى الله عليه وسلم يذكر الله على أحيانه	عن عائشة رضي الله عنها قالت كان النبي صلى الله عليه وسلم يذكر الله على كل أحيانه : لا تشترط الطهارة من الحدث الأصغر والأكبر لذكر الله تعالى فيجوز للمسلم أن يسبح الله تعالى ويحمده ويهلله ويستغفره ويقرأ القرآن ما لم يكن جنبا لورود السنة بذلك ـ عموم الحديث يدل على أن للحائض والنفساء قراءة القرآن لكن من غير مس له بل من وراء حائل كالقفاز ونحوه ـ مداومة النبي صلى الله عليه وسلم لذكر الله تعالى	عن عائشة رضي الله عنها قالت كان النبي صلى الله عليه وسلم يذكر الله على كل أحيانهْ آعن أبي هريرة رضي الله عنه قال كان رسول الله صلى الله علية وسلم يذكر ربه على كل أوقاته : الحث على الإكثار من ذكر الله تعالى في كل الأوقات ـ الإكثار من الذكر والدعاء والتسبيح والتهليل والتضرع إلى الله عز وجل ـ الحث على قراءة القرآن الكريم والمحافظة على الأذكار الواردة فيه

Continued on the next page

Table 3.6 : Data Dictionary

Hadith	Reference Benefits	Predicted Benefits
لعن رسول الله آكل الربا وموكله وكاتبه وشاهديه	لعن رسول الله آكل الربا وموكله وكاتبه وشاهديه : يدخل في عموم الربا المحرم ربا القروض بأن يقرض شخص آخر مبلغا من المال على أن يردها المقترض بعد سنة ـ دل الحديث على تحريم كل عمل فيه معونة على الربا ككتابة عقد الربا أو الشهادة عليه ونحو ذلك ـ لا يجوز لمسلم أن يعمل في بنك تعامله بالربا ولو كان العمل الذي يتولاه ذلك المسلم في البنك غير ربوي لتوفيره لموظفيه الذين يعملون في الربويات ما يحتاجونه ويستعينون به على أعمالهم الربوية وقد قال تعالى ولا تعاونوا على الإثم والعدوان	لعن رسول الله آكل الربا وموكله وكاتبه وشاهديه : تحريم أكل الربا وأكل أموال الناس بالباطل ـ تحريم أكل المال بالباطل وإخباره بما فيه من الحرام كذب الكذب والإنكار على ما فيه من الغش والخداع والكذب في الدين ـ الحث على تعلم العلم والعمل به وعدم إنكاره أو الإنكار عليه

TABLE 3.7 – Representation of benefits prediction with araGPT2 model ,excluding the inclusion of hadiths

Hadith	Original Benefits	Predicted Benefits
عن عائشة رضي الله عنها قالت كان النبي صلى الله عليه وسلم يذكر الله على كل أحيانه	لا تشترط الطهارة من الحدث الأصغر والأكبر لذكر الله تعالى فيجوز للمسلم أن يسبح الله تعالى ويحمده ويهلله ويستغفره ويقرأ القرآن ما لم يكن جنبا لورود السنة بذلك ـ عموم الحديث يدل على أن للحائض والنفساء قراءة القرآن لكن من غير مس له بل من وراء حائل كالقفاز ونحوه ـ مداومة النبي صلى الله عليه وسلم لذكر الله تعالى	الحث على الإكثار من ذكر الله تعالى في كل الأوقات ـ الإكثار من الذكر والدعاء والتسبيح والتهليل والتضرع إلى الله عز وجل ـ الحث على قراءة القرآن الكريم والمحافظة على الأذكار الواردة فيه
لعن رسول الله آكل الربا وموكله وكاتبه وشاهديه	يدخل في عموم الربا المحرم ربا القروض بأن يقرض شخص آخر مبلغا من المال على أن يردها المقترض بعد سنة ـ دل الحديث على تحريم كل عمل فيه معونة على الربا ككتابة عقد الربا أو الشهادة عليه ونحو ذلك ـ لا يجوز لمسلم أن يعمل في بنك تعامله بالربا ولو كان العمل الذي يتولاه ذلك المسلم في البنك غير ربوي لتوفيره لموظفيه الذين يعملون في الربويات ما يحتاجونه ويستعينون به على أعمالهم الربوية وقد قال تعالى ولا تعاونوا على الإثم والعدوان	تحريم أكل الربا وأكل أموال الناس بالباطل ـ تحريم أكل المال بالباطل وإخباره بما فيه من الحرام كذب الكذب والإنكار على ما فيه من الغش والخداع والكذب في الدين ـ الحث على تعلم العلم والعمل به وعدم إنكاره أو الإنكار عليه

3.6 Conclusion :

In summary, this chapter has provided a concise overview of the implementation of transformer models in a deep abstractive approach. We have presented our significant contributions in the areas of the abstractive section and benefits extraction. The detailed implementation and evaluation of our proposal will be discussed in the next chapter.

Chapitre 4

Implementation and Evaluation

4.1 Introduction

n the previous chapter, we delved into the exploration of different models and conducted thorough research to uncover valuable insights relatively to the content of Prophetic Hadith. This chapter presents the findings obtained from the research process, highlighting the details and deductions made along the way. Additionally, proper acknowledgment is given to the resources that facilitated the entire research endeavor.

4.2 Software Configuration for implementation

To implement the different models, we utilized the Kaggle platform and Google Colab Pro, both of which are popular environments for data scientists and machine learning practitioners. These platforms offer a cloud-based environment that supports programming languages such as Python and R.

4.2.1 Kaggle platform

On the Kaggle platform, we benefited from various resources, including access to NVIDIA P100 TESLA GPUs, 13 GB of RAM, and storage space per project of 73 GB. However, it's important to note that each session on Kaggle is limited to 12 hours.

4.2.2 google colab pro

Google Colab Pro, on the other hand, provides 100 computing units per month. These computing units expire after 90 days, and additional units can be purchased as per the user's requirements. Google Colab Pro

also offers faster GPUs, the ability to access machines with higher memory, and the option to use a terminal with the connected virtual machine.

4.2.3 Language and libraries

— **Python :** For our implementation, we chose Python as the programming language due to its extensive range of libraries that greatly facilitate deep learning tasks. We made use of the following libraries :

— **Pandas :** A Python library that provides efficient data processing and analysis capabilities, offering convenient data structures and data manipulation functions.

— **NumPy :** A scientific computing library that specializes in matrix manipulation. It also includes linear algebra functionalities, making it a valuable tool for numerical operations. NumPy is an open-source project freely available for use with Python.

— **PyTorch :** It is a popular deep learning framework in Python. It provides a dynamic computational graph, making it efficient for training and inference in neural networks. PyTorch offers a range of tools and functionalities for building and training deep learning models, including automatic differentiation, GPU acceleration, and support for various neural network architectures. It is known for its flexibility and ease of use, allowing researchers and developers to quickly prototype and experiment with different neural network architectures.

By leveraging these libraries and the computational resources provided by Kaggle and Google Colab Pro, we were able to implement different models effectively and efficiently.

4.3 Experimental Results

To assess the quality of our proposition, we employed the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric for automated evaluation. The ROUGE metric is widely used in natural language processing, text generation and summarization tasks. ROUGE evaluation concerns Hadith Section abstraction and Hadith Benefits extraction.
In our evaluation, we focused on four specific ROUGE scores : ROUGE-1, ROUGE-2, ROUGE-L. These scores allow us to evaluate different aspects of the generated information (Predicted Hadith sections, Predicted Hadith benefits) :

— **ROUGE-1 :** This score measures the overlap of 1-gram (individual words) between the reference text and the system-generated text. It provides insights into the informativeness of the generated text in capturing the key information from the reference.

— **ROUGE-2 :** This score focuses on bi-grams, which are pairs of consecutive words. It calculates the overlap of bi-grams between the system-generated text and the reference text. ROUGE-2 helps us assess how well the generated text captures the phrase-level information present in the reference.

— **ROUGE-L :** This score evaluates the longest common subsequence (LCS) between the system-generated text and the reference text. It measures the lengthiest word sequence that matches between the two texts. ROUGE-L provides insights into the fluency and coherence of the generated text.

The following example (Table 1) shows how to calculate the recall score of the three ROUGE metrics.
- Reference sentence : في السحور بركة دينية ودنيوية (In the pre-dawn meal, there is religious and worldly blessings)
-Candidate sentence (to be evaluated) : استحباب السحور لأن في السحور بركة (The recommendation of having the pre-dawn meal is because there is blessings in it)

Rouge metrics	Ref_Sentence	Cand_Sentence	Rouge Score
ROUGE-1	في ـ السحور ـ بركة ـ دينية ـ و ـ دنيوية	استحباب ـ السحور ـ لان ـ في ـ السحور ـ بركة	2/6
ROUGE-2	في السحور ـ السحور بركة ـ بركة دينية دينية و ـ و دنيوية	استحباب السحور ـ السحور لان ـ لان في ـ في السحور ـ السحور بركة	1/5
ROUGE-L	في السحور بركة دينية ودنيوية	السحور بركة	2/6

Table 4.1 – An example of how to calculate ROUGE metric

4.4 Section Abstraction Results

In this section, we provide an overview of the experimental results and evaluation of two models : AraT5 and mbert2mbert relatively to the Prophetic Hadith section abstraction. We evaluate the performance of these models based on various metrics, including training and validation loss, as well as ROUGE scores.The training loss is a measure of the disparity between the predicted and actual values during the training phase. It indicates how well the model is fitting the training data. On the other hand, the validation loss assesses the model's ability to generalize its predictions to unseen data. It provides insights into the model's performance on new, unseen examples.

4.4.1 AraT5 Training results

4.4.1.1 Training and validation loss

To track the model's learning progress, we recorded and plotted the training and validation loss values after each epoch.

This graph (Figure 1) allows us to analyze how the model's performance evolves over the training process and assess its convergence and generalization capabilities. The graph represents the training and validation

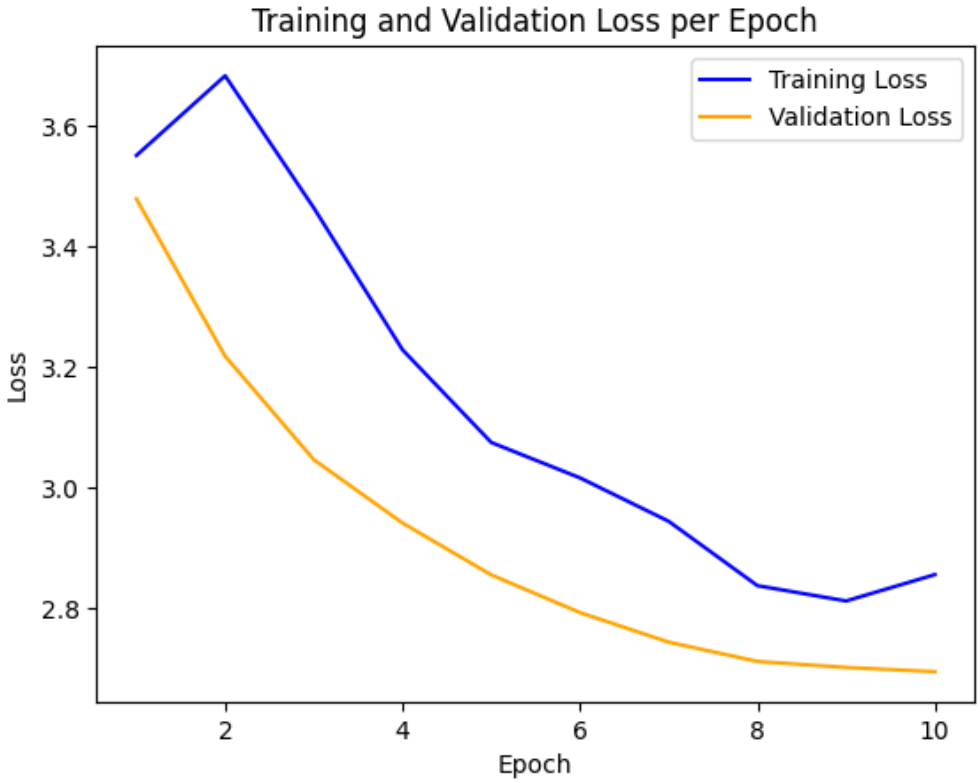

FIGURE 4.1 – Training and validation loss of Arat5

loss per epoch for the AraT5 training model. The x-axis corresponds to the number of epochs, while the y-axis represents the loss value.. The training loss steadily decreases from an initial value of 3.55, indicating that the model is effectively learning and refining its performance over time. Similarly, the validation loss decreases with each epoch, suggesting that the model is generalizing well to unseen data. The fact that the training loss consistently outperforms the validation loss is expected since the model is optimized based on the training data it has already seen. Both loss trends show convergence, indicating improved performance with additional epochs, although it is crucial to monitor these values to prevent overfitting. Overall, the graph reflects the positive progress of the AraT5 training model, showcasing its ability to learn and generalize effectively.

4.4.1.2 AraT5 Training Rouge score

In addition to evaluating the training and validation loss, we also assessed the performance of the AraT5 model using ROUGE scores The graph illustrates the ROUGE scores per epoch for the implemented model,

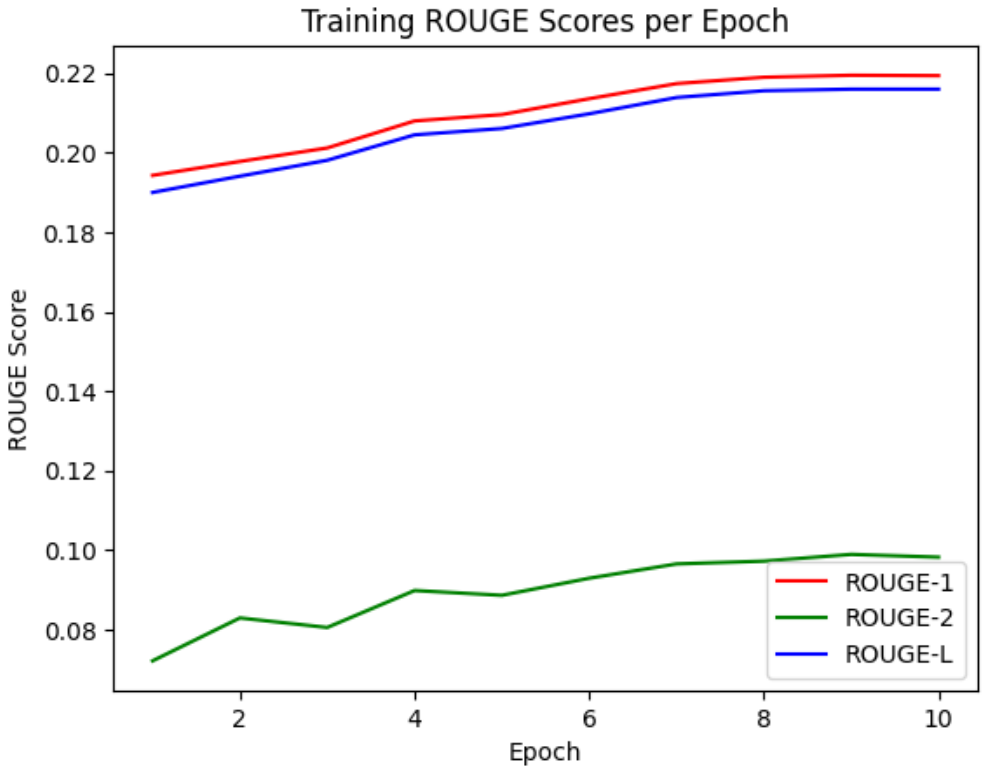

FIGURE 4.2 – Training Rouge score metric of araT5

which evaluates the quality of generated summaries. The ROUGE-1 scores, represented by the red bars, gradually increase from 0.1943 at Epoch 1 to 0.2194 at Epoch 10, indicating the model's improvement in capturing unigrams. Similarly, the ROUGE-2 scores, depicted by the green bars, show an increasing trend from 0.0721 to 0.102, highlighting the model's ability to capture bigrams. The blue bars represent the ROUGE-L scores, which exhibit a similar upward trend from 0.19 to 0.216, reflecting the model's improvement in capturing overall semantic meaning. Overall, the graph demonstrates that the implemented model steadily improves its ROUGE scores over the training process, indicating its ability to generate section titles that closely resemble the reference sections in terms of individual words and longer phrases.

4.4.2 mbert2mbert Training results

4.4.2.1 Training and validation loss :

Similarly, we also implemented the mbert2mbert model and evaluated its performance using training and validation loss as well as ROUGE scores. The graph represents the training and validation loss per epoch for

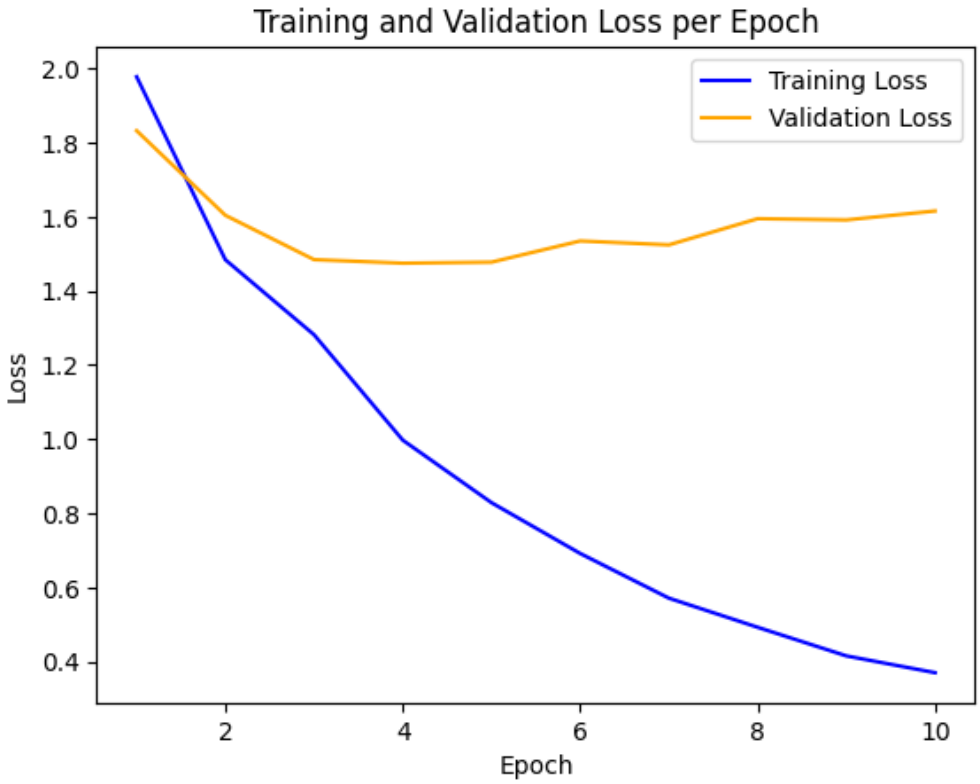

FIGURE 4.3 – training and validation loss of mbert2mbert

the mbert2mbert training model. The training loss starts at 1.9773 and consistently decreases with each epoch, reaching 0.3703 at Epoch 10. The validation loss follows a similar decreasing trend, indicating the model's ability to generalize to unseen data. The convergence of both loss curves suggests improved performance and effective learning of the mbert2mbert model.

4.4.2.2 mbert2mbert Training ROUGE SCORE

We evaluated the performance of the mbert2mbert model using ROUGE scores, which measure the quality of the generated summaries. The graph illustrates the ROUGE scores per epoch for three different metrics : ROUGE-1, ROUGE-2, and ROUGE-L.

Upon analyzing the graph, it is evident that the ROUGE scores exhibit a consistent pattern as the number of epochs increases. Initially, at epoch 1, all three ROUGE scoresROUGE-1, ROUGE-2, and ROUGE-Lstart at relatively low values. However, as the number of epochs progresses, the scores gradually increase.

Specifically, for ROUGE-1, the score begins at 0.0167 and steadily rises, with some fluctuations, until it reaches 0.0169 at epoch 10. Similarly, ROUGE-2 starts at 0.0127 and shows an upward trend, albeit with slight variations. As for ROUGE-L, it commences at 0.0168 and demonstrates a consistent upward trajectory, with minor fluctuations.

Therefore, the graph indicates that the ROUGE scores consistently increase as the number of epochs

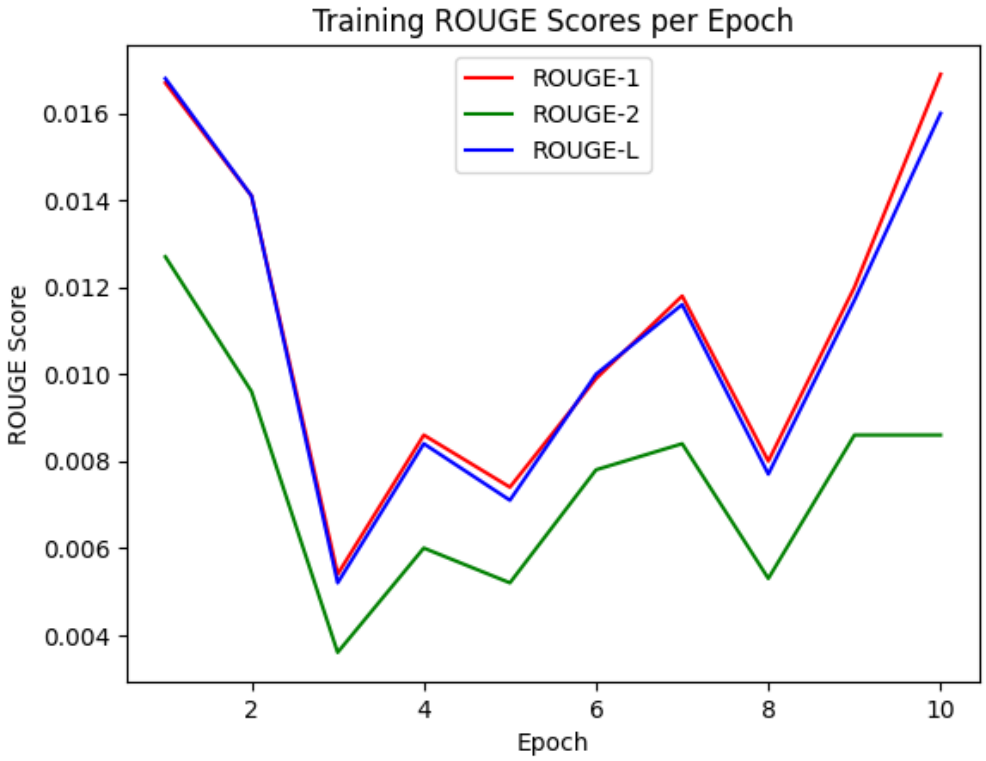

FIGURE 4.4 – Training Rouge Score metric of mbert2mbert

increases, implying improved model performance with extended training.

Both models show improvements in performance over the training process, as indicated by decreasing loss values and increasing ROUGE scores. However, the AraT5 model generally outperforms the mbert2mbert model in terms of both loss and ROUGE scores, suggesting its superior performance in generating sentences that closely resemble the reference sentences.

4.4.3 Section abstraction Test results

Figure 2 presents the test ROUGE scores for two models : AraT5 and mbert2mbert. The scores include ROUGE-1, ROUGE-2, and ROUGE-L, which assess the models' performance in capturing different aspects of section abstraction quality. The evaluation consists of a comparison between the predicted sections and the reference sections of the test set Hadiths. The test ROUGE scores reveal that the AraT5 model outperforms the mbert2mbert model in generating high-quality Section. The AraT5 model achieves higher scores across all three ROUGE metrics : ROUGE-1 (20.28 vs. 18.92), ROUGE-2 (10.66 vs. 8.6), and ROUGE-L (20.0 vs. 18.5). These results indicate that the AraT5 model excels in capturing unigrams, bigrams, and the overall semantic meaning of the reference sections , demonstrating its superior performance compared to the mbert2mbert model.

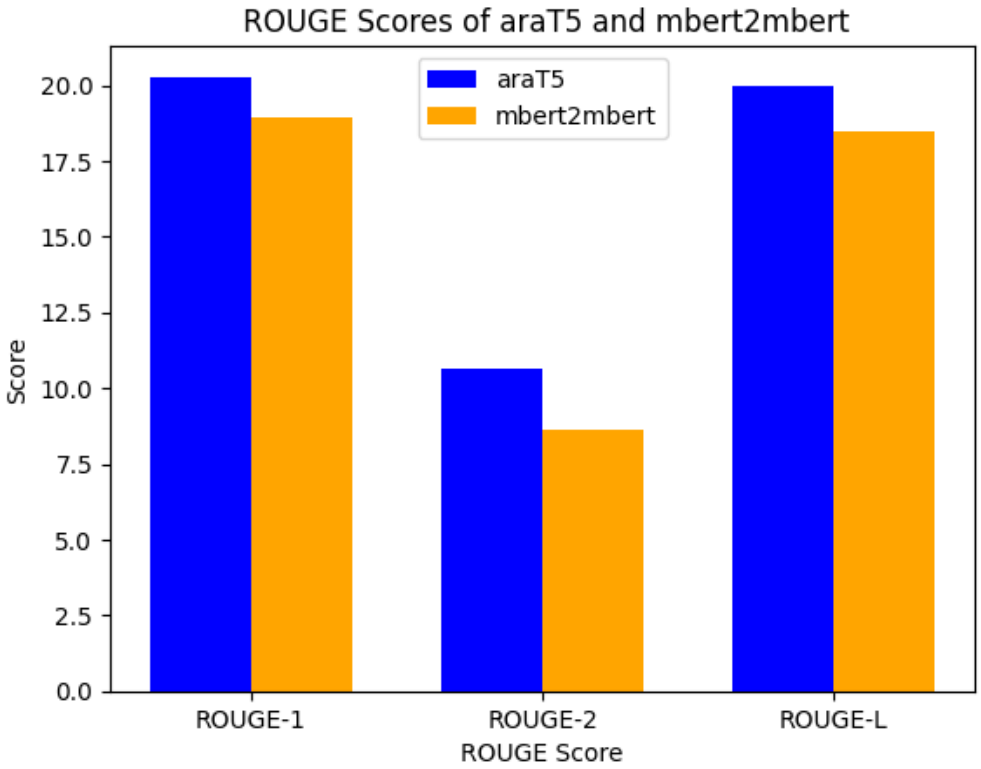

FIGURE 4.5 – ROUGE scores of araT5 and mbert2mbert model

4.5 Benefits Extraction Results

In this section, we discussed the obtained results for the two methods we have used for benefits extraction : Beam search and Text completion. Beam search concerns the two models AraT5 and mbert2mbert while text completion concerns AraGPT2 model.

4.5.1 Beam search

For beam search, we explored two techniques. First, we separated the three benefits of a hadith into different lines and generated all possible permutations of predicted benefits in the test set (Table 2)

Table 4.2 – All possible permutations of predicted benefits

Hadith	Benefits	Predicted Benefits	Permutations
ـ قال رسول الله صلى الله عليه وسلم اقتلوا الأسودين في الصلاة الحية والعقرب	ـ استحباب قتل الحية والعقرب في الصلاة أن هذه حركة قليلة محمودة فلا تبطل الصلاة ولا تنقصها ولو لم تكن من مصلحة الصلاة مشروعية قتل كل مؤذ من الهوام وغيرها في الصلاة أو خارجها فإنه إذا استحب قتل هذه الفواسق في الصلاة فقتلها خارجها يكون أولى	ـ النهي عن ما تبطل الصلاة ـ الحث على قتل الحية والعقرب في الصلاة ـ إيقاف ما يبطل الصلاة	— Predicted Benefit 1 — predicted Benefit 2 — predicted Benefit 3

Continued on the next page

Table 4.2 : Data Dictionary

Hadith	Benefits	Predicted Benefits	Permutations
ـ قال رسول الله صلى الله عليه وسلم اقتلوا الأسودين في الصلاة الحية والعقرب	ـ استحباب قتل الحية والعقرب في الصلاة أن هذه حركة قليلة محمودة فلا تبطل الصلاة ولا تنقصها ولو لم تكن من مصلحة الصلاة مشروعية قتل كل مؤذ من الهوام وغيرها في الصلاة أو خارجها فإنه إذا استحب قتل هذه الفواسق في الصلاة فقتلها خارجها يكون أولى	ـ النهي عن ما تبطل الصلاة ـ إيقاف ما يبطل الصلاة ـ الحث على قتل الحية والعقرب في الصلاة	— Predicted Benefit 1 — predicted Benefit 3 — predicted Benefit 2

Continued on the next page

Table 4.2 : Data Dictionary

Hadith	Benefits	Predicted Benefits	Permutations
ـ قال رسول الله صلى الله عليه وسلم اقتلوا الأسودين في الصلاة الحية والعقرب	ـ استحباب قتل الحية والعقرب في الصلاة أن هذه حركة قليلة محمودة فلا تبطل الصلاة ولا تنقصها ولو لم تكن من مصلحة الصلاة مشروعية قتل كل مؤذ من الهوام وغيرها في الصلاة أو خارجها فإنه إذا استحب قتل هذه الفواسق في الصلاة فقتلها خارجها يكون أولى	ـ الحث على قتل الحية والعقرب في الصلاة ـ النهي عن ما تبطل الصلاة ـ إيقاف ما يبطل الصلاة	— Predicted Benefit 2 — predicted Benefit 1 — predicted Benefit 3

Continued on the next page

Table 4.2 : Data Dictionary

Hadith	Benefits	Predicted Benefits	Permutations
ـ قال رسول الله صلى الله عليه وسلم اقتلوا الأسودين في الصلاة الحية والعقرب	ـ استحباب قتل الحية والعقرب في الصلاة أن هذه حركة قليلة محمودة فلا تبطل الصلاة ولا تنقصها ولو لم تكن من مصلحة الصلاة مشروعية قتل كل مؤذ من الهوام وغيرها في الصلاة أو خارجها فإنه إذا استحب قتل هذه الفواسق في الصلاة فقتلها خارجها يكون أولى	ـ الحث على قتل الحية والعقرب في الصلاة ـ إيقاف ما يبطل الصلاة ـ النهي عن ما تبطل الصلاة	— Predicted Benefit 2 — predicted Benefit 3 — predicted Benefit 1

Continued on the next page

Table 4.2 : Data Dictionary

Hadith	Benefits	Predicted Benefits	Permutations
ـ قال رسول الله صلى الله عليه وسلم اقتلوا الأسودين في الصلاة الحية والعقرب	ـ استحباب قتل الحية والعقرب في الصلاة أن هذه حركة قليلة محمودة فلا تبطل الصلاة ولا تنقصها ولو لم تكن من مصلحة الصلاة مشروعية قتل كل مؤذ من الهوام وغيرها في الصلاة أو خارجها فإنه إذا استحب قتل هذه الفواسق في الصلاة فقتلها خارجها يكون أولى	ـ إيقاف ما يبطل الصلاة ـ النهي عن ما تبطل الصلاة ـ الحث على قتل الحية والعقرب في الصلاة	— Predicted Benefit 3 — predicted Benefit 1 — predicted Benefit 2

Continued on the next page

Table 4.2 : Data Dictionary

Hadith	Benefits	Predicted Benefits	Permutations
ـ قال رسول الله صلى الله عليه وسلم اقتلوا الأسودين في الصلاة الحية والعقرب	ـ استحباب قتل الحية والعقرب في الصلاة أن هذه حركة قليلة محمودة فلا تبطل الصلاة ولا تنقصها ولو لم تكن من مصلحة الصلاة مشروعية قتل كل مؤذ من الهوام وغيرها في الصلاة أو خارجها فإنه إذا استحب قتل هذه الفواسق في الصلاة فقتلها خارجها يكون أولى	ـ إيقاف ما يبطل الصلاة ـ الحث على قتل الحية والعقرب في الصلاة ـ النهي عن ما تبطل الصلاة	— Predicted Benefit 3 — predicted Benefit 2 — predicted Benefit 1

Second, we concatenated the three benefits into the same line (Table3). We evaluated the results using the ROUGE metric.

Table 4.3 – Concatenation of three benefits

Hadith	Benefits	Predicted Benefits
Continued on the next page		

Table 4.3 : Data Dictionary

Hadith	Benefits	Predicted Benefits
ـ قال رسول الله صلى الله عليه وسلم اقتلوا الأسودين في الصلاة الحية والعقرب	استحباب قتل الحية والعقرب في الصلاة ـ أن هذه حركة قليلة محمودة فلا تبطل الصلاة ولا تنقصها ولو لم تكن من مصلحة الصلاة ـ مشروعية قتل كل مؤذ من الهوام وغيرها في الصلاة أو خارجها فإنه إذا استحب قتل هذه الفواسق في الصلاة فقتلها خارجها يكون أولى	النبي صلى الله عليه وسلم ـ يدعو إلى قتل الأسودين في الصلاه جواز قتل الحية ـ جواز قتل والعقرب

4.5.1.1 AraT5 Training results

Training and validation loss

The graph below illustrates the training and validation loss per epoch for the AraT5 training model :

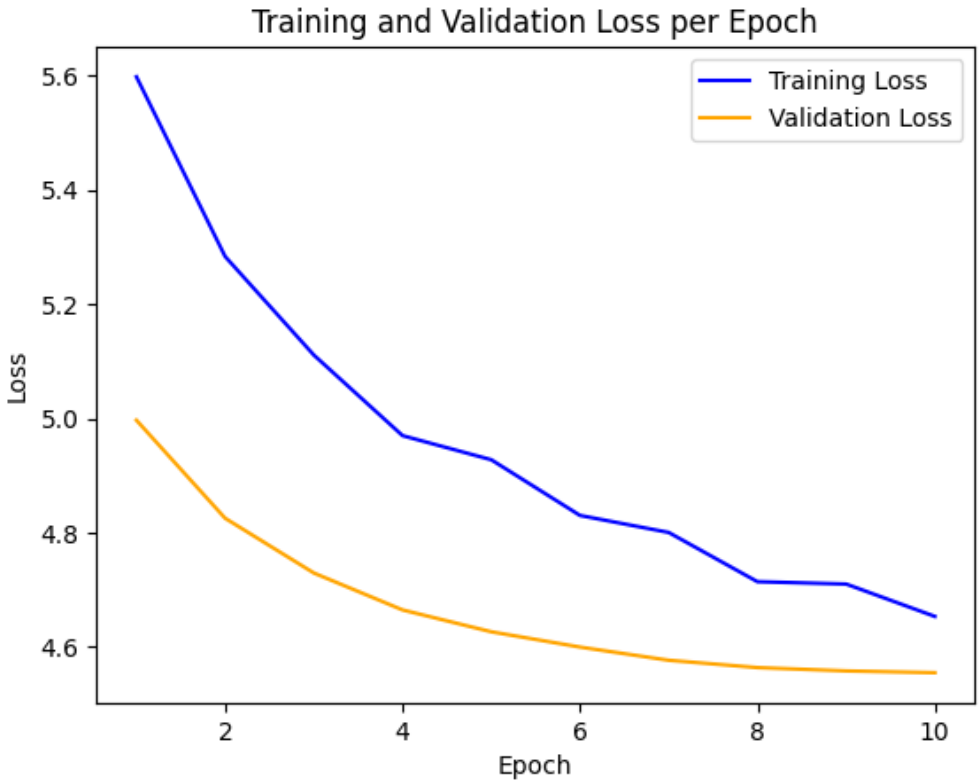

FIGURE 4.6 – Training and validation loss of AraT5

The training and validation loss per epoch for the T5 model are shown in the graph. The training loss steadily decreases from 5.598 to 4.6532, indicating the model's learning and improvement over time. Similarly, the validation loss shows a decreasing trend from 4.9968 to 4.5547, demonstrating the model's ability to generalize to unseen data. The convergence of both loss curves suggests improved performance with additional training epochs, although it is essential to monitor for signs of overfitting. Overall, the graph illustrates the positive progress of the T5 model in extracting benefits from the Hadith.

AraT5 Training ROUGE score

Apart from training and validation loss, we also evaluate the performance of both models using ROUGE scores.

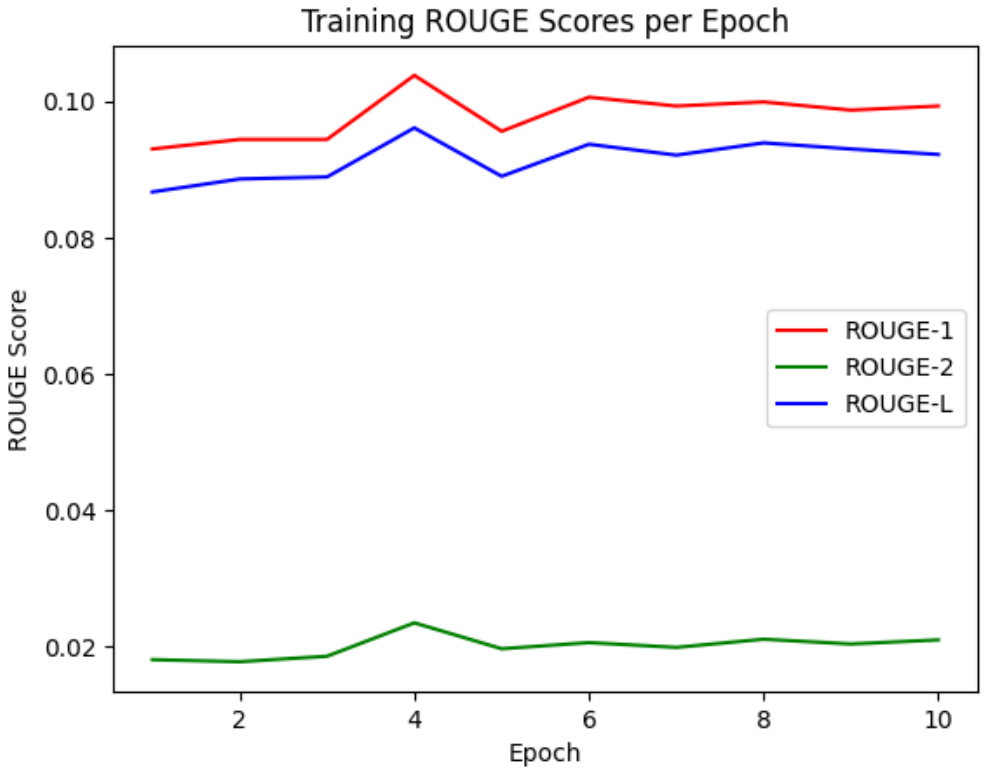

FIGURE 4.7 – Training rouge score of AraT5

The ROUGE scores per epoch for the T5 model are displayed in the graph. The ROUGE-1 scores range from 0.093 to 0.0993, showing an increasing trend. The ROUGE-2 scores range from 0.0181 to 0.021, indicating the model's ability to capture bigrams. The ROUGE-L scores range from 0.0867 to 0.0922, demonstrating the model's improvement in capturing the overall semantic meaning of the benefits. The upward trends in all three ROUGE metrics signify the T5 model's progress in predicted benefits that closely resemble the reference benefits in terms of unigrams, bigrams, and semantic coherence.

4.5.1.2 mbert2mbert Training results

Training and validation loss

Similarly, we tracked the training and validation loss for the mbert2mbert model. The graph below represents the loss values per epoch :

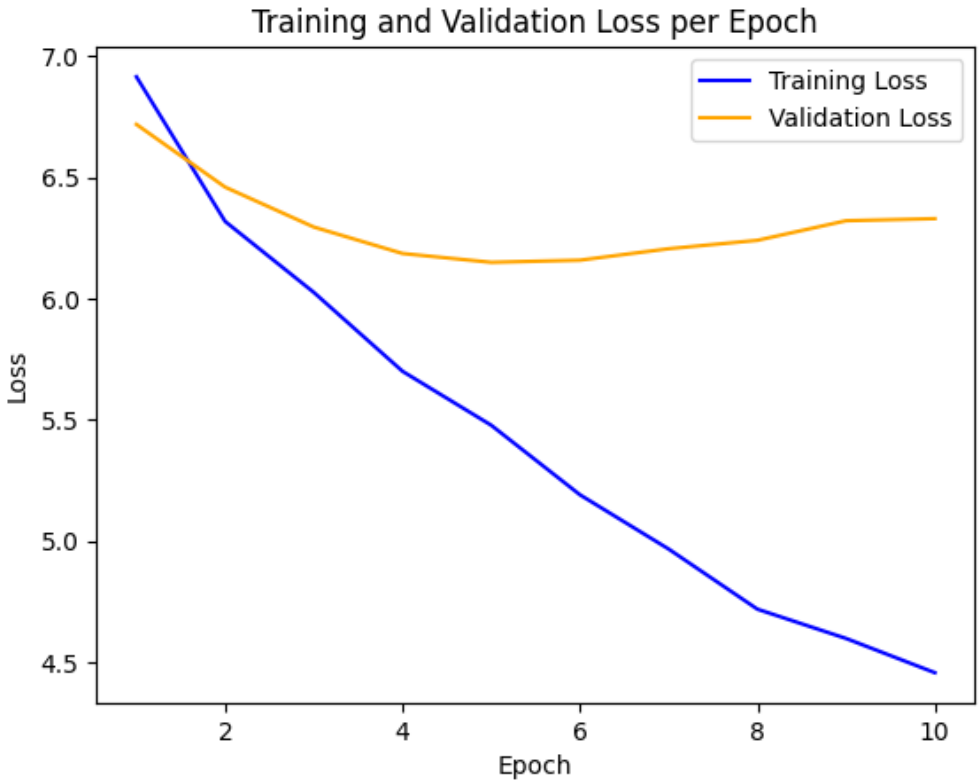

FIGURE 4.8 – Training and validation loss of mbert2mbert

For the mBERT2mBERT model, the training and validation loss per epoch are displayed in the first graph. The training loss steadily decreases from 6.9149 to 4.456, indicating the model's learning and improvement over time. Similarly, the validation loss shows a decreasing trend from 6.7184 to 6.3294, suggesting that the model can generalize well to unseen data. The convergence of the loss curves indicates improved performance with additional training epochs.

mbert2mbert Training ROUGE score

In addition to monitoring training and validation loss, we assess the performance of both models using ROUGE scores The second graph presents the ROUGE scores per epoch for the mBERT2mBERT model. The ROUGE-1 scores indicate an increasing trend, ranging from 0.0436 to 0.0791. This suggests that the model improves in generating benefits that closely match the reference benefits in terms of individual words. The ROUGE-2 scores range from 0.0014 to 0.0149, indicating the model's ability to capture bigrams (pairs of consecutive words). The ROUGE-L scores range from 0.0405 to 0.0675, demonstrating the model's enhancement in capturing the overall semantic meaning of the benefits.

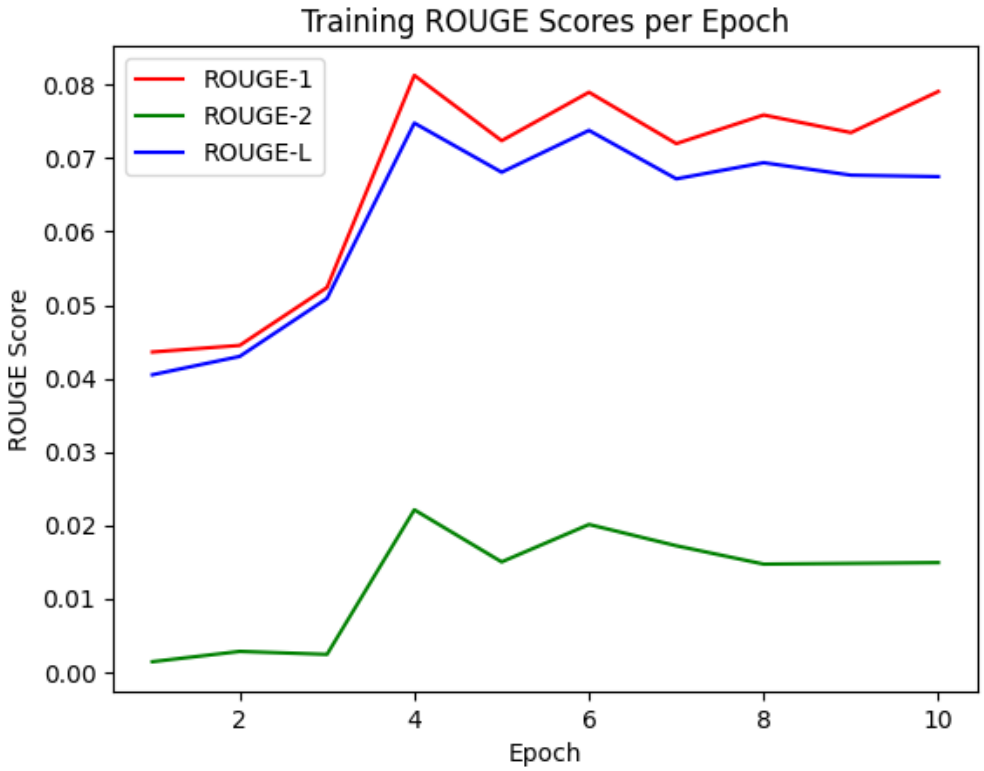

FIGURE 4.9 – Training rouge score of mbert2mbert

4.5.1.3 Beam search test results

We notice that the benefits generation is done accordingly to a single Hadith matn. For which, triple benefits are generated, however, we cant do an exact dependency between a reference benefit and a predicted one, as a result, and in order to do a worthy evaluation we have proposed two ways to calculate the Rouge scores, by permutation and by concatenation. In the permutation technique, a set of benefits permutations orders for each matn is used. Consequently, we have six permutations for each matn (table 2).
In the concatenation technique, we have concatenated the three benefits to form a single one text (table3).
The ROUGE scores achieved by the AraT5 and mbert2mbert models are as follows :

Benefits permutation

— **Permutation 1 :** In this order, the AraT5 model consistently achieved higher scores compared to the mbert2mbert model in all three ROUGE metrics (ROUGE-1, ROUGE-2, and ROUGE-L). This suggests that AraT5 performs better in capturing the similarity between the predicted benefits and the reference benefits for this specific order of benefits.

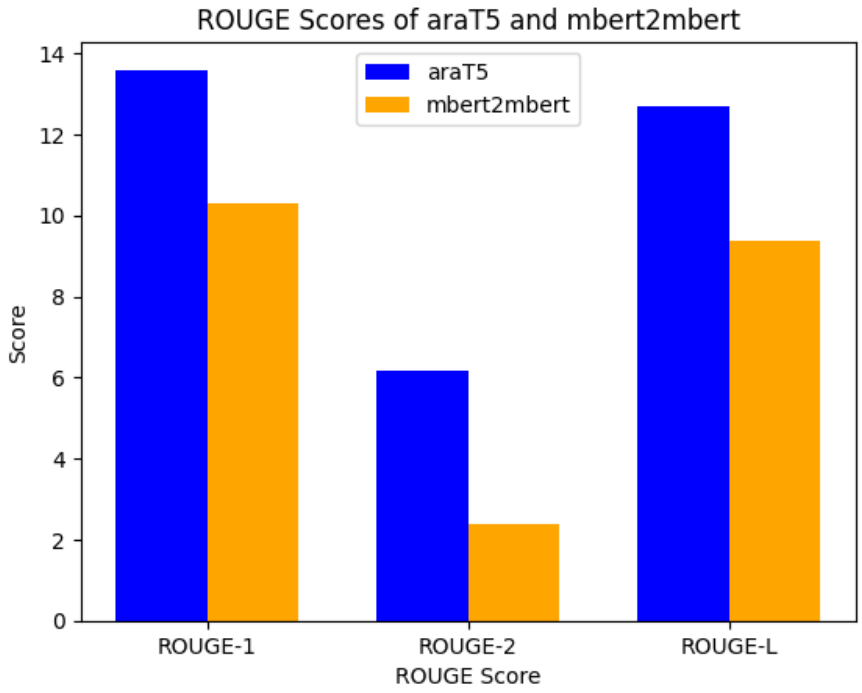

FIGURE 4.10 – ROUGE Scores for Permutation 1

— **Permutation 2 :** The ROUGE scores for both models decreased compared to Table 1. However, AraT5 still outperformed mbert2mbert in terms of all three metrics..

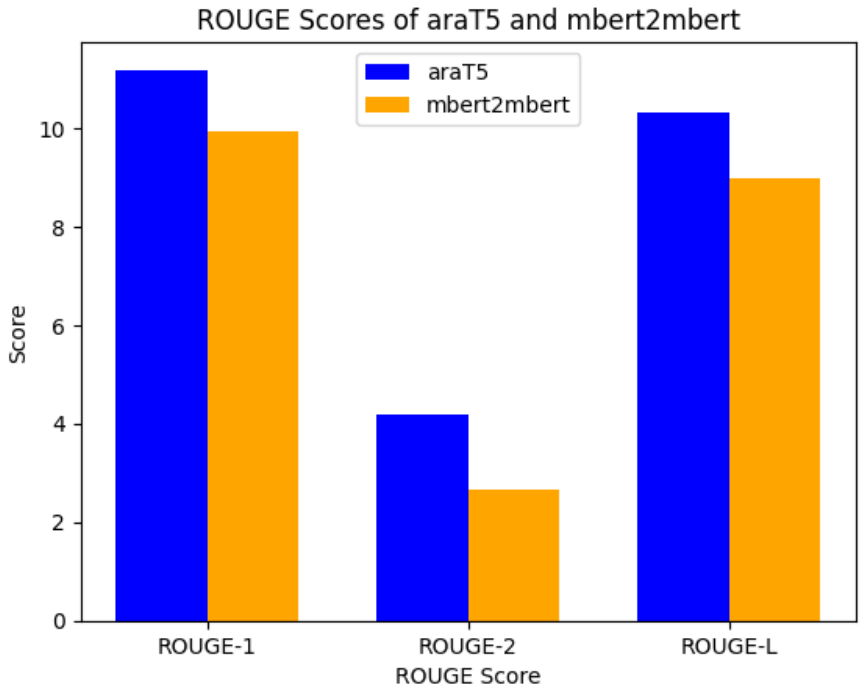

FIGURE 4.11 – ROUGE Scores for Permutation 2

— **Permutation 3 :** AraT5 achieved higher ROUGE scores compared to mbert2mbert across all metrics. However, the scores for both models decreased compared to Figure 1.

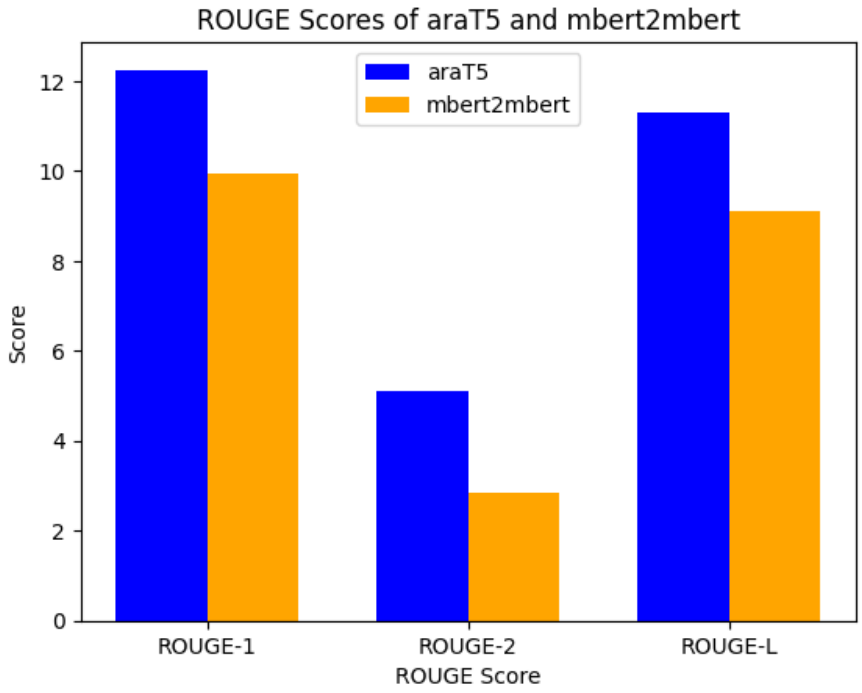

FIGURE 4.12 – ROUGE Scores for Permutation 3

— **Permutation 4 :** Both AraT5 and mbert2mbert models showed lower ROUGE scores compared to the previous tables. AraT5 performed slightly better than mbert2mbert in terms of all three metrics

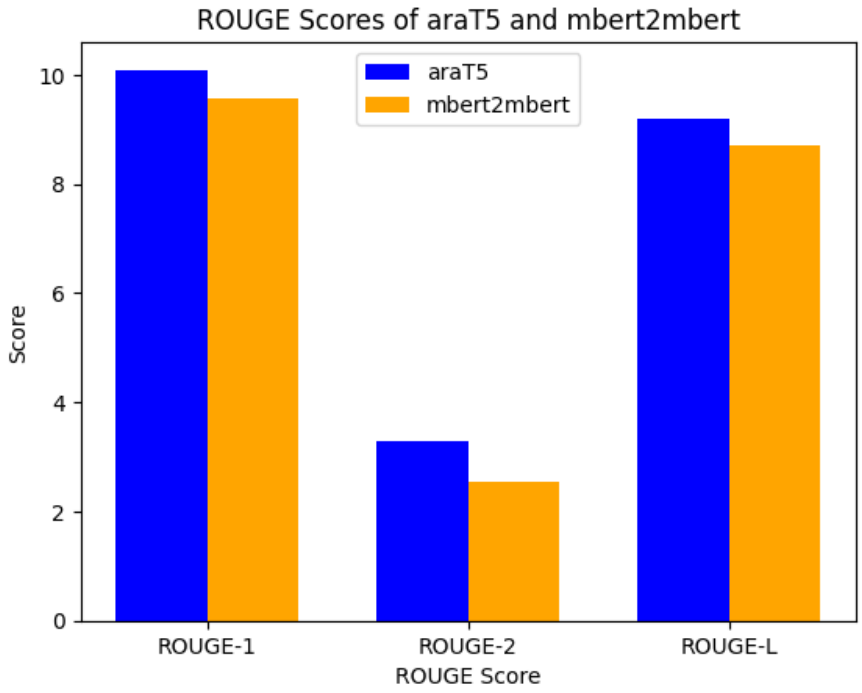

FIGURE 4.13 – ROUGE Scores for Permutation 4

— **Permutation 5 :** The ROUGE scores further decreased for both models compared to the previous tables. AraT5 consistently outperformed mbert2mbert in terms of all three metrics.

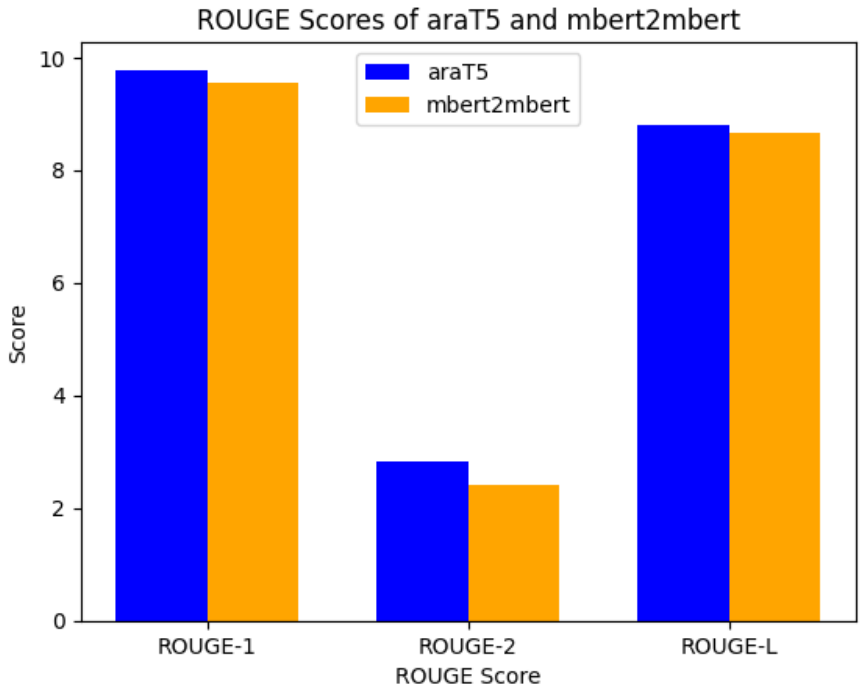

FIGURE 4.14 – ROUGE Scores for Permutation 5

— **Permutation 6** The ROUGE scores were similar to Table 5, with both models achieving relatively low scores. AraT5 consistently performed better than mbert2mbert across all metrics.

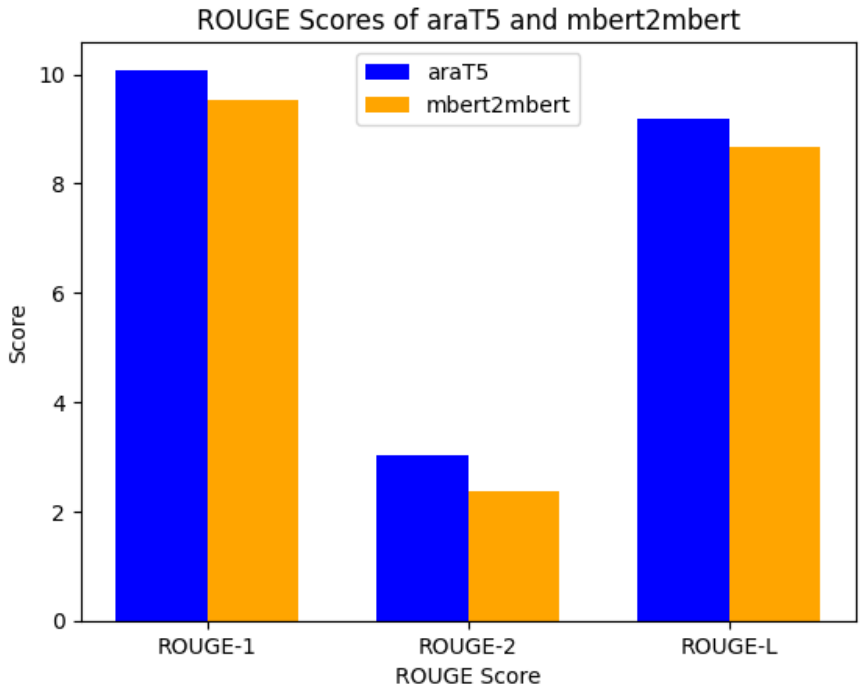

FIGURE 4.15 – ROUGE Scores for Permutation 6

Overall, Permutation 1 showed better performance than other orders for both models in terms of ROUGE scores.

Benefits concatenation

In the concatenation technique, all three benefits of the hadith are concatenated into a single line. The benefits extraction results using the two models, AraT5 and mbert2mbert, are as follows :

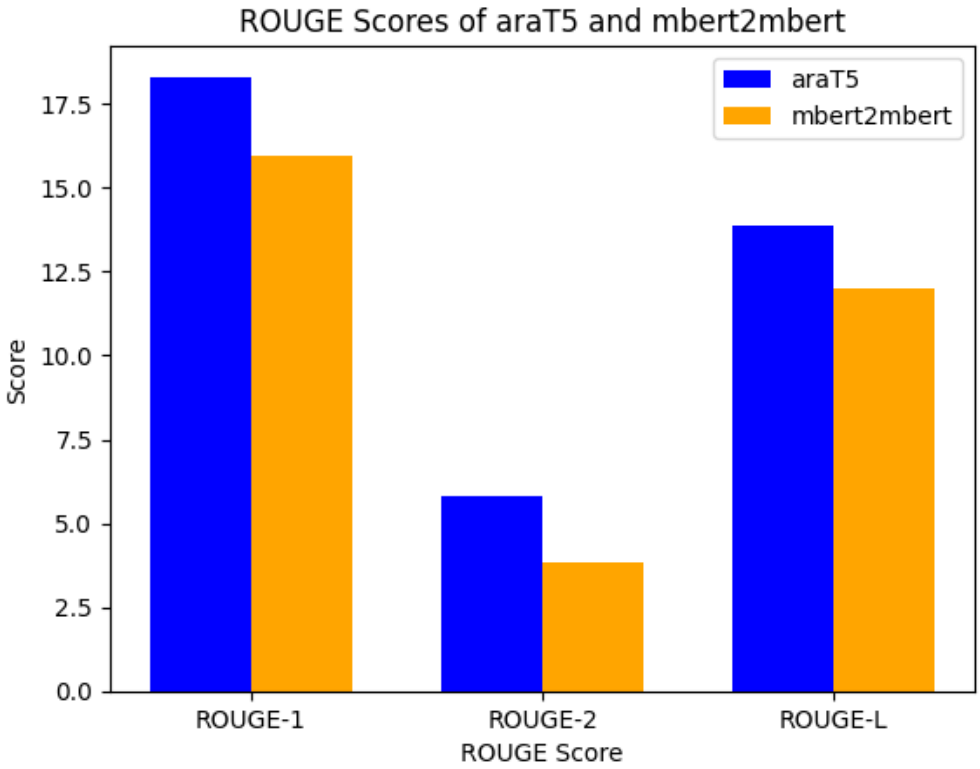

FIGURE 4.16 – ROUGE Scores for concatenation technique

The results demonstrate that the concatenation technique yields higher ROUGE scores compared to the beam search approach. Both AraT5 and mbert2mbert models achieve improved scores across all ROUGE metrics, indicating the effectiveness of concatenating benefits into a single line for extraction.

4.5.2 Text Completion

In the second method, we employed the AraGPT2 model for text completion. The aim was to evaluate how well the model performed in extracting benefits from Prophetic Hadiths. To gauge the effectiveness of the model, we analyzed various metrics, such as training and validation loss, along with ROUGE scores.

4.5.2.1 AraGPT2 Training results

Training and validation loss

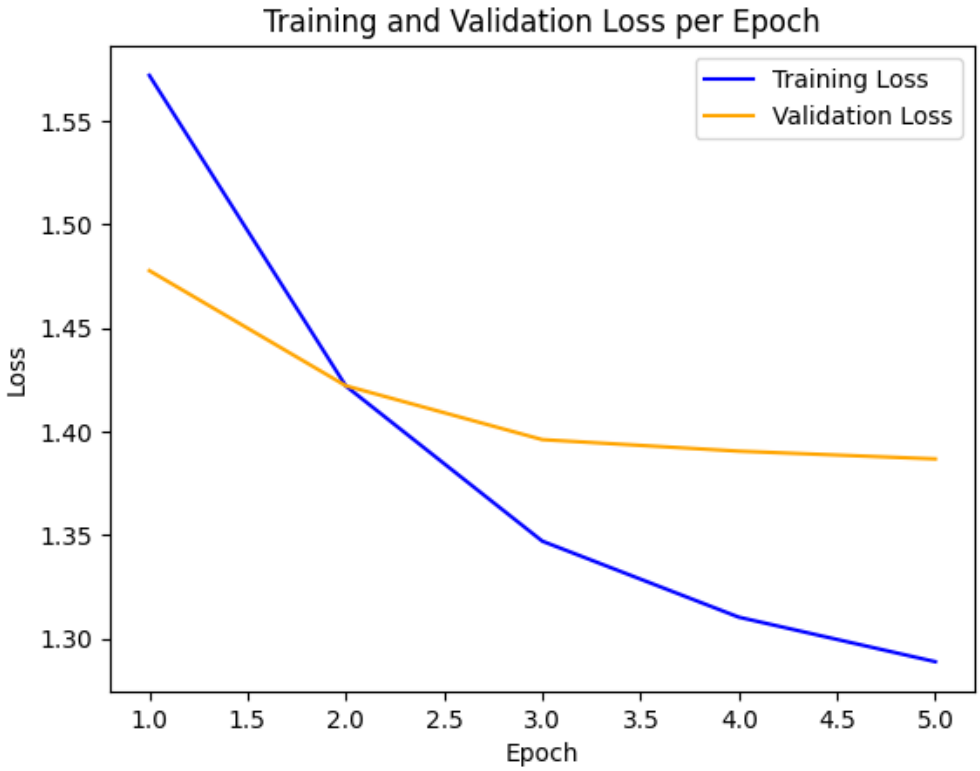

FIGURE 4.17 – Training and validation loss of AraGPT2

In the given graph, the training and validation loss per epoch demonstrate a decreasing trend, indicating that the model continues to learn and improve over time. It is worth noting that although the graph only includes five epochs, there is potential for further improvement by increasing the number of epochs. However, due to limited space and time constraints, the training was stopped at five epochs. Despite the abbreviated training duration, the model shows promising results with decreasing loss values, suggesting improved performance in predicting the desired outcomes.

4.5.2.2 AraGPT2 Test results

The table below presents the ROUGE scores achieved by the AraGPT2 model for benefits extraction using text completion :

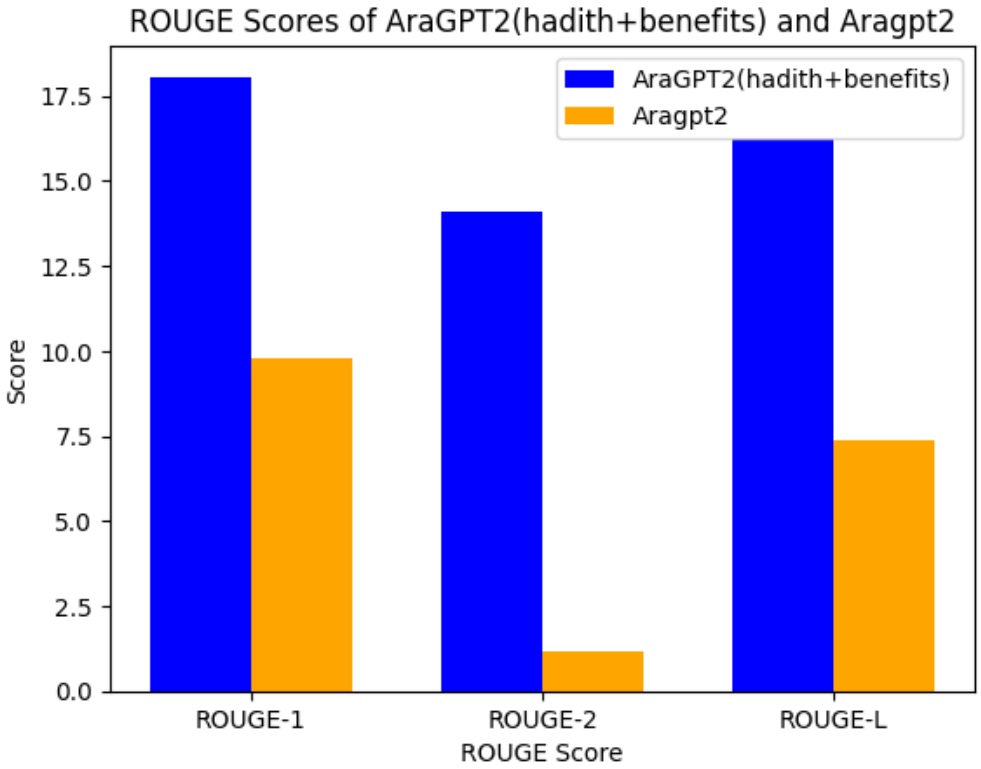

FIGURE 4.18 – Training and validation loss of AraGPT2

In our final investigation, we examined text completion as an alternative method for extracting benefits. To conduct this analysis, we employed the AraGPT2 model and assessed its performance using the ROUGE metric. The outcomes revealed that the AraGPT2 model obtained notably high ROUGE scores for ROUGE-1, ROUGE-2, and ROUGE-L. This suggests that the model is capable of generating benefits that closely align with the reference benefits. However, we observed a decline in ROUGE scores when Hadiths were excluded from both the actual and predicted benefits. This implies that the model heavily relies on the context provided by the Hadiths for accurate extraction of benefits.

4.6 Some related works results

This section provides a brief overview of the related works in the field of Arabic text summarization and evaluation. The studies mentioned in this table have focused on developing various techniques and approaches for summarizing Arabic text and evaluating the quality of the generated summaries using the ROUGE metric.

TABLE 4.4 – Comparison among different techniques for Arabic summarization

Ref	Year	Model	Methodology	ROUGE score	Dataset
[61]	2022	Seq2Seq Model	Using a seq2seq model,GRU and LSTM and BILSTM with global attention	Rouge1$\bar{0}$.5149 , Rouge2$\bar{0}$.12, RougeŁ=0.343	AHS and AMN
[29]	2020	BERT Model	Using pretrained BERT and encoder BERTSUM	Rouge1$\bar{0}$.42, Rouge2$\bar{0}$.2459 and RougeL$\bar{0}$.422	EASC and a KALIMAT
[31]	2020	Arabic Pretrained, and LSTM	AraVec is pretrained for word embedding and encoder-decoder sequences	Rouge1$\bar{0}$.38 ROUGE1 NOORDER $\bar{0}$.46	They collected the dataset from various sources like Reuters,Aljazeera
[43]	2020	Ara BERT Model and Clustering algorithm	Combining NLU (Ara BERT) and Clustering algorithms	Rouge1$\bar{0}$.54, Rouge2$\bar{0}$.54	They have prepared the data set (multiple articles in different domains and it's summary).
[28]	2020	Bert for Multi Lingual	BERT for multilingual and the word2vec model	ROUGE1$\bar{0}$.75 for the first algorithm ,ROUGE1$\bar{0}$.75 for the second one	EASC
[6]	2019	Linear Discriminant Analysis(LDA)	Using linear discriminant analysis(LDA) , modified Page Randk ,and the k-means clustering technique	ROUGE1$\bar{0}$.450 and ROUGE2 $\bar{0}$.307	EASC

In our project, we have not come across any existing research or prior work specifically focusing on Section abstraction or benefits extraction. Therefore, we have not been able to build upon previous methodologies or leverage established techniques in this particular domain.

To evaluate the effectiveness of our approach, we conducted a comprehensive comparison with prior works in the broader field of Arabic text summarization, using the rouge metric. We observed that the results from those works were somewhat more promising in terms of overall performance and quality compared to our own findings in extracting meaning and benefits from Hadith.

The difference in outcomes can be attributed to a significant disparity in the scale and nature of the datasets used to train the models. In the field of text summarization, previous works have had the advantage of training their models on large and diverse collections of textual data, enabling them to capture a broader range of language patterns, nuances, and context. In contrast, the models we utilized in our project were likely trained on a more limited set of data.

However, it is worth mentioning that models such as arat5, mbet2mbert, and gpt2 demonstrate the capability to generate benefits that closely align with the reference benefits, as well as generate sections that closely align with the reference benefits.

4.7 Conclusion

In conclusion, this chapter provided an overview of the implementation process for the proposed models, including araT5, mbert2mbert, and araGPT2. Through the use of different evaluation metrics, we thoroughly examined and discussed the results obtained from each model.

Conclusion

Hadith processing plays a vital role in understanding and interpreting the teachings of Islam. Hadiths, the sayings and actions of Prophet Muhammad (صلى الله عليه وسلم), provide valuable guidance for Muslims in various aspects of their lives, including matters of faith, ethics, and jurisprudence. However, the sheer volume and complexity of Hadith literature pose significant challenges in extracting relevant information and deriving meaningful insights. Effective Hadith processing techniques are essential for scholars, researchers, and the broader Muslim community to access, analyze, and apply the teachings of Prophet Muhammad (صلى الله عليه وسلم) accurately and comprehensively. By employing advanced computational methods and natural language processing techniques, Hadith processing enables us to unlock the insights contained within this vast corpus of Islamic traditions, promoting a deeper understanding of the religion and facilitating its practical application in contemporary contexts.

our research has centered around training multiple models to achieve the objective of section abstraction and extracting benefits from Hadith. We have employed three distinct models (AraT5, mbert2mbert, and Aragpt2) in order to generate meaningful sections from Hadith and identify the associated benefits.
Throughout our evaluation process, we utilized the ROUGE metric to assess the performance of these models. The results indicate that transformer models hold significant potential, outperforming traditional RNN-based approaches. This is primarily due to the attention-based nature of transformers, which allows for more efficient and precise processing, eliminating the risk of information loss that can occur with sequential RNN training. Our findings suggest that employing transformer models for section abstraction and benefit extraction from Hadith can yield superior outcomes. This not only enhances the quality of generated sections but also improves the overall efficiency of the process. By leveraging deep learning techniques, particularly within the domain of Islamic text processing and Hadith understanding, we can bridge the existing gap and better cater to the needs of the Muslim community.

Our future work involves expanding the size of the second dataset and enhancing the first dataset to incorporate additional hadiths and their associated benefits.We plan to leverage the capabilities of other large language models like GPT-3 and BLME.We also plan to utilize other evaluated metrics such as BLEU for improved evaluation.

Bibliography

[1] The science of hadith. https://www.islamicawareness.org/hadith/ulum/hadsciences.html.

[2] عبد الرحمن بن ناصر البراك. الجامع لفوائد بلوغ المرام. https://islamhouse.com/ar/books/2896832/, May 2023.

[3] S. M. AL-Ghuribi and S. A. M. Noah. Arabic text mining, 2022.

[4] H. K. Aldayel and A. M. Azmi. Arabic tweets sentiment analysis–a hybrid scheme. *Journal of Information Science*, 42(6) :782–797, 2016.

[5] M. Alhawarat, M. Hegazi, and A. Hilal. Processing the text of the holy quran : a text mining study. *International Journal of Advanced Computer Science Applications (IJACSA)*, 6(2) :262–267, 2015.

[6] Z. H. Ali and A. P. D. S. Malallah. Multilingual text summarization based on lda and modified pagerank. *Journal of Information Technology V*, 9(3) :2018, 2019.

[7] S. Altammami, E. Atwell, and A. Alsalka. Text segmentation using n-grams to annotate hadith corpus. In *Proc. 3rd Workshop on Arabic Corpus Linguist.*, pages 31–39, 2019.

[8] S. Altammami, E. Atwell, and A. Alsalka. Constructing a bilingual hadith corpus using a segmentation tool. In *Proc. 12th Lang. Resour. and Eval. Conf.*, 2020.

[9] alukah.net. علم فقه الحديث. https://www.alukah.net/fiqh-alhadith, February 2023.

[10] alukah.net. قواعد فقه الحديث. https://www.alukah.net/sharia/0/113745/, February 2023.

[11] N. N. Amirah, T. M. Rahim, Z. Mabni, H. M. Hanum, and N. A. Rahman. A malay hadith translated document retrieval using parallel latent semantic indexing (lsi). 2016.

[12] Wissam Antoun, Fadi Baly, and Hazem Hajj. Aragpt2 : Pre-trained transformer for arabic language generation. *arXiv preprint arXiv :2012.15520*, 2020.

[13] Noureddine Ater. *The Method of Criticism in the Sciences of Hadith*. Dar Al-Fikr, 1418 AH-1997 AD.

[14] atharah.net. الفوائد المنتخبة من أحاديث الصيام من صحيح البخاري. https://www.atharah.net/article/10184, May 2023.

[15] E. S. Atlam, K. Morita, M. Fuketa, and J. I. Aoe. A new approach for arabic text classification using arabic field-association terms. *Journal of the American Society for Information Science and Technology*, 62(11) :2266–2276, 2011.

[16] Judge Ayyad. *Al-Ilma'*.

[17] Judge Ayyad. *Familiarity with the knowledge of the origins of the narration and the restriction of hearing.* Investigated by Mr. Ahmed Saqr.

[18] A. Azmi and N.B. Badia. itree - automating the construction of the narration tree of hadiths (prophetic traditions). In *Proc. the 6th Int. Conf. Natural Language Processing and Knowledge Engineering (NLPKE-2010)*, Aug. 2010.

[19] Abu Al-Hussein Ahmed bin Faris bin Zakaria. *Dictionary of Language Standards.* Dar Revival of Arab Traditions, Beirut, 1st edition, 1422 AH.

[20] Abu Al-Hussein Ahmed bin Faris bin Zakaria. *Dictionary of Language Standards.* Dar Revival of Arab Traditions - Beirut, first edition, 1422 AH.

[21] Muhammad bin Manzur. *Lisan al-Arab (Fiqh).* Dar Sader, first edition.

[22] A. Brahmi, A. Ech-Cherif, and A. Benyettou. Arabic texts analysis for topic modeling evaluation. *Information Retrieval*, 15(1) :33–53, 2012.

[23] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, and D. ... Amodei. Language models are few-shot learners. *arXiv preprint arXiv :2005.14165*, 2020.

[24] W. Cherif, A. Madani, and M. Kissi. A new modeling approach for arabic opinion mining recognition. In *Intelligent Systems and Computer Vision (ISCV)*, pages 1–6. IEEE, March 2015.

[25] K. Clark, M. T. Luong, Q. V. Le, and C. D. Manning. Electra : Pre-training text encoders as discriminators rather than generators. 2020.

[26] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. 2018.

[27] R. M. Duwairi and I. Qarqaz. Arabic sentiment analysis using supervised classification. In *2014 International Conference on Future Internet of Things and Cloud (FiCloud)*, pages 579–583. IEEE, August 2014.

[28] B. Elayeb, A. Chouigui, M. Bounhas, and O. B. Khiroun. Automatic arabic text summarization using analogical proportions. *Cognitive Computation*, 12(5) :10431069, Sep 2020.

[29] K. N. Elmadani, M. Elgezouli, and A. Showk. Bert fine-tuning for arabic text summarization, 2020.

[30] Abdelrhman Elmadany and Muhammad Abdul-Mageed. Arat5 : Text-to-text transformers for arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 628–647, May 2022.

[31] S. Encoder. Deep learning based abstractive arabic text summarization using two layers encoder and one layer decoder. *Journal of Theoretical and Applied Information Technology*, 98(16) :32333244, 2020.

[32] (event). *Dictionary of Language Standards*.

[33] GeeksforGeeks. Introduction to recurrent neural network. https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/, April 2023.

[34] M. Ghazizadeh, M. Zahedi, M. Kahani, and B. Bidgoli. Fuzzy expert system in determining hadith validity. In *Advances in Computer and Information Sciences and Engineering*, pages 354–359.

[35] Ibn Hajar. *Al-Nokat on Ibn al-Salah's book*.

[36] F. Haque, A. Orthy, and S. Siddique. Hadith authenticity prediction using sentiment analysis and machine learning, 2020.

[37] L. He and X. Zhang. Deberta : Decoding-enhanced bert with disentangled attention. 2020.

[38] IslamHouse.com. الفوائد المستنبطة من الأربعين النووية وتتمتها الرجبية. https://islamhouse.com/ar/articles/13727/, May 2023.

[39] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert : A lite bert for self-supervised learning of language representations. 2019.

[40] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*, 2019.

[41] malmarjeh Hugging Face. mbert2mbert-arabic-text. https://github.com/malmarjeh/mbert2mbert-arabic-text-summarization, jun 2023.

[42] A. B. Muhammad. *Annotation of conceptual co-reference and text mining the Qur'an*. PhD thesis, University of Leeds, 2012.

[43] A. M. A. Nada, E. Alajrami, A. A. Al-Saqqa, and S. S. Abu-Naser. Arabic text summarization using arabert model using extractive text summarization approach. *International Journal of Academic Information Systems Research*, 4(8) :69, Aug 2020.

[44] M. Najeeb, A. Abdelkader, M. Al-Zghoul, and A. Osman. A lexicon for hadith science based on a corpus. *International Journal of Computer Science and Information Technologies*, 6(2) :1336–1340, 2015.

[45] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2) :1–135, 2008.

[46] F. N. Patel and N. R. Soni. Text mining : A brief survey. *International Journal of Advanced Computer Research*, 2(4) :243, 2012.

[47] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

[48] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. 2019.

[49] S. A. Salloum, A. Q. AlHamad, M. Al-Emran, and K. Shaalan. A survey of arabic text mining. In *Intelligent natural language processing : Trends and applications*, pages 417–431. 2018.

[50] A. Selamat and C. C. Ng. Arabic script web page language identifications using decision tree neural networks. *Pattern Recognition*, 44(1) :133–144, 2011.

[51] A. M. Sharaf. The qur'an annotation for text mining. first year transfer report. Technical report, School of Computing, Leeds University, December 2009.

[52] ShathaTm. LK-Hadith-Corpus. https://github.com/ShathaTm/LK-Hadith-Corpus, April 2023.

[53] Y. Shen, Y. Guo, Y. Kang, Z. Yang, Z. Liu, and M. Sun. Path-augmented language model. 2020.

[54] M.A. Siddiqui, M.E. Saleh, and A.A. Bagais. Extraction and visualization of the chain of narrators from hadiths using named entity recognition and classification. *Int. J. Comput. Linguist. Res*, 5(1) :14–25, 2014.

[55] Simplilearn. Recurrent neural network (rnn) tutorial : Types and examples [updated]. https://www.simplilearn.com/tutorials/deep-learning-tutorial/rnn, April 2023.

[56] T. H. Soliman, M. A. Elmasry, A. Hedar, and M. M. Doss. Sentiment analysis of arabic slang comments on facebook. *International Journal of Computer Technology*, 12(5) :3470–3478, 2014.

[57] R. Talib, M. K. Hanif, S. Ayesha, and F. Fatima. Text mining : techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, 7(11), 2016.

[58] the ruler. *Knowledge of the sciences of hadith.*

[59] UBC-NLPH ugging Face. Arat5-base-title-generation. https://github.com/UBC-NLP/AraT5-base-title-generation, May 2023.

[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[61] Y. M. Wazery, M. E. Saleh, A. Alharbi, and A. A. Ali. Abstractive arabic text summarization based on deep learning. *Computational Intelligence and Neuroscience*, 2022 :114, Jan 2022.

[62] Wisdom Geek. Recurrent neural networks. https://www.wisdomgeek.com/machine-learning/recurrent-neural-networks/, April 2023.

[63] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet : Generalized autoregressive pretraining for language understanding. 2019.

[64] F. Zaraket and J. Makhlouta. Arabic cross-document nlp for the hadith and biography literature. In *Proc. 25th Int. Florida Artificial Intell. Res. Soc. Conf.*, Jan. 2012.

Appendix

W&B visualizations (AraGPT-2)

— **validation :**

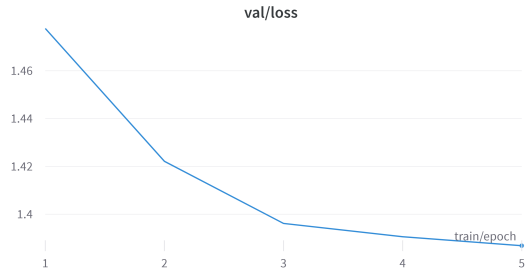

FIGURE 0.19 – Validation Loss

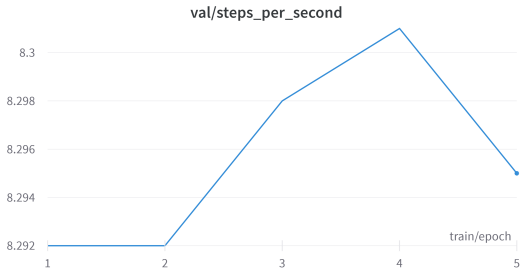

FIGURE 0.20 – Validation Steps per Second

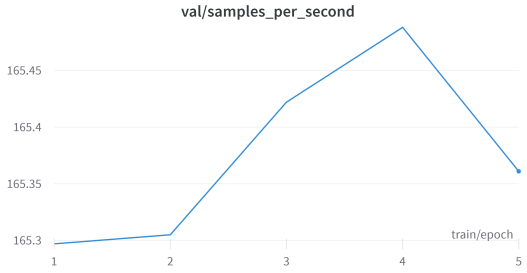

FIGURE 0.21 – validation samples per second

— **Training :**

FIGURE 0.22 – Training loss

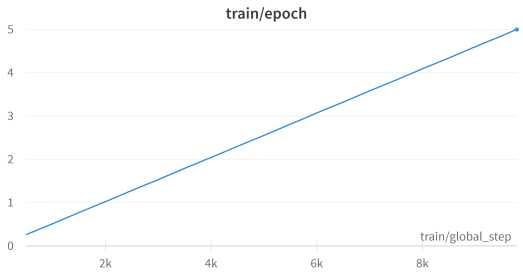

FIGURE 0.23 – Training epochs

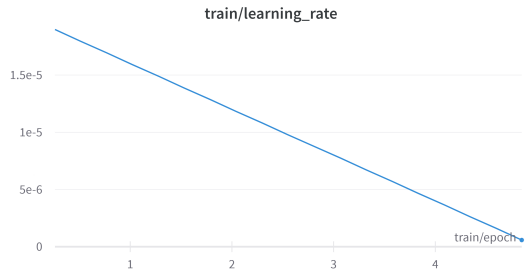

FIGURE 0.24 – Training learning rate

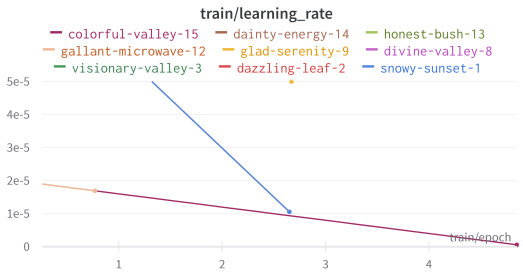

FIGURE 0.25 – Training learning rate (all sweeps)

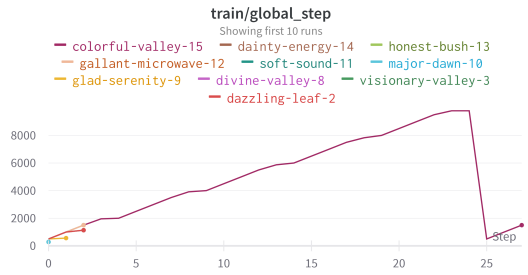

FIGURE 0.26 – train global steps

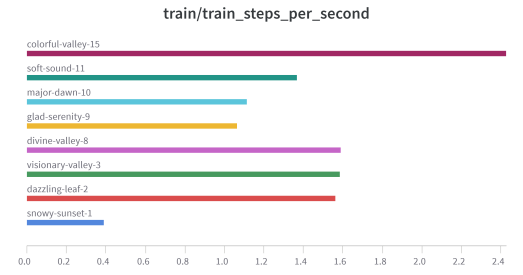

FIGURE 0.27 – Training Steps per second

— **GPU :**

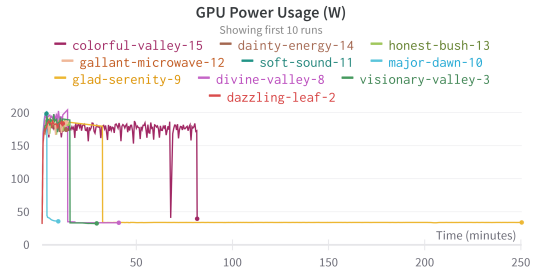

FIGURE 0.28 – GPU power usage (W)

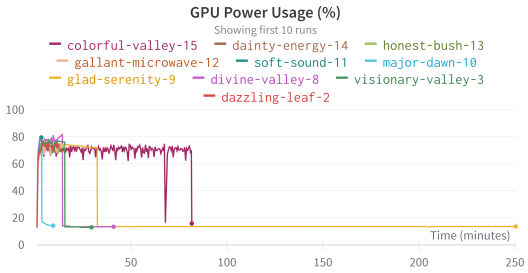

FIGURE 0.29 – GPU power usage (%)

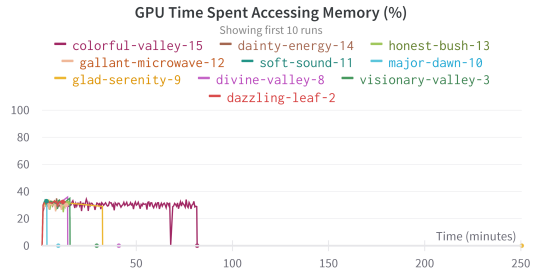

FIGURE 0.30 – GPU Time Spent Accessing Memory (%)

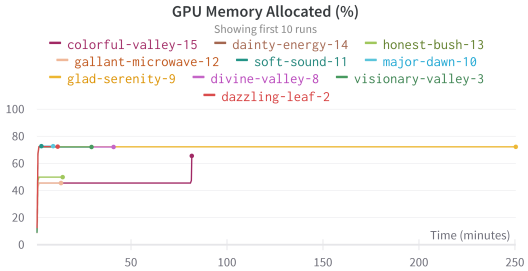

FIGURE 0.31 – GPU Memory Allocated (%)

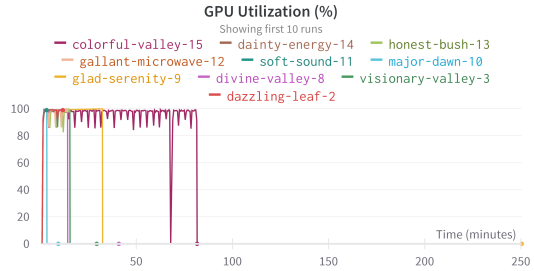

FIGURE 0.32 – GPU Utilization (%)

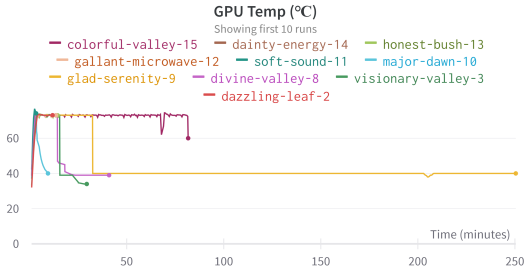

FIGURE 0.33 – GPU Temp (C)