Artificial Intelligence for Understanding the Hadith

UNIVERSITY OF LEEDS

Shatha Hamad Altammami

School of Computing

University of Leeds

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

January 2023

Publications

Chapter 4,5,6,7 and 8 of this thesis are based on jointly-authored publications. I was the lead author and the co-authors acted in an advisory capacity, providing supervision and review. All original contributions presented here are my own.

Chapter 4:
Altammami, S., Atwell, E. and Alsalka, A. 2019. Text Segmentation Using N-grams to Annotate Hadith Corpus. In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics (WACL-3)*, pp. 31–39, Cardiff, United Kingdom. Association for Computational Linguistics.

Altammami, S., Atwell, E. and Alsalka, A. 2020. Constructing a Bilingual Hadith Corpus Using a Segmentation Tool. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pp. 3390–3398, European Language Resources Association. Marseille, France. .

Chapter 5:
Altammami, S., Atwell, E. and Alsalka, A., 2020. The Arabic-English Parallel Corpus of Authentic Hadith. *International Journal on Islamic Applications in Computer Science And Technology*, 8(2).pp. 1-10

Chapter 6:
Altammami, S., Atwell, E. and Alsalka, A., 2021. Towards a Joint

Ontology of Quran and Hadith. *International Journal on Islamic Applications in Computer Science And Technology*, 9(2), pp. 1-12.

Chapter 7 and 8:

Altammami, S. and Atwell, E., 2022, June. Challenging the Transformer-based models with a Classical Arabic dataset: Quran and Hadith. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC).* pp. 1462-1471, European Language Resources Association. Marseille, France.

Dedication

To my parents, Hamad and Helah
for their unwavering love and support.

Acknowledgements

I am deeply grateful for the opportunity that Allah has granted me to pursue my PhD under the supervision of Professor Eric Atwell. He was an exceptional primary supervisor, providing invaluable support, guidance, and encouragement that helped me to develop intellectually, explore research areas that I am passionate about, and publish my work. My gratitude also extends to my second supervisor, Dr. Ammar Alsalka, for his support. I would also like to express my sincere thanks to my examiners, Professor James Dickens and Professor Paul Rayson, for the thoughtful feedback they provided. Additionally, I am thankful for the scholarship from King Saud University, which has made this endeavour possible.

I could not have undertaken this journey without my husband, Abdullah, who helped me pursue my dreams and celebrated my accomplishments. My heartfelt gratitude also goes to my wonderful children, Leena and Abdulrahman, for their understanding and patience during the long hours of research and for always lifting my spirits with their funny jokes. I am truly blessed to be their mother.

An immense thank you to my parents for always believing in me and keeping me in their prayers. Additionally, I am grateful to my siblings for their support, especially my brother Faisal, who sacrificed his first year in college to help me settle down in Leeds.

A special thanks to my friend Ahoud Aldhafeeri, for the enlightening conversations and for pointing me to valuable resources on the Quran. My gratitude also goes to members of the Artificial Intelligence for Language (AI4L) research group at the University of Leeds for their informative discussions. I am also grateful to all my amazing friends in Leeds, who have made this journey enjoyable and fun.

Last, but certainly not least, I would like to acknowledge my baby boy, Tameem, who came into this world in 2022 and brought my happiness and multitasking skills to the next level!

Abstract

My research aims to utilize Artificial Intelligence to model the meanings of Classical Arabic Hadith, which are the reports of the life and teachings of the Prophet Muhammad. The goal is to find similarities and relatedness between Hadith and other religious texts, specifically the Quran. These findings can facilitate downstream tasks, such as Islamic question-answering systems, and enhance understanding of these texts to shed light on new interpretations.

To achieve this goal, a well-structured Hadith corpus should be created, with the Matn (Hadith teaching) and Isnad (chain of narrators) segmented. Hence, a preliminary task is conducted to build a segmentation tool using machine learning models that automatically deconstruct the Hadith into Isnad and Matn with 92.5% accuracy. This tool is then used to create a well-structured corpus of the canonical Hadith books.

After building the Hadith corpus, Matns are extracted to investigate different methods of representing their meanings. Two main methods are tested: a knowledge-based approach and a deep-learning-based approach. To apply the former, existing Islamic ontologies are enumerated, most of which are intended for the Quran. Since the Quran and the Hadith are in the same domain, the extent to which these ontologies cover the Hadith is examined using a corpus-based evaluation. Results show that the most comprehensive Quran ontology covers only 26.8% of Hadith concepts, and extending it is expensive. Therefore, the second approach is investigated by building and evaluating various deep-learning models for a binary classification task of detecting relatedness between the Hadith and the Quran. Results show that the likelihood of the current models reaching a human-level understanding of such texts remains somewhat elusive.

Contents

1 Introduction **1**

1.1 Aims and Motivation 1

1.2 Research Questions 3

1.2.1 Is annotating Isnad and Matn feasible? 3

1.2.2 Is semantic similarity in Quran and Hadith detectable? . 4

1.3 Contributions 4

1.3.1 Hadith Segmenter 5

1.3.2 Hadith Parallel Corpus 6

1.3.3 Dataset of Related Quran and Hadith Pairs 7

1.3.4 Evaluating Quran Ontologies for Hadith 8

1.3.5 Semantic Similarity Between Quran and Hadith 8

1.4 Thesis Outline 9

2 Background **12**

2.1 Introduction 12

2.2 The Arabic Language 12

2.2.1 Classical Arabic 14

2.2.2 Modern Standard Arabic 16

2.2.3 Dialect Arabic 16

2.3 The Quran 18

2.3.1 Quran Structure 18

2.3.2 Quran's Linguistic Features 18

2.3.3 Quran Exegesis (Tafsir) 21

2.4 The Hadith 23

2.4.1 Hadith Structure 25

2.4.2 Hadith Books 25
2.4.3 Hadith Types 27
2.4.4 Useful Sources for Hadith 27
2.4.5 Hadith Commentaries 28
2.5 Conclusion 29

3 Literature Review 30
3.1 Introduction 30
3.2 Computational Hadith Research 31
3.2.1 Hadith Authentication 31
3.2.2 Hadith Topic Classification 32
3.2.3 Hadith Ontology 32
3.3 Text Segmentation 33
3.4 Corpus Building 34
3.5 Semantic Similarity 35
3.5.1 Knowledge-Based 37
3.5.2 Corpus-Based 38
3.5.3 Deep-learning Based 39
3.6 Conclusion 39

4 Building the Hadith Segmenter 40
4.1 Introduction 40
4.2 Related Work 42
4.3 Arabic Text Preprocessing 47
4.4 Data Preparation 48
4.4.1 Testing Data 48
4.4.2 Training Data 49
4.5 Look-up Approach 50
4.5.1 Look-up Lists 50
4.5.2 The N-gram Model 50
4.5.3 Experiments 54
4.5.4 Result 57
4.6 Machine Learning Approach 63
4.6.1 Choosing a Classifier Model 63

4.6.2 Experiments . . . 67
4.6.3 Results . . . 68
4.7 Hidden Markov Model . . . 71
4.7.1 Data Preparation . . . 72
4.7.2 Experiments . . . 72
4.7.3 Results . . . 75
4.8 Comparison and Discussion . . . 80
4.9 Conclusion . . . 82

5 Developing the Hadith Corpus 84
5.1 Introduction . . . 84
5.2 Related Work . . . 85
5.3 Defining Corpus Data . . . 88
5.3.1 Hadith Structure . . . 88
5.3.2 Existing Hadith Books . . . 88
5.4 Corpus Creation . . . 89
5.4.1 Resource Enumeration . . . 89
5.4.2 Data Collection . . . 90
5.4.3 Corpus Annotation . . . 91
5.5 Corpus Analysis . . . 95
5.5.1 Corpus Statistics . . . 95
5.5.2 Manual Correction and Verification . . . 98
5.6 Potential Uses . . . 100
5.7 Conclusion . . . 100

6 Evaluating Quran Ontologies for Hadith 102
6.1 Introduction . . . 102
6.2 What is an Ontology? . . . 103
6.3 Existing Quran Ontologies . . . 104
6.4 Ontology Evaluation Methods . . . 106
6.5 Quran Ontology Evaluation . . . 109
6.5.1 Data Sources . . . 110
6.5.2 Data Analysis . . . 111
6.5.3 Experiment . . . 111

6.6 Results . . . 113
6.7 Discussion . . . 115
6.8 Conclusion . . . 117

7 Creating the Quran-Hadith Dataset 118
7.1 Introduction . . . 118
7.2 Quran and Hadith . . . 118
7.2.1 The Quran . . . 122
7.2.2 The Hadith . . . 128
7.3 Existing Arabic Datasets . . . 131
7.4 Datasets . . . 132
7.4.1 Testing Dataset: Quran-Hadith (QH) Pairs . . . 133
7.4.2 Training Dataset: Quran-Quran (QQ) Pairs . . . 140
7.5 Measuring Data Leakage . . . 142
7.6 Conclusion . . . 144

8 Semantic Relatedness Using Deep Learning 146
8.1 Introduction . . . 146
8.2 Related Work . . . 147
8.3 Datasets . . . 148
8.4 Experimental Setup . . . 150
8.4.1 Evaluation Metrics . . . 150
8.4.2 Baseline Model . . . 151
8.4.3 Transformer-based Models . . . 153
8.5 Results . . . 156
8.5.1 Validation . . . 156
8.5.2 Testing . . . 158
8.6 Analysis and Discussion . . . 159
8.7 Conclusion . . . 165

9 Conclusion and Future Work 166
9.1 Introduction . . . 166
9.2 Overall Findings . . . 167
9.3 Limitations and Future Work . . . 169

9.4 Research Contributions 171
9.5 Concluding Summary 171

References **190**

List of Figures

1.1 A simple structure of Hadith showing Matn in bold 5

1.2 Hadith with comment after Matn 5

1.3 Hadith with a vague segmentation point 6

1.4 Thesis outline 11

2.1 Arabic language categories 13

2.2 Concordance from the arTenTen12 corpus 16

2.3 Arabic dialects are different within the same country[1] 17

2.4 Hadith example, Matn in bold 25

3.1 Various methods to approach textual semantic similarity 37

4.1 Hadith example, Isnad in bold face text, followed by Matn 41

4.2 Isnad example, Isnad lexical items underlined 49

4.3 Python script to perform the segmentation on Albukhari Hadith . 51

4.4 Example of Isnad and Matn extracted from the training data and tokenized to unigrams, bigrams and trigrams then added to the lists accordingly 52

4.5 Pipeline of Hadith segmenter using the ML approach 64

4.6 Data distribution of bigram tokens 65

4.7 Confusion matrix of the tested ML classifiers 66

4.8 Hadith annotation for training the HMM 73

4.9 Visualization of the HMM to illustrate the probability of transitioning to a new state conditioned on a present state 74

4.10 Example of a Hadith with regular structure, and the correct segmentation indicated by the horizontal line after token number 13 . 76

4.11 Example of a Hadith with a vague segmentation point, incorrectly segmented as shown by the horizontal line 77
4.12 Example of incorrect segmentation: it should be at token 30 . . . 78
4.13 Example of a Hadith with irregular structure: it has more than one Isnad and Matn segment 79
4.14 Segmenter performance with unigrams, bigrams, and trigrams . . 82

5.1 Structure of the canonical Hadith books 90
5.2 HTML of a Hadith page on *Sunnah.com* 91
5.3 Example of a Hadith record extracted from Tarmizi Chapter10.csv – Part 1 91
5.4 Continued example of Hadith record – Part 2 92
5.5 Continued example of Hadith record – Part 3 92
5.6 Structure of the LK Hadith corpus 94
5.7 A comparison between the Arabic and English parts of the LK Hadith corpus. Note the difference in the y-axis numbering to reflect the number of words 96
5.8 Most frequent words in the Isnad 97
5.9 Most frequent words in the Matn 99

6.1 An overview of ontology evaluation methods (Raad & Cruz, 2015) 107
6.2 A word from Chapter 106 in the Quran 112
6.3 Overlap between Hadith section-headings and the Quran ontologies 114

7.1 Quran and Hadith word clouds 121
7.2 The number of words and lexemes in Quran and Hadith 122
7.3 Quran and Hadith lexeme clouds 123
7.4 Fatwa page where the Quran, Hadith, and the tag is shown in red, green, and grey accordingly 135
7.5 Algorithm for creating positive QH samples 136
7.6 Some of the Quran symbols 138
7.7 Algorithm for creating negative QH samples 139
7.8 Algorithm for creating negative QQ samples 141
7.9 Algorithm to compute data leakage (Elangovan *et al.*, 2021) . . . 143

8.1 Baseline model architecture 152
8.2 Weights and Biases is used to determine the hyper-parameters . . 154
8.3 The number of tokens in each instance of the training data 154
8.4 Fine-tuning CAMeLBERT-CA for 8 epochs 155
8.5 How the WordPiece tokenizes a Quran-verse and a Hadith-teaching from Example 6 in Table 8.7 160
8.6 Sentence length distribution where x-axes represent the number of tokens and the y-axes reflect the number of instances 163
8.7 Models' tokenizers' fertility 164

List of Tables

2.1 Examples of orthographic challenges associated with the Arabic language . . . 14
2.2 Examples from Alzrkashi's book . . . 15
2.3 The different meanings of the word 'Ayn' . . . 21
2.4 Useful resources on the web that can be utilized in Hadith computational studies . . . 28

4.1 Review of previous research on Hadith segmentation into Isnad and Matn components . . . 45
4.2 Result of tokenizing and annotating the 500 Hadith testing data using look-up lists . . . 53
4.3 Example of incorrect Hadith segmentation using trigrams with back-off appraoch . . . 57
4.4 Correct segmentation, regular pattern . . . 59
4.5 Correct segmentation, introductory statement . . . 59
4.6 Correct segmentation, conversation of the Prophet . . . 60
4.7 Correct segmentation, no prophetic words . . . 60
4.8 Incorrectly segmented, parallel Isnad . . . 61
4.9 Incorrectly segmented, Isnad contains Matn lexical items . . . 61
4.10 Incorrectly segmented, names should be part of Matn . . . 62
4.11 Accuracy of ML models on the testing data to classify bigram of words as Isnad or Matn . . . 67
4.12 Correctly segmented Hadiths with parallel Isnads . . . 69
4.13 Correctly segmented Hadith with irregular pattern . . . 70
4.14 Incorrectly segmented Hadith with irregular patterns . . . 71

4.15 Tag set for HMM training 72
4.16 Comparison of results using the first two annotation approaches and normalization with Algorithm 3 81
4.17 Performance of the different approaches to build the Hadith segmenter 81

5.1 Comparison of CA corpora that includes Hadith 87
5.2 Corpus Annotation 93
5.3 Number of Hadiths in each book 95

6.1 Existing Quran ontologies 108
6.2 A Quran verse and concepts extracted from the two ontologies . . 110
6.3 Example of a Hadith Matn and the associated section-heading . . 111
6.4 Vocabulary statistics for ontologies and Hadith headings 112
6.5 Synsets obtained from Arabic Wordnet 115
6.6 Examples of lexemes extracted from each part illustrated in Figure 6.3 116

7.1 Quran and Hadith statistics 120
7.2 Quran and Hadith unique words 120
7.3 Unique lexemes in Quran and Hadiths 121
7.4 The available Arabic semantic similarity and relatedness dataset . 131
7.5 An example from Sahih Albukhari 134
7.6 The different training/fine-tuning dataset. QQ=Quran-Quran, QH=Quran-Hadith, Ar= Arabic, En= English, M= Multilingual (43 languages) 142
7.7 Measuring data leakage 144

8.1 Examples from the QQ and QH datasets 149
8.2 The training/fine-tuning dataset as introduced in Table 7.6 . . . 150
8.3 Models' performance on the QQ validation dataset shown in %. Overall best models based on F1 score are highlighted in bold. Ar = Arabic Quran, En = English Quran, M = Multilingual Quran (43 languages), Taf = Arabic Tafsir(commentaries) 157
8.4 Performance of SBERT multilingual models on the QQ dataset . . 157
8.5 Examples of QQ pairs classified by CAMeLBERT-CA 158

8.6 Performance of models on the Quran-Hadith testing dataset . . . 159
8.7 Examples of QH pairs classified by CAMeLBERT-CA 161

Abbreviations

CA	Classical Arabic
CNN	Convolutional Neural Networks
BERT	Bidirectional Encoder Representations from Transformers
BOW	Bag Of Words
DA	Dialect Arabic
DL	Deep Learning
FST	Finite State Transducers
IDF	Inverse Document Frequency
IR	Information Retrieval
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
ML	Machine Learning
MSA	Modern Standard Arabic
NB	Naive Bayes
NER	Named Entity Recognition
NLP	Natural Language Processing
OOV	Out of Vocabulary
POS	Part of Speech
PBUH	Peace Be Upon Him
QA	Question Answering
SOTA	State Of The Art
STS	Semantic Textual Similarity
SVM	Support Vector Machine
TF	Term Frequency

Chapter 1

Introduction

1.1 Aims and Motivation

Artificial Intelligence (AI) is a branch of computer science that seeks to design and develop systems capable of performing tasks requiring human-like intelligence. One of these tasks is understanding the human language, which is addressed by a subfield of AI known as Natural Language Processing (NLP)(Vajjala *et al.*, 2020). The recent advances in NLP can be attributed to the abundance of resources, including data and computing power (Goodfellow *et al.*, 2016). These resources enable NLP systems to utilize Machine Learning (ML) and Deep Learning (DL) methods, which require a significant amount of processing power and a large number of examples, in order to learn and automate tasks.

Currently, NLP is used across several domains to automate domain-specific tasks such as financial narrative summarization (El-Haj *et al.*, 2022b), medication extraction from clinical notes (Henry *et al.*, 2020), and social media hate speech detection (Mubarak *et al.*, 2020). One of the domains that has attracted interest in NLP is the modelling of Islamic religious texts as they are sources of knowledge, wisdom and law (Atwell *et al.*, 2011). Successfully modelling these texts can aid scholars, researchers, and laymen in extracting and inferring knowledge from these texts. For example, the Islamicate Digital Humanities Network

(IDHN)[2] is a group of researchers who are interested in utilizing NLP tools in their Islamic studies research. The main obstacle in this domain is the availability of annotated corpora and datasets suitable for training and evaluating NLP systems to perform religious-oriented tasks (Bounhas, 2019).

In this thesis, I intend to fill the gap by enriching the under-resourced religious text of Islamic Hadith, which is the set of narratives reporting the words, actions and habits of the prophet Muhammad. Hadith is considered Classical Arabic (CA), an ancient language that is used in Islamic religious texts, and is the direct ancestor of Modern Standard Arabic (MSA) used today (Habash, 2010). Although Hadith's importance is second to the Quran's (Muslims' holy book), most laws and legislation are obtained from the Hadith due to its larger scope and incorporated details (Brown, 2017). However, research in the area of Hadith computation is still in its infancy (Bounhas, 2019). However, here is an annual increase in the number of published papers in this field, indicating wider attention from multi-disciplinary researchers (Azmi *et al.*, 2019).

The main practical problem in Hadith computational studies is that researchers gather their own datasets from different sources and sometimes manually process them (Luthfi *et al.*, 2018). This suggests a lack of adequate language resources and re-usability is limited since the collected datasets are not published to be shared. Hence, it is unfeasible to establish benchmarks, compare results or set evaluation measures (Guellil *et al.*, 2019). This makes building a well-structured Hadith corpus an imperative. Therefore, one of the main goals of this thesis is a Hadith corpus which is built automatically while answering the first research question discussed in Section 1.2

After creating the Hadith corpus, the feasibility of detecting semantic similarity between Quran and Hadith texts is investigated as the second research question of this thesis. This is important for three reasons. First, Semantic Textual Similarity (STS) detection is at the core of many tasks like information retrieval (IR) and question answering (QA) (Cer *et al.*, 2017). Second, religious

[2] https://idhn.org/

texts are important to understand, but are hard to interpret with their layers of embedded meanings as described in Chapter 7. Third, the performance of recent AI advances is largely unexplored on CA texts (Guellil *et al.*, 2021).

1.2 Research Questions

1.2.1 Is annotating Isnad and Matn feasible?

Most Hadiths consists of two components, *Isnad*, which is the chain of narrators, followed by *Matn*, which is the actual Hadith teaching. The example below shows a Hadith[3] in the simplest form,‘ where the segmentation point between Isnad and Matn is clear as indicated by the long horizontal line (——).

> **Hadith Example:**
>
> حَدَّثَنَا عَلِيُّ بْنُ مُحَمَّدٍ، حَدَّثَنَا وَكِيعٌ، حَدَّثَنَا يُونُسُ بْنُ أَبِي إِسْحَاقَ، عَنْ
> مُجَاهِدٍ، عَنْ أَبِي هُرَيْرَةَ، قَالَ —— قَالَ رَسُولُ اللَّهِ صلى الله عليه وسلم مَا
> زَالَ جِبْرَائِيلُ يُوصِينِي بِالْجَارِ حَتَّى ظَنَنْتُ أَنَّهُ سَيُوَرِّثُهُ.
>
> Ali bin Mohammed told us, Wakia told him that Younis bin Abi Ishaq heard Mujahid, heard Abu Hurayrah said that —— the Messenger of Allah peace be upon him (PBUH) said Jibra'il kept enjoining good treatment of neighbours until I thought he would make neighbours heirs.

Hadith computational studies either focus on Isnad or Matn (Azmi *et al.*, 2019). This is because the Isnad is used for authentication and is considered a piece of meta-data that does not add any meaning to the actual Hadith teaching (Matn). Research studies that focus on Isnad usually aim to build systems for

[3] English translation of Hadiths mentioned throughout this thesis are obtained from Sunnah.com

Hadith authentication (Dalloul, 2013), and those that focus on Matn are concerned with enabling machines to capture the meaning of the Matn text (Saloot *et al.*, 2016). Hence, it is useful to build a Hadith corpus for the research community where Isnad is segmented from Matn to minimise manual annotation (Luthfi *et al.*, 2018).

The main obstacle to automatically segmenting the Isnad from the Matn is that Hadiths have different structures with vague endings of Isnad as illustrated in the examples in Section 1.3.1. Hence, automatically identifying and segmenting the Isnad from the Matn is a non-trivial task. This thesis investigates the feasibility of utilizing AI to automatically identify them to produce a well-structured Hadith corpus where Isnad is segmented from Matn.

1.2.2 Is semantic similarity in Quran and Hadith detectable?

Can AI tools and methods model the underlying meaning of CA religious texts? To answer this question, the task of detecting semantic similarity between Quran-verse and Hadith-Matn is chosen because it is at the core of many useful applications like QA and IR (Chandrasekaran & Mago, 2021). Detecting semantic similarity between religious texts is a challenging task which requires deep human understanding considering they are different in style and structure, as elaborated in Chapter 7. However, the main obstacle to this experiment is the lack of an evaluation dataset. Therefore, before answering this question, I developed a methodology that follows rigorous heuristics to automatically create a dataset of Quran-verse and Hadith-Matn pairs.

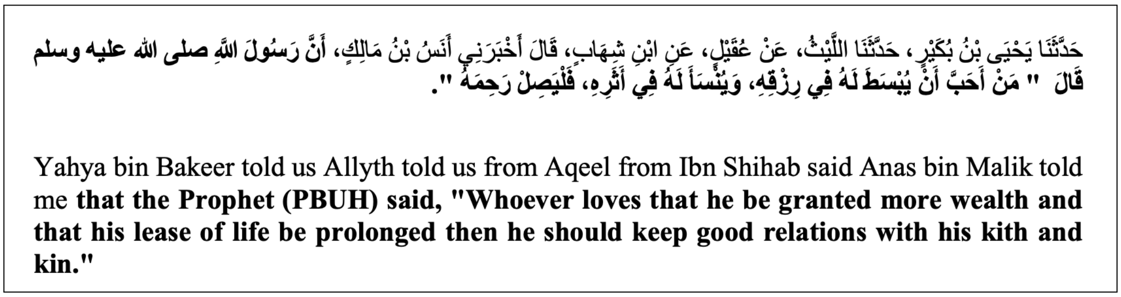
The following sections lists the contributions made to answer these two questions and enrich the Classical Arabic NLP research.

1.3 Contributions

The originality and novelty of contributions presented in this thesis is discussed below.

1.3.1 Hadith Segmenter

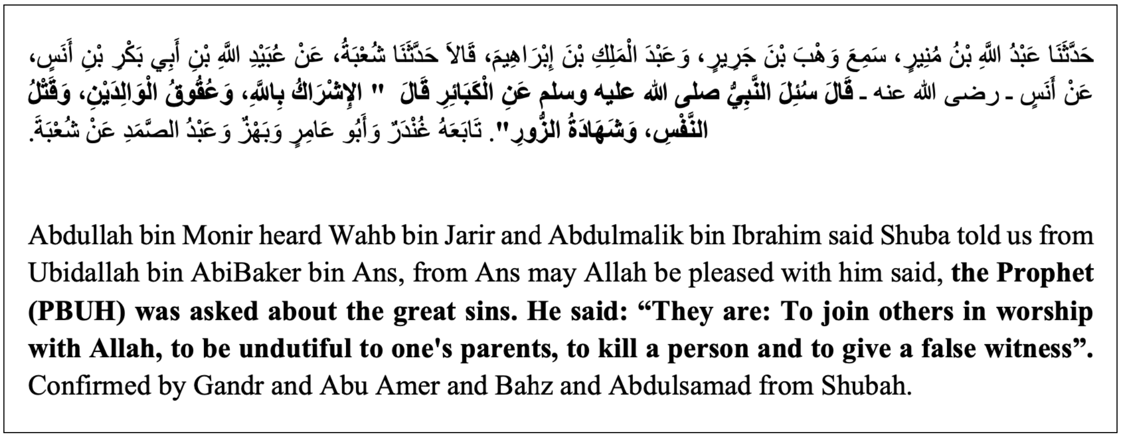
The first contribution is a novel Hadith segmentation tool that identifies and segments the Isnad from the Matn in the Hadith. This is a non-trivial task since not all Hadiths[4] follow consistent patterns. For example, some Isnad segments contain Matn patterns and vice versa. The following examples show Hadiths with Matn in bold. They are extracted from the most famous Hadith books. The first two examples in Figure 1.1 and 1.2 are extracted from *Sahih Albukhari's*, which consists of many Hadiths that follow this consistent structure, Isnad followed by Matn and an optional comment at the end. However, it contains some Hadiths with vague segmentation point as in Figure 1.3. Chapter 4 discusses the creation of the Hadith segmenter.



حَدَّثَنَا يَحْيَى بْنُ بُكَيْرٍ، حَدَّثَنَا اللَّيْثُ، عَنْ عُقَيْلٍ، عَنِ ابْنِ شِهَابٍ، قَالَ أَخْبَرَنِي أَنَسُ بْنُ مَالِكٍ، **أَنَّ رَسُولَ اللَّهِ صلى الله عليه وسلم**
قَالَ " مَنْ أَحَبَّ أَنْ يُبْسَطَ لَهُ فِي رِزْقِهِ، وَيُنْسَأَ لَهُ فِي أَثَرِهِ، فَلْيَصِلْ رَحِمَهُ ".

Yahya bin Bakeer told us Allyth told us from Aqeel from Ibn Shihab said Anas bin Malik told me **that the Prophet (PBUH) said, "Whoever loves that he be granted more wealth and that his lease of life be prolonged then he should keep good relations with his kith and kin."**

Figure 1.1: A simple structure of Hadith showing Matn in bold

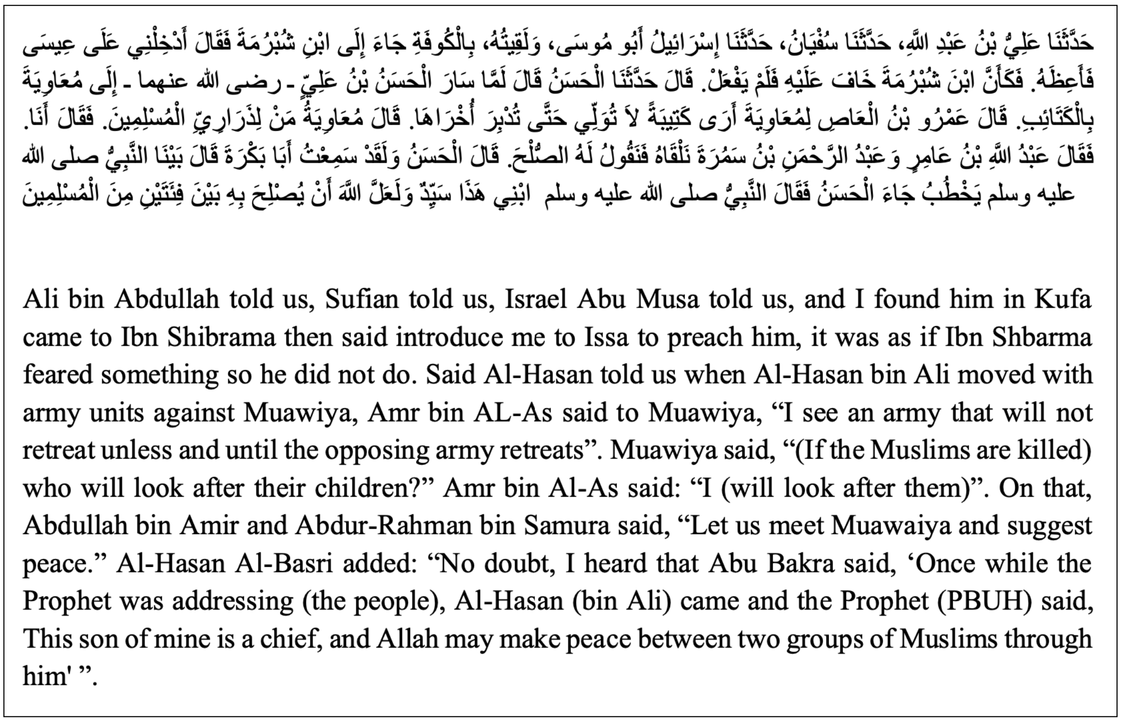


حَدَّثَنَا عَبْدُ اللَّهِ بْنُ مُنِيرٍ، سَمِعَ وَهْبَ بْنَ جَرِيرٍ، وَعَبْدَ الْمَلِكِ بْنَ إِبْرَاهِيمَ، قَالاَ حَدَّثَنَا شُعْبَةُ، عَنْ عُبَيْدِ اللَّهِ بْنِ أَبِي بَكْرِ بْنِ أَنَسٍ،
عَنْ أَنَسٍ ـ رضى الله عنه ـ **قَالَ سُئِلَ النَّبِيُّ صلى الله عليه وسلم عَنِ الْكَبَائِرِ قَالَ " الإِشْرَاكُ بِاللَّهِ، وَعُقُوقُ الْوَالِدَيْنِ، وَقَتْلُ**
النَّفْسِ، وَشَهَادَةُ الزُّورِ". تَابَعَهُ غُنْدَرٌ وَأَبُو عَامِرٍ وَبَهْزٌ وَعَبْدُ الصَّمَدِ عَنْ شُعْبَةَ.

Abdullah bin Monir heard Wahb bin Jarir and Abdulmalik bin Ibrahim said Shuba told us from Ubidallah bin AbiBaker bin Ans, from Ans may Allah be pleased with him said, **the Prophet (PBUH) was asked about the great sins. He said: "They are: To join others in worship with Allah, to be undutiful to one's parents, to kill a person and to give a false witness".** Confirmed by Gandr and Abu Amer and Bahz and Abdulsamad from Shubah.

Figure 1.2: Hadith with comment after Matn

[4] The plural of Hadith in Arabic is AHadith, but I will use Hadiths



حَدَّثَنَا عَلِيُّ بْنُ عَبْدِ اللَّهِ، حَدَّثَنَا سُفْيَانُ، حَدَّثَنَا إِسْرَائِيلُ أَبُو مُوسَى، وَلَقِيتُهُ، بِالْكُوفَةِ جَاءَ إِلَى ابْنِ شُبْرُمَةَ فَقَالَ أَدْخِلْنِي عَلَى عِيسَى
فَأَعِظَهُ. فَكَأَنَّ ابْنَ شُبْرُمَةَ خَافَ عَلَيْهِ فَلَمْ يَفْعَلْ. قَالَ حَدَّثَنَا الْحَسَنُ قَالَ لَمَّا سَارَ الْحَسَنُ بْنُ عَلِيٍّ ـ رضى الله عنهما ـ إِلَى مُعَاوِيَةَ
بِالْكَتَائِبِ. قَالَ عَمْرُو بْنُ الْعَاصِ لِمُعَاوِيَةَ أَرَى كَتِيبَةً لاَ تُوَلِّي حَتَّى تُدْبِرَ أُخْرَاهَا. قَالَ مُعَاوِيَةُ مَنْ لِذَرَارِيِّ الْمُسْلِمِينَ. فَقَالَ أَنَا.
فَقَالَ عَبْدُ اللَّهِ بْنُ عَامِرٍ وَعَبْدُ الرَّحْمَنِ بْنُ سَمُرَةَ نَلْقَاهُ فَنَقُولُ لَهُ الصُّلْحَ. قَالَ الْحَسَنُ وَلَقَدْ سَمِعْتُ أَبَا بَكْرَةَ قَالَ بَيْنَا النَّبِيُّ صلى الله
عليه وسلم يَخْطُبُ جَاءَ الْحَسَنُ فَقَالَ النَّبِيُّ صلى الله عليه وسلم ابْنِي هَذَا سَيِّدٌ وَلَعَلَّ اللَّهَ أَنْ يُصْلِحَ بِهِ بَيْنَ فِئَتَيْنِ مِنَ الْمُسْلِمِينَ

Ali bin Abdullah told us, Sufian told us, Israel Abu Musa told us, and I found him in Kufa came to Ibn Shibrama then said introduce me to Issa to preach him, it was as if Ibn Shbarma feared something so he did not do. Said Al-Hasan told us when Al-Hasan bin Ali moved with army units against Muawiya, Amr bin AL-As said to Muawiya, "I see an army that will not retreat unless and until the opposing army retreats". Muawiya said, "(If the Muslims are killed) who will look after their children?" Amr bin Al-As said: "I (will look after them)". On that, Abdullah bin Amir and Abdur-Rahman bin Samura said, "Let us meet Muawaiya and suggest peace." Al-Hasan Al-Basri added: "No doubt, I heard that Abu Bakra said, 'Once while the Prophet was addressing (the people), Al-Hasan (bin Ali) came and the Prophet (PBUH) said, This son of mine is a chief, and Allah may make peace between two groups of Muslims through him' ".

Figure 1.3: Hadith with a vague segmentation point

1.3.2 Hadith Parallel Corpus

The second contribution is the first well-structured corpus of Hadith with data obtained from Hadith books known for their high degree of authenticity [5]. The corpus captures Hadiths in Arabic and their corresponding English translations at the narrative level. This is because not all Muslims or those studying Hadith know Arabic, but they are more likely to know English since it is a global language.

The corpus is named "Leeds and KSU (LK) Hadith Corpus" to reflect the collaboration between Leeds University where I am doing my PhD, and King Saud University (KSU) where I am a lecturer. Potentially, this project will be extended to a larger scope that I plan to work on at KSU after finishing my PhD.

[5] From a Sunni perspective.

To the best of my knowledge, there is no corpus for Hadith that is well-structured and freely available for the research community. This new corpus can be accessed through my GitHub (LK-Hadith-Corpus)[6]. The availability of the corpus has proved to be useful for the research community as it is already being used (Habash *et al.*, 2022; Tarmom *et al.*, 2021). Chapter 5 introduces the methodology of creating the corpus and discusses potential uses.

1.3.3 Dataset of Related Quran and Hadith Pairs

This thesis investigates the ability of AI to model the meaning of Hadith. Hence, a downstream task of Semantic Textual Similarity (STS) is defined to measure the extent AI can capture meaning in Hadith texts by detecting their relatedness to Quran texts. Such a task requires an evaluation dataset of related Quran and Hadith texts, which shall be created as no such dataset, to my best knowledge, exists. However, following the traditional approach of using human annotators and taking the Kappa agreement is not appropriate. This is because religious texts are complex and require religious scholars to interpret them. Hence, I developed a novel framework to create Quran-Hadith pairs by extracting relatedness information from a reliable source: an archived Fatwa of an Islamic scholar as discussed in Chapter 7. The new Quran-Hadith dataset is available on my GitHub (Quran_Hadith_Datasets)[7].

The usefulness of this dataset is two-fold. First, provides the NLP research community with a challenging dataset of CA that is considered low-resourced in terms of available datasets. Second, finding semantic similarity in religious scriptures is especially useful to the Islamic studies field where various religious texts are processed manually. For example, Quran exegesis is a living activity that can be done by several means (Saeed, 2018). One approach is by finding related Quran verses, Hadiths, or reasons of revelation to help understand a specific verse better or shed light on new insights. Other means include Biblical, Rabbinical, Syriac etymology, or various modern ideologies (Saeed, 2018). Hence, AI tools

[6] https://github.com/ShathaTm/LK-Hadith-Corpus

[7] https://github.com/ShathaTm/Quran_Hadith_Datasets

that can process low-resource languages will be useful to identify similarities in the different scriptures, thus, assisting researchers by limiting the time spent in sifting through texts, and detecting underlying meanings that are not yet identified.

1.3.4 Evaluating Quran Ontologies for Hadith

Once the Quran-Hadith dataset was created, I studied the different approaches to semantic similarity detection. There are two main approaches, one of which is a knowledge-based approach using ontologies, while the other involves deep learning. The former approach is feasible if appropriate ontologies exist. Since there are many Quran ontologies, I have enumerated, compared and evaluated their 'fit' for Hadith as discussed in Chapter 6. This is the first evaluation of existing Quran ontologies' appropriateness for the Hadith. I have identified the most comprehensive Quran ontology, which covers 26.8% of Hadith topics. Therefore, I decided to try the second approach using state-of-the-art (SOTA) DL models which has shown remarkable results on many NLP downstream tasks (Devlin *et al.*, 2018; Conneau *et al.*, 2019).

1.3.5 Semantic Similarity Between Quran and Hadith

To the best of my knowledge, automatically linking Hadith to the Quran is a novelty. The existing links between Quran and Hadith are explicitly mentioned by the Prophet and Islamic scholars. Hence, new interpretations could be inferred if current SOTA tools succeed in modelling the meaning of these CA texts. Consider the example below. The Quran-verse and Hadith-Matn cover the same topic, but the wording is completely different. Chapter 8 evaluates SOTA models' ability to detect semantic similarity in these texts.

> يَا أَيُهَا الَذِينَ آمَنُوا لاَ تَأْكُلُوا الرِبا أَضْعَافًا مُضَاعَفَةً وَاتَقوا اللَه لَعَلَكمْ تُفْلِحُونَ.

> O you who have believed, do not consume usury, doubled and multiplied, but fear Allah that you may be successful.
>
> -*The Quran* [3:130]

> لَيَأْتِيَنَّ عَلَى النَّاسِ زَمَانٌ لاَ يُبَالِي الْمَرْءُ بِمَا أَخَذَ الْمَالَ، أَمِنْ حَلاَلٍ أَمْ مِنْ حَرَامٍ.
>
> Certainly a time will come when people will not bother to know from where they earned the money, by lawful means or unlawful means.
>
> - *The Hadith* [Matn from Sahih Albukhari]

1.4 Thesis Outline

This thesis is divided into four parts as shown in Figure 1.4. The first introduces necessary background, Chapter 2 contains the relevant information about the Arabic language, Quran and Hadith. It explains challenges associated with the Arabic language and all its variants. Then it demonstrates the obstacles to applying NLP on the Quran and Hadith texts in particular. Chapter 3 presents the literature review, which gives a general overview of topics tackled in this thesis, while the more focused work is discussed in the associated chapters.

Part 2 answers the first research question of the thesis and introduces the first two contributions. Chapter 4 describes the creation of the Hadith segmenter. This is followed by Chapter 5, which presents the creation of the LK Hadith corpus using the aforementioned Hadith segmenter.

Part 3 aims to answer the second research question of this thesis. Chapter 7 describes the creation of the evaluation dataset of Quran-Hadith pairs which is used for the downstream task of detecting semantic similarity in Islamic religious texts. Since there are several approaches to semantic similarity, Chapter 6 focuses on enumerating and evaluating Quran ontologies' 'fit' for Hadith to ascertain if a knowledge-based approach is feasible. Then Chapter 8 demonstrates the evaluation of DL based approaches to semantic similarity.

The thesis concludes with Part 4, in which Chapter 9 summarizes the main contributions, discusses challenges and limitations of the work, and presents recommendations for future directions.

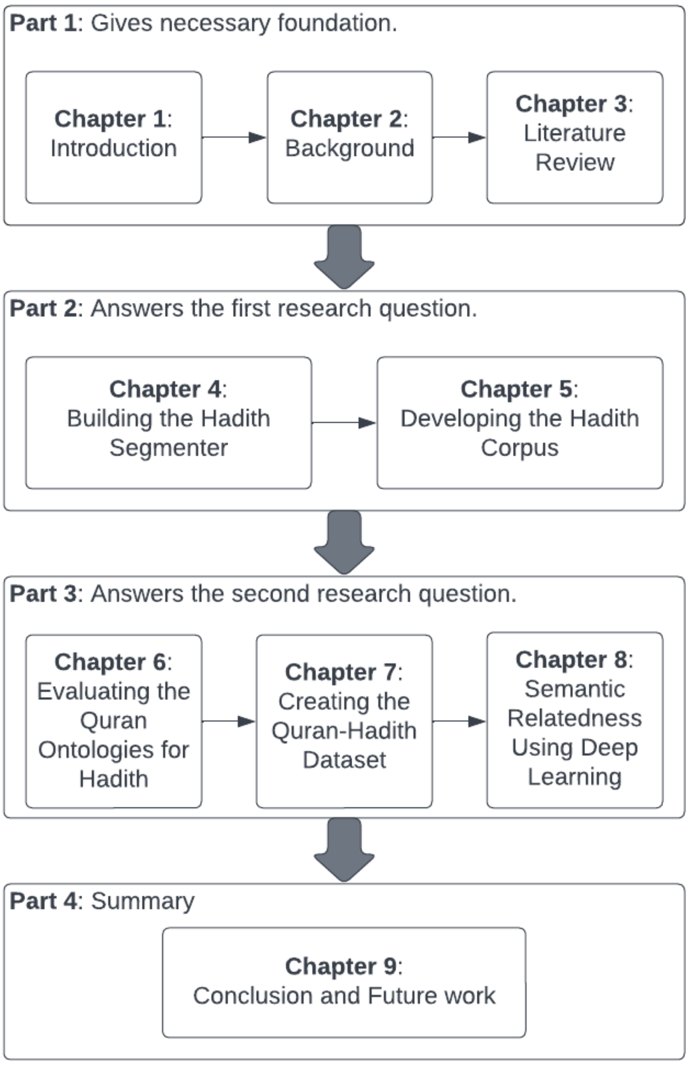

Part 1: Gives necessary foundation.
Chapter 1: Introduction
Chapter 2: Background
Chapter 3: Literature Review
Part 2: Answers the first research question.
Chapter 4: Building the Hadith Segmenter
Chapter 5: Developing the Hadith Corpus
Part 3: Answers the second research question.
Chapter 6: Evaluating the Quran Ontologies for Hadith
Chapter 7: Creating the Quran-Hadith Dataset
Chapter 8: Semantic Relatedness Using Deep Learning
Part 4: Summary
Chapter 9: Conclusion and Future work

Figure 1.4: Thesis outline

Chapter 2

Background

2.1 Introduction

This chapter provides a brief historical and linguistic background to the Arabic language and points to references that elaborate on the challenges associated with it. It also explains the Islamic religious texts, with a particular focus on the Quran and Hadith. These texts hold great significance for Muslims around the world, but they also present challenges in understanding and interpreting their meanings due to cultural and linguistic barriers. The chapter discusses these difficulties and highlights helpful resources for further study.

2.2 The Arabic Language

Arabic is one of the most widely spoken languages today and possesses a great importance due to its association with Islam. Muslims use it in their daily prayers five times a day while reciting the Quran, which was revealed and must be read in Arabic. However, today the term 'Arabic' is multivalent and can be divided into three categories, Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialect (or Colloquial) Arabic (DA) (Habash, 2010). The relatedness between the different categories is depicted in Figure 2.1. These Arabic variants can be used daily by the same person depending on the different social context in a linguistic situation called diglossia (Ferguson, 1959). For example, DA is used at home, CA is used in prayer, and MSA is used at work.

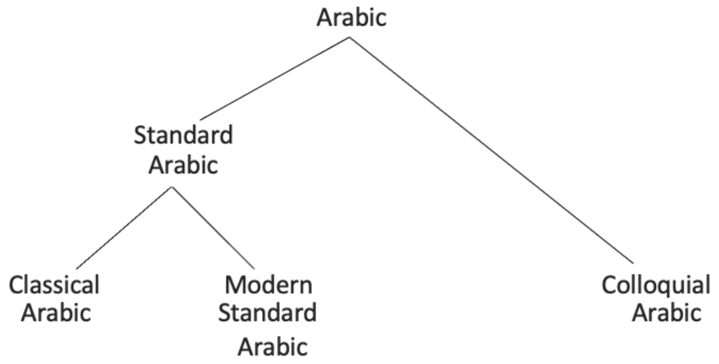


Figure 2.1: Arabic language categories

The Arabic language with all its forms shares basic properties that impose challenges to NLP tasks. Arabic is written from right to left in Arabic scripts, which is also used in other languages like Urdu. Arabic letters have different forms according to the position of the letter in a word as shown in the first example in Table 2.1. Although once a problem, this is no longer an issue as many of the current NLP tools can process Arabic letters.

Arabic letters, almost without exception, have a one-to-one mapping to their sound, but short vowels are represented as diacritics which are optionally written to disambiguate words. These diacritics often change the meaning of a word as shown in the second example in Table 2.1 where the word عقد with different diacritics can mean different things like ‘hold’, ‘necklace’, ‘complex’, and ‘contract’. Most of the available Arabic texts are generally undiacritized as in Wikipedia, news websites, and social media. This is because Arabic readers are able to disambiguate words based on context.

Another Arabic property is agglutination, where a sentence can be writing in a single complex word that contains affixes and clitics representing various parts of speech as shown in the third example in Table 2.1 (Habash, 2010). Not only complex morphological, structural and grammatical nature of Arabic that impose a challenge, but also there is a great level of ambiguity due to lack of capitalization and minimal punctuation in texts (Farghaly & Shaalan, 2009). Moreover, there are challenges specific to each form as discussed below.

Ex.	Property	Example
1	Different letter-forms depending on position, like the letter ع.	باع - عن - بعد - مع
2	Short vowels are represented using diacritics to disambiguate meanings.	عَقْد - عقَّدَ - عِقْد - عَقَدَ
3	Agglutination is common in Arabic words.	وسيكفيكهم

Table 2.1: Examples of orthographic challenges associated with the Arabic language

2.2.1 Classical Arabic

CA is the archaic version of the Arabic language that was spoken during the Prophet's time in the seventh century. It is believed that this form of Arabic disappeared from prose during the second half of the eighth century AD (Fischer, 2013). However, it remained the educational language of the Muslim elites. Today, it represents the highest form of Arabic, found in the Quran and Hadith, and it is thus the ritual language used by Muslim communities around the world.

The importance of CA stems from the requirement of scholars to be erudite in CA to explain the meaning of the Quran. Ibn Abbas was the Prophet's cousin and pioneering Quran exegete who is quoted to explain Quranic words by referring to their usage in CA poetry and referring to the usage of the أعراب 'A'rab' (Bedouin)

[8] since their language represented the purest form of CA [9].

Table 2.2 shows examples from the renowned book البرهان في علوم القران by the famous Islamic scholar *Alzrkashi* where he quoted Ibn Abbas explaining Quranic words by referring to incidents with an أعرابي ‘A’eraby’. Example 1 shows the meaning of ‘فاطر’ as quoted in the incident documented in the next lines:

> ابن عباس: ما كنت أدري ما فاطر السموات والأرض حتى أتاني إعرابيان يختصمان في بئر فقال أحدهما: أنا فطرتها ، يعني ابتدأتها.
>
> ibn Abbas: I did not know what fātir of the heavens and the earth was until two Arabs came to me arguing in a well, and one of them said: I fatartuhā, meaning ‘I initiated it’ .

Examples 2 and 3 in Table 2.2 involves similar stories in which ibn Abbas derives the meaning of some Quranic words by refer to incidents where they were used by an A’raby. [10].

Example	**Quran text**	**Actual Meaning**
1	فَاطِرِ السَّمَاوَاتِ وَالْأَرْضِ	فاطر يعني ابتدأها
	The Creator of the Heavens and the Earth	‘fātir’ means start it
2	مَتَىٰ هَذا ٱلْفَتْحُ إِن كُنتُمْ صَدِقِينَ	الفتح يعني القضاء
	When will this promise be fulfilled, if ye are truthful?	‘Alfathu’ means judgment
3	فَبَشَّرْنَهَا بِإِسْحَق وَمِن وَرَآءِ إِسْحَق يَعْقُوبَ	وراء يعني ولد الولد
	then We gave her the good news of Isaac and Jacob after Isaac	‘warā’ is the grandson

Table 2.2: Examples from Alzrkashi’s book

[8] Also known as the Bedouins, they are the inhabitants of the deserts, as contrasted with the settled people in the villages.

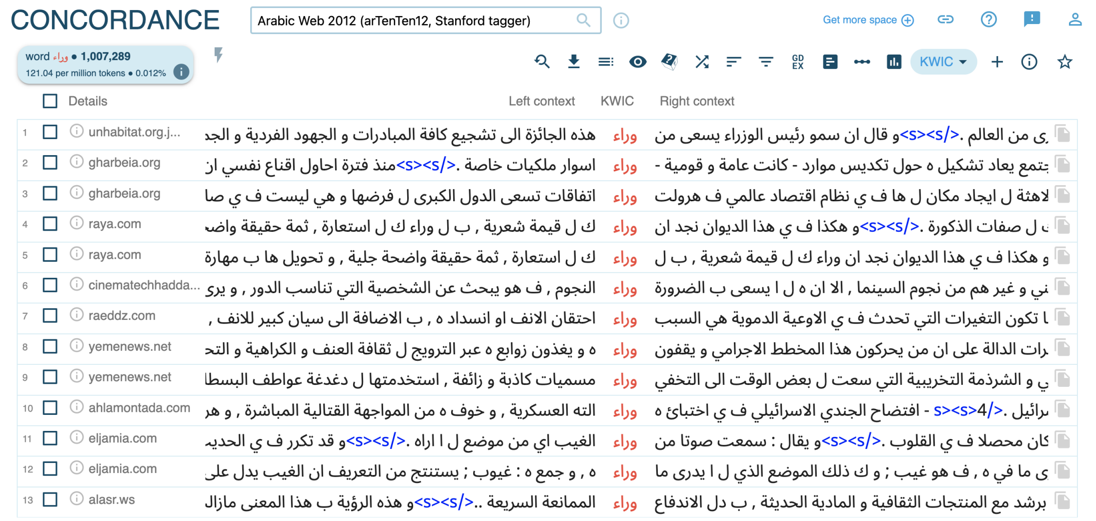
[9] البرهان في علوم القرآن للزركشي

[10] For more details refer to Alzrkashi’s book page 293

The most challenging aspect of CA is language change, where some vocabulary items became obsolete or experienced a semantic shift (Fischer, 2013; Manning, 2015). This is especially problematic with the rise of deep learning (DL) models that require large data with hundreds of examples to facilitate dealing with ambiguity by learning contextual representations of words.

2.2.2 Modern Standard Arabic

MSA is derived from CA and is not any Arab's native language. However, it is the official language of all Arab countries, since it is used in formal settings such as education, news and political interviews. MSA is different from CA in terms of some vocabulary and stylistic features while the morphology and syntactic features remain almost entirely the same (Fischer, 2013). For example, Figure 2.2 shows the concordance of the word وراء mentioned in Example 3 in Table 2.2. It is evident that the meaning in all the examples is 'behind' and not 'grandson'.


CONCORDANCE
Arabic Web 2012 (arTenTen12, Stanford tagger)
Get more space
word وراء • 1,007,289
121.04 per million tokens • 0.012%
GD EX
KWIC

	☐ Details	Left context	KWIC	Right context
1	☐ unhabitat.org.j...	هذه الجائزة الى تشجيع كافة المبادرات و الجهود الفردية و الجا	وراء	ى من العالم .<s></s>و قال ان سمو رئيس الوزراء يسعى من
2	☐ gharbeia.org	اسوار ملكيات خاصة .<s></s>منذ فترة احاول اقناع نفسي از	وراء	تمع يعاد تشكيل ه حول تكديس موارد - كانت عامة و قومية -
3	☐ gharbeia.org	اتفاقات تسعى الدول الكبرى ل فرضها و هي ليست ف ي صا	وراء	لاهثة ل ايجاد مكان ل ها ف ي نظام اقتصاد عالمي ف هرولت
4	☐ raya.com	ك ل قيمة شعرية , ب ل وراء ك ل استعارة , ثمة حقيقة واضـ	وراء	ل ل صفات الذكورة .<s></s>و هكذا ف ي هذا الديوان نجد ان
5	☐ raya.com	ك ل استعارة , ثمة حقيقة واضحة جلية , و تحويل ها ب مهارة	وراء	و هكذا ف ي هذا الديوان نجد ان وراء ك ل قيمة شعرية , ب ل
6	☐ cinematechhadda...	النجوم , ف هو يبحث عن الشخصية التي تناسب الدور , و يرى	وراء	ني و غير هم من نجوم السينما , الا ان ه ل ا يسعى ب الضرورة
7	☐ raeddz.com	احتقان الانف او انسداد ه , ب الاضافة الى سيان كبير للانف ,	وراء	ا تكون التغيرات التي تحدث ف ي الاوعية الدموية هي السبب
8	☐ yemenews.net	ه و يغذون زوابع ه عبر الترويج ل ثقافة العنف و الكراهية و التح	وراء	رات الدالة على ان من يحركون هذا المخطط الاجرامي و يقفون
9	☐ yemenews.net	مسميات كاذبة و زائفة , استخدمتها ل دغدغة عواطف البسطا	وراء	ي و الشرذمة التخريبية التي سعت ل بعض الوقت الى التخفي
10	☐ ahlamontada.com	الته العسكرية , و خوف ه من المواجهة القتالية المباشرة , و هر	وراء	رائيل .s><s>4/> - افتضاح الجندي الاسرائيلي ف ي اختباء ه
11	☐ eljamia.com	الغيب اي من موضع ل ا اراه .<s></s>و قد تكرر ف ي الحديـ	وراء	كان محصلا ف ي القلوب .<s></s>و يقال : سمعت صوتا من
12	☐ eljamia.com	ه , و جمع ه : غيوب ; يستنتج من التعريف ان الغيب يدل على	وراء	ى ما في ه , ف هو غيب ; و ك ذلك الموضع الذي ل ا يدرى ما
13	☐ alasr.ws	الممانعة السريعة ..<s></s>و هذه الرؤية ب هذا المعنى مازال	وراء	برشد مع المنتجات الثقافية و المادية الحديثة , ب دل الاندفاع

Figure 2.2: Concordance from the arTenTen12 corpus

2.2.3 Dialect Arabic

Dialect Arabic, i.e. Arabic dialects are regional languages and the native tongue of Arabic speakers. Arabic dialect can be divided into five main groups on linguistic grounds: Egyptian, Levantine, Gulf, Iraqi and Moroccan (Habash, 2010). These can be further sub-divided into fine-grained dialects. For example, the Gulf alone has several dialects such as Hijazi, Najdi, Bahraini, and Kuwaiti, as shown in Figure 2.3. With the rise of social media platforms, these dialects became pervasive in written form posing a challenge to NLP tools. This is partially because they lack standard orthography which lead to many spelling variations and inconsistencies (Habash *et al.*, 2012a).

Dialects contain some words that are unique and different from MSA. For example, 'I want' is expressed differently depending on the dialect. It is 'ابي' in Najdi, 'عايز' in Egyptian and 'داير' in Sudani . This suggests that dialects could be considered different languages [11]. For this reason, it is crucial to be careful when building ML or DL models since training data should be chosen depending on the language and downstream task (Inoue *et al.*, 2021).

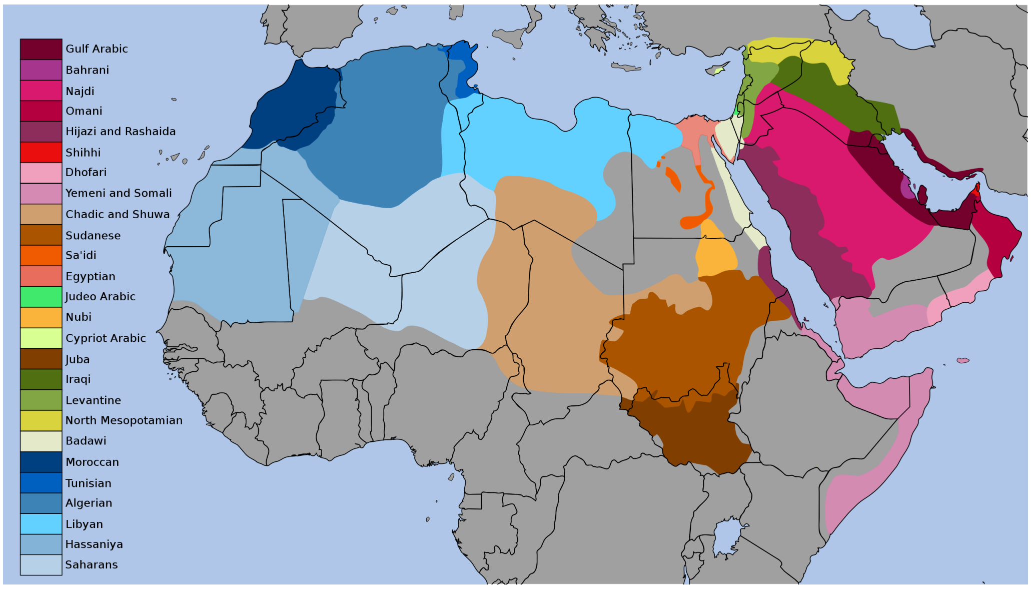


Figure 2.3: Arabic dialects are different within the same country[12]

[11] International Organization for Standardization (ISO) considers them different languages: https://iso639-3.sil.org/code/ara

[12] Source: https://en.wikipedia.org/wiki/Varieties_of_Arabic

After discussing the different variants of Arabic, the CA Quran and Hadith are explained in the next paragraphs.

2.3 The Quran

Muslims believe the Quran is God's divine words transmitted to the Prophet Muhammad by the angel Gabriel over a period of 23 years. It contains a variety of topics including guidance and historical narrative. However, the Quran is not a detailed legal manual for every life aspect. In fact, only about five hundred of the book's verses provide legal injunctions (Brown, 2017). Hence, Muslims believe the Prophet's Hadith is essential for elaborating the Quranic message to unlock its manifold meanings to an evolving community.

2.3.1 Quran Structure

The Quran has 114 chapters (Surahs), each including a varying number of verses (Ayas). The total number of verses in the Quran is 6,236, and an individual verse can be a few letters long or consist of sentences that span several lines. So a verse can cover many topics, or one topic can be covered by short consecutive verses. Moreover, a topic can be covered by verses scattered in different chapters in the Quran. Hence, many studies have attempted to build an ontology for the Quran to enable connecting these topics that can be used in information retrieval systems (Alqahtani & Atwell, 2018).

2.3.2 Quran's Linguistic Features

The Quran is considered the purest form of Arabic. Hence, it is used to study the Arabic grammar and its linguistic features. The famous grammarian *Sibawayh* derived the earliest books on Arabic grammar, Al-Kitab, by relying heavily on the Quran. This is because the Arabic language was codified primarily in the Quran (Watson, 2007).

Every letter and diacritic in the Quran contributes to the meaning. For example, verse [2:61] shown below contains the phrase اهْبِطُوا مِصْراً and verse [12:99] contain a similar phrase ادْخُلُوا مِصْرَ so both contain the word مصر ‘Misr’ as an object مفعول به. Although مصر is in the same part of speech in both examples, it does not have the same diacritics. The first has تنوين ‘Tanween’ to denote a city (any city outside the desert), while the second is ممنوع من الصرف ‘indeclinable’ to indicate it is a foreign name أعجمي and refers to the city, or country, we know now as Egypt [13].

> وَإِذْ قُلْتُمْ يَـٰمُوسَىٰ لَن نَّصْبِرَ عَلَىٰ طَعَامٍ وَٰحِدٍ فَادْعُ لَنَا رَبَّكَ يُخْرِجْ لَنَا مِمَّا
> تُنۢبِتُ ٱلْأَرْضُ مِنۢ بَقْلِهَا وَقِثَّآئِهَا وَفُومِهَا وَعَدَسِهَا وَبَصَلِهَا قَالَ أَتَسْتَبْدِلُونَ
> ٱلَّذِى هُوَ أَدْنَىٰ بِذى هُوَ خَيْرٌ ٱهْبِطُوا مِصْرًا فَإِنَّ لَكُم مَّا سَأَلْتُمْ وَضُرِبَتْ
> عَلَيْهِمُ ٱلذِّلَّةُ وَمَسْكَنَةُ وَبَآءُو بِغَضَبٍ مِّنَ ٱللَّهِ ذَٰلِكَ بِأَنَّهُمْ كَانُوا يَكْفُرُونَ
> بِـَٔايَـٰتِ ٱللَّهِ وَيَقْتُلُونَ ٱلنَّبِيِّـۧنَ بِغَيْرِ ٱلْحَقِّ ذَٰلِكَ بِمَا عَصَوا وَّكَانُوا يَعْتَدُونَ
>
> And [recall] when you said, “O Moses, we can never endure one [kind of] food. So call upon your Lord to bring forth for us from the earth its green herbs and its cucumbers and its garlic and its lentils and its onions.” [Moses] said, “Would you exchange what is better for what is less? Go into [any] settlement and indeed, you will have what you have asked.” And they were covered with humiliation and poverty and returned with anger from Allah [upon

[13] https://www.ahewar.org/debat/show.art.asp?aid=168080

> them]. That was because they [repeatedly] disbelieved in the signs of Allah and killed the prophets without right. That was because they disobeyed and were [habitually] transgressing.
>
> -*The Quran* [2:61]

> فَلَمَّا دَخَلُوا عَلَىٰ يُوسُفَ ءَاوَىٰ إِلَيْهِ أَبَوَيْهِ وَقَالَ ٱدْخُلُوا مِصْرَ إِن شَآءَ ٱللَّهُ ءَامِنِينَ
>
> And when they entered upon Joseph, he took his parents to himself and said, “Enter Egypt, Allah willing, safe [and secure]”.
>
> -*The Quran* [12:99]

The Quran has been studied extensively to explore it linguistically and spiritually. One of these fields is علم الوجوه والنظائر ‘The science of polysemy’ where famous scholars like Ibn al-Jawzi, Nisaburi, Al-Suyuti, and many others have studied polysemic words in the Quran to explain the different meanings of the same word in different contexts. For example, ibn Aljawzi’s book نزهة الاعين النواظر في علم الوجوه والنظائر ‘The excursion of watchful eyes in the field of polysemy’ discusses the different meanings of Quranic words as shown in Table 2.3.

Another common feature in the Quran is the abstract imagery as in the verse [17:24] shown below. It contains the words ‘lower’أخفض and ‘wing’ جناح, which

Meaning of the word ‘عين’	English Translation
العين اللتي يبصر بها	The eye(s) which the person uses to see
الذهب	Gold
كثرة المطر	Heavy rain
نفس الشيء	The same

Table 2.3: The different meanings of the word ‘Ayn’

have completely different meanings in this phrase as illustrated by the translation. This is due to the phenomenon that frequently a “combination of words creates a meaning that they do not have in isolation” (Dickins *et al.*, 2016, p. 97).

> وَٱخْفِضْ لَهُمَا جَنَاحَ ٱلذُّلِّ مِنَ ٱلرَّحْمَةِ وَقُل رَّبِّ ٱرْحَمْهُمَا كَمَا رَبَّيَانِى صَغِيرًا
>
> And be humble with them out of mercy, and pray, “My Lord! Be merciful to them as they raised me when I was young.”
>
> -*The Quran* [17:24]

More Quranic features are discussed in Chapter 7.

2.3.3 Quran Exegesis (Tafsir)

There are many books of Tafsir aimed at explaining the meanings of Quran’s verses , clarifying their importance and indicating their significance. This is because “Tafsir is an ongoing activity rather than something settled in the distant past” (Saeed, 2018, p. 17). Hence, many Tafsir books have been written throughout the past 1400 years and Tafsir is still developing to accommodate the development of societies that ask new and different questions.

Tafsir of the Quran can be done by referring to the Quran itself (تفسير القران بالقران ‘Interpretation of the Qur’an by the Qur’an’). This is common in many situations as illustrated in verses [6:82] and [31:13] below. The first verse contains the word ظُلْمٌ ‘Injustice’ which implies different meanings, but the meaning is clarified in verse [31:13] to indicate the meaning in this context is ٱلشِّرْكَ ‘Polytheism’.

Another approach to conducting Tafsir is through the Hadith. For example, verse [2:173] below indicates which food if is permissible to eat and which is forbidden. More details about this topic are obtained from the Hadith that follows. It gives further details to indicate that anything from the sea is permissible to eat.

> ٱلَّذِينَ ءَامَنُوا وَلَمْ يَلْبِسُوٓا۟ إِيمَـٰنَهُم بِظُلْمٍ أُو۟لَـٰٓئِكَ لَهُمُ ٱلْأَمْنُ وَهُم مُّهْتَدُونَ
>
> It is those who believe and confuse not their beliefs with wrong - that are (truly) in security, for they are on (right) guidance.
>
> -*The Quran* [6:82]

> وَإِذْ قَالَ لُقْمَـٰنُ لِٱبْنِهِۦ وَهُوَ يَعِظُهُۥ يَـٰبُنَىَّ لَا تُشْرِكْ بِٱللَّهِ إِنَّ ٱلشِّرْكَ لَظُلْمٌ عَظِيمٌ
>
> And [mention, O Muhammad], when Luqman said to his son while he was instructing him, “O my son, do not associate [anything] with Allah . Indeed, association [with him] is great injustice.”
>
> -*The Quran* [31:13]

> إِنَّمَا حَرَّمَ عَلَيْكُمُ الْمَيْتَةَ وَالدَّمَ وَلَحْمَ الْخِنزِيرِ وَمَا أُهِلَّ بِهِ لِغَيْرِ اللَّـهِ فَمَنِ
> اضْطُرَّ غَيْرَ بَاغٍ وَلَا عَادٍ فَلَا إِثْمَ عَلَيْهِ إِنَّ اللَّـهَ غَفُورٌ رَّحِيمٌ
>
> He has only forbidden to you dead animals, blood, the flesh of swine, and that which has been dedicated to other than Allah. But whoever is forced [by necessity], neither desiring [it] nor transgressing [its limit], there is no sin upon him. Indeed, Allah is Forgiving and Merciful.
>
> -*The Quran* [2:173]

> سأَلَ رجلٌ النَبْيَ صلَىْ اللهُ علَيهِ وسلَمِ فقالَ يا رسولَ اللَه إنَاْ نركبُ البحرَ
> ونحمِلُ معنا القَليلَ منَ الماءِ فإن توضَأنا بِهِ عطَشنا أفنتَوضَأ بماءِ البحرِ فقالَ
> رسولُ اللَه صلَىْ اللهُ علَيهِ وسلَم هوَ الطَهْورُ ماؤُهُ الحلُ ميتتُهُ
>
> "A man asked the Prophet (PBUH): O Messenger of Allah, we travel by sea and we take a little water with us, but if we use it for Wudu, we will go thirsty. Can we perform Wudu with seawater? The Messenger of Allah (PBUH) said: Its water is a means of purification and its dead meat is permissible."
>
> -*The Hadith*

2.4 The Hadith

Muslims believe the Quran is God's divine words, which enjoined them to follow the guidance of the Prophet Muhammad in their laws, legislations, and moral guidance. This is evident in the Quran as shown in the below verse.

> قُلْ أَطِيعُوا۟ ٱللَّهَ وَأَطِيعُوا۟ ٱلرَّسُولَ فَإِن تَوَلَّوْا۟ فَإِنَّمَا عَلَيْهِ مَا حُمِّلَ وَعَلَيْكُم مَّا
> حُمِّلْتُمْ ۖ وَإِن تُطِيعُوهُ تَهْتَدُوا۟ ۚ وَمَا عَلَى ٱلرَّسُولِ إِلَّا ٱلْبَلَٰغُ ٱلْمُبِينُ
>
> Say, “Obey Allah and obey the Messenger; but if you turn away - then upon him is only that [duty] with which he has been charged, and upon you is that with which you have been charged. And if you obey him, you will be [rightly] guided. And there is not upon the Messenger except the [responsibility for] clear notification.”
>
> -*The Quran* [24:54]

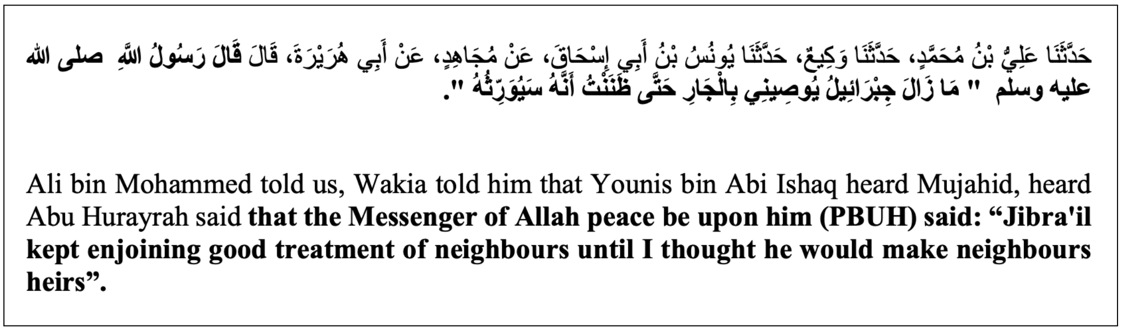
This clear instruction to emulate the Prophet and follow his judgements is necessary because not all Islamic laws and regulations are mentioned in the Quran. For example, Islamic prayer, including the calling to prayer, the *Adhaan*, is obtained from the Prophet’s reported actions. Ritual prayers are stated in the Quran as an obligation without the exact details of practice, which shows that most Islamic practice is obtained from the Prophet. Therefore, recording the Prophet’s words and actions is of great importance.

The act of reporting the different aspects of the Prophet’s life became known as Hadith, which is an Arabic word for ‘speech’, ‘report’, or ‘narrative’. Hadith types vary. They could be a short sentence or long paragraph describing what the Prophet said in a specific incident, a dialogue of the Prophet’s conversation with someone, or a story told by the Prophet’s companions that explains the Prophet’s actions in a specific matter like prayers.

Unlike the Quran, Hadith was not documented immediately after the Prophet’s death. Instead, it was passed down the generations verbally by scholars each mentioning the person from whom they heard the Hadith. However, some dishonest individuals deliberately fabricated material and ascribed it to the Prophet. This led to the development of Hadith science, in which scholars study the chain of narrators and their biographies to accept or reject the Hadith teaching. The process of this formed the unique structure of Hadith (Brown, 2017).

2.4.1 Hadith Structure

Hadith consists of two parts, as shown in Figure 2.4, the **Isnad** which is a chain of narrators followed by the **Matn** in bold, which is the actual teaching. The term ‘Isnad’ can be translated to mean ‘support’, since it is used to identify the authenticity of Hadith following the narrator’s genealogy. It is meta-data that is useful for authenticity but does not add useful information to the context of the actual narration (Matn). Therefore, in designing the Hadith corpus, it is crucial to separate the Isnad from the Matn to allow researchers to focus on their text of interest.



حَدَّثَنَا عَلِيُّ بْنُ مُحَمَّدٍ، حَدَّثَنَا وَكِيعٌ، حَدَّثَنَا يُونُسُ بْنُ أَبِي إِسْحَاقَ، عَنْ مُجَاهِدٍ، عَنْ أَبِي هُرَيْرَةَ، قَالَ **قَالَ رَسُولُ اللَّهِ صلى الله**
عليه وسلم " مَا زَالَ جِبْرَائِيلُ يُوصِينِي بِالْجَارِ حَتَّى ظَنَنْتُ أَنَّهُ سَيُوَرِّثُهُ ".

Ali bin Mohammed told us, Wakia told him that Younis bin Abi Ishaq heard Mujahid, heard Abu Hurayrah said **that the Messenger of Allah peace be upon him (PBUH) said: “Jibra'il kept enjoining good treatment of neighbours until I thought he would make neighbours heirs”.**

Figure 2.4: Hadith example, Matn in bold

Unlike the Quran, which consists of a number of verses in one sura, each Hadith is a stand-alone statement or act by the Prophet that was later written, collected, and compiled into books. Scholars have categorized Hadiths into topics by relying on their deep knowledge and understanding of Hadith. The following section gives a general idea about the different types of Hadith books.

2.4.2 Hadith Books

Not all Hadiths are considered authentic. For this reason, early Islamic scholars identified the need to compile authentic Hadiths for later generations. The pioneering work was conducted by an Islamic scholar named Muhammad Albukhari who died in 870 C.E. He collected Hadiths that met the most rigorous standards of authenticity into a book commonly known as ‘Sahih Albukhari’. The word ’Sahih ‘is an Arabic word which means correct, sound and authentic.

Albukari’s book is considered a hybrid of two genres of Hadith books, *Musannaf* and *Musnad* (Brown, 2017). The former include books that categorize Hadiths into topics and does not emphasis the authenticity of Hadith. Some books do not include the Isnad of the Hadith. Musnad books organize Hadith based on chains of narrators. This hybrid genre of Hadith became known as Sahih or Sunan books where Hadiths are organized under subtitles that deal with Islamic law, dogma, and the legal implication the reader should derive from the subsequent Hadiths. This is a useful feature that is exploited in this research as shown in Chapter 6.

Following Albukhari’s lead, other Islamic scholars compiled more Sahih/Sunan books including Albukhari’s student, Muslim ibn al-Hajjaj, who compiled a Hadith book known as ‘Sahih Muslim’. Currently, there are six recognised Sahih/Sunan books collectively referred to as “Al-Sihah al-Sittah”, which translates as “The Authentic Six” or canonical books. These books are *Sahih Albukhari*[14], *Sahih Muslim*[15], *Sunan Abu Dawood*[16], *Sunan Altarmithi*[17], *Sunan Ibn Maja*[18], and *Sunan Al-Nasai* [19] [20]. They form the bases for Islamic Hadith books generally. Although they are called the authentic six, not all incorporated Hadiths possess the same degree of authenticity, but they were named the authentic six based on the majority of Hadiths in these books (Khan, 1987). The six books contain approximately 19,600 Hadiths, or around 35,000 with repetition (Brown, 2017).

[14] Mohammad Albukhari, Sahih Albukhari (2002). Damascus: Dar Ibn Kathir.

[15] Muslim Ibn Al-Hajjaj, Sahih Muslim (2017). Damascus: Dar Ibn Kathir.

[16] Abu Dawood, Sunan Ibi Dawood (2017). Damascus: Dar Ibn Kathir.

[17] Altarmithi , Sunan Altarmithi (2016). Damascus: Dar Ibn Kathir.

[18] Ibn Majah , Sunan ibn Majah (2016). Damascus: Dar Ibn Kathir.

[19] Imam Nasai, Sunan Al-Nasai (2017). Damascus: Dar Ibn Kathir.

[20] The English spelling of the books names is adapted from Azmi *et al.* (2019), a survey of Hadith computational research .

2.4.3 Hadith Types

There are different types of Hadiths, which Brown (2017) classifies into a number of categories. First, there are historical narratives, which incorporate stories about past events. These are usually about previous Prophets and their people, which the archangel Gabriel transmitted to the Prophet Mohammed (PBUH). Second, there are the Prophet's personal narratives, which include stories about his life with his family, companions, and his actions in relation to specific incidents and major events. These were recorded and disseminated to teach Islamic morals and best behaviour. Third, there are metaphysical narratives about the future that usually involve the Day of Judgement and events of the hereafter. Fourth, there are Hadiths that contain facts instead of Prophetic sayings and teachings. These are usually used in authenticating Hadiths and studying narrators' genealogy, as shown in the example below.

> حَدَّثَنَا أَحْمَدُ ـ هُوَ ابْنُ صَالِحٍ ـ حَدَّثَنَا عَنْبَسَةُ، حَدَّثَنَا يُونُسُ، قَالَ ابْنُ
> شِهَابٍ ثُمَّ سَأَلْتُ الْحُصَيْنَ بْنَ مُحَمَّدٍ ـ وَهْوَ أَحَدُ بَنِي سَالِمٍ وَهْوَ مِنْ سَرَاتِهِمْ ـ
> عَنْ حَدِيثِ، مَحْمُودِ بْنِ الرَّبِيعِ عَنْ عِتْبَانَ بْنِ مَالِكٍ، فَصَدَّقَةُ.
>
> Ahmad told us, he is son of Saleh, Anbasa told us, Abu-yunus told us, Ibn Shihab said, I asked Alhusain bin Muhammed, he is one of Bani Salim family, about Hadith Mahmood bin Rabi from Utban bin Malik, he confirmed it.
>
> -*The Hadith*

2.4.4 Useful Sources for Hadith

The Hadith and Hadith literature have became prevalent online through several websites like *sunnah.com*, *hadithcollection.com*, and *ahadith.co.uk*. To create a parallel corpus of Arabic Hadith and its English translations aligned at the narrative level, an exhaustive search and comparison of available web resources was conducted as discussed in chapter 5. Through the search, I stumbled upon very useful resources for computational Hadith research in general, which I enumerate in Table 2.4. These include موسوعة الحديث [21] , الدرر السنية [22] , موسوعة رواة الحديث [23]قاعدة بيانات علماء المسلمين [24]معجم الدوحة التاريخي للغة العربية [25]. I think these can be utilized in various types of computational Hadith research such as authentication and classification.

source	Description
موسوعة الحديث	A database of narrators and Hadith categorized into topics.
الدرر السنية	A large encyclopedia of Islamic topics including Quran exegesis and Hadith.
موسوعة رواة الحديث	The Hadith Transmitters Encyclopedia is a comprehensive biographical dictionary of Hadith transmitters.
قاعدة بيانات علماء المسلمين	This contains biographies of Muslim scholars.
معجم الدوحة التاريخي للغة العربية	Historical dictionary of Classical Arabic that shows the different meanings of a word and its usage and the date it was first used with that meaning.

Table 2.4: Useful resources on the web that can be utilized in Hadith computational studies

2.4.5 Hadith Commentaries

Hadith literature is huge, consisting of the actual Hadith books like the canonical books mentioned in section 2.4.2 and the supporting work to help grade a Hadith's authenticity or explain its contents. Such books are known as Hadith commentaries, where scholars with in-depth knowledge record their insights into the

[21] https://hadith.islam-db.com/

[22] https://www.dorar.net/

[23] http://hadithtransmitters.hawramani.com

[24] https://muslimscholars.info

[25] https://www.dohadictionary.org

Hadith for the ordinary layman. Commentators devote a lifetime to interpret and elucidating Hadiths by employing their knowledge of the Quran, Quranic commentary (Tafsir), Arabic grammar and rhetoric. The number of commentaries is large because many with different types of explanation have been developed for the same set of Hadiths. For example, Sahih Albukhari alone has fifty-six commentaries produced by different scholars, the most renowned one being Ibn Hajar al-Asqalani's *Fath Al-bari* 'Victory of the Creator' (Blecher, 2016). Some of these commentaries along with the Quran and the Hadith books were used to train and build domain-specific word embeddings as described in Chapter 8.

2.5 Conclusion

In this chapter, the required background is covered for the reader to gain an understanding of the Arabic language and its variants with the associated features that impose a challenge to NLP. Furthermore, the necessary background to Islamic religious texts and their features are discussed to demonstrate the obstacles they may cause in modelling the meaning of Islamic text using AI.

Chapter 3

Literature Review

3.1 Introduction

This chapter gives a general overview of the main research areas that the thesis builds upon and points to the forthcoming chapters where a deeper discussion of the associated literature is provided. Section 3.2 presents a review of the computational Hadith research and concludes with the need for a well-structured Hadith corpus as previous studies have emphasized (Bounhas, 2019; Azmi *et al.*, 2019).

To build such a resource automatically, it is essential to first create a tool that segments and annotates the two main components of Hadith, Isnad and Matn. Therefore, the literature on text segmentation and how to build on established methods are discussed in Section 3.3. Then a brief overview of corpus creation methods is given in Section 3.4 to demonstrate how these shaped the choices made in building the Hadith corpus.

Once the corpus is built, the next goal is to use AI methods to model the meaning of these Classical Arabic (CA) sacred texts to serve those studying them. For example, an AI system can be used to consolidate and process a large body of textual data to enable tasks such as information retrieval, or grouping Hadiths and Quran-verses based on topics. Hence, a downstream task is needed to measure how well the meaning in such complex texts is captured. The task of finding semantic similarities between Quran and Hadith texts is chosen because these texts are different in structure and linguistic style but cover the same domain. Section 3.5 discusses methods of semantic similarity measures and why particular paths were chosen.

3.2 Computational Hadith Research

In the past decade an extensive amount of computational research has been done on the Quran to enhance understanding of its meaning and the message it conveys (Safeena & Kammani, 2013; Bashir *et al.*, 2022). Researchers produced Quranic tools, corpora, and ontologies (Alrehaili & Atwell, 2018; Hamoud & Atwell, 2016; Alqahtani & Atwell, 2018; Hakkoum & Raghay, 2016). Such valuable research can be extended to cover Hadith since there are limited advances in this field despite the importance of Hadith (Bounhas, 2019). The computational research done on Hadith can be categorized into three areas which are discussed in the following subsections.

3.2.1 Hadith Authentication

The first area focusses on classifying Hadith based on authenticity by particularly examining the chain of narrators (Isnad). Several researchers contributed to this area as shown in the survey papers covering the field (Hakak *et al.*, 2022; Binbeshr *et al.*, 2021; Luthfi *et al.*, 2018). This is because Isnad is the proof of authenticity and as a famous scholar named Bin Mubarak stated, “If not for the Isnad anyone could say anything they wanted” and claim it is a Prophet saying (Hasan, 1994). Some work attempts to simply categorise Hadith as authentic or not (Bilal & Mohsin, 2012; Dalloul, 2013; Baraka & Dalloul, 2014). Other work assigns a degree of validity known as ‘Takhreej Al-Hadith’, which is the process of categorizing Hadith based on its validity degree. In other words, it classifies Hadith into groups ranging from authentic to a complete fabrication, known as Sahih (authentic), Hasan (Good), Da’ef(weak) and Maudo (fabricated)(Aldhaln *et al.*, 2012; Ghazizadeh *et al.*, 2008). Another method of Hadith authentication proposed building visual network of narrators using Named Entity Recognition

(NER) and Machine Learning (ML) (Muazzam Siddiqui, 2014; Azmi & Badia, 2010). Since most of these studies required a dataset of narrators (Isnad), the Hadith corpus presented Chapter 5 facilitates using the Isnad since it is segmented from the Matn.

3.2.2 Hadith Topic Classification

The second type of research focuses on the actual narration, the Matn. For example, Matn was used to develop a Hadith WordNet by using tools to identify POS tags and Classical Arabic dictionaries to extract meaning (Alkhatib *et al.*, 2017). Another type of research that focuses on Matn aims to categorize Hadith (Saloot *et al.*, 2016). For example, Jbara (2010) used a supervised approach to classify Hadith based on section titles extracted from Hadith books. Another research work treated Hadith as a life coach that contains suggestions or guidance for Muslims attempting to classify Hadith into three categories: dos, don'ts, and information (Al Faraby *et al.*, 2018).

3.2.3 Hadith Ontology

The third type of Hadith research focuses on creating ontologies to enhance Hadith understanding and information retrieval. This is because ontologies are usually built to represent domain knowledge by identifying a set of concepts and the relationships between them. Ideally, an ontology is represented by the following tuple:

$$O = <C,\ H,\ R,\ A>$$

Here O represents an ontology, C represents a set of classes or concepts, H represents taxonomic relations, which are hierarchical links between the concepts, R represents non-taxonomic relation that include the set of conceptual links, and A represents the set of rules and axioms (Al-Aswadi *et al.*, 2020). Creating an ontology with such features is challenging since fully automatic construction is not feasible and manual creation using experts is expensive. Therefore, the ontology learning approach is used, which involves building ontologies semi-automatically where a combination of computational tools and experts are utilized.

Previous studies in the area of Islamic ontologies have almost exclusively focused on the Quran (Atwell, 2018), while research in Hadith ontologies remains limited (Bounhas, 2019). Existing ontologies cover a specific topic in Hadith like the one presented by Al-Sanasleh & Hammo (2017), who manually produced an ontology of prophets and messengers from the Quran and Hadith by consulting experts in the domain. Another example is the ontology created for Salat (Islamic prayer) by consulting the Qur'an, Hadith, and scholars' books about Salat (Saad *et al.*, 2011).

Some research works have utilized Hadith as a case study to test their proposed computational methods. For example, Lahbib *et al.* (2014) introduced an approach to extract domain terminology and applied it to Sahih Albukhari and its English translation. They used GIZA++ (Och & Ney, 2003) to align the parallel corpus at the word level, which was then used to extract the terms using TF-IDF. However, the extracted terms are not available.

I am not aware of any Hadith ontology that covers all the Hadiths topics.The existing ones model the structure or meta-data of Hadith. For example, Jaafar & Che Pa (2017) built on Dalloul (2013) ontology to link Hadith to commentaries and answer questions such as "What is the commentary of a particular Hadith?". Hence, the area of Hadith ontology is relatively unexplored and requires more work to reach a mature status.

3.3 Text Segmentation

Text segmentation involves splitting texts into meaningful units called segments, which can be a word, sentence, topic, or a phrase (Pak & Teh, 2018). One type of text segmentation is Hadith segmentation,which aims to split a Hadith into two components, Isnad and Matn. The first part is the chain of narrators which exists for the sake of authenticity but does not add meaning to the actual Hadith teaching, the Matn. Hence, the task of Hadith segmentation can be considered a special type of topic segmentation.

Topic segmentation is a well-studied area of research since it is often considered a pre-requisite for tasks like information retrieval or text summarization (Purver, 2011). One of the seminal works in this area is Texttiling (Hearst, 1997), which uses a vector space model to measures cosine similarity between neighbouring units to detect topic change and assign boundaries to dissimilar units. Developments based on the Texttiling method has been proposed including utilizing BERT to obtain units' semantic embeddings (Solbiati *et al.*, 2021) since lexical semantics are better captured with distributed representations (Goldberg, 2017).

Although the performance of the aforementioned methods is encouraging, the task of Hadith segmentation aims to detect the Isnad span, which is a chain of narrators that usually follows a pattern. Therefore it is more logical to first try the classical methods mentioned in Reynar (1998), where features like word cues, word n-grams, word repetition, and word co-occurrence are utilized to identify topic shift.

Special features of Isnad including narrators' names and transmission words like 'said', 'heard', and 'told' have been utilized in previous attempts to segment Hadith. A more specific literature review of Hadith segmentation is found in section 4.2, which discusses how a Hadith segmentation tool presented in this thesis builds on previous research. After creating the Hadith segmenter, the Hadith corpus is created. The next section gives a brief overview of corpus building.

3.4 Corpus Building

'Corpus' is a word derived from Latin meaning 'body', which is used to refer to any collection of texts. The earlies corpus created was in 1897 consisting of 11 million words to study spelling conventions in the German language (McEnery

& Wilson, 1996). Currently, a 'corpus'[26] refers to any body of text available in a machine-readable format that can be used in a wide range of research that covers general topics such as lexicography, syntax, semantics, and language pedagogy. Furthermore, a corpus may answer specific questions such as authorship attribution and change in word meaning through time (Kennedy, 2014).

To make the corpus more useful, it is common practice to add a meaningful annotation. For example, POS tagging a corpus increases its utility and enables better analysis of the text in question. Ideally, corpus annotation should follow Leech's maxims (Leech, 1993), which state that it should be possible to revert to the raw corpus by removing annotation. Furthermore, the end user should be aware of for whom and how the annotation was conducted, which means it is a potentially useful but not infallible tool. Hence, in creating the Hadith corpus, these were considered, as stated in Chapter 5.

The Hadith corpus serves a dual purpose. Firstly, it aims to provide a common resource for Hadith computational studies. Secondly, it aims to improve the availability of data as Classical Arabic is considered under-resourced in terms of available datasets (Alyafeai *et al.*, 2021). Hence, a bilingual parallel corpus of Classical Arabic and the English translations aligned at the Hadith level is created as discussed in Chapter 5. This can be beneficial for various fields of research such as machine translation. In Section 5.2, the existing Arabic corpora are listed and discussed to highlight the usefulness of this new Hadith corpus, in addition to its intended use for detecting semantic similarities in religious CA texts.

3.5 Semantic Similarity

Automatically detecting semantic similarity between texts is useful for general NLP tasks including spelling correction, information retrieval, and plagiarism

[26] The plural form of 'corpus' is 'corpora'.

detection (Hadj Taieb *et al.*, 2020). Moreover, it can be beneficial for domain-specific tasks. For example, researchers at KITAB created the Passim[27] algorithm to detect how extensively pieces of Classical Arabic writing spread across space and time. This is because in Classical Arabic books, it was common practice for authors to copy large chunks from previous books without referring to them. Although the proposed system is not designed to detect paraphrasing, it could be developed by utilizing recent advances in semantic similarity detection methods.

The field of semantic similarity has evolved through several stages starting from lexical similarity by counting similar words, to recent advances in using deep neural network-based methods (Chandrasekaran & Mago, 2021). Figure 3.1 shows a diagram of the various approaches to semantic similarity which are discussed in the following sections

[27] https://kitab-project.org/methods/text-reuse

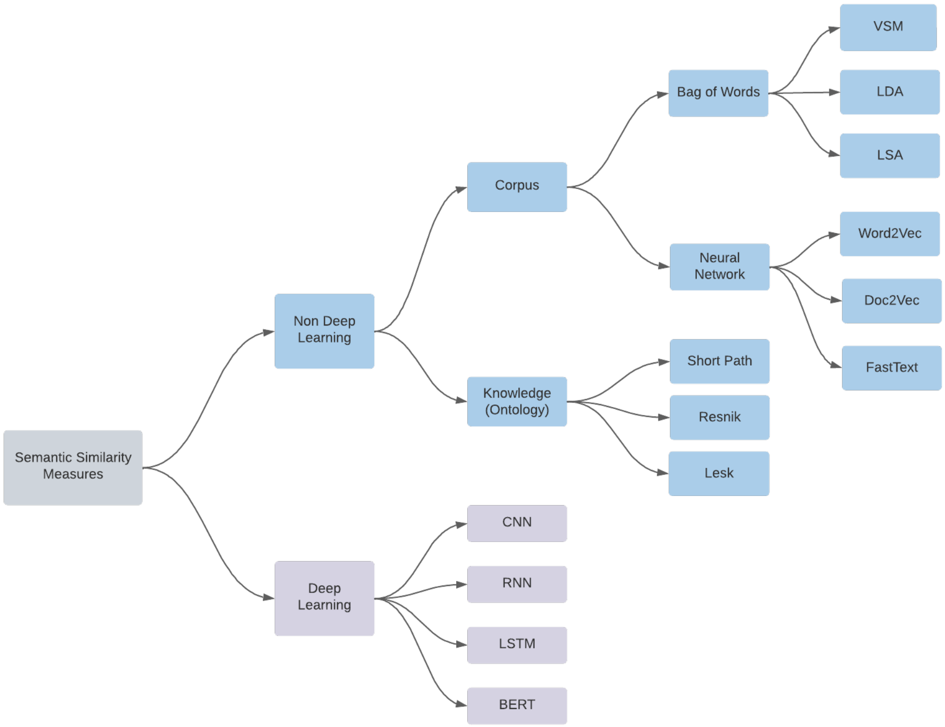


Figure 3.1: Various methods to approach textual semantic similarity

3.5.1 Knowledge-Based

Knowledge-based semantic similarity methods utilize existing resources, such as domain-specific ontologies or general ones like WordNet (Miller, 1995), a lexical database of a handcrafted semantic network of English words. The approach relies on the structure of the ontology, which is typically a graph connecting words taxonomically. The semantic similarity between concepts can be measured using the ‘Short Path’ method, which involves identifying the concepts in the ontology graph and counting the edges of the shortest path between them. The similarity score is inversely proportional to the shortest path length between the two terms; the shorter the path, the more similar the concepts are (Rada *et al.*, 1989). Another method, called Lesk (Banerjee *et al.*, 2003), takes advantage of glossary information provided by language resources. For example, BabelNet (Navigli & Ponzetto, 2012), the largest multilingual semantic ontology, provides meanings of words (glosses). The Lesk algorithm assigns a relatedness value between two words based on the overlap in their glosses. For more information on knowledge-based approaches, refer to Hadj Taieb *et al.* (2020) and Chandrasekaran & Mago (2021). It is worth noting that these semantic similarity measures are only possible if a suitable ontology exists. Therefore, I have listed, compared, and evaluated the 'fit' of existing Quran ontologies for Hadith in Chapter 6 to determine the feasibility of this approach.

3.5.2 Corpus-Based

The corpus-based approach uses statistical analysis to convert textual data to numerical representations, which allows for the mathematical measurement of semantic similarity. The earliest and most basic representation is the Bag of Words (BOW), along with dimensionality reduction methods such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). However, this method ignores the relationship between words in a sentence (Mihalcea *et al.*, 2006). To address this limitation, Mikolov *et al.* (2013) introduced Word2Vec, which is a vector representation of words that captures the underlying linguistic relationship between words. These vectors are called word embeddings, and are produced using neural networks trained on large corpora to learn word associations by observing a word and its neighbouring words. Currently, the most widely used word embeddings are Word2vec and Glove (Pennington *et al.*, 2014). Another widely used model is the character n-gram embeddings called FastText (Bojanowski *et al.*, 2017). This was introduced to solve the problem of out-of-vocabulary (OOV) words by obtaining the embedding of the syllabus or consecutive letters that are the building blocks of the word, then combining the embeddings to obtain one embedding of the OOV word. These models produced promising results on several tasks (Alharbi & Lee, 2020; Nagoudi *et al.*, 2017). An extension of word embeddings research aimed to produce sentence embeddings using averaging or concatenation of word vectors in an approach known as Doc2Vec (Le & Mikolov, 2014). However, these embeddings suffer from meaning conflation deficiency, because one word with all its possible meanings is represented as a single vector. This is addressed by the contextual embeddings introduced by Devlin *et al.* (2018).

3.5.3 Deep-learning Based

Several methods have been proposed to measure semantic similarity in textual data using deep neural networks, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. One proposed method replaces words in a sentence with their word embeddings, which are produced by GloVe. These word embeddings are used as features and are fed into a CNN that predicts the semantic similarity values between sentences (Shao, 2017). Recently, transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) have been used to produce contextual embeddings and have become the best-performing method on downstream tasks (Seelawi *et al.*, 2019). Although deep-neural network methods outperform traditional corpus-based methods, they suffer from a lack of interpretability.

In Chapter 8, I have used a corpus-based method where I trained a fastText model on different corpora consisting of Hadith, Quran and their associated literature to obtain domain-specific word embeddings. Then, I compared them to a deep-learning method by using contextual embeddings on a semantic similarity downstream task to ascertain the extent to which these models represent the complex Quran and Hadith texts.

3.6 Conclusion

This chapter gives a brief overview of the main research topics this thesis builds upon, and points to the deeper literature discussed in the other chapters within this thesis.

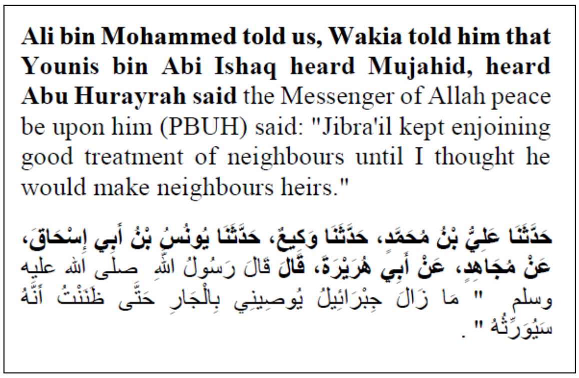
Chapter 4

Building the Hadith Segmenter

4.1 Introduction

This chapter is largely based on my two publications Altammami *et al.* (2019) presented at the 3rd Workshop on Arabic Corpus Linguistics (WACL-3), and Altammami *et al.* (2020a) presented at the 12th Language Resources and Evaluation Conference (LREC2020). It demonstrates the experiments undertaken to build the Hadith segmenter, which deconstructs the Hadith into its two main components Isnad and Matn.

Before conducting the experiments, it is worth recalling some essential background information. The Hadith and its literature are thoroughly explained in Chapter 2. This chapter specifically focuses on the structure of the Hadith. Most Hadiths consist of two parts: the Isnad, which is a chain of narrators, and the Matn, which is the actual teaching. These two parts are illustrated in Figure 4.1. Note that in this study, the phrase قال رسول الله صلى الله عليه وسلم '*The Prophet PBUH said*' is considered part of the Matn, which means the Hadith corpus discussed in Chapter 5 is built according to this assumption. This will facilitate determining if a Matn consists of prophetic words, or words of his companions without the need to refer to Isnad. More discussion on this is given in the next chapter.



Ali bin Mohammed told us, Wakia told him that Younis bin Abi Ishaq heard Mujahid, heard Abu Hurayrah said the Messenger of Allah peace be upon him (PBUH) said: "Jibra'il kept enjoining good treatment of neighbours until I thought he would make neighbours heirs."

حَدَّثَنَا عَلِيُّ بْنُ مُحَمَّدٍ، حَدَّثَنَا وَكِيعٌ، حَدَّثَنَا يُونُسُ بْنُ أَبِي إِسْحَاقَ،
عَنْ مُجَاهِدٍ، عَنْ أَبِي هُرَيْرَةَ، قَالَ قَالَ رَسُولُ اللَّهِ صلى الله عليه
وسلم " مَا زَالَ جِبْرَائِيلُ يُوصِينِي بِالْجَارِ حَتَّى ظَنَنْتُ أَنَّهُ
سَيُوَرِّثُهُ " .

Figure 4.1: Hadith example, Isnad in bold face text, followed by Matn

It is important to note that in Hadith computational studies, researchers often focus on one specific aspect of the Hadith, such as the Isnad or the Matn (Luthfi *et al.*, 2018). For example, some researchers may analyse the Isnad in order to visualize the chain of narrators (Siddiqui *et al.*, 2014) and tracing the transmission of Hadith over time. Other researchers may focus on the Matn with the goal of categorizing Hadith into subtopics (Saloot *et al.*, 2016). Having a dataset of Hadith where the Isnad and Matn are clearly segmented can greatly assist these researchers in their studies. However, manually segmenting the Hadith can be a tedious and error-prone task. Automating the process using a segmentation tool can provide more coherent and consistent results.

The following section enumerates the previous attempts to create a Hadith segmenter. Then the subsequent sections discuss the different approaches undertaken in this study to build the Hadith segmenter. The initial approach uses a dictionary look-up method (Section 4.5). Although its performance is acceptable, it is developed further to improve its accuracy to 92.5% (Section 4.6). This is achieved by adding a machine learning (ML) component and changing the algorithm to detect Isnad with irregular patterns. Additionally, another approach was tested that involves using a Hidden Markov Model (HMM), which is known to be effective in pattern recognition (Section 4.7). Finally, the various approaches are compared and discussed in order to identify potential future directions and areas for improvement.

4.2 Related Work

There have been a number of attempts to detect Isnad patterns. Table 4.1 shows summaries between such studies and the following paragraphs provides an overview of these papers.

Siddiqui *et al.* (2014) attempted to segment Isnad from Matn by using supervised ML algorithms that require an annotated corpus. Thus, native Arabic speaker-annotated Hadith tokens were extracted from Sahih Albukhari divided into five classes (beginning of Person, inside Person, beginning of Narrator, inside Narrator, and Other) where 'Narrator' corresponds to names in the Isnad, and 'Person' corresponds to names in the Matn. After annotating the corpus, Siddiqui et al studied contextual patterns of Hadith to identify features beneficial for classification. For example, the word حدثنا 'told us' is followed by a narrator's name in the Hadith Isnad, and the word بن 'son of' is part of a name. Another example is the honorific phrases that always follow a person's name. The classifier takes the training data and features in the form of 'feature, class' where each word is classified as 'beginning/inside Person', 'beginning/ inside Narrator', or 'Other'. The system classifies new Hadith tokens and segments the Hadith by finding the end of the consecutive list of narrators.

The system's performance was evaluated based on two factors: firstly, its ability to accurately assign each token to the correct type, regardless of the boundaries, as long as there is an overlap, and secondly, its ability to correctly identify the boundary of each name, regardless of the type assigned (narrator or person). The average F1 score was calculated for both results, which was 85% in the training phase. In the testing phase, another manually labelled Hadith book titled 'Musnad Ahmed' containing 5K tokens was used, which resulted in a 71% F1 score. It is important to note that these scores reflect the system's performance in annotating tokens, rather than the segmentation of Hadith. Unfortunately, the annotated data is not available to reproduce the experiments and evaluate the segmentation performance.

Harrag (2014) built a Finite State Transducers (FST) system to extract Hadith segments and meta data which includes Num-Kitab, Title-Kitab, Num-Bab, Title-Bab, Num-Hadith, Isnad, Matn, Taalik, and Atraf. This is accomplished by identifying beginning of words like 'K' for Kitab to extract the book title. Furthermore, punctuations are used to identify other parts of the Hadith assuming that all the Matn is surrounded by parenthesis. These features depend on the Hadith book used and how well and correctly it is punctuated. Thus, it cannot be applied to all kinds of Hadith books. His work measures the system's performance to identify several components of Hadith that range from Isnad, to going deeper, identifying the narrator's names. However, for the purpose of this study, only the result of Isnad extraction is reported, which is 44% precision.

Azmi & Badia (2010) developed a system that aimed to construct a tree of narrators, but first, the Isnad needed to be extracted. To extract the Isnad, they implemented several pre-processing steps, including removing diacritics and punctuation, and applied shallow parsing to handle noise and exclude non-parsable words. Using the output of the shallow parsing, noun phrases were considered as potential names of narrators. After pre-processing the data, they employed a context-free grammar to identify each segment in the Hadith by comparing the tokens to a list of Hadith lexicon they had compiled earlier. Since the goal of their study was to construct the tree, their results reflect the system's success in building the tree, rather than the segmentation of Hadith.

Maraoui *et al.* (2018) compiled a list of trigger words that come before, after, and between narrator's names. Furthermore, they identified words that mark the termination of each Hadith, which are أطرافه or تحفه. Using these comprehensive lists of words, they were able to segment Isnad from Matn for Sahih Albukhari. However, it is not clear whether it can be used to segment other Hadith books, considering that Sahih Albukhari is well structured.

Boella (2011) proposed the HedExtractor system, which uses regular expressions (Regex) to extract Hadith. The system first extracts each Hadith from the book by identifying the number of each Hadith. Then, the Arabic text is converted to its transliteration to locate the words of transmission based on a list that the researchers compiled. The system assumes that words between these transmission words are the names of narrators. Once no transmission terms are detected, the system marks the end of the Isnad. However, the exact point of Hadith segmentation can be ambiguous, even for humans. To address this issue, they set a threshold of 100 characters, which was chosen based on trial and error. This threshold tells the system that if no other transmission word is detected within the next 100 characters, then the Matn starts.

Mahmood *et al.* (2018) aimed to build a multilingual Hadith corpus by extracting Hadiths from different websites then segmenting the Hadith into Isnad and Matn using regular expressions. However, the examples presented in their paper indicate that trigger words were encoded in regular expression statements tailored for specific types of Hadiths, such as a particular Hadith book. A potential drawback of this method is its lack of versatility in application to different types of Hadith books.

Limitation of Previous Studies

In the previous work, Hadith segmentation was done using three approaches. The first was rule-based that consist of allow list (or gazetteers) to identify names and Isnad specific words, a filtration mechanism (or deny list) to identify Matn words, and grammar rules (as a set of regular expressions) to identify the segmentation point. The second was the ML approach which consists of data annotation, feature and algorithm selection, training and classification. The third was the FST, which depends on the degree of consistency in a well-structured text.

Observing Table 4.1, it is evident that the rule-based approach produced better results. However, it is not clear if the rule-based approach designed for one book can be applied to other Hadith books. As demonstrated by the work of

Paper	**Approach**	**Technique**	**Pre-processing**	**Manual annotation**	**Data**	**Result**
Siddiqui *et al.* (2014)	Machine Learning	Naïve Base, KNN, Decision tree	Remove diacritics, stemming	Person, Narrator, other	Albukhari Musnad Ahmed	71%
Harrag (2014)	Finite State Transducer	-	Tokenize,	None	Albukhari	44%
Azmi & Badia (2010)	Rule Based	Context Free Grammar	Shallow parsing, Remove diacritic and punctuation	Hadith Lexicons	Albukhari	87%
Maraoui *et al.* (2018)	Rule Based	Dictionary Lookup	None	Hadith Lexicons	Albukhari	96%
Boella (2011)	Rule Based	Regular Expressions	Transliteration	Hadith Lexicons	Albukhari	97%
Mahmood *et al.* (2018)	Rule Based	Regular Expressions	None	None	Muslim, Albukhari, Abu Dawud, Imam Malik	98%

Table 4.1: Review of previous research on Hadith segmentation into Isnad and Matn components

Mahmood *et al.* (2018), who developed distinct regular expressions for various Hadith books, it is evident that relying solely on rule-based methods may not be suitable for universal application. On the other hand, the ML approach is no better since it requires training data that represent all kinds of Hadith to make its performance acceptable when applied to the different Hadith books. For example, the study presented in Siddiqui *et al.* (2014) reported a drop in performance by 14 points once the model was tested with a different book. Another problem associated with FST is that segmentation will not work if the Hadith book does not use unique punctuation that surrounds each segment, e.g. parenthesis around the Matn.

Although an attempt was made to compare system performance in Table 4.1, it is crucial to clarify that the approaches are not comparable for two reasons. First, the data used to test the systems are different in terms of size and type. Second, the way system performance was measured is different in every study. For example, the study by Siddiqui *et al.* (2014) measured the precision of the system's ability to annotate the person's name as Narrator or not: that is, whether each name is part of the Isnad or Matn. Therefore, their system goal is not to segment but rather to identify narrators. To sum up, the results column in table 4.1 for papers Harrag (2014); Maraoui *et al.* (2018); Boella (2011), and Mahmood *et al.* (2018) reflects the segmentation performance, while the other studies report the performance of named entity recognition(NER) of narrators in Hadith. Overall, it would have been beneficial if previous studies had shared their datasets to enable better analysis and comparison. This work ensures that all created data is shared with the research community.

Before discussing the experiments for creating a Hadith segmenter, a brief overview of Arabic text preprocessing is given. This is necessary because many Arabic NLP tasks require some form of preprocessing, depending on the type of text used and the intended task.

4.3 Arabic Text Preprocessing

Text preprocessing is an important step in Arabic NLP. Previous studies measured the impact of various preprocessing steps on a range of downstream tasks such as sentiment analysis (Duwairi & El-Orfali, 2014), document classification (Alhaj *et al.*, 2019), and semantic similarity (Alhawarat *et al.*, 2021). These studies suggest that the impact of preprocessing steps may vary depending on the specific task and data used. The following lines discuss common preprocessing steps and which ones are more suitable for identifying Isnad and Matn in Hadith texts to aid in building the Hadith segmenter.

Remove punctuation marks

The use of punctuation marks is not consistent throughout the different Hadith books. Hence, it is not a reliable approach to rely on them in identifying the Isnad and Matn segments. Therefore, the removal of punctuation marks is considered in the text preprocessing pipeline.

Remove diacritics

Sometimes Classical Arabic texts have, in addition to the basic letter, diacritics that are located above or below the letters. The same letter-form may take various diacritics to represent the different words and avoid ambiguity as in عِقْد 'necklace', and عَقَدَ 'complicate'. However, considering diacritics leads to data sparsity, and this task aims to identify narrators' names and transition words, which are less likely to be ambiguous. Hence, removing diacritics is considered in the text preprocessing pipeline.

Remove Stop-words

Stop-words are frequent words that have no significant meaning for the text, such as prepositions, pronouns, and conjunctions. However, such Arabic words like أن 'that', and عن 'from' are cornerstones of the Isnad structure and therefore are essential to identify the Isnad pattern. Hence, no stop-word removal is incorporated in the preprocessing pipeline.

Stemming and Lemmatization

One of the common preprocessing steps in Arabic NLP is stemming and lemmatization. However, this is not useful for the current task because Isnad mostly consists of narrator names and transmission words. Hence, unique Isnad words such as حدثنا (‘[he] told me’, and أخبرنا ‘[he] informed me’) will be converted to words that are not Isnad-specific (خبر، حدث). Therefore, the preprocessing pipeline does not include this step.

Normalization

Normalization is the process of reducing orthographic ambiguity by normalizing the differences in spelling to minimize data sparsity. For example, the letters آ، أ، إ are all normalized to ا, and ة is normalized to ه. The effect of normalization on this task cannot be predicted. Therefore, it is added to the pipeline of data preprocessing to test its effect.

4.4 Data Preparation

4.4.1 Testing Data

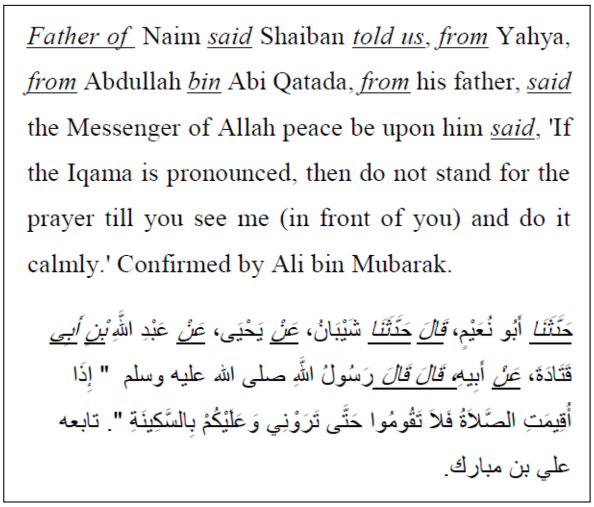
Testing data must be prepared before building the segmenter. There are a countless number of Hadith books with a varying degree of authenticity. For the purpose of this project, the six famous books are included. These are commonly referred to *The Authentic Six* or canonical Hadith books. These books are *Sahih Albukhari*, *Sahih Muslim*, *Sunan Abi Dawood*, *Sunan Al-Nasai*, *Sunan Altarmithi*, and *Sunan Ibn Maja*. From each book, around 80 Hadiths were carefully chosen to form 500 Hadiths that include Hadiths with irregular patterns. These can be downloaded from my repository[28]. This is to ensure two goals are achieved. First, I needed to overcome the limitations of previous studies that relied on one book, and second, to produce a realistic performance of a segmenter that can deal with various types of Hadiths.

[28] https://github.com/ShathaTm/Hadith_segmenter_testing_data

4.4.2 Training Data

Previous research works were consulted to determine the most effective way to obtain training data for a Hadith segmenter. Although, their data is not available, it is clear that Isnad specific lexicons are used in regular expressions and rule-based approaches to identify the Isnad segments. To extract these terms, Sahih Albukhari is used for two reasons. First, it is the most commonly used book in previous studies and it is well-structured.

To automate this task, Isnads in Hadiths are scrutinized to ascertain that they contain of a closed set of words, as indicated by previous studies (Maraoui *et al.*, 2018; Boella, 2011). The example of Isnad in Figure 4.2 underlines these words. A common pattern in Isnad is the narrator's name which takes the form of *first name - son of- father's name*, so this involves two names connected by a 'relation' word. Narrators' names are usually followed by 'transmission' words that reflect how the Hadith was reported e.g. *x heard y* or *x said*. Hence, transmission words usually appear four words apart.



Father of Naim *said* Shaiban *told us*, *from* Yahya, *from* Abdullah *bin* Abi Qatada, *from* his father, *said* the Messenger of Allah peace be upon him *said*, 'If the Iqama is pronounced, then do not stand for the prayer till you see me (in front of you) and do it calmly.' Confirmed by Ali bin Mubarak.

حَدَّثَنَا أَبُو نُعَيْمٍ، قَالَ حَدَّثَنَا شَيْبَانُ، عَنْ يَحْيَى، عَنْ عَبْدِ اللَّهِ بْنِ أَبِي
قَتَادَةَ، عَنْ أَبِيهِ، قَالَ قَالَ رَسُولُ اللَّهِ صلى الله عليه وسلم " إِذَا
أُقِيمَتِ الصَّلاَةُ فَلاَ تَقُومُوا حَتَّى تَرَوْنِي وَعَلَيْكُمْ بِالسَّكِينَةِ ". تابعه
علي بن مبارك.

Figure 4.2: Isnad example, Isnad lexical items underlined

Using this information, a list of Isnad lexical items (or lexemes) is created that consists of 'transmission' and 'relation' words similar to those mentioned in

Siddiqui *et al.* (2014); Maraoui *et al.* (2018); Boella (2011). Then a python script is built as illustrated in Figure 4.3 which tokenizes a Hadith into space-separated words. Then it takes four tokens at a time to check if an Isnad lexical item is present. Once it detects a group of four words with no Isnad lexical item, it assumes the beginning of the Matn text and separates the Hadith at that point. This approach automatically detects Isnads with regular patterns only, so it aims to collect the various names of narrators. The final step is performed manually to verify the results of this bootstrapping approach. This produced a collection of more than four thousand segmented Hadiths to be the training data of Isnad and Matn segments.

4.5 Look-up Approach

4.5.1 Look-up Lists

The first approach to build the segmenter relies on the simplest model that uses dictionary look-up lists. Hence, to create these lists, the training data of Isnad and Matn segments created previously as discussed in Section 4.4.2 are extracted. Then they are tokenized to trigrams, bigrams and unigrams, which are added to the look-up lists accordingly as shown in the example in figure 4.4.

4.5.2 The N-gram Model

This section empirically tests the performance of different n-gram models for Hadith token labelling, which is similar to part-of-speech (POS) tagging tasks. However, in POS tagging, words can be ambiguous if POS information is obtained from a dictionary, as some words can have different POS tags depending on the context. The bigram model, which takes into account the context of words, has been found to be effective in POS tagging tasks (Atwell, 1983). Therefore, it is speculated that the same applies to Hadith token labelling, as many words can exist in both the Isnad and Matn, making them ambiguous without considering the context.

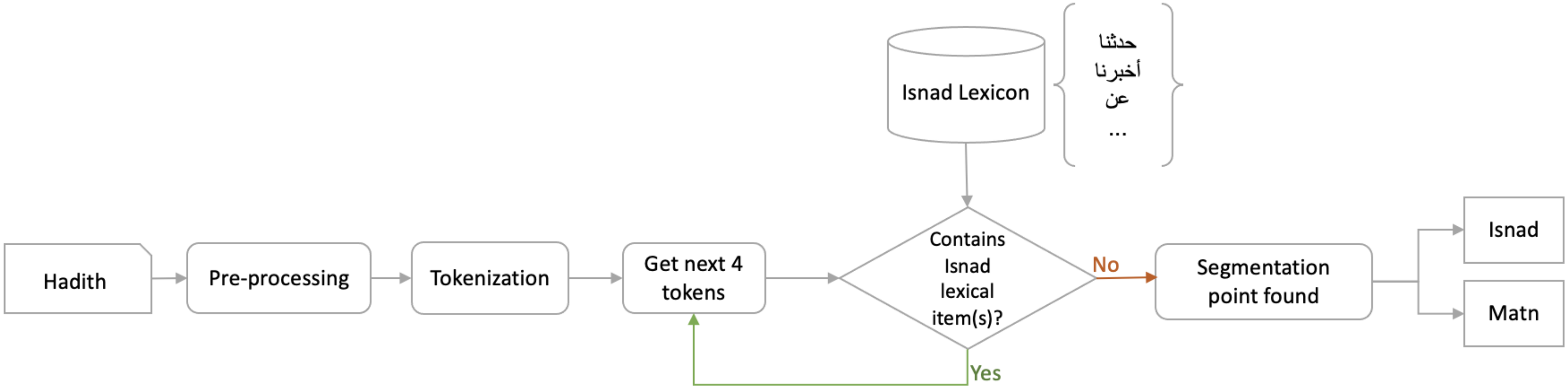


Figure 4.3: Python script to perform the segmentation on Albukhari Hadith

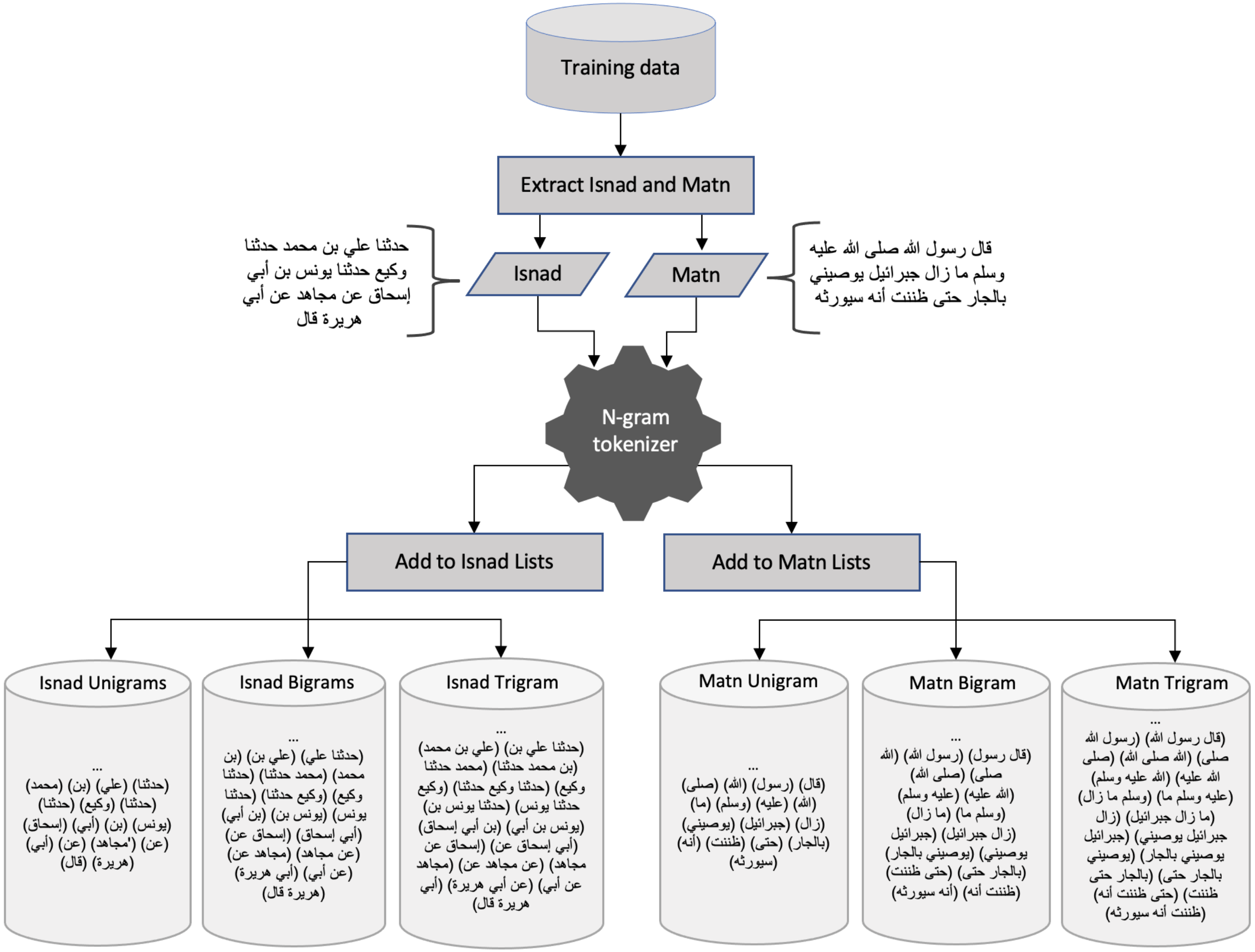


Figure 4.4: Example of Isnad and Matn extracted from the training data and tokenized to unigrams, bigrams and trigrams then added to the lists accordingly

To test this hypothesis, an empirical study is conducted using the training and testing datasets discussed in Section 4.4. The study follows these steps:

1. Extract the trigram, bigram, and unigram lists from the training data to be used as dictionary look-up lists.

2. Tokenize the 500 Hadiths in the testing data using the different n-gram models.

3. Perform a dictionary look-up task on the testing n-grams to identify whether each token is part of Isnad or Matn.

4. If the token is found in both, Isnad and Matn, then label it ‘ambiguous’.

5. If the token is not found in any list, then label it ‘token with no match’

As shown in Table 4.2, when the bigram model is used, 40% of the tokens are recognized. On the other hand, the unigrams model produces the most ambiguous words, because most unigrams, e.g. words, exist in both the Isnad and the Matn. Although the trigram model produced the lowest rate of ambiguous tokens, it introduced a major drawback. It requires large training data, as is evident by the high number of unrecognized trigram tokens.

Ngram	Number of Tokens	Isnad Tokens	Matn Tokens	Ambiguous tokens	Tokens with no match
Trigrams	34,267	13%	16%	14%	54%
Bigrams	35,272	8%	32%	23%	34%
Unigrams	36,277	1%	26%	65%	6%

Table 4.2: Result of tokenizing and annotating the 500 Hadith testing data using look-up lists

4.5.3 Experiments

Trigram with back-off

Instead of relying solely on the bigram representation model, a better approach is to combine the three n-gram representation models using a back-off approach that can handle irregularity and missing information. For example, if an encountered trigram token has no match in the trigram lists of the lookup training data, it can be annotated according to its components by converting it to two bigram tokens and looking it up in the bigram lists. If no match is found it is converted to unigram tokens and looked up in the unigram lists. Consider a case in which a narrator's full name is not captured in the lists, then Hadith lexical item like بن ('son of') will enable the system to identify this trigram as part of the narrator's chain and labels it *Isnad*. This approach is detailed in Algorithm 1. Once every token is labelled, the system finds the segmentation point as detailed in Algorithm 2.

Algorithm 1 Annotate Trigram tokens

Tokenize Hadith into Trigrams “T”

```
for $t \in T$ do
    if $t \in IsnadTrigramList$ then
        $t\_Label$ = Isnad
    else if $t \in MatnTrigramList$ then
        $t\_Label$ = Matn
    else
        Convert $t$ to Bigrams $B$
        for $b \in B$ do
            if $b \in IsnadBigramList$ then
                $t\_Label$ = Isnad
                Break
            else if $b \in MatnBigramList$ then
                $t\_Label$ = Matn
                Break
            else if $b$ last token in $B$ then
                Convert $t$ to Unigrams $U$
                for $u \in U$ do
                    if $u \in IsnadUnigramList$ then
                        $t\_Label$ = Isnad
                        Break
                    else if $u \in MatnUnigramList$ then
                        $t\_Label$ = Matn
                        Break
                    else
                        $t\_Label$ = Niether
                    end if
                end for
            end if
        end for
    end if
    $output\_list \leftarrow [t, t\_Label]$
end for
```

Algorithm 2 Find Segmentation point

```
for [t, t_Label] in output_list do
    if t_Label == Matn then
        if followed by t_Label == (Matn or Neither) then
            Mark Segmentation Point at t
            Break
        else if t_Label == Neither then
            if followed by 2 t_Label== (Matn or Neither) then
                Mark Segmentation Point at t
                Break
            end if
        end if
    end if
end for
```

This approach produced 48% accuracy. To understand this disappointing result, the incorrectly segmented Hadiths are inspected. It appears that the system rarely used the trigram feature, but rather relied on the bi-gram and uni-gram features to annotate tokens. Consider the example in Table 4.3. Feeding this Hadith to the system produces 79 trigrams, of which only 15 found a match in the trigram training set. The remaining 64 trigrams relied on the bi-gram and uni-gram training set to be annotated. This dependency on bi-gram/uni-gram features to annotate Hadith trigrams produced unreliable results as illustrated in the example. The phrase أن رجلا قال 'that a man said' should mark the beginning of the Matn, instead it was labelled as Isnad. This is because when the system did not find a match in the trigram training set, it applied the back-off approach and searched in the bi-gram and uni-gram lists. Since it found a match for the term قال 'he said' in the Isnad lists, it labelled the phrase accordingly. Therefore, using trigrams did not prove useful in this case for two reasons. First, the training data is not large enough to cover all known narrators. Second, it is obtained from only one Hadith book, which does not include all Hadith lexical items and patterns.

Isnad	Matn
حدثنا قتيبة حدثنا مروان بن معاوية الفزاري عن أبي يعفور عن الوليد بن العيزار عن أبي عمرو الشيباني أن رجلا قال لابن مسعود أي العمل أفضل قال سألت عنه رسول الله صلى الله عليه وسلم فقال الصلاة على مواقيتها قلت وماذا	يا رسول الله قال وبر الوالدين
Qutaiba told us Marwan bin Muawiya al-Fizari from Abu Yafour from Al-Walid bin Al-Azar from Abu Amr AlShibani that a man said to Ibn Masood, which work is better? He said I asked the Messenger of Allah (PBUH): "Which action is dearest to Allah?" He (PBUH) replied, "Performing the prayer at its earliest fixed time." I asked, "What is next ?"	O Prophet, He said, "Kindness towards parents."

Table 4.3: Example of incorrect Hadith segmentation using trigrams with back-off appraoch

Bigram with back-off

To improve the system performance, the trigram features are omitted, and instead, bigrams and unigrams are used. The bigram technique produced better results as expected.

4.5.4 Result

The segmenter performance is measured using accuracy. The segmentation is considered correct if it is off by three words as shown in formula 4.1. This is based on the narrators' names in the Isnad, which usually do not exceed three tokens. Hence, it assumes the segmentation is correct even if it is off by three words. This measurement of accuracy is used with the various approaches discussed in this chapter to build the Hadiths segmenter.

$$|\texttt{\# of actual Isnad tokens} - \texttt{\# of produced Isnad tokens}| < 4 \qquad (4.1)$$

The accuracy of the segmenter is 87.1% , which is improved further after normalization where إ ، أ and آ are all normalized to ا, and ة is normalized to ه. This made the accuracy reach 88.3%.

This segmenter is able to segment Hadiths with different structures. For example, the traditional ones where a Matn start with a prophetic saying as shown in Table 4.4. Other Hadith structures include those containing irregular patterns where a Matn starts with an introductory phrase followed by the prophetic saying as shown in Table 4.5, a dialogue with the Prophet as shown in Table 4.6, or an explanation of a prophetic deed as in Table 4.7.

While performance is improved, the incorrectly segmented Hadiths are inspected and it appears they are the ones with irregular patterns. For example, a Hadith may contain a parallel Isnad, which is a chain of narrators that ends at the Prophet followed by another chain of narrators that ends at the Prophet again, as shown in Table 4.8. Another example of an irregular pattern in Isnad is shown in Table 4.9, which illustrates that an Isnad may contain different Matn patterns. Finally, Table 4.10 shows that some Hadiths possess a vague segmentation point. Note that due to space constraint some Hadiths in the examples were truncated as indicated by '...'.

Isnad	Matn
حدثنا كثير بن عبيد الحمصي حدثنا محمد بن خالد عن عبيد الله بن الوليد الوصافي عن محارب بن دثار عن عبد الله بن عمر قال Kathir bin Obeid Al-Homsi told us Mohammed bin Khalid from Obidallah bin Walid Al-Wasafi from Moharib bin dathar from Abdullah bin Omar said	قال رسول الله صلى الله عليه وسلم أبغض الحلال إلى الله الطلاق The Prophet (PBUH) said, “Of all the lawful acts the most detestable to Allah is divorce”.

Table 4.4: Correct segmentation, regular pattern

Isnad	Matn
حدثنا أبو معمر قال حدثنا عبد الوارث عن عبد العزيز قال أنس Abu Muammar told us that Abdul Warith told us from Abdul Aziz said that Anas said	إنه ليمنعني أن أحدثكم حديثا كثيرا أن النبي صلى الله عليه وسلم قال من تعمد علي كذبا فليتبوأ مقعده من النار I refrain from telling you many things about the Prophet because I heard the Prophet (PBUH) said, “He who deliberately forges a lie against me let him have his abode in the Hell.”

Table 4.5: Correct segmentation, introductory statement

Isnad	Matn
حدثنا قتيبه قال حدثنا الليث عن يزيد بن ابي حبيب عن ابي الخير عن عبد الله بن عمرو Qaytibah told us Alith from Yazid ibn Abi Habib from Abi Al-Khair from Abdullah bin Amr	أن رجلا سال رسول الله صلى الله عليه وسلم أي الاسلام خير قال تطعم الطعام وتقرا السلام على من عرفت ومن لم تعرف A man asked the Messenger of Allah (PBUH): “Which act in Islam is the best?” He (PBUH) replied, “To give food, and to greet everyone, whether you know or you do not.”

Table 4.6: Correct segmentation, conversation of the Prophet

Isnad	Matn
حدثنا إسمعيل بن موسى الفزاري حدثنا شريك عن أبي إسحق عن الحارث عن على بن أبي طالب قال Ismail bin Musa al-Fazari told us Sharik said Abu Ishaq from AlHarith from Ali bin Abi Talib said	من السنة أن تخرج إلى العيد ماشيا وأن تأكل شيئا قبل أن تخرج It is the Sunnah (prophetic tradition) to go out to the Eid prayer walking and eat something before you go out.

Table 4.7: Correct segmentation, no prophetic words

Isnad	Matn
حدثنا مسدد قال حدثنا يحى عن شعبة عن قتادة عن أنس رضي آلله عنه عن النبي	صلى الله عليه وسلم وعن حسين المعلم قال حدثنا قتادة عن أنس عن النبي صلى الله عليه وسلم قال لا يؤمن أحدكم حتى يحب لأخيه ما يحب لنفس
Mosadad said Yahya told us Shoba heard Qatada from Anas may Allah be pleased with him, the Prophet	(PBUH), and from Husayn al-Muallim said Qatada told us from Anas that the Prophet (PBUH) said: "No one of you becomes a true believer until he likes for his brother what he likes for himself".

Table 4.8: Incorrectly segmented, parallel Isnad

Isnad	Matn
حدثنا نصر بن علي الجهضمي وأبو عمار والمعنى	واحد واللفظ لفظ أبي عمار قالا أخبرنا سفيان بن عيينة عن الزهري عن حميد بن عبد الرحمن عن أبي هريرة قال أتاه رجل فقال يا رسول الله هلكت...
Nasser bin Ali Juhadhmi and Abu Ammar told us and the meaning	Is the same but the words are of Ammar they said, Sufian bin Aayneh from Alzahri from Hamid bin Abdul Rahman on the authority of Abu Hurayrah said a man came and said, "O Allah's Apostle! I have been ruined..."

Table 4.9: Incorrectly segmented, Isnad contains Matn lexical items

Isnad	Matn
أخبرنا محمد بن منصور قال حدثنا سفيان قال حدثنا يحيى بن سعيد عن مسلم بن أبي مريم شيخ من أهل المدينة ثم لقيت الشيخ فقال سمعت علي بن عبد الرحمن يقول صليت إلى جنب ابن عمر فقلبت الحصى فقال لي ابن عمر	لا تقلب الحصى فإن تقليب الحصى من الشيطان وافعل كما رأيت رسول الله صلى الله عليه وسلم يفعل قلت وكيف رأيت رسول الله صلى الله عليه وسلم يفعل قال هكذ ...
Muhammad bin Mansour told us, that Sufian said Yahya bin Said told us about Muslim bin Abi Maryam a Sheikh from Madinah then I met the Sheikh and he said he heard Ali bin Abdul Rahman say I prayed beside Ibn Omar, while I turned the gravel he said	Do not turn the gravel, turning the gravel is from the devil and do as I saw the Messenger of Allah peace be upon him do...

Table 4.10: Incorrectly segmented, names should be part of Matn

4.6 Machine Learning Approach

Although the look-up lists approach to build a Hadith segmenter produced acceptable results, its performance is improved in this experiment by incorporating a machine learning (ML) model into the pipeline. Furthermore, the segmentation algorithm is modified to deal with irregular Hadith structures. The proposed Hadith segmenter pipeline is shown in Figure 4.5, where it applies the following steps:

1. First, it takes the Hadith input and pre-processes it to remove diacritics, punctuation marks and extra white spaces. Diacritics were removed to overcome data sparseness and enhance term weighting. The experiment was run with and without normalization to understand its effect[29].

2. Then it tokenizes the pre-processed Hadith into bigrams of words. Bigrams were chosen based on their better performance compared to the other n-gram features as explained in Section 4.5.2.

3. After that it labels every token as 'Isnad' or 'Matn' by using an ML classifier which is described in Section 4.6.1.

4. Once every token is labelled, a rule-based algorithm is applied to find the exact segmentation point as detailed in Section 4.6.2.

5. Finally, the segmentation point is applied on the original Hadith to produce Isnad and Matn segments with diacritics and punctuations intact.

4.6.1 Choosing a Classifier Model

Training a supervised ML model requires labelled training data to learn the mapping function that takes an input variables (x) and produces an output variable (y). In other words, it solves for f in equation 4.2.

$$y = f(x) \tag{4.2}$$

[29] Pre-processing was implemented from scratch in python as CAMeL tools were not published at the time of running this experiment

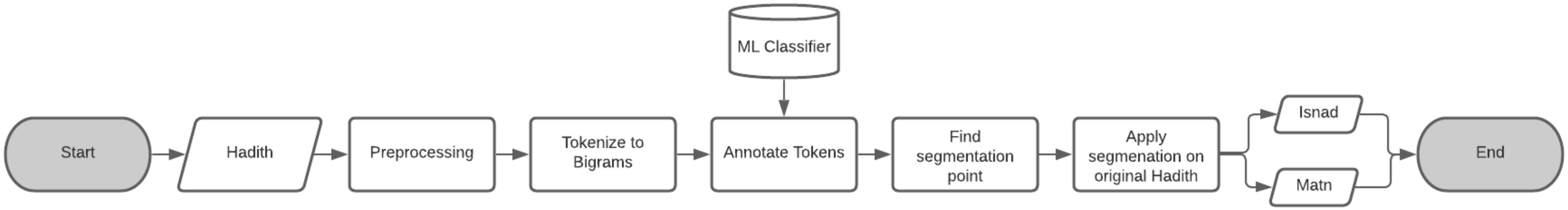


Figure 4.5: Pipeline of Hadith segmenter using the ML approach

A ML Classifier takes in a Hadith bigram and classifies it as ‘Isnad’ or ‘Matn’. It is trained on 4,686 segmented Hadiths extracted from Sahih Albukhari. Additionally, a list of narrator names was tokenized to bigrams and added to the training data. These were collected form a website[30] and can be downloaded from my repository[31]. The training data includes 314,340 bigram instances as shown in Figure 4.6 which are divided into 70% training and 30% validation. The testing bigrams are obtained from the 500 Hadiths extracted from the six Hadith books as explained in Section 4.4.

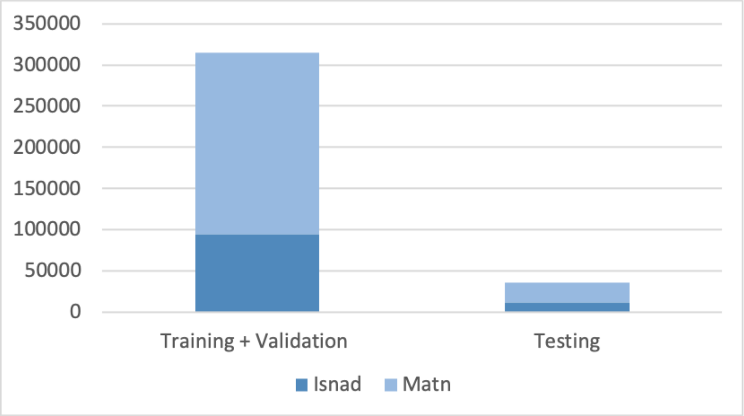


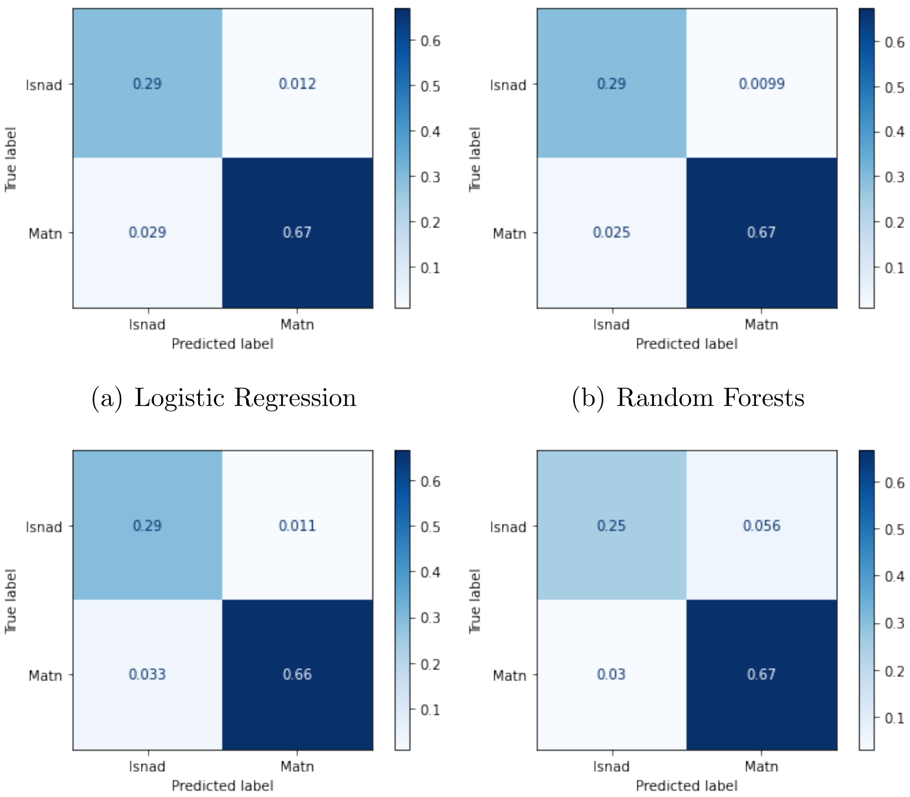
Figure 4.6: Data distribution of bigram tokens

Ideally, all suitable ML algorithms should be tested. However, this experiment only focuses on the most widely used ones. Various classical ML algorithms were trained using the scikit-learn library. These include multinomial Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), and Random Forests. Each training instance is a bigram token with a label that is fed into a model as [word1 word2, label]. An example of an Isnad training instance is [حدثنا قتيبه, Isnad]. These data are converted to numerical form using TF-IDF encoding. The performance is measured based on the classifier’s correct predication of each token’s label e.g. ‘Isnad’ or ‘Matn’.

[30] موسوعة رواة الحديث

[31] https://github.com/ShathaTm/Hadith_Narrator_Names

Figure 4.7 shows the confusion matrix of the various ML models. It is clear that their performance is relatively similar, with Random Forests showing the best performance as indicated in Figure 4.7(b). Although training data consists of more Matn bigrams than Isnad bigrams as shown in Figure 4.6, the performance of the classifiers is satisfactory for both categories. This could be due to the fact that the Isnad usually consists of words that are either proper names or transmission words e.g. ‘said’, ‘heard’.



(a) Logistic Regression

(b) Random Forests

(c) Naive Bayes

(d) SVM

Figure 4.7: Confusion matrix of the tested ML classifiers

The classifiers are then evaluated on the testing data which includes 33,290 bigrams. Table 4.11 highlights the performance of these classifiers in labelling testing data bigrams.

Algorithm	Accuracy
Multinomial Naive Bayes	95.6%
Support Vector Machine	91.4%
Logistic Regression	95.9 %
Random Forest	96.9%

Table 4.11: Accuracy of ML models on the testing data to classify bigram of words as Isnad or Matn

The best performing model is Random Forest. Therefore, the bigram tokens (t_1,t_2,..,t_n) annotated by this model as ([t_1, t_Label_1],[t_2, t_Label_2],...,[t_n, t_Label_n]) are stored in a list (*output_list*) which is used to segment Hadiths as discussed in the next section.

It is worth noting that the typical out of vocabulary (OOV) problem, which is usually associated with proper nouns, does not necessarily apply to Hadith. This is because the names of narrators in the Hadith literature are a relatively closed set.

4.6.2 Experiments

Once the bigram tokens are annotated by the ML model as [t, t_Label] and stored in a list (*output_list*), the segmenter finds the exact segmentation point. Hadith segmentation is a domain specific task, and as shown in the previous survey in Section 4.2, rule-based approaches produced the highest accuracies. The rule-based Hadith segmentation approach is simplified in Algorithm 3. This algorithm is able to identify Isnads with irregular patterns, and Hadiths that contain parallel Isnads. These were the limitations in the previous approach in Algorithm 2.

Algorithm 3 : Find Segmentation Point

```
for [t, t_Label] in output_list do
    if t_Label == Matn then
        if next three t_Label == Matn then
            Segmentation point A found at t
            output_list = output_list [Segmentation point A: End]
        end if
    end if
end for
if Segmentation point A is found then
    Check if another Isnad exists
    for [t, t_Label] in output_list do
        if 5 consecutive t_Label == Isnad then
            parallel Isnad found
            Find next set of t_Label == Matn
            for [t, t_Label] in output_list do
                if t_Label == Matn then
                    if next three t_Label == Matn then
                        Segmentation Point B found at t
                    end if
                end if
            end for
        end if
    end for
end if
if no segmentation point found then
    Hadith does not contain Matn
end if
```

4.6.3 Results

The result using normalization is 92.5% accuracy, a slightly better performance than not normalizing data. The effect of normalization is discussed further in Section 4.16. The following lines explains how the segmenter performance is improved with examples.

Table 4.12 shows an example of Hadith with parallel Isnad. The first chain of narrators is followed by the Prophet's name, which is followed by another chain of narrators that ends with the Prophet's name as well. These two chains of narrators are followed by the *Matn*, where the segmentation point should be detected.

To segment this Hadith, the tool uses the ML classifier to label the first 14 tokens as *Isnad*, followed by 5 tokens as 'Matn', then another set of 9 tokens as 'Isnad', and finally 15 tokens as 'Matn'. It determines the segmentation point by detecting that the first set of 'Matn' tokens is followed by another set of 'Isnad' tokens. Therefore, it segmented the Hadith after the second set of 'Isnad' tokens as indicated in table 4.12

Isnad	Matn
حدثنا مسدد قال حدثنا يحيى عن شعبة عن قتادة عن أنس رضي الله عنه عن النبي صلى الله عليه وسلم وعن حسين المعلم قال حدثنا قتادة عن أنس	عن النبي صلى الله عليه وسلم قال لا يؤمن أحدكم حتى يحب لأخيه ما يحب لنفسه.
Mosadad said Yahya told us Shoba heard Qatada from Anas may Allah be pleased with him, that he heard the Prophet (PBUH), and from Husayn al-Muallim said Qatada told us that Anas said that	the Prophet (PBUH) said: "No one of you becomes a true believer until he likes for his brother what he likes for himself".

Table 4.12: Correctly segmented Hadiths with parallel Isnads

Another limitation of the first approach is dealing with Isnads that contain irregular patterns. This is addressed in this approach as illustrated by the example in Table 4.13. However, the example in table 4.14 demonstrates that there is room for improvement, specially in dealing with Hadiths that contain vague segmentation points. Note that due to space constraints some Hadiths in the examples were truncated, as indicated by '...'.

Isnad	Matn
حدثنا نصر بن علي الجهضمي وأبو عمار والمعنى واحد واللفظ لفظ أبي عمار قالا أخبرنا سفيان بن عيينة عن الزهري عن حميد بن عبد الرحمن عن أبي هريرة قال أتاه	رجل فقال يا رسول الله هلكت...
Nasser bin Ali Juhadhmi and Abu Ammar told us and the meaning is the same but the words are of Ammar they said, Sufian bin Aayneh from Alzahri from Hamid bin Abdul Rahman on the authority of Abu Hurayrah said	a man came and said, "O Allah's Apostle! I have been ruined." ...

Table 4.13: Correctly segmented Hadith with irregular pattern

Isnad	Matn
أخبرنا محمد بن منصور قال حدثنا سفيان قال حدثنا يحيى بن سعيد عن مسلم بن أبي مريم	شيخ من أهل المدينة ثم لقيت الشيخ فقال سمعت علي بن عبد الرحمن يقول صليت إلى جنب ابن عمر فقلبت الحصى فقال لي ابن عمر لا تقلب الحصى فإن تقليب الحصى من الشيطان وافعل كما رأيت رسول الله صلى الله عليه وسلم يفعل قلت وكيف رأيت رسول الله صلى الله عليه وسلم يفعل قال هكذا ...
Muhammad bin Mansour told us, that Sufian said Yahya bin Said told us about Muslim bin Abi Maryam	a Sheikh from Madinah then I met the Sheikh and he said he heard Ali bin Abdul Rahman say I prayed beside Ibn Omar, while I turned the gravel he said Do not fluctuate the gravel, turning the gravel is from the devil and do as I saw the Messenger of Allah peace be upon him do ...

Table 4.14: Incorrectly segmented Hadith with irregular patterns

4.7 Hidden Markov Model

Instead of labelling bigram tokens independently from the previous token, the Hidden Markov Model (HMM) is investigated. HMM is a class of probabilistic models that allows prediction of sequence of unknown (hidden) variables from a set of observed variables. It has been used to predict the POS tags (hidden variable) based on the word (observed variable) (Jurafsky & Martin, 2020). In this experiment, the HMM predicts whether each word in the Hadith is part of the Isnad or Matn.

4.7.1 Data Preparation

As in the previous experiments, Albukhari's Hadiths are used as training data for the HMM. However, instead of using bi-grams of words as tokens, every Hadith word is considered a token where the model labels it as part of the Isnad or Matn using the tag set in Table 4.15. Thus, the HMM computes the joint probability of a set of hidden states (labels) given a set of observed states (words). The labels tag set is used following the approach of Siddiqui *et al.* (2014), who defined a similar tag set to identify narrators' names in the Isnad, as discussed in Section 4.2. An example of how each training record is labelled is shown in Figure 4.8.

Label	Token Type
/B_I,	begin Isnad
/I_I	inside Isnad
/E_I	end Isnad
/B_M	begin Matn
/I_M	inside Matn
/E_M	end Matn

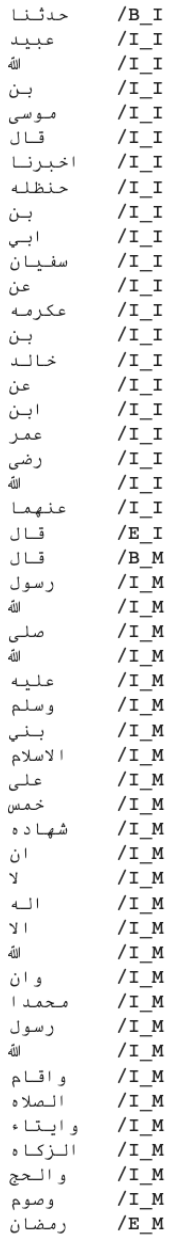
Table 4.15: Tag set for HMM training

4.7.2 Experiments

The HMM is built using the pomegranate[32] python library by training it on 6,278 Hadiths, consisting of more than 6000 words in each of the following `/B_I`, `/E_I`, `/B_M` and `/E_M`. The `/I_I` consists of more words exceeding 100,000, and the `/I_M` contains the largest share exceeding 300,000 words.

After training, the model was validated on 1,570 Hadiths, which produced 97% accuracy for annotating tokens. To illustrate how the model captured the transition probability between the different states, a visualization of the model is

[32] https://pomegranate.readthedocs.io/en/latest/HiddenMarkovModel.html



حدثنا	/B_I
عبيد	/I_I
الله	/I_I
بن	/I_I
موسى	/I_I
قال	/I_I
اخبرنا	/I_I
حنظله	/I_I
بن	/I_I
ابي	/I_I
سفيان	/I_I
عن	/I_I
عكرمه	/I_I
بن	/I_I
خالد	/I_I
عن	/I_I
ابن	/I_I
عمر	/I_I
رضى	/I_I
الله	/I_I
عنهما	/I_I
قال	/E_I
قال	/B_M
رسول	/I_M
الله	/I_M
صلى	/I_M
الله	/I_M
عليه	/I_M
وسلم	/I_M
بني	/I_M
الاسلام	/I_M
على	/I_M
خمس	/I_M
شهاده	/I_M
ان	/I_M
لا	/I_M
اله	/I_M
الا	/I_M
الله	/I_M
وان	/I_M
محمدا	/I_M
رسول	/I_M
الله	/I_M
واقام	/I_M
الصلاه	/I_M
وايتاء	/I_M
الزكاه	/I_M
والحج	/I_M
وصوم	/I_M
رمضان	/E_M

Figure 4.8: Hadith annotation for training the HMM

depicted in Figure 4.9.

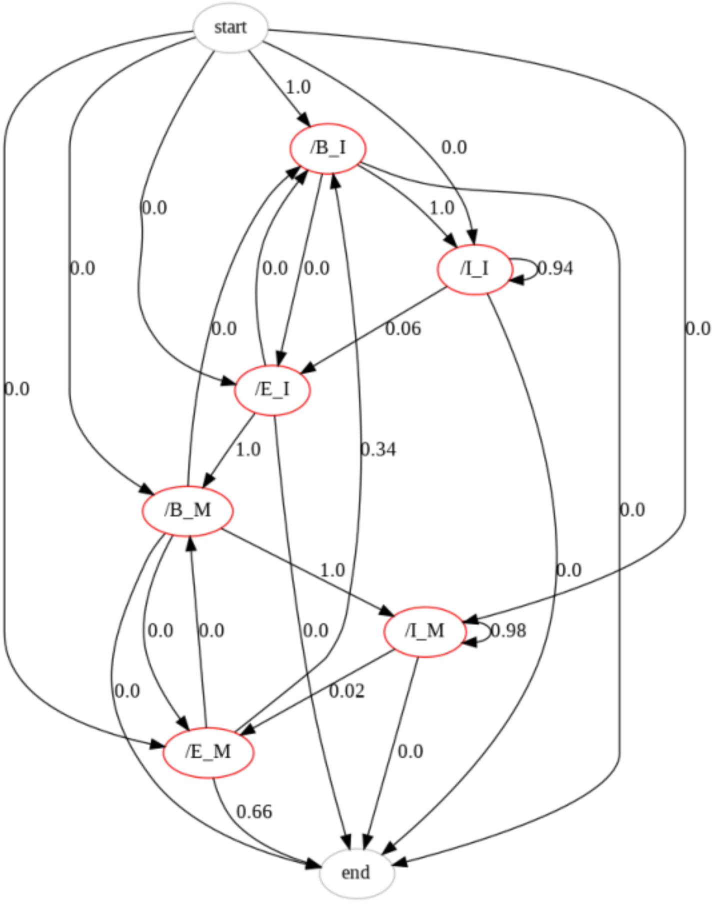


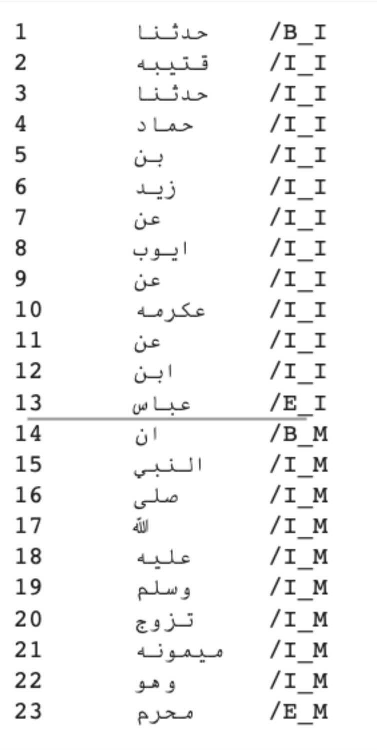
Figure 4.9: Visualization of the HMM to illustrate the probability of transitioning to a new state conditioned on a present state

Once the HMM model is validated, it is used as a segmentation tool for testing data of 500 Hadiths from the 6 different Hadith books. First, the model annotates every token in the Hadith. Then the segmentation algorithm finds the last `/E_I` and `/B_M` tokens to segment the Hadith at that point.

4.7.3 Results

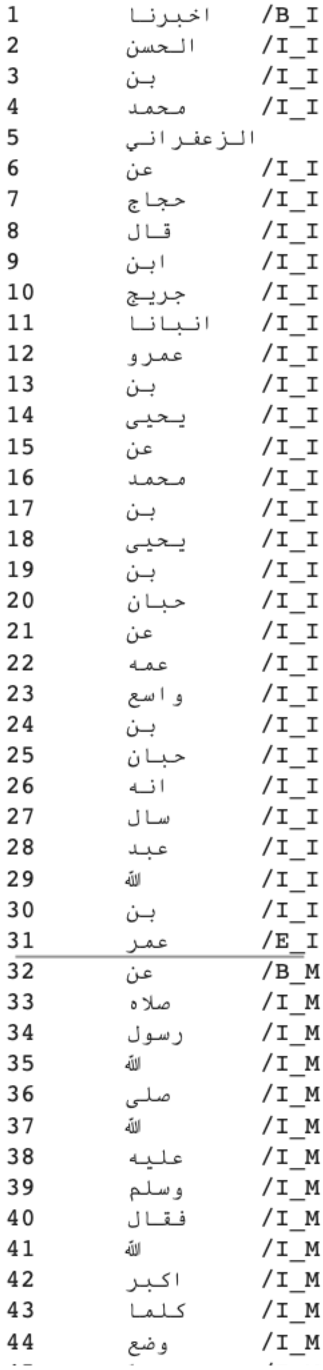
The accuracy of annotating words in the testing data is 79.15%. However, the accuracy of segmentation is 90.25%. The segmenter using HMM identified the Isnad and Matn in Hadiths with regular structures successfully as shown in the example in Figure 4.10. In the other hand, the incorrectly segmented Hadiths fall into three categories as discussed next with examples to illustrate. Bear in mind that some Hadiths in the examples are truncated for space purposes.

The first type of incorrectly segmented Hadiths contain vague segmentation points as in Figure 4.11. The segmentation should be at token 25 instead of 31. However, this result is understandable since the vague tokens consist of names which have a greater probability of being part of the Isnad. Second, some Hadiths Isnad contain description of a narrator which the model identified as part of the Matn as shown in Figure 4.12. In this example segmentation should be at token 30. Again, this annotation is understandable since the words are not names of narrators. Third, there are Hadiths with several Isnad and Matn segments, as shown in Figures 4.13. This is because the segmenter algorithm finds the last `/E_I` and `/B_M` tokens to segment the Hadith at that point without considering that there is more than one Isnad and Matn segments. Hence, the HMM could be the answer to annotating Hadith to fine-grained segments.



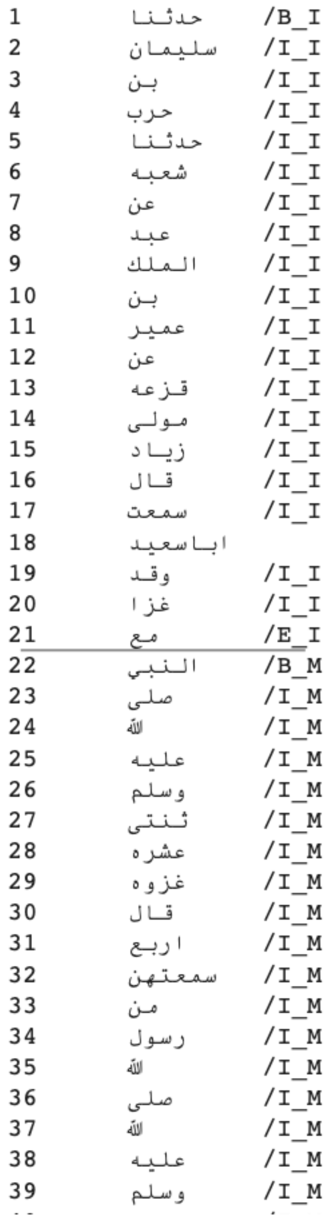
1	حدثنا	/B_I
2	قتيبه	/I_I
3	حدثنا	/I_I
4	حماد	/I_I
5	بن	/I_I
6	زيد	/I_I
7	عن	/I_I
8	ايوب	/I_I
9	عن	/I_I
10	عكرمه	/I_I
11	عن	/I_I
12	ابن	/I_I
13	عباس	/E_I
14	ان	/B_M
15	النبي	/I_M
16	صلى	/I_M
17	ﷲ	/I_M
18	عليه	/I_M
19	وسلم	/I_M
20	تزوج	/I_M
21	ميمونه	/I_M
22	وهو	/I_M
23	محرم	/E_M

Figure 4.10: Example of a Hadith with regular structure, and the correct segmentation indicated by the horizontal line after token number 13



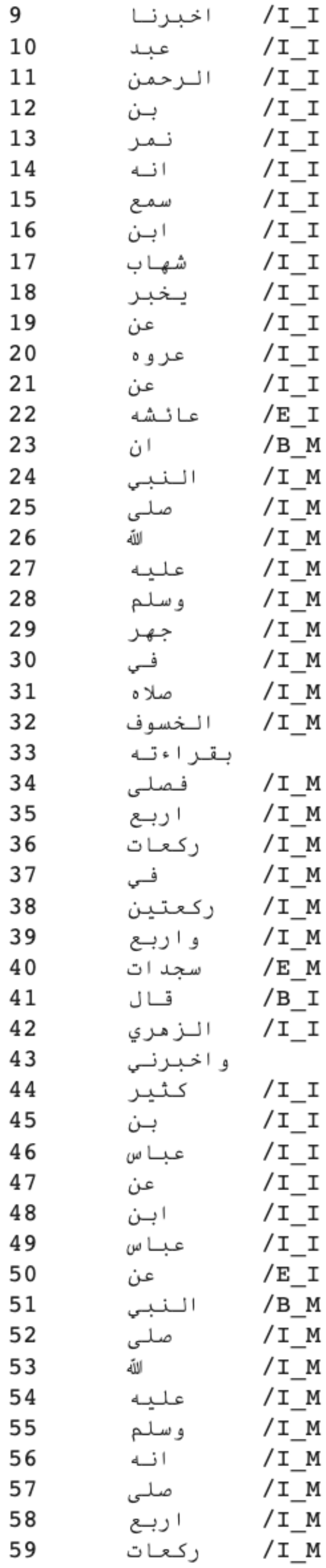
1	اخبرنا	/B_I
2	الحسن	/I_I
3	بن	/I_I
4	محمد	/I_I
5	الزعفراني	
6	عن	/I_I
7	حجاج	/I_I
8	قال	/I_I
9	ابن	/I_I
10	جريج	/I_I
11	انبانا	/I_I
12	عمرو	/I_I
13	بن	/I_I
14	يحيى	/I_I
15	عن	/I_I
16	محمد	/I_I
17	بن	/I_I
18	يحيى	/I_I
19	بن	/I_I
20	حبان	/I_I
21	عن	/I_I
22	عمه	/I_I
23	واسع	/I_I
24	بن	/I_I
25	حبان	/I_I
26	انه	/I_I
27	سال	/I_I
28	عبد	/I_I
29	ﷲ	/I_I
30	بن	/I_I
31	عمر	/E_I
32	عن	/B_M
33	صلاه	/I_M
34	رسول	/I_M
35	ﷲ	/I_M
36	صلى	/I_M
37	ﷲ	/I_M
38	عليه	/I_M
39	وسلم	/I_M
40	فقال	/I_M
41	ﷲ	/I_M
42	اكبر	/I_M
43	كلما	/I_M
44	وضع	/I_M

Figure 4.11: Example of a Hadith with a vague segmentation point, incorrectly segmented as shown by the horizontal line



```
1     حدثنا     /B_I
2     سليمان    /I_I
3     بن        /I_I
4     حرب       /I_I
5     حدثنا     /I_I
6     شعبه      /I_I
7     عن        /I_I
8     عبد       /I_I
9     الملك     /I_I
10    بن        /I_I
11    عمير      /I_I
12    عن        /I_I
13    قزعه      /I_I
14    مولى      /I_I
15    زياد      /I_I
16    قال       /I_I
17    سمعت      /I_I
18    اباسعيد
19    وقد       /I_I
20    غزا       /I_I
21    مع        /E_I
22    النبي     /B_M
23    صلى       /I_M
24    ﷲ         /I_M
25    عليه      /I_M
26    وسلم      /I_M
27    ثنتى      /I_M
28    عشره      /I_M
29    غزوه      /I_M
30    قال       /I_M
31    اربع      /I_M
32    سمعتهن    /I_M
33    من        /I_M
34    رسول      /I_M
35    ﷲ         /I_M
36    صلى       /I_M
37    ﷲ         /I_M
38    عليه      /I_M
39    وسلم      /I_M
```

Figure 4.12: Example of incorrect segmentation: it should be at token 30



```
9    اخبرنا    /I_I
10   عبد    /I_I
11   الرحمن    /I_I
12   بن    /I_I
13   نمر    /I_I
14   انه    /I_I
15   سمع    /I_I
16   ابن    /I_I
17   شهاب    /I_I
18   يخبر    /I_I
19   عن    /I_I
20   عروه    /I_I
21   عن    /I_I
22   عائشه    /E_I
23   ان    /B_M
24   النبي    /I_M
25   صلى    /I_M
26   ﷲ    /I_M
27   عليه    /I_M
28   وسلم    /I_M
29   جهر    /I_M
30   في    /I_M
31   صلاه    /I_M
32   الخسوف    /I_M
33   بقراءته
34   فصلى    /I_M
35   اربع    /I_M
36   ركعات    /I_M
37   في    /I_M
38   ركعتين    /I_M
39   واربع    /I_M
40   سجدات    /E_M
41   قال    /B_I
42   الزهري    /I_I
43   واخبرني
44   كثير    /I_I
45   بن    /I_I
46   عباس    /I_I
47   عن    /I_I
48   ابن    /I_I
49   عباس    /I_I
50   عن    /E_I
51   النبي    /B_M
52   صلى    /I_M
53   ﷲ    /I_M
54   عليه    /I_M
55   وسلم    /I_M
56   انه    /I_M
57   صلى    /I_M
58   اربع    /I_M
59   ركعات    /I_M
```

Figure 4.13: Example of a Hadith with irregular structure: it has more than one Isnad and Matn segment

4.8 Comparison and Discussion

This section discusses the different approaches to building the Hadith segmenter. First, it compares the look-up and ML approaches considering they were published in two different papers where the second approach (Altammami *et al.*, 2020a) is an improvement of the first (Altammami *et al.*, 2019). It investigates whether the approach used to annotate Hadith tokens contributes to the accuracy, or if the improvement is attributed to using Algorithm 3 for segmentation, as described in Section 4.6, which detects Isnads with irregular patterns.

In the following experiment, several factors are compared. First, the ML approach described in Section 4.6 is compared to the look-up approach described in Section 4.5. Also Algorithm 3 is used with both annotation approaches to segment Hadiths, since it is better than Algorithm 2 which assumes all Hadiths have a well-structured Isnad. Furthermore, data normalization is tested to understand if this factor contributes to better performance. The testing data of 500 Hadiths is divided into 450 regular and 50 irregular Hadiths to investigate how each approach handles the different kinds of Hadiths.

Table 4.16 shows the accuracy of using the two approaches along with Algorithm 3 to segment Hadiths. Results show that the ML model used for annotating tokens has an impact on performance. Also, normalization has a positive effect on the result. Hence it is recommended to normalize text before feeding it into the Hadith segmenter.

Another feature this section discusses is the use of bigrams as tokens. The experiments described in Section 4.5.2 show that using bigrams for Hadith segmentation is better than trigrams, specifically because the training data is relatively limited. Furthermore, annotating unigrams of words without considering context impose a challenge as many words exist in both the Isnad and the Matn. To illustrate this, a comparison of segmenter performance is shown in Figure 4.14, which depicts the accuracy of the Hadith segmenter built using the different n-gram representation models, token annotation approaches and segmentation

Dataset	Approach	Normalized	Accuracy
All 500 Hadith	Look-up lists	Yes	88.3%
	Look-up lists	No	87.1%
	ML	Yes	92.5%
	ML	No	92.3%
450 regular Hadith	Look-up lists	Yes	92.0%
	Look-up lists	No	90.5%
	ML	Yes	95.1%
	ML	No	95.1%
50 irregular Hadith	Look-up lists	Yes	56.0%
	Look-up lists	No	56.9%
	ML	Yes	79.0%
	ML	No	68.6%

Table 4.16: Comparison of results using the first two annotation approaches and normalization with Algorithm 3

algorithms. It is clear that among the different n-gram representation models, bigrams scores the highest.

The last topic to discuss in this section is the performance of the HMM compared to the other approaches. The table below shows their performance on the testing data of 500 Hadiths.

Approach	Accuracy
Look-up list	88.3%
ML	92.5%
HMM	90.25

Table 4.17: Performance of the different approaches to build the Hadith segmenter

Although the ML approach to build the Hadith segmenter produced the highest accuracy, not all Hadiths with irregular patterns were correctly segmented. In fact, this can be vague even for human annotators who are not experts in Hadith studies. For this reason, I argue that Hadith can be segmented to fine-grained

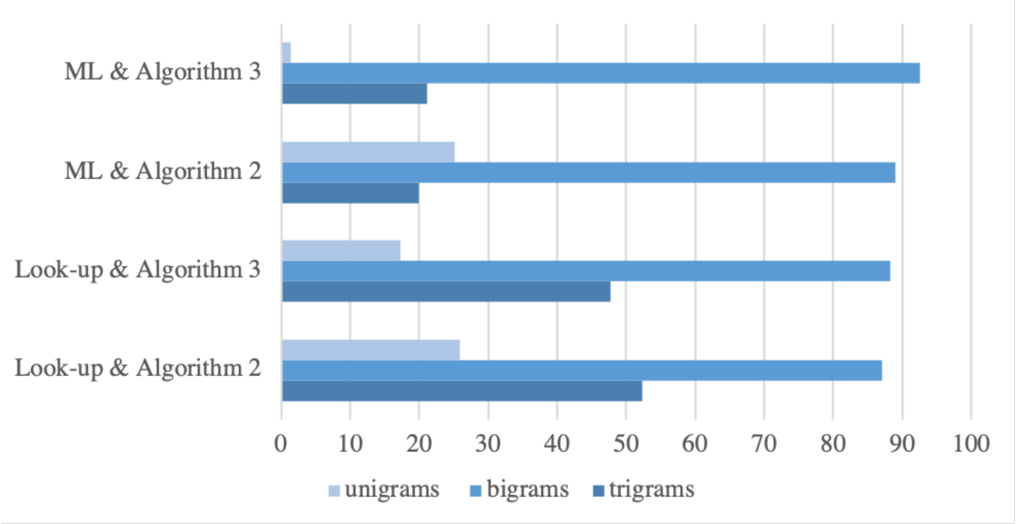


Figure 4.14: Segmenter performance with unigrams, bigrams, and trigrams

segments that go beyond two segments. This is because some Hadiths contain information in the Isnad that was identified as Matn segments by the proposed systems. For example, a Hadith Isnad may include information about a specific narrator as shown in the example in Figure 4.12 . Another example is a Hadith that starts a Matn segment with a piece of introductory information containing names of people, which is identified as an Isnad pattern by the segmenter as in Figure 4.11. Thus, the HMM could be the answer to such irregular Hadiths where the output is several segments instead of two as in Figure 4.13.

4.9 Conclusion

The main goal of this chapter is to create a system that can segment and annotate the components of Hadiths, the Isnad and the Matn. Previous attempts at segmenting Hadiths have used hand-crafted tools and were only tested on Hadiths from one book. To improve upon these limitations, the proposed system utilizes NLP techniques, and is tested on Hadiths from six different books to ensure a diverse range of Hadith types are covered. Despite achieving a 92.5% accuracy rate in segmenting Hadiths, the system struggles with some Hadiths that have irregular structures. However, it is still able to effectively segment and build a corpus of the six canonical Hadith books. In the future, the proposed HMM Hadith segmenter could be used to develop a tool that can deconstruct Hadiths with irregular patterns into fine-grained segments.

Chapter 5

Developing the Hadith Corpus

5.1 Introduction

Research in the area of Hadith computation is still in its infancy (Bounhas, 2019). However, there is an annual increase in the number of published papers indicating it is gaining wider attention from multi-disciplinary researchers (Azmi *et al.*, 2019). In such work, researchers gather their own datasets from different sources and sometimes manually process them (Luthfi *et al.*, 2018). This indicates a lack of adequate language resources and reusability is limited since most of the collected datasets are not published. Hence, it is unfeasible to establish benchmarks, compare results or set evaluation measures (Guellil *et al.*, 2019), which makes creating an accessible Hadith dataset an imperative.

This chapter, which is based on my publication Altammami *et al.* (2020b), describes the process of gathering and constructing a bilingual parallel corpus of Islamic Hadiths in their original Classical Arabic (CA) text and corresponding English translations. The corpus data is gathered from the six canonical Hadith collections then processed using a segmentation tool discussed in the previous chapter. The tool automatically segments and annotates the Hadith Isnad and Matn with 92.5% accuracy to produce this well-structured corpus containing more than 10M tokens [33].

[33] https://github.com/ShathaTm/LK-Hadith-Corpus

To the best of my knowledge, no parallel corpus of Hadith is freely available to the research community. The accessible data is scattered around the web in an unstructured format. Resources for CA constitutes only 11% of available Arabic resources (Guellil *et al.*, 2019), and Arabic parallel corpora in non-news genre are limited (Darwish *et al.*, 2021).

This language resource is named the *Leeds University and King Saud University (LK) Hadith corpus* to represent the collaboration between the two universities as I am pursuing my PhD at Leeds and working as a lecturer at King Saud University. The following section provides a detailed comparison and analysis of existing corpora, highlighting the unique features incorporated within the LK Hadith Corpus.

5.2 Related Work

There are many existing Arabic corpora (Atwell, 2018), but this section focuses on those that include Hadiths. Although there are several CA corpora, most existing ones are not dedicated to computational Hadith studies such as Al-Thubaity (2015). The KSU 50-million-word corpus of CA is designed to help researchers understand the use of words during the period of the Quran's revelation (Alrabiah *et al.*, 2013). Such corpora are not designed for researchers focused on Hadith as the Hadith components (Isnad and Matn) are not annotated.

There are other corpora which incorporate Hadith books, namely the Historical Arabic Corpus, or HAC (Hammo *et al.*, 2016), which contains 45 million words from different time periods. Tashkeela (Zerrouki & Balla, 2017) is a 76-million-word vocalised corpus of texts that represents classical and modern Arabic books. Again, these are not focused on the Hadith.

Another notable project is the Open Islamicate Texts Initiative (OpenITI), which is an international collaboration that incorporates various other projects, such as KITAB, under its umbrella. OpenITI uses an open-source OCR tool called Kraken ibn Ocropus to convert Arabic books into digital form. This tool relies on neural networks to recognize Arabic letters within an entire line, rather than trying to segment the text into words and then letters. This approach has an accuracy rate of 97.56% (Kiessling *et al.*, 2017). The goal of OpenITI is to eventually incorporate Persian and other languages, creating a vast Islamic corpus that includes all Hadith books, regardless of authenticity (Belinkov *et al.*, 2018). The meta-data included in this corpus includes the author's name, book title, and the author's date of death.

One of the recently published corpora is dedicated to Hadith. It includes four Hadith books scraped from different websites that cover several languages like Arabic, English and Urdu (Mahmood *et al.*, 2018). The LK Hadith corpus is different since it is an Arabic–English parallel corpus of the six canonical Hadith books. However, it might be worth investigating merging their corpus with the LK Hadith corpus by applying AI to align the Urdu translations.

Alosaimy & Atwell (2017) presented a Sunnah Arabic Corpus, which comprises 144,000 words extracted from the Hadith book Riyadu Assalihin . The corpus is annotated with POS tags. However, it does not segment Isnads from Matns.

A recent study surveyed and enumerated the freely available Arabic corpora. It mentions the existence of one Hadith corpus; however, this was not accessible or used in the literature (Zaghouani, 2017). This indicates a common problem where a dataset is lost. It occurs when researchers share data on personal websites that become obsolete after time. Therefore, to mitigate this problem, the LK Hadith corpus is shared on GitHub and has been indexed by The Arabic NLP Data Catalogue [34]. Table 5.1 highlights the difference between this corpus and previously existing ones.

[34] https://arbml.github.io/masader/

Corpus	**Hadith only**	**All canonical books**	**Isnad segmented from Matn**	**Parallel**	**Available**
Zerrouki & Balla (2017)		x			x
Hammo *et al.* (2016),		x			x
Alrabiah *et al.* (2013)		x			x
Belinkov *et al.* (2018)		x			x
Alosaimy & Atwell (2017)	x			x	
Mahmood *et al.* (2018)	x		x		x
LK Hadith Corpus	x	x	x	x	x

Table 5.1: Comparison of CA corpora that includes Hadith

5.3 Defining Corpus Data

This section provides an overview of the structural elements of Hadith and how they were considered in designing the LK Hadith Corpus. Furthermore, a discussion clarifying the reasons specific Hadith books were included in the corpus and the criteria used for their selection.

5.3.1 Hadith Structure

As mentioned in Chapter 2, a Hadith consists of two parts, Isnad and Matn. The first representing the reverse chronological chain of narrators and is followed by the latter, which is the actual teaching. 'Isnad' can be translated as 'support', since it is used to identify the authenticity of Hadith following the narrator's biography (علم الرجال). It is a meta-data that is useful for authenticity, but does not add useful information to the context of the actual narration (Matn). Therefore, in designing LK Hadith corpus, it is crucial to separate the Isnad from the Matn to enable researchers to access the different components.

5.3.2 Existing Hadith Books

As noted in Section 2.4.2, in the Islamic literature, there are six canonical Hadith books which are considered authentic. They are a hybrid of two book genres, *Musannaf* and *Musnnad* (Brown, 2017). The former includes books that categorizes Hadiths into topics and does not emphasise on the authenticity. On the other hand, Musnnad books organize Hadith based on chains of narrators to place more emphasise on authenticity. This hybrid genre became known as 'Sahih' or 'Sunan', where authentic Hadiths are organized under subtitles that indicate the legal implications or rulings the reader should derive from the subsequent Hadiths.

Nowadays, these canonical books are collectively called 'Al-Sihah al-Sittah', which translates as 'The Authentic Six', and they include *Sahih Albukhari*, *Sahih Muslim*, *Sunan Abu Daud*, *Sunan Tirmizi*, *Sunan Ibn Maja* and *Sunan Nesa'i*. These form the base for Islamic Hadith books. It is worth noting that these books are named after the scholars who compiled them. For example, Sahih Albukhari was compiled by Muhammad Albukhari who dedicated years of his life studying the authenticity of Hadiths before adding them to his book.

Despite their collective name ‘The Authentic Six’, not all incorporated Hadiths possess the same degree of authenticity. Rather, they were named on the basis of on the dominance of authentic Hadiths found in them (Khan, 1987).

5.4 Corpus Creation

5.4.1 Resource Enumeration

The canonical books organise Hadiths into a topology of topics on three levels, (Book, Chapter, Section). Each chapter is dedicated to one theme. Within the chapter, there are several sections that the author used to indicate a ruling on specific matters, given the incorporated Hadiths as evidence. The structure of these books is illustrated in Figure 5.1. Each Hadith consists of two parts, an Isnad and a Matn, and some books add a comment by the author, usually regarding the authenticity of the Hadith.

To maintain this structure in the corpus, electronic sources of Hadith that followed this structure are sought. Several websites host Hadith books; however, they did not meet this requirement. For example, *ahadith.co.uk* contains the English version of the Hadith with the section and chapter titles removed. Another valuable website is *islamweb.net*, which hosts a huge number of Islamic resources, including Hadiths. However, it does not fulfil the requirement of a parallel corpus (English and Arabic aligned Hadiths). Only *sunnah.com* provides the required features. It maintains the structure of the books, and the English translation is aligned in parallel with the Arabic Hadith at the narrative level.

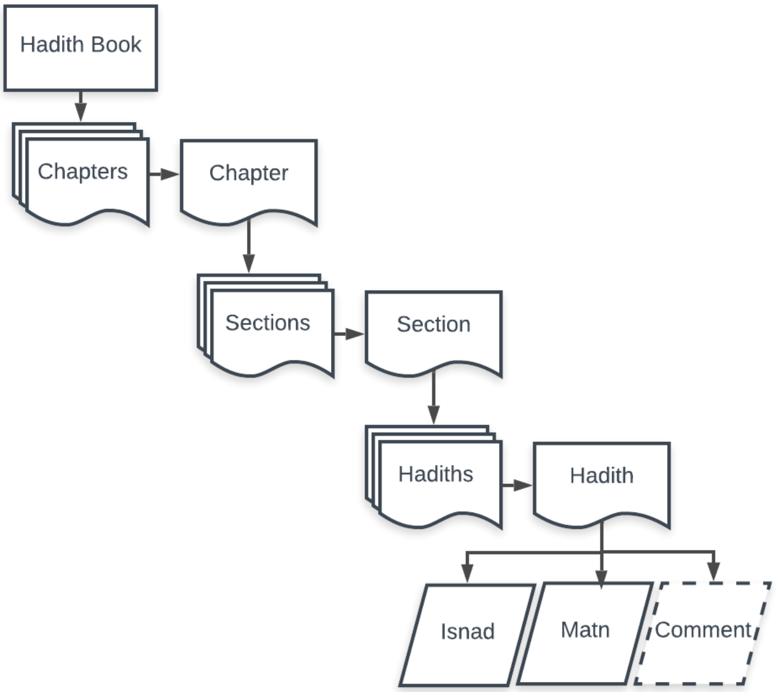
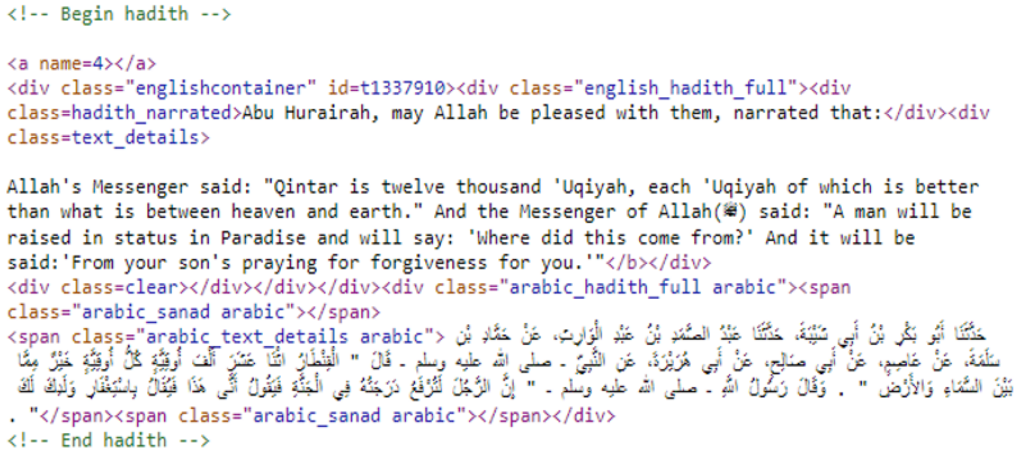


Figure 5.1: Structure of the canonical Hadith books

5.4.2 Data Collection

A tool is developed to scrape *sunnah.com*[35] pages and extract information from every Hadith. However, HTML tags are not used consistently on the website. This could be due to its being built by a group of web developers. For example, the Arabic Isnad is not separated from the Matn in most Hadiths, despite the existence of an HTML tag `<Arabic_sanad arabic>` dedicated to the Isnad, as shown in Figure 5.2. To overcome this, a Hadith segmentation tool, described in Chapter 4, is used to automatically maintain a consistent segmentation of Isnad and Matn.

[35] Sunnah.com permits using their data for education

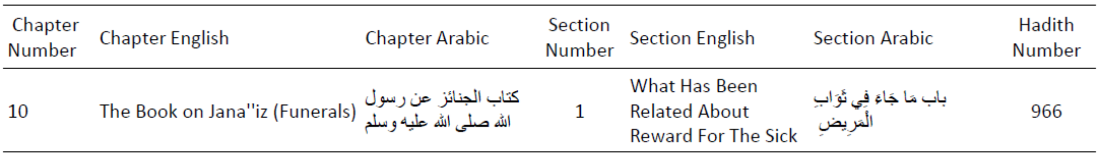


```
<!-- Begin hadith -->

<a name=4></a>
<div class="englishcontainer" id=t1337910><div class="english_hadith_full"><div
class=hadith_narrated>Abu Hurairah, may Allah be pleased with them, narrated that:</div><div
class=text_details>

Allah's Messenger said: "Qintar is twelve thousand 'Uqiyah, each 'Uqiyah of which is better
than what is between heaven and earth." And the Messenger of Allah(ﷺ) said: "A man will be
raised in status in Paradise and will say: 'Where did this come from?' And it will be
said:'From your son's praying for forgiveness for you.'"</b></div>
<div class=clear></div></div></div><div class="arabic_hadith_full arabic"><span
class="arabic_sanad arabic"></span>
<span class="arabic_text_details arabic"> حَدَّثَنَا أَبُو بَكْرِ بْنُ أَبِي شَيْبَةَ، حَدَّثَنَا عَبْدُ الصَّمَدِ بْنُ عَبْدِ الْوَارِثِ، عَنْ حَمَّادِ بْنِ
 سَلَمَةَ، عَنْ عَاصِمٍ، عَنْ أَبِي صَالِحٍ، عَنْ أَبِي هُرَيْرَةَ، عَنِ النَّبِيِّ - صلى الله عليه وسلم - قَالَ " الْقِنْطَارُ اثْنَا عَشَرَ أَلْفَ أُوقِيَّةٍ كُلُّ أُوقِيَّةٍ خَيْرٌ مِمَّا
 بَيْنَ السَّمَاءِ وَالأَرْضِ " . وَقَالَ رَسُولُ اللَّهِ - صلى الله عليه وسلم - " إِنَّ الرَّجُلَ لَتُرْفَعُ دَرَجَتُهُ فِي الْجَنَّةِ فَيَقُولُ أَنَّى هَذَا فَيُقَالُ بِاسْتِغْفَارِ وَلَدِكَ لَكَ
. "</span><span class="arabic_sanad arabic"></span></div>
<!-- End hadith -->
```

Figure 5.2: HTML of a Hadith page on *Sunnah.com*

5.4.3 Corpus Annotation

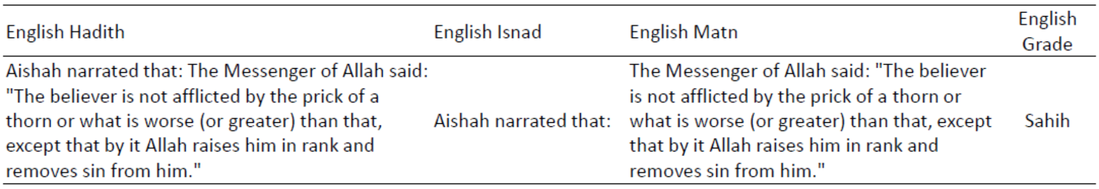
The scrapped information from *sunnah.com* is processed to be added to the corpus. The Hadith segmenter processes every Arabic Hadith to extract the Isnad and Matn. Then the Hadith, Isnad, Matn and the meta data including, chapter, section, and Hadith number are saved as a record where information are separated by commas. Hence, the CSV (comma separated values) files are used with UTF-8 encoding. This annotation can be easily converted to XML format that can be used across different systems. Every CSV file contains the following information listed in Table 5.2. An example of how one Hadith record is represented in a CSV file is broken down for readability is given in figures 5.3, 5.4 and 5.5.



Chapter Number	Chapter English	Chapter Arabic	Section Number	Section English	Section Arabic	Hadith Number
10	The Book on Jana''iz (Funerals)	كتاب الجنائز عن رسول الله صلى الله عليه وسلم	1	What Has Been Related About Reward For The Sick	باب مَا جَاءَ فِي ثَوَابِ الْمَرِيضِ	966

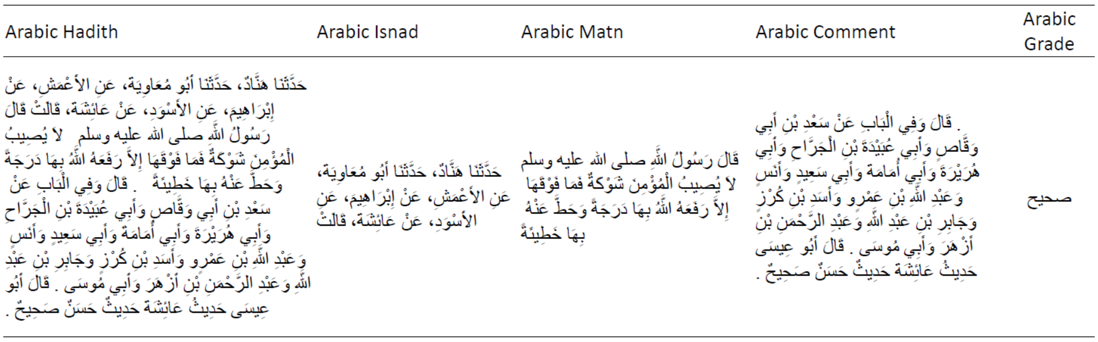
Figure 5.3: Example of a Hadith record extracted from Tarmizi Chapter10.csv – Part 1

An illustration of the LK Corpus structure is shown in Figure 5.6. It is a simple structure that corresponds to the original structure of the books. The LK



English Hadith	English Isnad	English Matn	English Grade
Aishah narrated that: The Messenger of Allah said: "The believer is not afflicted by the prick of a thorn or what is worse (or greater) than that, except that by it Allah raises him in rank and removes sin from him."	Aishah narrated that:	The Messenger of Allah said: "The believer is not afflicted by the prick of a thorn or what is worse (or greater) than that, except that by it Allah raises him in rank and removes sin from him."	Sahih

Figure 5.4: Continued example of Hadith record – Part 2



Arabic Hadith	Arabic Isnad	Arabic Matn	Arabic Comment	Arabic Grade
حَدَّثَنَا هَنَّادٌ، حَدَّثَنَا أَبُو مُعَاوِيَةَ، عَنِ الأَعْمَشِ، عَنْ إِبْرَاهِيمَ، عَنِ الأَسْوَدِ، عَنْ عَائِشَةَ، قَالَتْ قَالَ رَسُولُ اللَّهِ صلى الله عليه وسلم لاَ يُصِيبُ الْمُؤْمِنَ شَوْكَةٌ فَمَا فَوْقَهَا إِلاَّ رَفَعَهُ اللَّهُ بِهَا دَرَجَةً وَحَطَّ عَنْهُ بِهَا خَطِيئَةً . قَالَ وَفِي الْبَابِ عَنْ سَعْدِ بْنِ أَبِي وَقَّاصٍ وَأَبِي عُبَيْدَةَ بْنِ الْجَرَّاحِ وَأَبِي هُرَيْرَةَ وَأَبِي أُمَامَةَ وَأَبِي سَعِيدٍ وَأَنَسٍ وَعَبْدِ اللَّهِ بْنِ عَمْرٍو وَأَسَدِ بْنِ كُرْزٍ وَجَابِرِ بْنِ عَبْدِ اللَّهِ وَعَبْدِ الرَّحْمَنِ بْنِ أَزْهَرَ وَأَبِي مُوسَى . قَالَ أَبُو عِيسَى حَدِيثُ عَائِشَةَ حَدِيثٌ حَسَنٌ صَحِيحٌ .	حَدَّثَنَا هَنَّادٌ، حَدَّثَنَا أَبُو مُعَاوِيَةَ، عَنِ الأَعْمَشِ، عَنْ إِبْرَاهِيمَ، عَنِ الأَسْوَدِ، عَنْ عَائِشَةَ، قَالَتْ	قَالَ رَسُولُ اللَّهِ صلى الله عليه وسلم لاَ يُصِيبُ الْمُؤْمِنَ شَوْكَةٌ فَمَا فَوْقَهَا إِلاَّ رَفَعَهُ اللَّهُ بِهَا دَرَجَةً وَحَطَّ عَنْهُ بِهَا خَطِيئَةً	. قَالَ وَفِي الْبَابِ عَنْ سَعْدِ بْنِ أَبِي وَقَّاصٍ وَأَبِي عُبَيْدَةَ بْنِ الْجَرَّاحِ وَأَبِي هُرَيْرَةَ وَأَبِي أُمَامَةَ وَأَبِي سَعِيدٍ وَأَنَسٍ وَعَبْدِ اللَّهِ بْنِ عَمْرٍو وَأَسَدِ بْنِ كُرْزٍ وَجَابِرِ بْنِ عَبْدِ اللَّهِ وَعَبْدِ الرَّحْمَنِ بْنِ أَزْهَرَ وَأَبِي مُوسَى . قَالَ أَبُو عِيسَى حَدِيثُ عَائِشَةَ حَدِيثٌ حَسَنٌ صَحِيحٌ .	صحيح

Figure 5.5: Continued example of Hadith record – Part 3

Hadith corpus folder contains six folders representing the six canonical Hadith books. Within these folders the CSV files represent the chapters in the book. For example, 97 CSV files were created in the Sahih Albukhari folder, which represent the number of chapters in Sahih Albukhari. The first CSV file is named ‘Chapter1.csv’ and it contains seven Hadith records.

.

Annotation	Description
Chapter Number	The chapter number where the Hadith is listed.
Chapter English	Title of the chapter in English.
Chapter Arabic	Title of the chapter in Arabic.
Section Number	The section number where the Hadith is listed.
Section English	Title of the section in English.
Section Arabic	Title of the section in Arabic.
Hadith number	The sequential number of the Hadith.
English Hadith	The whole English Hadith consisting of Isnad and Matn.
English Isnad	The name of the first narrator in English.
English Matn	The actual Hadith teaching in English.
Arabic Hadith	The whole Arabic Hadith consisting of Isnad and Matn.
Arabic Isnad	The chain of narrators in Arabic.
Arabic Matn	The actual Hadith teaching in Arabic.
Arabic Comment	An optional value that contains the scholar's comment on the authenticity of the Hadith.
English Grade	The degree of authenticity in the transliteration.
Arabic Grade	The degree of authenticity in Arabic.

Table 5.2: Corpus Annotation

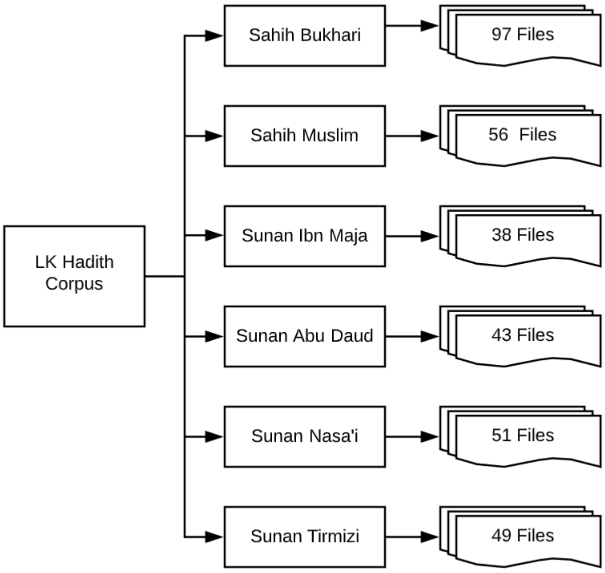


Figure 5.6: Structure of the LK Hadith corpus

5.5 Corpus Analysis

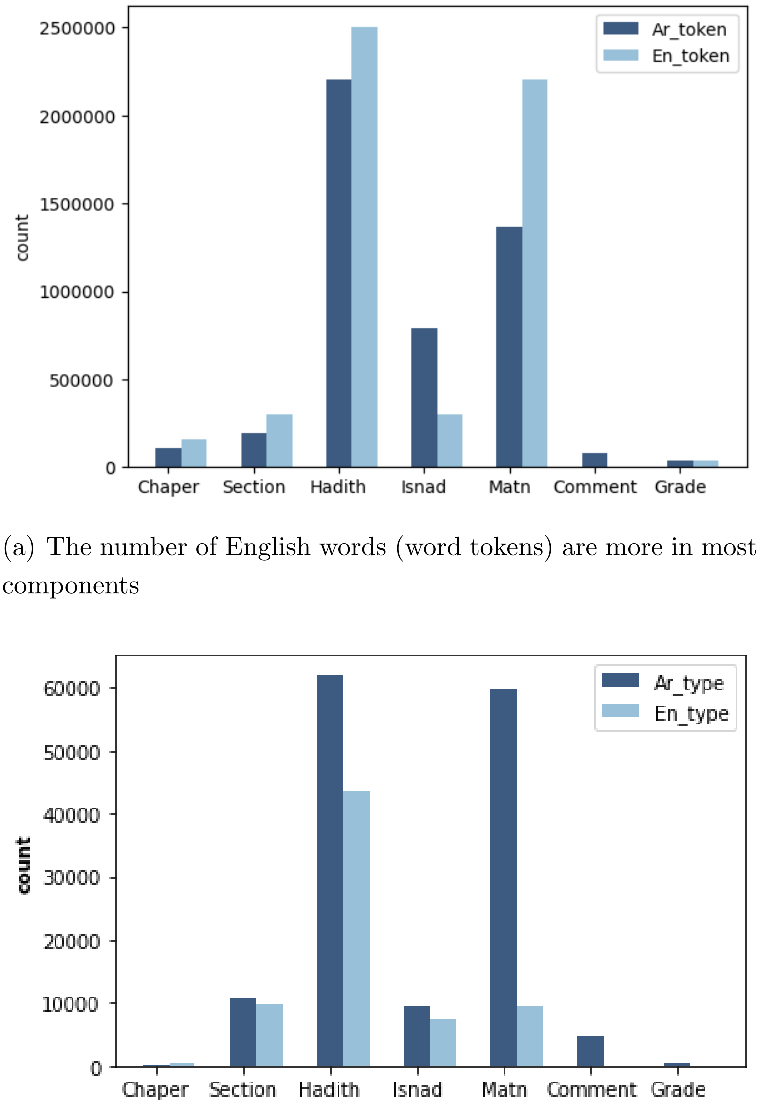
5.5.1 Corpus Statistics

The corpus contains 39,038 Hadith records distributed among the six books as illustrated by Table 5.3. Each record consists of the original Arabic and an aligned English translation, making more than 10 million tokens. A comparison between the Arabic and English texts in the corpus is shown in Figure 5.7. The number of word tokens in English Hadiths is larger than that in Arabic, as shown in Figure 5.7(a). This is because some Arabic words represent a phrase or a sentence. For example وسيبعثون is translated as 'and they will send'. The English Isnad has less words than the Arabic Isnad, because the English translation contains only the primary narrator of the Hadith rather than the chain. In the other hand, the Arabic Hadith is richer in word types, as shown in Figure 5.7(b).

Book	Number of Hadiths
Albukhari	6,633
Muslim	7,293
Nesa'i	5,680
Ibn Maja	10,082
Tirmizi	4,209
Abu Daud	5,141
Total	39,038

Table 5.3: Number of Hadiths in each book

To further analyse the corpus, the most frequent words in the Arabic and English Isnads are obtained, as shown in Figure 5.8. It is clear that the Isnad involves transmission words such as 'said', 'narrated', and 'reported'in addition to the narrators names. Furthermore, some narrators' names are spelled differently in different places in the English Translations.g. 'Hrairah', 'Hraira'. Hence, normalization should be considered when using the English Isnad data.



(a) The number of English words (word tokens) are more in most components

(b) The number of unique Arabic words (word type) are more in most components

Figure 5.7: A comparison between the Arabic and English parts of the LK Hadith corpus. Note the difference in the y-axis numbering to reflect the number of words

Since the honorific phrase صلى الله عليه وسلم Peace be upon him (PBUH) is incorporated in the Matn, these words appear as the most frequent one in Figure 5.9. They are considered to be part of the Matn to facilitate distinguishing the Prophet's sayings from sayings by his companions. However, this phrase could be removed easily using regular expressions if a researcher wishes to focus on the teachings alone.

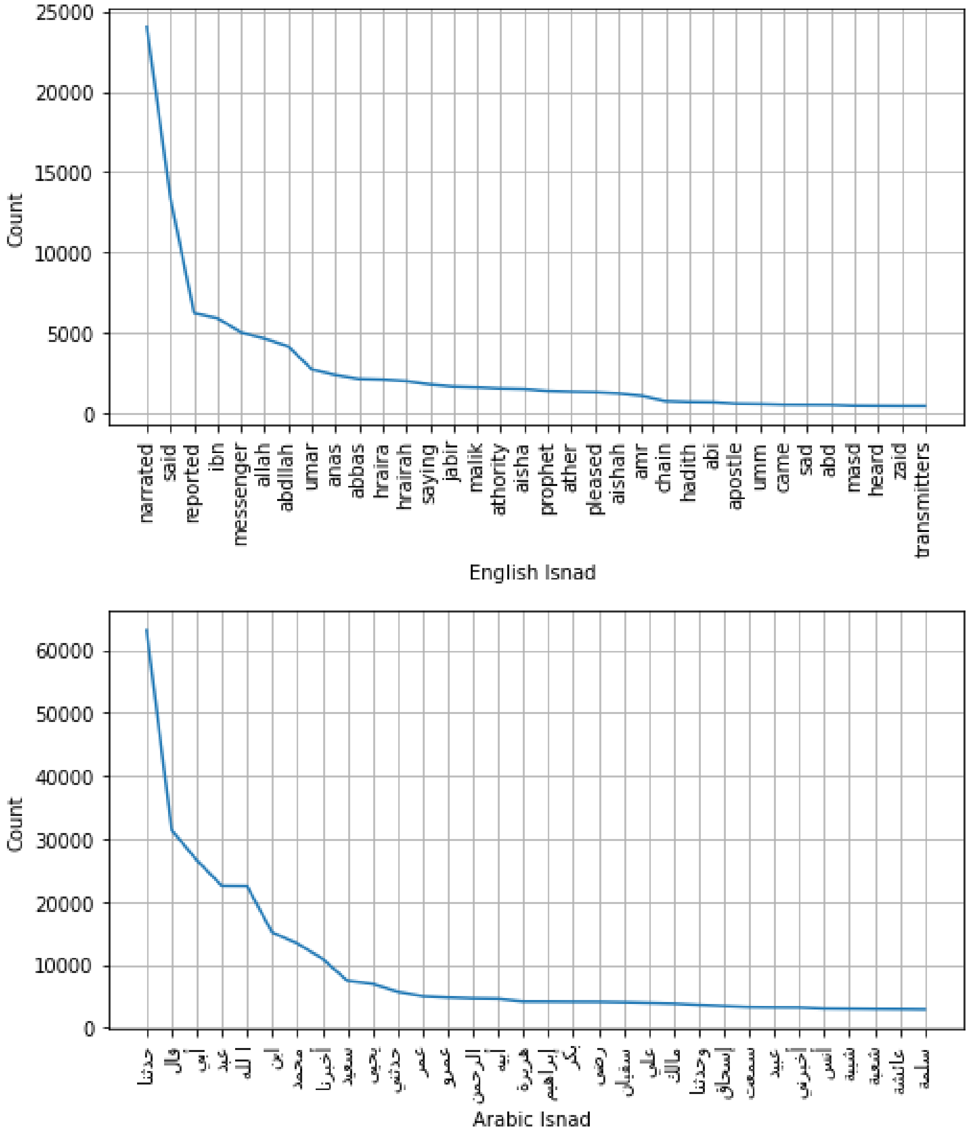


Figure 5.8: Most frequent words in the Isnad

5.5.2 Manual Correction and Verification

Following the initial compilation of the dataset, manual intervention was necessary to identify and fix inconsistencies, if any. I checked the Sahih Albukhari data against the PDF version of the book. Minor mistakes have been found e.g. a Hadith is placed under the wrong section, or the English translation was of another Hadith, which is normal since human efforts are susceptible to mistakes.

Therefore, LK Hadith corpus relies on the source. In other words, missing values or inconsistencies with the original book are dependent on *Sunnah.com*. Currently, Sahih Albukhari data has been thoroughly checked against the book and the mistakes produced by the segmenter has been fixed. This makes Sahih Albukhari the gold standard of the LK Hadith corpus. The other Hadith books were not fully checked, but the first 50 Hadiths of each book were manually scrutinized. The result show that the annotation of Isnad and Matn segments in the corpus has an error rate of 7.6%.

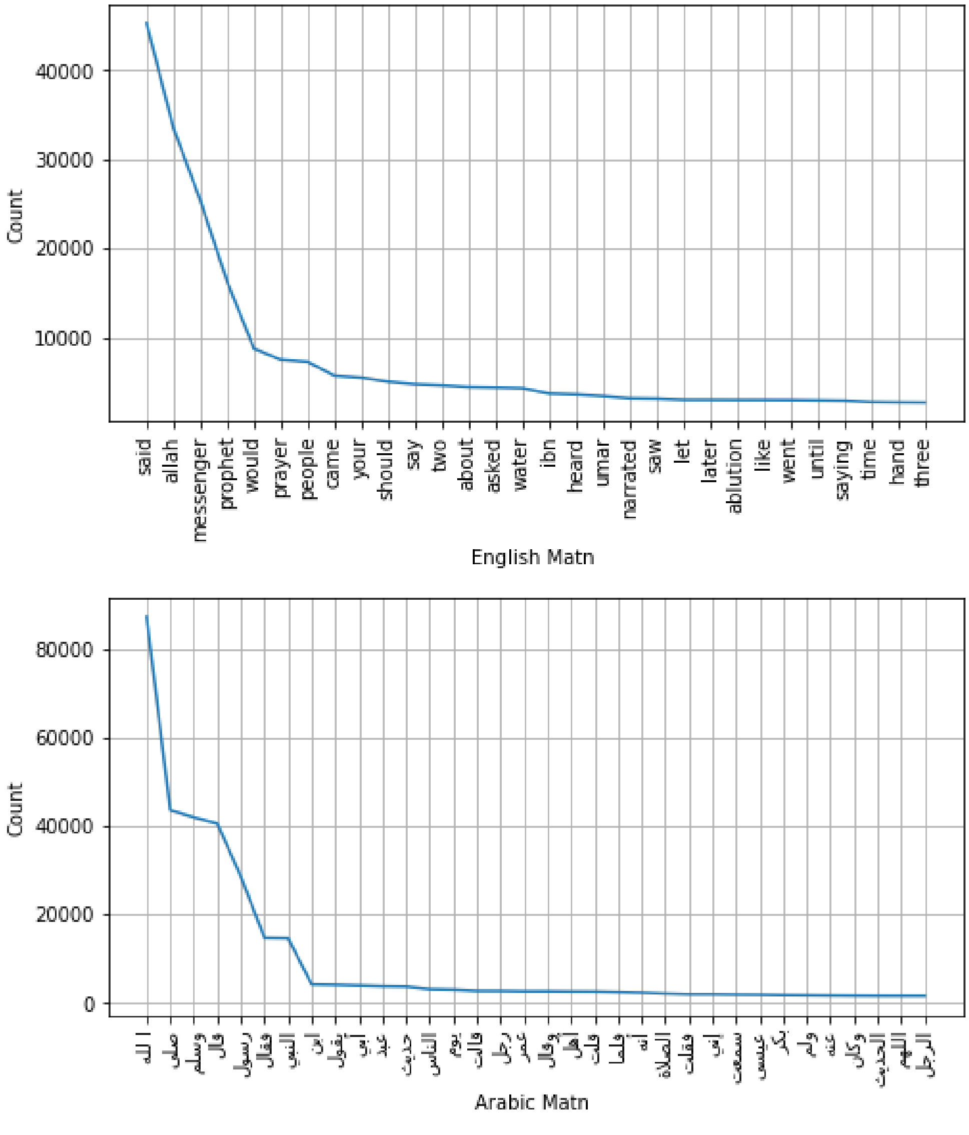


Figure 5.9: Most frequent words in the Matn

5.6 Potential Uses

One of the main objectives of this research is to provide the research community with a Hadith corpus that is well-annotated for diverse research purposes. Hence, every component is annotated in such a way as to be easily extracted. Once researchers have access to a common dataset, it is possible to set benchmarks and compare results.

This Hadith corpus is already being used (Tarmom *et al.*, 2021) to build AI systems that classify authentic Hadiths and the non-authentic Hadiths collected by Tarmom *et al.* (2019). Also Habash *et al.* (2022) included parts of the LK Hadith corpus in their annotated open-source dependency treebank. Potential uses for the corpus include:

1. To build ontologies that support Hadith authenticity by focusing on the Isnad. Such systems could be tested using the Isnads extracted from this corpus.

2. To study the words of the Prophet by extracting the Matn which start with 'The Prophet (PBUH) said'.

3. To apply AI methods to the Matn and automatically link Hadiths to the Quran without the Isnad affecting the results.

4. To use the parallel corpus for training machine translation models for CA.

5. To utilize section headings to derive Hadith topics that can be used as the terminology for a Hadith ontology.

5.7 Conclusion

This chapter presents the creation of the LK Hadith parallel corpus in its original Classical Arabic form with English translations. It incorporates more than 39 K Hadiths where every component and meta data is annotated accordingly. These include Hadiths and the book, chapter and section headings it is extracted from.

In addition, each Hadith is deconstructed into its two main components (Isnad and Matn) using a segmentation tool described in Chapter 4. This parallel corpus opens new avenues to various research areas such as machine translation of Classical Arabic.

In the subsequent chapters, the LK Hadith corpus plays a crucial role in the experiments aimed at identifying the relationships between Hadith teachings (Matn) and the Quran. The well-structured nature of the corpus allows for the convenient extraction of the Matn component, which enables an in-depth analysis of the actual Hadith teaching without the influence of the Isnad on the results.

Chapter 6

Evaluating Quran Ontologies for Hadith

6.1 Introduction

As discussed in Chapter 2, the Quran and the Hadith are two religious texts used by Muslims as a reference to perform their daily religious obligations. Hence, developing a computational tool to better understand and link these different texts will be useful for Islamic scholars, learners, and laymen. I hypothesize that linking them using a knowledge-based approach by using an existing Quran ontology that covers the Hadiths is possible. However, there are several Quran ontologies, none of which have been rigorously evaluated using a standard ontology evaluation method. For this reason, the experiment in this chapter starts with enumerating and discussing the existing Quran ontologies. Then the top candidates are evaluated using a corpus-based approach to determine the level of overlap between the ontologies and the Hadith texts. The experiment shows that one Quran ontology could be used as a starting point for a larger Islamic ontology that covers both the Quran and Hadiths.

6.2 What is an Ontology?

The fascinating idea of capturing human expert knowledge in a machine-readable format is a common goal of ontologies, which is defined by Gruber as ‘ a specification of a conceptualization’(Gruber, 1993). The word “conceptualization” refers to the abstract and simplified description of an area of concern one wishes to represent for a specific purpose. Hence, the ontology identifies concepts and relationships to infer associations that form the bases for complex semantic knowledge. In other words, a domain ontology is the backbone knowledge of a particular field, or an expert’s knowledge captured in a machine-readable format. This captivating idea sparked research interest in the knowledge representation of Islamic teachings.

Ontologies are especially important for religious texts as they provide a way to better understand them. This is because such ancient texts can be challenging for non-experts to interpret. Additionally, simple string matching searches are not sufficient for computers as they do not capture the underlying moral behind a story told by the Prophet. For instance, currently available Hadith search tools on the web like *Sunnah.com*, only return Hadiths with an exact match with the search terms, without considering the more general meaning of the text. This approach frequently fails to provide the desired information. Therefore, the goal is to model the meaning of Hadith texts to enhance information retrieval efficiency and aid in uncovering potentially obscured knowledge by connecting Hadiths to the Quran.

Instead of creating a new ontology for Hadiths from scratch, existing ontologies for the Quran are considered, because both cover the same domain of Islamic teachings. In recent years, various attempts have been made to construct ontologies that capture the structural and semantic connections within and between the verses of the Quran (Alqahtani & Atwell, 2018). Researchers have explored different types of knowledge representation models and attempted to combine different Quran ontologies through data aggregation (Abbas *et al.*, 2013). This chapter aims to investigate which Quran ontology is most suitable for Hadiths.

To the best of my knowledge, computationally linking Hadiths to the Quran is a novel activity and no prior research has explored using Quran ontologies as a foundation for creating an ontology for Hadith texts. Currently, the existing links between the Quran and Hadiths are created manually, through explicit references made by the Prophet or Islamic scholars.

6.3 Existing Quran Ontologies

A number of Quran ontologies have been developed; some cover the whole Quran while others focus on specific topics. Although this research area has been explored for a decade, it is not yet mature with a unified, validated, available Quranic ontology that covers the whole Quran with its various topics. The following paragraphs enumerate the existing Quran ontologies and discuss the most appropriate candidate to be the base for an Islamic ontology that covers the Hadiths.

The most frequently cited Quran ontologies aim to cover the whole Quran and were developed at Leeds University (Alqahtani & Atwell, 2018). These were at different levels and for different purposes. These include the verse level to encode the conceptual meaning of each verse, which was done by Abbas (2009) who incorporates more than 1,000 concepts linked to all verses in the Quran. She developed this using an Islamic book Mushaf Al-Tajweed, where the meaning of each Quranic verse is elucidated by scholars and the topic of the verse is encoded in the book index.

The other widely used ontology is the Quranic Arabic Corpus (QAC) developed by Dukes (2015), who used the famous book Tafsir Ibn Kathir to extract 300 concepts. Similarly, Tafsir Ibn Kathir has been used by Sharaf & Atwell (2012b) to build Qursim ontology that consists of 7,600 related verses. Sharaf & Atwell (2012a) also introduced an ontology that describes pronoun antecedents in the Quran. This consists of 1,050 concepts and 2,700 relations that were manually built during Sharaf's PhD.

Other researchers created ontologies for specific domains of the Quran. For example, Khan *et al.* (2013) presented an ontology based on an Islamic book describing animals mentioned in the Quran. Another ontology that focuses on a specific domain presents a model that captures the meaning of time nouns in the Quran (Al-Yahya *et al.*, 2010). A similar approach was used by Alromima *et al.* (2015) to build an ontology dedicated to place nouns in the Quran. Their work is based on Alam Almakan, Fe Al Quran, which is an MSc thesis involving a linguistic computational study. An interesting model to build a Quran ontology specific to human relations concepts was introduced by Tashtoush *et al.* (2017). For each class, they defined a set of synonyms in Arabic, then manually linked the concepts or classes to each verse. The Semantic Quran (Sherif & Ngonga Ngomo, 2015) focuses on concepts in terms of Quran text structure parsing: every chapter, verse, word and lexical item is treated as a "concept", linked by ontology relations such as 'isPartOf'.

Although the above ontologies are built using different technologies, there have been some attempts to unify them (Alrehaili & Atwell, 2018). However, not all ontologies are designed to be reused. Many proposed ontologies are not presented in an applicable and reusable way, but rather to prove a concept in the research environment (Ahmad *et al.*, 2013).

Some researchers went beyond aligning ontologies by not only combining ontologies, but extending them. Hakkoum & Raghay (2016) integrated Semantic Quran (Sherif & Ngonga Ngomo, 2015) and QAC ontologies using the Owl property 'Same-as' to indicate the two concepts representing the same thing, which can be used interchangeably. Then they extended it by manually extracted concepts from Tafsir (exegesis) books.

The discussed ontologies are further summarized in Table 6.1, which illustrates that four ontologies attempt to cover all topics across the whole Quran. However, the ontology introduced by Hakkoum & Raghay (2016) is composed of the ontology of Dukes (2015) and that of Sherif & Ngonga Ngomo (2015). Therefore, only the two distinctive ontologies are evaluated, Abbas' Qurany ontology and the

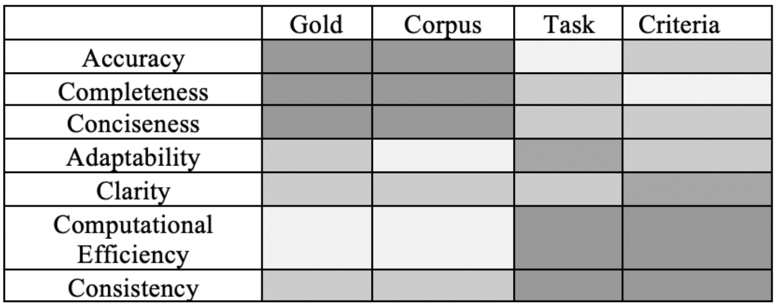
Quran ontology of Hakkoum & Raghay (2016). To avoid any confusion between the names of the evaluated ontologies, the ontology of Abbas (2009) is referred to as 'Ontology A' and the one by Hakkoum & Raghay (2016) as 'Ontology B' throughout this chapter.

6.4 Ontology Evaluation Methods

Ontology evaluation typically involves assessing existing ontologies for a particular purpose. In this research, the Quran ontologies are evaluated to find whether or not they cover the Hadith corpus and can be used to identify semantic relatedness between Hadiths and the Quran that go beyond lexical similarity. There is a plethora of research on ontology evaluation methods and techniques (McDaniel & Storey, 2019). The consensus is that ontology evaluation techniques can be grouped into four categories: gold-based, corpus-based, task-based and criteria based approaches (Raad & Cruz, 2015). These aim is to evaluate the ontology's quality and correctness by assessing several metrics including accuracy, completeness, conciseness, consistency, adaptability, expandability and clarity (Hlomani & Stacey, 2014). Each approach evaluates the ontology differently as shown in Figure 6.1, where a darker colour in the table represents a better coverage for the corresponding criterion.

The most appropriate approach to evaluating the Quran ontologies 'fit' for Hadiths is the corpus-based, for two reasons. First, they are created based on a textual corpus, the Quran. Second, these ontologies are evaluated for the Hadith, which is also a textual corpus. The other approaches could be more useful for other types of ontologies that were built using knowledge elicitation from experts.

A corpus-based approach, also known as data driven approach, is used to evaluate how far an ontology covers a given domain (Raad & Cruz, 2015). A widely used method of corpus-based evaluation is established by Brewster *et al.* (2004). It relies on three main steps. The first is extracting keywords from a corpus either using clustering methods or topic modelling. Then, query expansion is applied to capture synonyms using Wordnet or other information retrieval (IR)



	Gold	Corpus	Task	Criteria
Accuracy	▓	▓	░	▒
Completeness	▓	▓	▒	░
Conciseness	▓	▓	▒	▒
Adaptability	▒	░	▓	▒
Clarity	▒	▒	▒	▓
Computational Efficiency	░	░	▓	▓
Consistency	▒	▒	▓	▓

Figure 6.1: An overview of ontology evaluation methods (Raad & Cruz, 2015)

Paper	**Ontology Name**	**Source**	**Language**	**Topics**	**Format**	**Number of Concepts**	**Available**
Dukes (2015)	QAC	Ibn Kathir	En,Ar	Entire Quran	XML	300	Yes
Hakkoum & Raghay (2016)	Quran	QAC, Semantic Quran	En, Ar	Entire Quran	OWL	1181	Yes
Abbas (2009)	Qurany	Mushaf Al-Tajweed	En, Ar	All Quran	XML	1100	Yes
Sharaf & Atwell (2012b)	QurSim	Ibn Kathir	Ar	Similar Verses	SQL	7600	Yes
Sharaf & Atwell (2012a)	Qurana	Ibn Kathir	Ar	Pronoun Anacedent	SQL	1,050	Yes
Khan *et al.* (2013)	N/A	Hewanat-E-Qurani	En	Animals	OWL	167	No
Al-Yahya *et al.* (2010)	N/A	Arabic dictionaries and lexicons	Ar	Time Nouns	OWL	53	Yes
Alromima *et al.* (2015)	N/A	Alam Almakan Fe Al Quran	En, Ar	Places Nouns	OWL	99	No
Tashtoush *et al.* (2017)	N/A	Tafsir al-Jalalayn	Ar	Social Relations	OWL	15	No
Saad *et al.* (2011)	Solat	Quran + Hadith	En	Salat	OWL	48	No
Sherif & Ngonga Ngomo (2015)	Semantic Quran	Tanzil project, QAC	42 Lang.	Text structure parsing	RDF	26,735	Yes

Table 6.1: Existing Quran ontologies

techniques. Finally, the ontology is evaluated by mapping the set of keywords identified in the corpus to the concepts of an ontology being evaluated.

6.5 Quran Ontology Evaluation

In this experiment, two ontologies are evaluated to investigate their appropriateness for being the basis of an Islamic ontology that covers the Hadiths in addition to the Quran. Table 6.2 below shows a Quran verse with the associated concepts extracted from the two ontologies. It is worth noting that Ontology B not only captures topics, but also the structure of the Quran and the links between similar verses. Moreover, it includes the Tafsir (exegesis) extracted from two books, Jalalayn and Muyasser.

Since a corpus-based evaluation of the ontologies is conducted, the Albukhari Hadith book is used as the evaluating corpus. This book is the most famous Hadith compilation and it organizes and categorizes Hadiths under headings that indicate the legal implications or ruling the scholar (Albukhari) derived from the subsequent Hadiths based on his deep knowledge of and expertise in the Hadiths.

The first step in the corpus-based approach for ontology evaluation is keyword extraction. However, instead of extracting keywords from the Hadith Matn (the part of the Hadith that contains the actual teaching), the Albukhari Hadith book section-headings are used. I speculate that since the evaluated Quran ontologies are based on books like Mushaf Altajweed, where scholars categorise Quran verses into topics, section-headings in Albukhari are sensible keywords that capture the topic of the incorporated Hadiths.

I hypothesize that there is a larger degree of overlap between the ontology concepts and Albukhari's section-headings than the actual Hadith Matn and Quran verses. This assumption is based on a previous study(Sayoud, 2012) that identified how the Prophet Muhammed's style in Hadith statements (Matn) is frequently hyperboles (Brown, 2017), which are dramatic statements different from

	Arabic	English Translation
Quran verse (2:27)	الَّذِينَ يَنقُضُونَ عَهْدَ اللَّهِ مِن بَعْدِ مِيثَاقِهِ وَيَقْطَعُونَ مَا أَمَرَ اللَّهُ بِهِ أَن يُوصَلَ وَيُفْسِدُونَ فِي الأَرْضِ أُولَـئِكَ هُمُ الْخَاسِرُونَ	sever what Allah has ordered to be joined (as regards Allah's Religion of Islamic Monotheism, and to practise its legal laws on the earth and also as regards keeping good relations with kith and kin), and do mischief on earth, it is they who are the losers.
Ontology-A Abbas (2009)	أركان الإسلام، التوحيد	Pillars of Islam, Islamic
	وعيد المفسدين	The Threatening of the Mischievous
	والمجرمين والفاسقين	The Criminals and the Impious
	العمل، العمل الطالح	Action (Work), Bad Deeds
	وعيد المفسدين	Threatening the Spoilers
	الأخلاق الحميدة	Good Morals
	الأخلاق الذميمة	The Dispraised Morals
	الوفاء بالعهد	Keeping Promise
	الإفساد، الخبث، نقض العهد	Corruption, Cunning Break Promise
	الإنسان والعلاقات الاجتماعية	Human and The Social Relations
	صلة ذوي القربى	Kinship
Ontology-B Hakkoum & Raghay (2016)	أركان الإسلام، التوحيد	Pillars of Islam, Islamic
	وعيد المفسدين	The Threatening of the Mischievous
	والمجرمين والفاسقين	The Criminals and the Impious
	العمل، العمل الطالح	Action (Work), Bad Deeds
	وعيد المفسدين	Threatening the Spoilers
	الأخلاق الحميدة	Good Morals
	الأخلاق الذميمة	The Dispraised Morals
	الوفاء بالعهد	Keeping Promise
	الإفساد، الخبث، نقض العهد	Corruption, Cunning, Break Promise
	الإنسان والعلاقات الاجتماعية	Human and The Social Relations
	صلة ذوي القربى	Kinship
	ايه ٧٢، سورة ٢، حزب ١، صفحة ٥	Verse27, Chapter2, Hizb1, Page5
	متشابه بشدة	Strongly similar (link to one verse)
	متشابه قليلاً	Slightly similar (links to five different verses)
	تفسير الميسر، تفسير الجلالين	DescByJalalayn, DescByMuyaser

Table 6.2: A Quran verse and concepts extracted from the two ontologies

their suffice meaning. For example, the Hadith Matn in Table 6.3 does not literally mean that such a person is not considered a Muslim, but that a certain action or characteristic is not the conduct of a good Muslim, as is captured in the associated section-heading.

6.5.1 Data Sources

In this section, the sources of the data used in the experiment is discussed. Ontology A and Ontology B can be accessed from their dedicated websites. The Hadith section-headings were obtained from the LK-Hadith corpus, described in Chapter 5. This corpus consists of the six canonical Hadith books where every Hadith component and pieces of meta data is annotated. Using it, the researcher can

	Arabic	English Translation
Hadith Matn	منْ غشَّنا فليسَ منَّا	One who cheats is not from among us
Section-Heading	النهي عن الغش والخداع	Prohibition of deceiving others

Table 6.3: Example of a Hadith Matn and the associated section-heading

extract specific information; for example, the Hadiths Isnads, Matns, or section-headings. This corpus was built automatically using a Hadith segmentation tool as described in Chapter 4 that was created to deconstruct every Hadith to its two main components, the Isnad and the Matn. Furthermore, the GitHub[36] repository of this corpus comes with a python script that can be used to extract any component of the Hadith and its metadata.

6.5.2 Data Analysis

Before delving into the experiment, a brief overview of the two ontologies and the Hadith section-headings is given. Table 6.4 shows the number of word tokens in each. A ‘word’ resents a token consisting of consecutive letters separated by space, like the one in Figure 6.2, which is translated as an English sentence. Moreover, the table shows the number of unrepeated words, which is called a ‘word type’. Arabic being a highly inflected language, words lexemes are used. Lexemes form the basic abstract unit of meaning and represent set of words that are related through inflection. Therefore, lexemes are extracted using the CAMeL tool (Obeid *et al.*, 2020) to flatten the different forms of a word to its basic lexeme. This will make it possible to measure the ‘fit’ as precisely as possible where a word like والإسلام ‘and Islam’, للإسلام ‘for Islam’, are collapsed into the same lexeme إسلام ‘Islam’.

6.5.3 Experiment

As discussed earlier, a data-driven approach to ontology evaluation is used. The idea is to find which ontology from the two candidates covers the Hadiths better,

[36] https://github.com/ShathaTm/LK-Hadith-Corpus

	Word Tokens	Word Types	Lexeme
Headings	39,959	6,110	2,913
Ontology A	41,782	978	746
Ontology B	103,603	2,009	1,370

Table 6.4: Vocabulary statistics for ontologies and Hadith headings

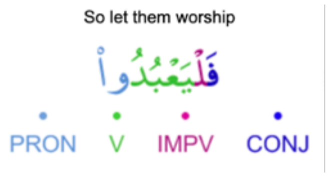


Figure 6.2: A word from Chapter 106 in the Quran

and which areas of the Hadiths are not covered. The following steps are followed to evaluate the ontologies.

1. Extract the concepts from Ontology A (Abbas, 2009) and Ontology B (Hakkoum & Raghay, 2016). Then extract the Hadith section-headings from LK Hadith corpus.

2. Remove punctuation marks and diacritics, then tokenize the text by white space. This step produces three lists of words. One list contains the words of concepts from Ontology A, the second list contains words of concepts from Ontology B, and the final list contains words in Hadith section-headings.

3. To remove stop-words, all the Matns from the LK Hadith corpus are extracted to get the most frequent terms in Hadith statements. Then the frequent terms are manually inspect to ensure they can be considered stop-words. These words are added to the list of Arabic stop-words obtained from the NLTK, resulting in 225 stop-words.

4. After removing stop-words from the three lists, the lexeme of each orthographic word is obtained . Then the overlap between the ontologies and Hadith section-headings lexeme are evaluated.

6.6 Results

To measure the overlap between the two different ontologies and Hadith section headings, the number of lexeme in the ontologies that are also found in the Hadith heading lexeme is counted. In addition, the number of lexeme that are found in both ontologies are counted.

- 61.5% of the Ontology A concepts lexemes are present in the Hadith headings lexemes.
- 56.9% of the Ontology B lexemes are found in the Hadith headings lexemes.
- 100% of the Ontology A lexemes are present in the Ontology B lexemes. This implies that ontology A is incorporated within ontology B, although it was not mentioned in their paper Hakkoum & Raghay (2016). However, the dedicated website for Ontology B states that it is an ongoing research work with constant updates.
- 26.8% of the Hadith headings lexemes are found in the Ontology B lexemes.

The overlap is demonstrated in Figure 6.3. Ontology B is 47% the size of the Hadith headings. While Ontology A is 25% the size of Hadith headings as is evident from the number of unique lexemes in each category as depicted in Table 6.4.

Since some parts of the ontologies do not have a match in the Hadith headings, the synonyms of the ontology concepts that do not have a match is extracted as proposed by the methodology (Brewster *et al.*, 2004). Arabic Wordnet (Abouenour *et al.*, 2013) is used to extract synsets of the ontology concepts lexemes that have no match in the Hadith headings lexeme list.

From 590 Quran lexemes, only 15% were found in Arabic Wordnet. Examples of the resulting synsets are shown in Table 6.5. Arabic Wordnet returns the synsets exactly as shown in the table, a pair of English and corresponding Arabic word or phrase (the author has not produced any translation). After scrutinizing the results, it is concluded that the Arabic Wordnet is more appropriate for MSA than CA. This is because important Islamic were not found like زكاة ‘Zakat’, فردوس

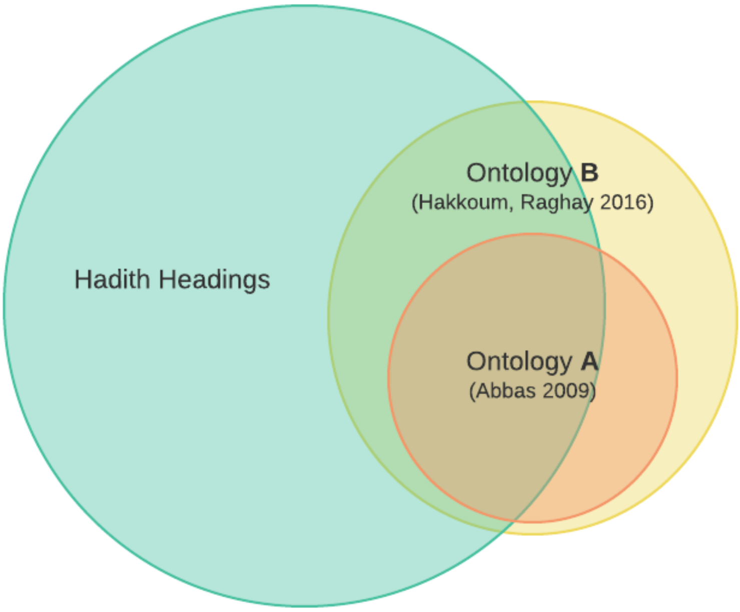


Figure 6.3: Overlap between Hadith section-headings and the Quran ontologies

‘paradise’, طهارة ‘purity’, and صوم ‘fasting’. Therefore, it has been decided to not use Arabic Wordnet.

Term	synsets	Term	synsets
سقر	Imaginary_place مكان خيالِي	حشرة	Arthropod مفْصِلِيّ الأرْجُل
تطابق	Change استبدل	تحدى	Argue تصالح
	Change إنتقل_من		Set about هاجم
	Adjust حضن_البيض		Hurt غيظ
قرار	Choice اِخْتِيار	ندم	Sadness تعاسة
	Act شاط_بشري		Sorrow حُزْن
	Due_process طُرُق_قانُونِيّة		Compunction تأنِيب_الضمِير

Table 6.5: Synsets obtained from Arabic Wordnet

6.7 Discussion

To understand the overlap illustrated in Figure 6.3, some lexeme examples from each part are enumerated and shown in Table 6.6. As one might predict, the lexemes that are found in all lists are the most important concepts, while the lexemes that are found in the ontologies are related to historical stories of prophets, messengers, and animals, with Ontology B containing more details. This observation was statistically studied by Sayoud (2012), who remarked that several animal names are not cited in the Albukhari Hadiths. For example, the name عجل 'calf', is cited ten times in the Quran and is completely absent in the Albukhari Hadiths.

Similarly, the lexemes in Hadith headings which were not covered in the ontologies are mostly about specific incidents or instructions that are only found in Hadiths. For example, the complete instructions for Islamic washing that must be preformed before prayer are obtained from the Hadith; the Quran states the abstract obligation of the act without the details of how to perform it. Table 6.6 shows these words like مضمضة 'mouth rinse'.

Overlap in Ontology A, B and Hadith headings	قيامة - يوم - غزا - شيطان - ساجد - ملأك - آدم - أكل - محرم - إسرائيل - فرعون - بحر - مريم - الله - جبرئيل - إبراهيم - يعقوب - مسجد - حج - نصراني - يهودي - موت - رجل - توحيد
Overlap in Ontology A and B	هابيل - قابيل - إلياس - عمران - حواري - نفاق - عقيدة - أخلاقي - ذميم - وعيد - فاسق - محيط ربوبية - وحداني - تحدى - جاحد - خلود - هداية - طالح - فلاح - سعادة - تشبيه - حشرة - تذكير
Ontology B only	عجل - آدم - عصي - حواء - بعوض - قدس - عدس - سينا - معتد - بابل - إسحاق - إرجاع - تابوت - جالوت - شريعة - طائر - أنعام - حصان - ولادة - زكريا - مهد - طير - قربان - غراب
Overlap in Ontology B and Hadith headings	إبليس - شجرة - نزول - شخص - ألام - قثاء - ثوم - بقل - بصل - طور - رفع - قرد - بقر - ذبح - قتيل - إسماعيل - كعبة - مكة - مروة - صفا - خنزير - رمضان - فجر - قريش
Hadith headings only	مضمضة مسح - قدم - وتر - تيمن - حان - قارئ - غشي - توضأ - كعب - نصيحة - شاة - لبن - أعرابي - نفاس - صبي - قاعد - فرك - جنابة - إبل - سمن - إمام - رأس - إفشاء

Table 6.6: Examples of lexemes extracted from each part illustrated in Figure 6.3

6.8 Conclusion

One of the main objectives of this thesis is to model the meaning of the Quran and Hadiths to discover semantic similarities within these texts. This chapter explores the possibility of using an ontology-based approach to semantic similarity detection. Since there are various Quran ontologies, they are reviewed, discussed, and evaluated using a corpus-based approach to identify which is best to serve as the foundation for an Islamic ontology that encompasses both the Quran and Hadiths. The experiment is conducted by extracting and comparing keywords from the Albukhari Hadith section-headings to the Quran ontology concepts. The study finds that Ontology B (Hakkoum & Raghay, 2016) is the most suitable candidate for an Islamic ontology. However, this ontology covers only 26.8% of Hadith topics, which calls for other means of modelling and linking the Hadiths to the Quran. Therefore, an alternative approach using DL to model the meaning of the Quran and Hadiths is explored in the following chapters.

Chapter 7

Creating the Quran-Hadith Dataset

7.1 Introduction

The previous chapter illustrates that a knowledge-based approach to finding relatedness between the Quran and Hadiths is not feasible using existing Quran ontologies. Hence, it is proposed that state-of-the-art Deep Learning (DL) models may capture the underlying meaning of these texts to find semantic similarities. However, DL models require a benchmark to guage their performance. This chapter discusses the methodology of creating a Quran-verse and Hadith-teaching related and non-related pairs dataset. Also the Qursim (Sharaf & Atwell, 2012b) dataset, which consists of Quran-verse related pairs, is extended to incorporate negative samples of non-related Quran-verse pairs.

7.2 Quran and Hadith

As stated in Chapter 2, Muslims believe the Quran is God's divine words transmitted to the Prophet Muhammad by the angel Gabriel over a period of 23 years. This holy book enjoins Muslims to follow the guidance of the Prophet Muhammad in their laws, legislation, and moral guidance. In fact, most laws and legislation are obtained from the Hadith, which is the reports of the Prophet Muhammed's statements, actions, approval, or criticism of something said or done in his presence. The importance of Hadith is due to its larger scope and incorporated details which is not present in the Quran. Consequently, many Islamic rulings (Fatwa) use Quran and Hadith together as evidence.

Although the Quran and Hadith are both Classical Arabic (CA) and cover the same domain of Islamic teachings, they are distinctive in structure, style, and orthography (Bashir *et al.*, 2022). This is aligned with an authorship attribution study using text-mining-based investigation, which shows that 62% of the Hadith words in Sahih Albukhari (the most famous Hadith book) do not occur in the Quran. Additionally, 83% of the Quran's words do not occur in the Sahih Albukhari Hadiths (Sayoud, 2012).

To further illustrate the differences, I compiled a corpus of Hadith Matns from the LK Hadith corpus[37] described in Chapter 5, which consists of the six canonical Hadith books, and compared it to the Quran corpus obtained from Tanzil[38]. Table 7.1 shows the number of tokens, which are words separated by spaces. The word type in each corpus is also counted. This refers to unique tokens without repetition. Other features are shown, including the number of unique words not present in the other corpus, and the number of words that are found in both corpora. Examples of these unique words are given in Table 7.2. It is evident that Hadith words are more specific and concern daily things while Quran words are more generic. This is also shown in the word clouds in Figure 7.1. The most common words in both is الله and قال as the size of the words indicates. However, the Quran and Hadith word clouds contain different smaller sized words which are derivative of the same word such as قالت ، وقال ، قال. Hence, lexemes are considered next to compare Quran and Hadith.

As noted in Section 6.5.2, a lexeme is the core representation of a word, stripped of any inflectional variations. It is the basic unit of meaning and is used

[37] https://github.com/ShathaTm/LKHadithCorpus

[38] https://tanzil.net/docs/tanzil_projec

	Quran	**Hadith**
Tokens	78,245	1,362,050
Word type	14,870	59,944
Unique words	6,484	51,558
Shared words	8,386	8,386
Lexemes	76,602	1,338,910
Lexeme type	4,222	9,597
Unique lexeme	497	5,872
Shared lexeme	3,725	3,725

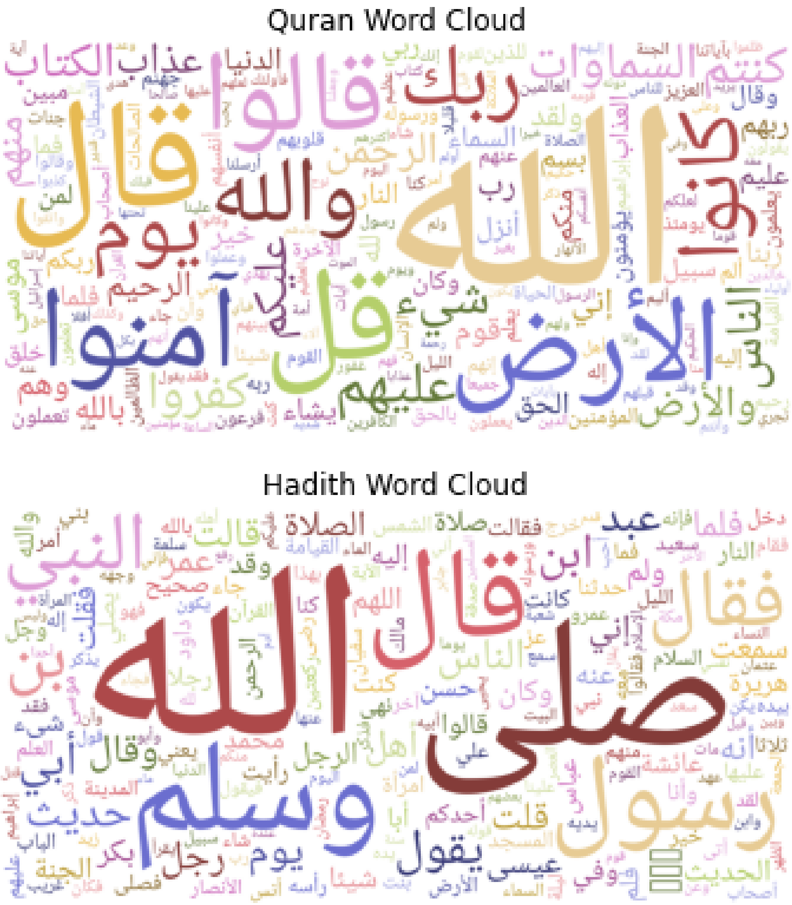
Table 7.1: Quran and Hadith statistics

Quran				**Hadith**			
داوود	الظلمات	المفسدين	السماوات	الأنصار	بالمدينة	القبر	ركعة
الظلمات	استكبروا	أهواءهم	الفلك	المرأة	الرجل	سجد	غسل
للمكذبين	الصابرين	آمنتم	بأسنا	الظهر	نهى	الدجال	الوجه
استكبروا	المفسدين	جاءتهم	تشكرون	صليت	الركوع	ثلاثا	فتوضأ
الصابرين	الفلك	يفترون	نجزي	الإمام	ركعتين	خيبر	العصر

Table 7.2: Quran and Hadith unique words

to identify the common form of a set of related words that have undergone inflection. The lexeme statistics of the Quran and Hadith are shown in the last four rows in Table 7.1. From the data, it is evident that the Hadith corpus is larger and more diverse, containing words and lexemes that do not exist in the Quran, as shown in Figure 7.2. Additionally, the lexemes found in the Hadith are more closely related to contemporary obligations and stories, as demonstrated in the examples in Table 7.3. This is further supported by the lexeme cloud comparison in Figure 7.3. It is worth noting that stop-words were removed and CAMeL Tool (Obeid *et al.*, 2020) was used to extract the lexemes.

Based on this information, finding semantic similarity by simple word match may not be feasible. Therefore, an approach that considers meaning is more likely to succeed. However, the Quran and Hadith have unique characteristics



Quran Word Cloud

Hadith Word Cloud

Figure 7.1: Quran and Hadith word clouds

Quran				Hadith			
استضعف	طغيان	رغد	غشاوة	وضوء	غسل	أخبر	ركعة
إنباء	طوعا	فائز	صاغر	منبر	تزوج	دجال	مسكر
مثوى	منتصر	أجرم	مدحور	جنازة	سجدة	مضمض	بيعة
دمر	استنكف	برزخ	خوض	أفطر	أمير	جنابة	مسألة
مثوبة	حطام	قسطاس	عنكبوت	نجاشي	زمزم	فطرة	بكاء

Table 7.3: Unique lexemes in Quran and Hadiths

that makes modelling their underlying meaning a grand challenge for AI tools. The following text will explore these features in greater detail.

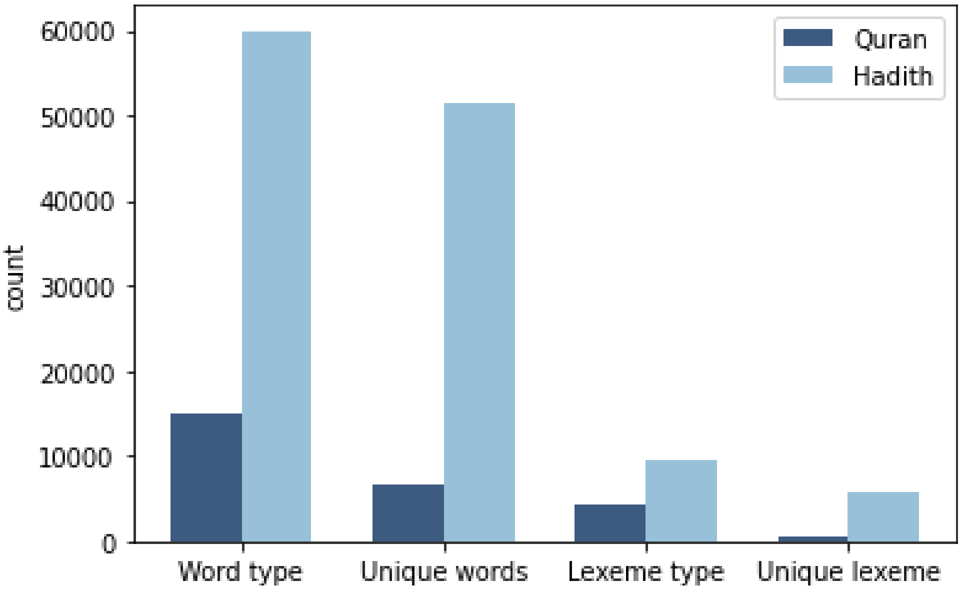
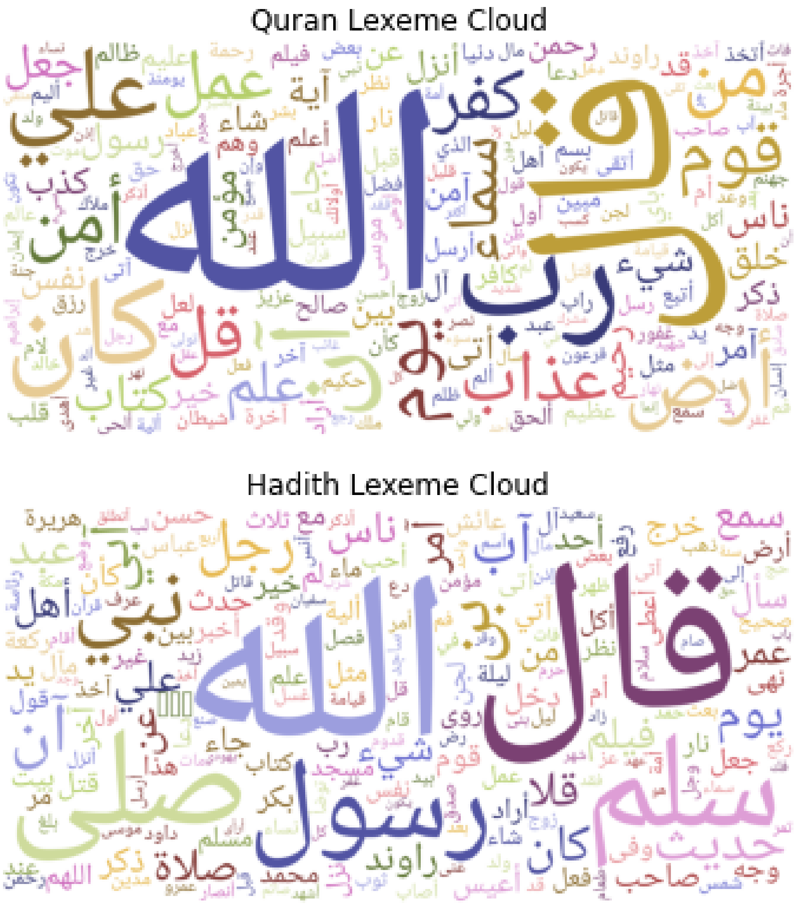


Figure 7.2: The number of words and lexemes in Quran and Hadith

7.2.1 The Quran

Modelling the meaning of Quran using AI methods is a grand challenge. This is because the Quranic discourse is believed to be inimitable. Muslim scholars claimed that the Quran cannot be reproduced in other languages which led to the delay in translating it for centuries. The 'meaning' of the Quran was first translated into English by Alexander Ross in 1649, but the first English Muslim-translated version was by Dr. Mohammad Abdul Hakim Khan in 1904 (Faqeer, 2017). Now even more translations are produced in English, each typically claiming to do a better job at revealing the true message of the Quran. So what features contribute to the difficulty of conveying the meaning of this text?

The Quran has striking syntactic, semantic, phonetic, and rhetorical features that makes it challenging for the machine to model. This is because Quranic discourse is different to that of other texts, such that words can have layers of embedded meanings that require human beings to consult major Quran exegeses and dictionaries in order to derive and provide the accurate underlying meaning of a given Quranic expression or even a preposition (Abdul-Raof, 2013). The


Quran Lexeme Cloud

Hadith Lexeme Cloud

Figure 7.3: Quran and Hadith lexeme clouds

following lines discusses some of these features.

Orthography

There are orthographical challenges specific to the Quran. For example, it has its spelling conventions which are different to all varianties of Arabic (MSA, non-Quranic CA, DA) as shown in the example below. The Quran word آلْحَيَوٰةُ is written in Arabic as الْحَيَاةُ. Another challenge is the different meaning of Quranic words from Arabic in general. For example, the word آلْحَيَوَانُ in Arabic means ‘the animal’ while in this verse it means ‘(is) the life’.

> وَمَا هَذهِ ٱلْحَيَوٰةُ ٱلدُّنْيَآ إِلَّا لَهْوٌ وَلَعِبٌ وَإِنَّ ٱلدَّارَ ٱلْءَاخِرَةَ لَهِىَ ٱلْحَيَوَانُ
> لَوْ كَانُوا يَعْلَمُونَ.
>
> And this worldly life is not but diversion and amusement. And indeed, the home of the Hereafter - that is the [eternal] life, if only they knew. [26:64]

Figurative Language

The Quran includes many figurative words such as **فتيلا** ، **نقيرا**, and **قطميرا** which are literally parts of the date, but in the Quranic context refer to size. The translation of Pickthall is used since it maintained these figurative words.

> أَلَمْ تَرَ إِلَى الَذينَ يُزَكُوْنَ أَنْفُسَهُمْ، بَلِ اللَه يُزَكِي مَنْ يَشاءُ وَلا يُظْلَمُونَ **فَتِيلًا** .
>
> Hast thou not seen those who praise themselves for purity? Nay, Allah purifieth whom He will, and they will not be wronged even **the hair upon a date stone**.[4:49]
>
> وَمَن يَعْمَلْ مِنَ الصَالْحَاتِ مِن ذَكَرٍ أَوْ أُنثَىٰ وَهُوَ مُؤْمِنٌ فَأُولَئكَ يَدْخُلُونَ الْجَنَة وَلَا
> يُظْلَمُــونَ **نَقِيرًا** .
>
> And whoso doeth good works, whether of male or female, and he (or she) is a believer, such will enter paradise and they will not be wronged the **dint in a date stone**. [4:124]
>
> يُولِجُ اللَّيْلَ فِي النَّهَارِ وَيُولِجُ النَّهَارَ فِي اللَّيْلِ وَسَخَّرَ الشَّمْسَ وَالْقَمَرَ كُلٌّ يَجْرِي
> لِأَجَلٍ مُّسَمًّى ذَلكُمُ اللَّهُ رَبُّكُمْ لَهُ الْمُلْكُ وَالَّذِينَ تَدْعُونَ مِن دُونِهِ مَا يَمْلِكُونَ
> مِن **قِطْمِيرٍ** .
>
> He maketh the night to pass into the day and He maketh the day to pass into the night. He hath subdued the sun and moon to service. Each runneth unto an appointed term. Such is Allah, your lord; His is the

> Sovereignty; and those unto whom ye pray instead of Him own not so much as **the white spot on a date stone**. [35:13]

More examples of figurative usage in the Quran is illustrated below which contains the word أُمُّهُ. This i.e. the lexeme أم does not mean 'mother' as it does literally, but rather figuratively denotes something that embraces or enfolds. The second verse has the use of الْمُعْصِرَاتِ that literally means 'presses' to refer to clouds.

> فَأُمُّهُ هَاوِيَةٌ
>
> Shall be engulfed by an abyss.[101:9]
>
> وَأَنزَلْنَا مِنَ الْمُعْصِرَاتِ مَاءً ثَجَّاجًا
>
> And from the wind-driven clouds We send down waters pouring in abundance. [87:14]

Polysemy

Polysemy is the phenomenon where words have different but related meanings. The Quran incorporates many words that have different meanings, for example, in the first verse below the word ظلم does not mean 'wrong doing' as it does in other contexts, but rather, it refers to 'worshipping others beside God' as is evident in the second verse. Another example is illustrated in the third verse that contains the word الجمل. This generally means a 'camel', but some Quran exegesis state that it may also refer to twisted rope [39], which is not a prevalent use of the word. The last example have the same word in different contexts denoting different meanings. The first ساعة means 'Day of Judgment' and the second means 'hour'.

[39] Tafsi Altabray https://quran.ksu.edu.sa/tafseer/tabary/sura7-aya40.html

> الَّذِينَ آمَنُوا وَلَمْ يَلْبِسُوا إِيمَانَهُمْ بِظُلْمٍ أُولَئِكَ لَهُمُ الْأَمْنُ وَهُمْ مُهْتَدُونَ.
>
> Those who believe and obscure not their belief by wrong doing, theirs is safety; and they are rightly guided. [6:82]
>
> وَإِذْ قَالَ لُقْمَانُ لِابْنِهِ وَهُوَ يَعِظُهُ يَا بُنَيَّ لَا تُشْرِكْ بِاللَّهِ إِنَّ الشِّرْكَ لَظُلْمٌ عَظِيمٌ.
>
> And (remember) when Luqman said unto his son, when he was exhorting him: O my dear son! Ascribe no partners unto Allah. Lo! to ascribe partners (unto Him) is a tremendous wrong. [13:31]
>
> إِنَّ الَّذِينَ كَذَّبُوا بِآيَاتِنَا وَاسْتَكْبَرُوا عَنْهَا لَا تُفَتَّحُ لَهُمْ أَبْوَابُ السَّمَاءِ وَلَا يَدْخُلُونَ
> الْجَنَّةَ حَتَّىٰ يَلِجَ الْجَمَلُ فِي سَمِّ الْخِيَاطِ وَكَذَلكَ نَجْزِي الْمُجْرِمِينَ.
>
> Lo! they who deny Our revelations and scorn them, for them the gates of Heaven will not be opened nor will they enter the Garden until the camel goeth through the needle's eye. Thus do We requite the guilty. [7:40]

> وَيَوْمَ تَقُومُ السَّاعَةُ يُقْسِمُ الْمُجْرِمُونَ مَا لَبِثُوا غَيْرَ سَاعَةٍ كَذَلكَ كَانُوا يُؤْفَكُونَ.
>
> Upon the Day when the Hour has come, the harmdoers will swear that they had stayed no more than an hour. As such they are deceived.[30:55]

Coherence

Another challenge is the Qur'anic structural coherence and propositional cohesion where a verse meaning is embedded in the context where it appears in the Quran. For example, verse [2:186] is placed between [2:185] and [2:187] which both refer to Ramadan, in order to highlight the significance of supplication during the holy month of Ramadan.

شَهْرُ رَمَضَانَ الَّذِي أُنْزِلَ فِيهِ الْقُرْآنُ هُدًى لِلنَّاسِ وَبَيِّنَاتٍ مِنَ الْهُدَىٰ وَالْفُرْقَانِ ۚ فَمَنْ
شَهِدَ مِنْكُمُ الشَّهْرَ فَلْيَصُمْهُ ۖ وَمَنْ كَانَ مَرِيضًا أَوْ عَلَىٰ سَفَرٍ فَعِدَّةٌ مِنْ أَيَّامٍ أُخَرَ ۗ
يُرِيدُ اللَّهُ بِكُمُ الْيُسْرَ وَلَا يُرِيدُ بِكُمُ الْعُسْرَ وَلِتُكْمِلُوا الْعِدَّةَ وَلِتُكَبِّرُوا اللَّهَ عَلَىٰ مَا
هَدَاكُمْ وَلَعَلَّكُمْ تَشْكُرُونَ.

Ramadhan is the (month) in which was sent down the Qur'an, as a guide to mankind, also clear (Signs) for guidance and judgment (Between right and wrong). So every one of you who is present (at his home) during that month should spend it in fasting, but if any one is ill, or on a journey, the prescribed period (Should be made up) by days later. Allah intends every facility for you; He does not want to put to difficulties. (He wants you) to complete the prescribed period, and to glorify Him in that He has guided you; and perchance ye shall be grateful.[2:185]

وَإِذَا سَأَلَكَ عِبَادِي عَنِّي فَإِنِّي قَرِيبٌ ۖ أُجِيبُ دَعْوَةَ الدَّاعِ إِذَا دَعَانِ ۖ فَلْيَسْتَجِيبُوا لِي
وَلْيُؤْمِنُوا بِي لَعَلَّهُمْ يَرْشُدُونَ.

When My servants ask thee concerning Me, I am indeed close (to them): I listen to the prayer of every suppliant when he calleth on Me: Let them also, with a will, Listen to My call, and believe in Me: That they may walk in the right way. [2:186]

أُحِلَّ لَكُمْ لَيْلَةَ الصِّيَامِ الرَّفَثُ إِلَىٰ نِسَائِكُمْ ۚ هُنَّ لِبَاسٌ لَكُمْ وَأَنْتُمْ لِبَاسٌ لَهُنَّ ۗ عَلِمَ
اللَّهُ أَنَّكُمْ كُنْتُمْ تَخْتَانُونَ أَنْفُسَكُمْ فَتَابَ عَلَيْكُمْ وَعَفَا عَنْكُمْ ۖ فَالْآنَ بَاشِرُوهُنَّ
وَابْتَغُوا مَا كَتَبَ اللَّهُ لَكُمْ ۚ وَكُلُوا وَاشْرَبُوا حَتَّىٰ يَتَبَيَّنَ لَكُمُ الْخَيْطُ الْأَبْيَضُ مِنَ
الْخَيْطِ الْأَسْوَدِ مِنَ الْفَجْرِ ۖ ثُمَّ أَتِمُّوا الصِّيَامَ إِلَى اللَّيْلِ ۚ وَلَا تُبَاشِرُوهُنَّ وَأَنْتُمْ
عَاكِفُونَ فِي الْمَسَاجِدِ ۗ تِلْكَ حُدُودُ اللَّهِ فَلَا تَقْرَبُوهَا ۗ كَذَٰلِكَ يُبَيِّنُ اللَّهُ آيَاتِهِ لِلنَّاسِ
لَعَلَّهُمْ يَتَّقُونَ.

Permitted to you, on the night of the fasts, is the approach to your wives. They are your garments and ye are their garments. Allah knoweth what

> ye used to do secretly among yourselves; but He turned to you and forgave you; so now associate with them, and seek what Allah Hath ordained for you, and eat and drink, until the white thread of dawn appear to you distinct from its black thread; then complete your fast Till the night appears; but do not associate with your wives while ye are in retreat in the mosques. Those are Limits (set by) Allah. Approach not nigh thereto. Thus doth Allah make clear His Signs to men: that they may learn self-restraint. [2:187]

Agglutination

One thing did make the Quranic texts complex is the Arabic language's inherent challenges. One of the main challenges is its agglutinative property where a sentence can be expressed in one word by adding affixes and clitics that represent various parts of speech (Habash, 2010). The examples below show such Quranic words and their translations.

> So will suffice you against them - فَسَيَكْفِيكَهُمُ
>
> Shall we compel you (to accept) it) - أَنُلْزِمُكُمُوهَا
>
> and We gave it to you to drink - فَأَسْقَيْنَاكُمُوهُ

7.2.2 The Hadith

Although the Hadith is different from the Quran, it possesses linguistic features that makes it challenging as well to model its underlying meaning. The following lines discuss some of these features. It is worth noting that since this study focuses on the Hadith teachings, the forthcoming examples show the Hadith Matn only.

Hyperbole

One of the most common features of the prophetic style in Hadiths is the frequency of hyperbole Brown (2017). The Hadith examples below are dramatic statements, but the way in which Muslim scholars have understood them is different from their literal meaning. The first example below simply means that cheating is prohibited, but it does not render the offender a non-Muslim.

> مَن غَشَنَا فليسَ مِنَاْ.
>
> One who cheats is not from among us.
>
> لاَ يَدْخُلُ الْجَنَّةَ أَحَدٌ فِي قَلْبِهِ مِثْقَالُ حَبَّةِ خَرْدَلٍ مِنْ كِبْرِيَاءَ.
>
> No one will enter heaven who has even a grain's weight of pride in his heart.
>
> سِبَابُ الْمُسْلِمِ فُسوقٌ وَقِتَالُهُ كُفْرٌ.
>
> Cursing a Muslim is iniquity and fighting one is unbelief (kufr).

Simile

Another feature found in Hadith is the use of simile, which the Prophet often uses to give vivid descriptions.

> المؤمن لِلْمؤْمن كالبُنْيان يَشُدُّ بَعْضُه بَعْضا، ثُمَّ شَبّك بين أَصابعه.
>
> "The relationship of the believer with another believer is like (the bricks of) a building, each strengthens the other." He (PBUH) illustrated this by interlacing the fingers of both his hands.
>
> مَثَلُ المُؤْمِنِينَ في تَوَادِّهِمْ وتَرَاحُمِهِمْ وتَعَاطُفِهِمْ، مَثَلُ الجَسَدِ إذا اشْتَكَى مِنْهُ عُضْوٌ تَدَاعَى له سَائِرُ الجَسَدِ بالسَّهَرِ والحُمَّى.

> Messenger of Allah (PBUH) said, "The believers in their mutual kindness, compassion and sympathy are just like one body. When one of the limbs suffers, the whole body responds to it with wakefulness and fever".

Metaphor

The Hadith includes metaphor in which one word or phrase is used in place of another action to suggest a likeness or analogy between them.

> رُوَيْدَكَ يَا أَنْجَشَةُ، لاَ تَكْسِرِ الْقَوَارِيرَ . قَالَ قَتَادَةُ يَعْنِي ضَعَفَةَ النِّسَاءِ.
>
> The Prophet (PBUH) said to him, "(Drive) slowly, O Anjasha! Do not break the glass vessels!" And Qatada said, "(By vessels) he meant the weak women."
>
> اليَدُ العُلْيَا خَيْرٌ مِنَ اليَدِ السُّفْلَى، وابْدَأْ بمَن تَعُولُ، وخَيْرُ الصَّدَقَةِ عن ظَهْرِ غِنًى،
> ومَن يَسْتَعْفِفْ يُعِفَّهُ اللَّهُ، ومَن يَسْتَغْنِ يُغْنِهِ اللَّهُ.
> "The upper hand is better than the lower one (i.e., the spending hand is better than the receiving hand); and begin (charity) with those who are under your care; and the best charity is that which given out of surplus; and he who asks (Allah) to help him abstain from the unlawful and the forbidden, Allah will fulfill his wish; and he who seeks self-sufficiency will be made self-sufficient by Allah"
>
> ورَجُلٌ تَصَدَّقَ بصَدَقَةٍ فأَخْفَاهَا حتَّى لا تَعْلَمَ شِمَالُهُ ما تُنْفِقُ يَمِينُهُ
>
> and a man who gives charity so secretly that his left hand does not know what his right hand has given.

Despite the differences in these texts and the challenging linguistic features, I hypothesize that finding semantic similarity between the Quran and Hadith might be possible using State Of The Art (SOTA) DL models. However, to gauge the performance of such models, it is necessary to create a dataset (benchmark) to evaluate the performance of different models for a specific NLP task. The following sections enumerate the available datasets that can be used for the task of semantic similarity detection between the Quran and Hadith.

7.3 Existing Arabic Datasets

Many studies have been conducted on Semantic Textual Similarity (STS) and relatedness. However, limited research focused on CA, which is considered a low resource language in terms of available datasets. This is a major obstacle since gold standard datasets are essential to gauge the performance of any given system on a downstream task.

Recently, many Arabic resources have been published [40] for different NLP downstream tasks including question answering (Mozannar *et al.*, 2019), offensive language detection (Mubarak *et al.*, 2020) and datasets for sentiment analysis, machine translation, and topic classification (Alyafeai *et al.*, 2021). However, only a few are dedicated to the semantic similarity task, and only one is for CA. Table 7.4 shows the existing Arabic datasets for the semantic similarity, which are discussed in the following lines.

Dataset	**Arabic Variant**	**num of pairs**	**Year**
Qursim	CA	7,679	2012
Q2Q	MSA	15,712	2019
STS	MSA	1104	2017

Table 7.4: The available Arabic semantic similarity and relatedness dataset

Qursim (Sharaf & Atwell, 2012b) is a semantic relatedness dataset consisting of 7,679 related Quran-verse pairs. It is extracted from the well-known Quran commentary of Ibn Kathir, an Islamic scholar who died in 1373. His methodology is clearly stated in the introduction of his book, where each verse is discussed and explained by referring to other verses in the Quran that contain more details or explain other aspects of the same topic. This is the only available CA dataset I

[40] https://arbml.github.io/masader/

am aware of, and is discussed further in Section 7.4.2.

The second dataset is dedicated to a Semantic Question Similarity shared task that was conducted during the NLP Solutions for Under Resourced Languages (NSURL) workshop (Seelawi *et al.*, 2019). The dataset consists of three fields, question1, question2, and label. The label is 1 if both questions have a similar answer, or 0 otherwise. The third available dataset was used at the SemEval-2017 Semantic Textual Similarity (STS) shared task (Cer *et al.*, 2017). It consists of Arabic pairs machine translated from English then checked by human translators. The label of each pair can be in a range from 0 to 5 with five being the most similar.

The last two datasets are not comparable to the introduced dataset of related Quran and Hadith texts. This is because finding the semantic relatedness between Quran and Hadith is different from standard STS tasks. The aim is to detect similarity in the underlying meaning and message conveyed by these scared texts, which is more complex than simple similarity measurement. The next section discusses the automatic benchmark construction methodology having observed the lack of publicly available CA resources.

7.4 Datasets

This section explains the methodology of creating the Quran-Hadith dataset, which is the main contribution of this chapter. Then a description is given of the process for extending the Qursim (Sharaf & Atwell, 2012b) to produce the datasets used in the training/fine-tuning and validation phase across the various models. These datasets are available in my GitHub[41]. They are intended to be beneficial for the wider NLP community, particularly those working on under-resourced languages since there is an increased interest in applying NLP tools on such texts like religious scripture, but challenges are still unresolved (Bounhas, 2019; Bashir *et al.*, 2022).

[41] https://github.com/ShathaTm/Quran_Hadith_Datasets

7.4.1 Testing Dataset: Quran-Hadith (QH) Pairs

To build this dataset, two methods were designed to create the positive and negative samples. The main tasks in these methods can be summarized in the following five steps:

1. Select the sources where a reputable Islamic scholar explicitly stated the relatedness.
2. Collect the text, then extract the Quran and the Hadith as a pair.
3. Ensure the whole Quran verse or Hadith teaching is obtained by cross-referencing the original corpora.
4. Create samples of non-related pairs.
5. Process the dataset to remove punctuation marks and diacritics.

The following paragraphs discuss these steps.

Step 1: Select sources of the datasets

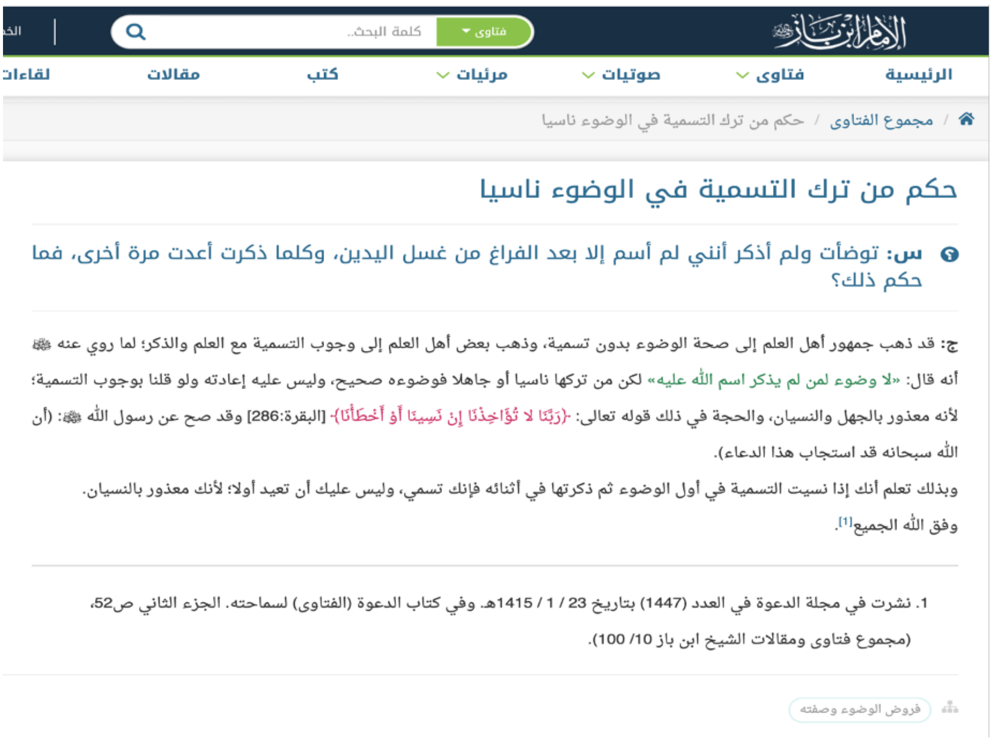
The traditional approach to building a dataset using crowd-sourcing is not possible for religious texts since it requires domain experts. Hence, two reliable sources involving reputable Islamic scholars are identified where they explicitly mentioned the relatedness between the pairs. The first is Sahih Albukhari, which incorporates a collection of Hadith-teachings organized into topics by the well-known scholar Muhammed Albukhari who died in 870. The book's structure is exploited to create the dataset. In many cases, section headings consist of a Quran-verse, which the scholar used to indicate that it is related to the Hadiths within the section. Table 7.5 shows an instance of such a case. The LK-Hadith-Corpus [42] is used as it provides a well-structured version of the canonical Hadith collections. These section-headings and their associated Hadith Matn are extracted as a related Quran-Hadith pair.

[42] https://github.com/ShathaTm/LK-Hadith-Corpus

Chapter	Sales and Trade كتاب البيوع
Section	O you who have believed, do not consume usury, doubled and multiplied, but fear Allah that you may be successful. [3:130] يَا أَيُها الَذينَ آمَنُوا لاَ تَأْكُلُوا الرِبا أَضْعَافًا مُضَاعَفَةً وَاتَقوا اللَه لَعَلَكمْ تُفْلِحُونَ.
Hadith	Adam told us, Ibn Abi Dhib told us, Saeed Al-Maqbari told us, Abu Hurairah said the Prophet (PBUH) said "Certainly a time will come when people will not bother to know from where they earned the money, by lawful means or unlawful means." . حَدَّثَنَا آدَمُ، حَدَّثَنَا ابْنُ أَبِي ذِئْبٍ، حَدَّثَنَا سَعِيدٌ الْمَقْبُرِيُّ، عَنْ أَبِي هُرَيْرَةَ، عَنِ النَّبِيِّ صلى الله عليه وسلم قَالَ: لَيَأْتِيَنَّ عَلَى النَّاسِ زَمَانٌ لاَ يُبَالِي الْمَرْءُ بِمَا أَخَذَ الْمَالَ، أَمِنْ حَلاَلٍ أَمْ مِنْ حَرَامٍ.

Table 7.5: An example from Sahih Albukhari

The second source is a website dedicated to Abdul-Aziz ibn Baz (died in 1999) who was a reputed Islamic scholar who answered religious questions on mainstream media which were later collected and archived on a website [43]. Most of these archived Fatwas (a ruling on an Islamic law given by a recognized authority) contains an answer to a specific question supported by a Quran-verse and a Hadith-teaching as shown in Figure 7.4. However, some Fatwas contain several Quran verses and Hadith teachings to answer complex questions consisting of various topics. Hence, the relatedness is not clear if taken out of context. Therefore, only Fatwas which contain one Quran-verse and one Hadith-teaching are collected to ensure they are related to a distinct topic. The methodology is depicted in Figures 7.5



مجموع الفتاوى / حكم من ترك التسمية في الوضوء ناسيا

حكم من ترك التسمية في الوضوء ناسيا

س: توضأت ولم أذكر أنني لم أسم إلا بعد الفراغ من غسل اليدين، وكلما ذكرت أعدت مرة أخرى، فما حكم ذلك؟

ج: قد ذهب جمهور أهل العلم إلى صحة الوضوء بدون تسمية، وذهب بعض أهل العلم إلى وجوب التسمية مع العلم والذكر؛ لما روي عنه ﷺ أنه قال: «لا وضوء لمن لم يذكر اسم الله عليه» لكن من تركها ناسيا أو جاهلا فوضوءه صحيح، وليس عليه إعادته ولو قلنا بوجوب التسمية؛ لأنه معذور بالجهل والنسيان، والحجة في ذلك قوله تعالى: ﴿رَبَّنَا لا تُؤَاخِذْنَا إِنْ نَسِينَا أَوْ أَخْطَأْنَا﴾ [البقرة:286] وقد صح عن رسول الله ﷺ: (أن الله سبحانه قد استجاب هذا الدعاء).

وبذلك تعلم أنك إذا نسيت التسمية في أول الوضوء ثم ذكرتها في أثنائه فإنك تسمي، وليس عليك أن تعيد أولا؛ لأنك معذور بالنسيان.

وفق الله الجميع[1].

1. نشرت في مجلة الدعوة في العدد (1447) بتاريخ 23 / 1 / 1415هـ. وفي كتاب الدعوة (الفتاوى) لسماحته. الجزء الثاني ص52، (مجموع فتاوى ومقالات الشيخ ابن باز 10/ 100).

فروض الوضوء وصفته

Figure 7.4: Fatwa page where the Quran, Hadith, and the tag is shown in red, green, and grey accordingly

Step 2: Extract the Quran and Hadith related pairs

[43] https://binbaz.org.sa/fatwas

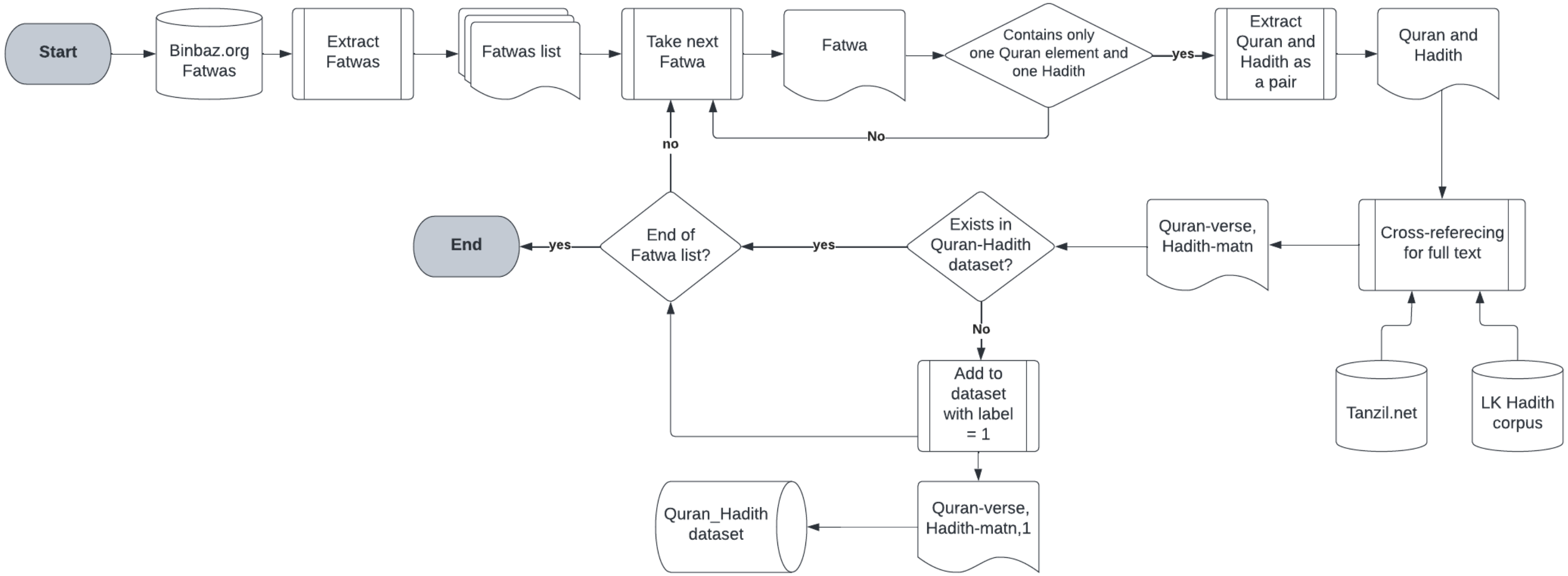

Start
Binbaz.org Fatwas
Extract Fatwas
Fatwas list
Take next Fatwa
Fatwa
Contains only one Quran element and one Hadith
yes
Extract Quran and Hadith as a pair
Quran and Hadith
No
no
Cross-referecing for full text
Quran-verse, Hadith-matn
Exists in Quran-Hadith dataset?
yes
End of Fatwa list?
yes
End
No
Add to dataset with label = 1
Tanzil.net
LK Hadith corpus
Quran-verse, Hadith-matn,1
Quran_Hadith dataset

Figure 7.5: Algorithm for creating positive QH samples

The Quran-verse and the Hadith-teaching are extracted from the data collected from Albukhari and Binbaz to form the related pairs.

Step 3: Create Quran-Hadith Non-related pairs

In the Binbaz website, each Fatwa is tagged with one or more topics. This feature is exploited to create the negative samples. A Quran-verse is extracted from a random Fatwa, and a Hadith-teaching is extracted from another Fatwa, given they contain different topic tags. This process is depicted in Figure 7.7.

Step 4: Dataset cross-referencing and filtration

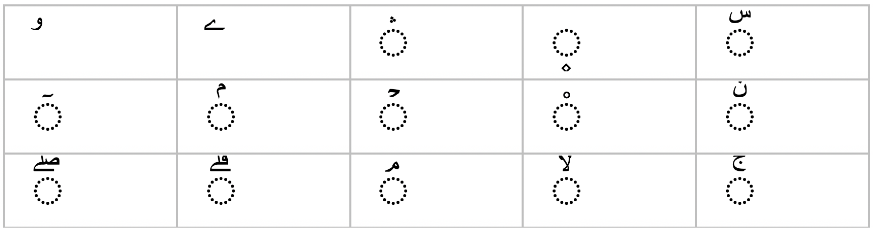
The collected pairs were further processed to ensure the full text is included, because the scholar may mention part of the Quran-verse or the Hadith-teaching. So, the extracted text is cross-referenced to find its complete instance on Tanzil for the Quran, and the LK Hadith Corpus[44] for the Hadith Matn. This was done by first checking that the number of words in the extracted verse is at least three. Then the Levenshtein distance is calculated between the verse and every verse in the Quran (Tanzil corpus) to get the best match where the lower the number the more similar the two texts are. Specifically, the Fuzzywuzzy Package[45] and the partial-ratio command are used to conduct the substring matching. The same process is applied to the extracted Hadith-teachings to find their match in the LK Hadith corpus and get the full Matn. Finally, the duplicates are removed to form a balanced dataset of 310 related and not-related Quran-Hadith pairs.

Step 5: Data Preprocessing

The Quran and Hadith contain diacritics which are important for readability. Furthermore, the Quran contains other special symbols and signs that indicate how the verse should be read. For example, the sign on the bottom right of Figure 7.6 is a small ج that is found on the top of words indicating that a reciter of the Quran is allowed to pause at that instance. However, since DL models were pre-trained on Arabic corpora without these diacritics and symbols, they were removed.

[44] Created in Chapter 5

[45] https://pypi.org/project/fuzzywuzzy/



و	ے	◌ۨ	◌۪	◌ۜ
◌ٓ	◌ۢ	◌ۡ	◌۟	◌ۨ
◌ۖ	◌ۗ	◌ۘ	◌ۙ	◌ۚ

Figure 7.6: Some of the Quran symbols

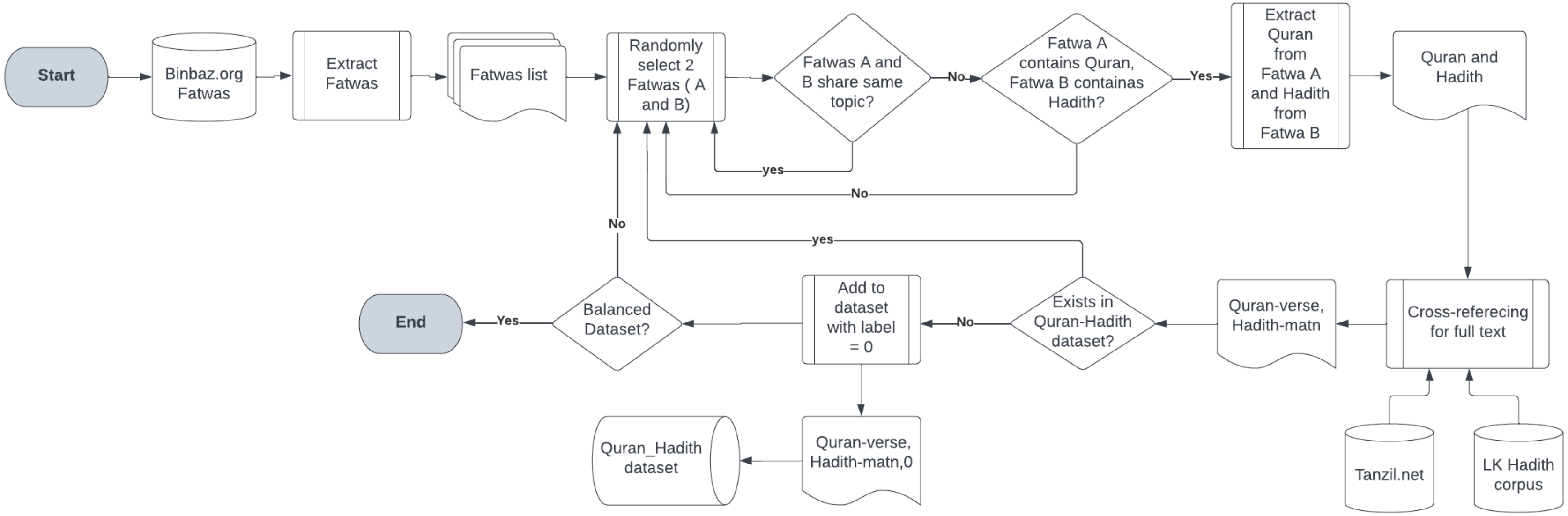

Start
Binbaz.org Fatwas
Extract Fatwas
Fatwas list
Randomly select 2 Fatwas (A and B)
Fatwas A and B share same topic?
No
yes
Fatwa A contains Quran, Fatwa B containas Hadith?
Yes
No
Extract Quran from Fatwa A and Hadith from Fatwa B
Quran and Hadith
Cross-referecing for full text
Tanzil.net
LK Hadith corpus
Quran-verse, Hadith-matn
Exists in Quran-Hadith dataset?
yes
No
Add to dataset with label = 0
Quran-verse, Hadith-matn,0
Quran_Hadith dataset
Balanced Dataset?
No
Yes
End

Figure 7.7: Algorithm for creating negative QH samples

7.4.2 Training Dataset: Quran-Quran (QQ) Pairs

To create the training dataset, part of the Qursim dataset (Sharaf & Atwell, 2012b) is used. This is because the authors of Qursim analysed the pairs manually to ensure their relatedness is clear regardless of Ibn Kathir's comments. They found that not all pairs showed clear relatedness out of context. Therefore, each pair is classified into one of three categories: strong relation, weak relation, or no-obvious relation. Since this was done by one annotator, only the pairs with strong relations are used. To further process the dataset, the methodology by Alsaleh *et al.* (2021) is used to remove duplicates.

QQ Non-related Pairs

To create the QQ negative samples of non-related pairs, a different approach to that of Alsaleh *et al.* (2021) is used. Instead of randomly extracting two Quran verses and assuming they are not related, the Quran ontology of Hakkoum & Raghay (2016) is used to extract pairs that do not share the same ontological concepts. This ontology is used because it is the most comprehensive Quran ontology (Altammami *et al.*, 2021).

Figure 7.8 shows the algorithm for creating the negative samples. First, two Quran verses are selected from Tanzil dataset (x, y). Then the associated Quran ontology concepts (Cx, Cy) are extracted from the Quran ontology. The extracted concepts are tokenized into words $c_1, c_2, ...c_n \in Cx$ and $c_1, c_2, ...c_n \in Cy$. After that a comparison is conducted to ensure there is no match in verse1 concepts and verse2 concepts$(Cx \cap Cy = \emptyset)$. Otherwise, these pairs are discarded and the system restarts to extract new pairs of Quran verses until it finds a pair with no intersection of concepts. Then the algorithm ensures the pair in both orders is not already in the negative sample list $(x, y) \notin (X, Y)$ and $(y, x) \notin (X, Y)$. Finally it is added to the list of negative samples $(x, y) \in (X, Y)$ and the process is repeated until the number of collected negative samples is 2600. This formed the balanced dataset of 5,096 related and non-related QQ pairs.

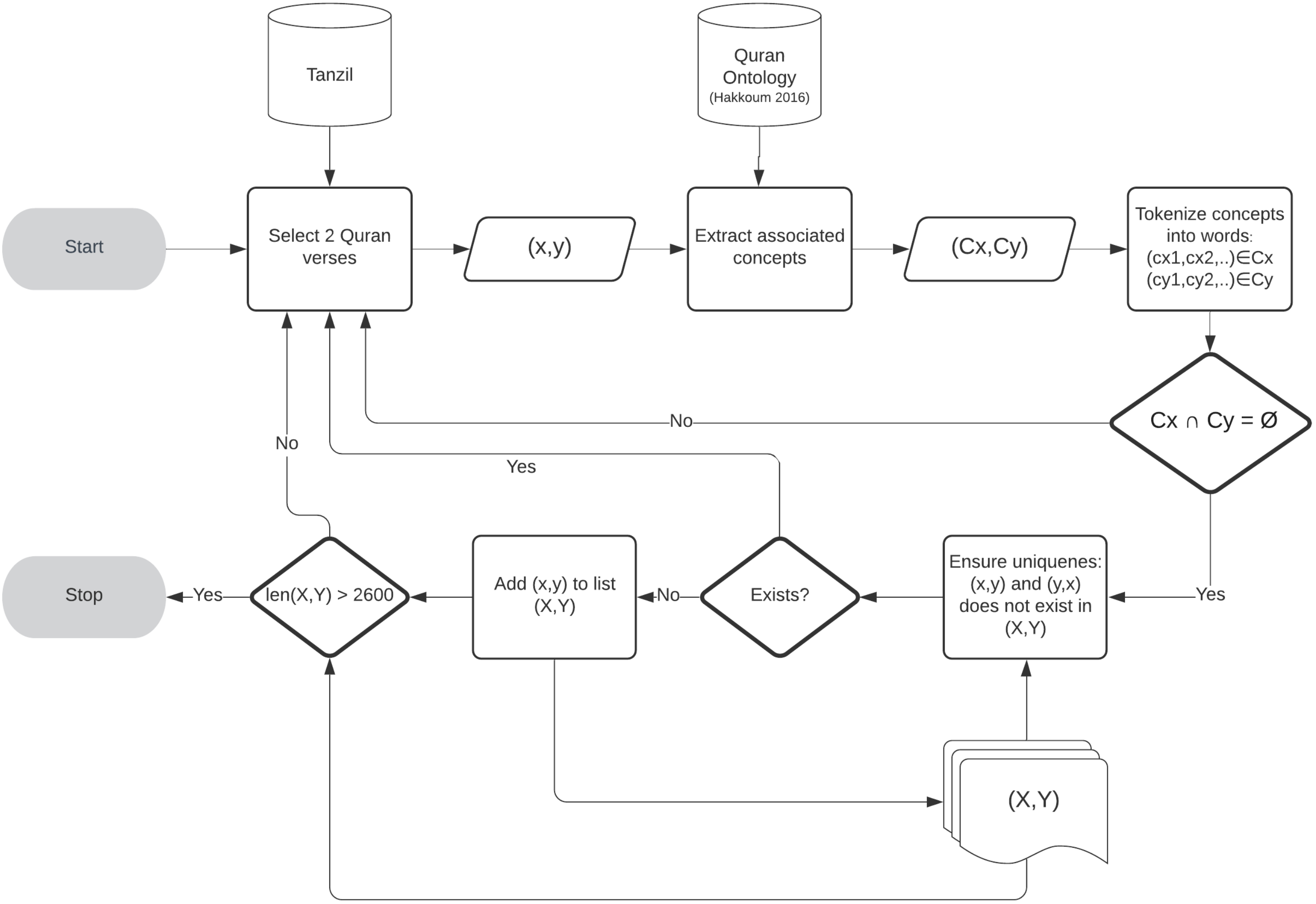

Tanzil
Quran Ontology (Hakkoum 2016)
Start
Select 2 Quran verses
(x,y)
Extract associated concepts
(Cx,Cy)
Tokenize concepts into words: (cx1,cx2,..)∈Cx (cy1,cy2,..)∈Cy
Cx ∩ Cy = Ø
No
Yes
No
Yes
Ensure uniquenes: (x,y) and (y,x) does not exist in (X,Y)
Exists?
No
Yes
Add (x,y) to list (X,Y)
len(X,Y) > 2600
No
Yes
Stop
(X,Y)

Figure 7.8: Algorithm for creating negative QQ samples

QQ Augmentation

The Quran is translated into various languages at the verse level, which can be utilized for data augmentation to increase the size of the Quran-Quran (QQ) training data. The 43 languages available on the Tanzil project can be used for zero-shot transfer learning by fine-tuning multilingual models on the translated QQ training pairs, and then testing them on Arabic QQ data. Additionally, the 17 English translations on Tanzil can be used to fine-tune an English-Arabic cross-lingual model. The Tanzil project also provides electronic versions of Arabic Quran commentaries (Aljalalayn and Almuyasser) that are aligned at the verse level, which can be used to increase the Arabic training data. These translations and commentaries are extracted from Tanzil to form the training, validation and testing data detailed in Table 7.6

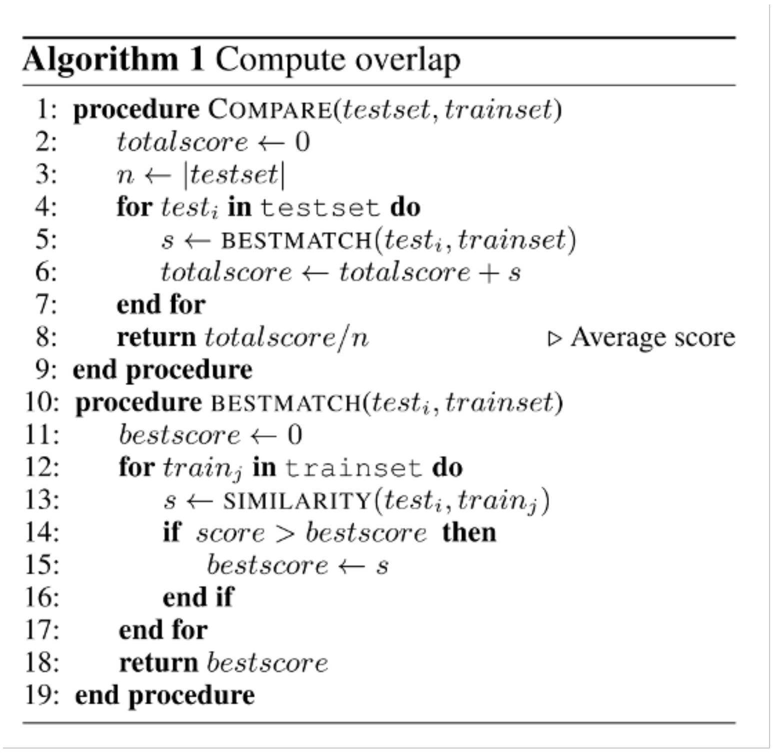
Dataset	**Pairs type**	**# of pairs**
Training	Ar QQ	4,072
	Ar QQ + Tafsir	12,216
	En QQ	20,360
	M QQ	256,536
Validation	Ar QQ	1,019
Testing	Ar QH	310

Table 7.6: The different training/fine-tuning dataset. QQ=Quran-Quran, QH=Quran-Hadith, Ar= Arabic, En= English, M= Multilingual (43 languages)

7.5 Measuring Data Leakage

Recent research have revealed that many benchmark datasets possess a degree of overlap between training and testing data. This can inadvertently lead to evaluating the model's ability to memorize rather than generalize. This section applies the framework introduced by Elangovan *et al.* (2021) to measure the data leakage from the QQ training to QQ validation dataset, and from QQ training to QH testing dataset.

The algorithm for measuring similarity is shown in Figure 7.9. Given a test instance $test_i$ compute its cosine similarity with the most similar instance in the training set. Hence the algorithm goes through every training instance and computes the cosine similarity with the test instance to find the $BESTMATCH$. This is facilitated by the function $similarity(test_i, train_j)$ which converts the two texts into a bag-of-words. Then removes stop words (a list obtained from NLTK) to compute the cosine similarity between the two bag of words as shown in equation 7.1. The algorithm keeps the cosine similarity score of the instance that is most similar with $test_i$. Finally, it computes the average cosine similarity over all the test instances as an indicator to measure the extent of train/test overlap.



Algorithm 1 Compute overlap

```
procedure COMPARE($testset, trainset$)
    $totalscore \leftarrow 0$
    $n \leftarrow |testset|$
    for $test_i$ in testset do
        $s \leftarrow$ BESTMATCH($test_i, trainset$)
        $totalscore \leftarrow totalscore + s$
    end for
    return $totalscore/n$                          ▷ Average score
end procedure
procedure BESTMATCH($test_i, trainset$)
    $bestscore \leftarrow 0$
    for $train_j$ in trainset do
        $s \leftarrow$ SIMILARITY($test_i, train_j$)
        if $score > bestscore$ then
            $bestscore \leftarrow s$
        end if
    end for
    return $bestscore$
end procedure
```

Figure 7.9: Algorithm to compute data leakage (Elangovan *et al.*, 2021)

$$Cosine(A, B) = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{7.1}$$

It is clear that the data leakage in the QQ validation dataset is more significant than that in the QH testing dataset. This is predictable since the testing dataset consists of Hadiths which are not present in the training dataset. However, the Quran data is static and limited which means it is unfeasible to avoid such data leakage. Hence, using the QQ for validation and QH for testing in the next chapter ensures that the models are measured on their ability to generalize rather than memorize.

Evaluation Dataset	**Training Dataset**	**Avg. cosine similarity**
QH Testing	QQ training	49.3
QQ Validation	QQ training	62.4

Table 7.7: Measuring data leakage

7.6 Conclusion

This chapter presents a statistical overview of the differences between Quran and Hadith texts and the unique challenges they pose for AI-based modelling of their underlying meaning. Despite these challenges, recent advances in deep-learning models have shown promising results in achieving human-level performance for different languages. Thus, the chapter suggests investigating the ability of these models to capture the meaning of Quran and Hadith texts, which requires training and evaluation datasets. The main contribution of this chapter is a framework for creating two datasets, the Quran-Hadith (QH) dataset consisting of 310 balanced pairs of related and non-related Quran verses and Hadith teachings, and the Quran-Quran (QQ) dataset containing 5,096 balanced pairs of related and non-related Quran-verse pairs. After creating the datasets, the data leakage is investigated and quantified to ensure that the evaluated models in the next chapter are measured on their ability to generalise rather than memorise data.

Chapter 8

Semantic Relatedness Using Deep Learning

8.1 Introduction

This chapter is largely based on my publication Altammami & Atwell (2022), which describes evaluating deep learning (DL) models' effectiveness in identifying relatedness between the Quran and the Hadith Classical Arabic (CA) texts. This is a binary classification task of identifying whether two pieces of CA text convey the same underlying message, which is a special case of semantic similarity.

The idea of using DL models on CA text is inspired by the models' near-perfect result on other languages and downstream tasks. Moreover, the application of deep learning models on CA texts remains largely unexplored, offering a significant opportunity for further research and exploration. Hence, this study aims to fill the gap by evaluating monolingual, bilingual, and multilingual state-of-the-art (SOTA) models to detect relatedness between the Quran and the Hadith. However, before delving into the experiments, the literature points to interesting observations that shall be considered in this study.

Recently, transformer-based DL models have shown unprecedented performance that surpasses human scores on common benchmarks like SICK (Marelli *et al.*, 2014), which created the hype that such models could "understand" textual data (Bender & Koller, 2020). However, an emerging trend of probing these models yields striking results that suggest models are not able to generalize, but rather are more likely to memorize (Elangovan *et al.*, 2021). This is due to the common practice of generating training and testing data using random split, which inadvertently leads to data leakage from the training set to the testing set reaching 70% of instance overlap (Lewis *et al.*, 2020). Moreover, existing benchmark datasets possess a low readability index which does not reflect real-world complex data (Chandrasekaran & Mago, 2020). This is aggravated for CA since it is a low-resource language and inherently challenging for natural language processing (NLP) tasks (Habash, 2010).

To address the low-readability index limitation, a new CA dataset of related and non-related religious texts from the Quran and the Hadith is created as explained in Chapter 7. These texts involve embedded meanings which require reasoning and deep human understanding. Moreover, they contain complex syntactic and rhetorical features including, verbal idioms, irony, hyperbole, rhetorical questions (Abdul-Raof, 2013) as discussed in Chapter 7.

To mitigate the limitation of data-leakage, the models are fine-tuned on the extended version of the Qursim (Sharaf & Atwell, 2012b) dataset consisting of related and non-related Quran-verse pairs described in Chapter 7. Then the best performing models are tested on the more challenging dataset of Quran-Hadith pairs. Throughout this chapter, the phrase 'Quran-Hadith pairs' is used where 'Hadith' refers to an instance of a Hadith teaching (Matn), and the term 'Quran' refers to a Quran verse.

8.2 Related Work

This section discusses research, mostly in the digital humanities, that aims to utilize advancements in AI to identify similarities in sacred scriptures. There have been several attempts to automatically detect semantic similarity and relatedness between religious text, ranging from within the same book (Saeed *et al.*,

2020) to different religious scriptures (Verma, 2017; Varghese & Punithavalli, 2019; Peuriekeu *et al.*, 2021; Qahl, 2014). The studies also differ in their scope, such that some measure the semantic similarity at the corpus level (Qahl, 2014) and others at the verse level (Alshammeri *et al.*, 2020; Alsaleh *et al.*, 2021).

The glaring weakness of the studies which compare different religious books (e.g., the Quran and the Bible) is where translations are used instead of the original texts. Ideally, such studies should use the texts in their original languages (e.g. CA, Hebrew, Greek) to keep their true meaning, which could be lost in the translations. This is due to the inherent biases or misinterpretations of the translators. For example, the Quran's meaning is translated into more than 60 English translations, each one typically claiming to rectify deficiencies in the previous versions (Kidwai, 1987). Hence, multilingual models with zero-shot learning could be the answer to overcome such problem.

One of the studies that focused on the Quran and is more related to this work is Alshammeri *et al.* (2020). They trained a Doc2Vec model (Le & Mikolov, 2014) on the Quran corpus to obtain embeddings for each verse. Then they calculated the cosine similarity between the Quran pairs. To verify their results, they examined whether the pair of verses with high cosine similarity falls within the same concept in a Quran ontology (Abbas, 2009). Another study on the Quran presented by Alsaleh *et al.* (2021) utilizes AraBERT (Antoun *et al.*, 2020), a transformer-based model, to identify semantic similarity between Quran verses. Their findings show promising results. Contrarily, the experiments in this chapter explore if such SOTA models fine-tuned on the Quran pairs dataset can be generalized to perform as well when presented with the Quran-Hadith pairs dataset. Hence, this work is not comparable to the previous research, because they used the same dataset for both training and testing by utilizing random split.

8.3 Datasets

Throughout this experiment, the dataset explained in Chapter 7 is used. The training/fine-tuning is conducted with Quran-Quran (QQ) verse pairs where every training example consists of a pair of verses along with a label, "1" for related and "0" for not related. Then the testing is conducted on a dataset of Quran-Hadith (QH) pairs. An example from each dataset is shown in Table 8.1. Using the QQ dataset for training and the QH dataset for testing ensures that the models are measured on how well they generalize to identify relatedness in CA texts instead of memorizing the training data.

Dataset	Label	Text 1	Text 2
QQ	1	لكل امرئ منهم يومئذ شأن يغنيه	ولا يسأل حميم حميما
QQ	0	أولئك الذين طبع الله على قلوبهم وسمعهم وأبصارهم وأولئك هم الغافلون	إني وجدت امرأة تملكهم وأوتيت من كل شيء ولها عرش عظيم
QH	1	يا أيها الذين آمنوا لا تأكلوا الربا أضعافا مضاعفة واتقوا الله لعلكم تفلحون	النبي صلى الله عليه وسلم قال ليأتين على الناس زمان لا يبالي المرء بما أخذ المال أمن حلال أم من حرام.
QH	0	وإن تطع أكثر من في الأرض يضلوك عن سبيل الله إن يتبعون إلا الظن وإن هم إلا يخرصون	قال رسول الله صلى الله عليه وسلم نفس المؤمن معلقة بدينه حتى يقضى عنه

Table 8.1: Examples from the QQ and QH datasets

The QQ dataset is a balanced dataset of 5,096 related and non-related verse pairs which is shuffled and divided into 80% training and 20% validation. The datasets used throughout the experiments in the next section are shown in Table 8.2. The Quran translations in 43 languages provided electronically on the Tanzil[46] project are used to fine-tune multilingual models, while the 17 different English translations on Tanzil were used to fine-tune the English-Arabic bilingual model. Also, the Arabic Tafsir for each pair is extracted to augment the Arabic training data. After fine-tuning, the models are validated on the 20% of QQ Arabic pairs. Finally, the QH dataset, consisting of 310 balanced positive and negative pairs, is used in the testing phase [47].

[46] https://tanzil.net/docs/tanzil_project

[47] Refer to Chapter 7 for more details on the automatic construction of these datasets

Dataset	Pairs type	# of pairs
Training	Ar QQ	4,072
	Ar QQ + Tafsir	12,216
	En QQ	20,360
	M QQ	256,536
Validation	Ar QQ	1,019
Testing	Ar QH	310

Table 8.2: The training/fine-tuning dataset as introduced in Table 7.6

8.4 Experimental Setup

This section introduces the different models trained/fine-tuned on the QQ dataset and tested for identifying relatedness in the QH dataset. This is not semantic similarity in the traditional sense where being synonymous or directly equivalent is what is measured, but rather identifies relatedness in the underlying religious teaching. The models are given $[text1, text2, label]$ where label "1" indicates their relatedness and "0" their non-relatedness.

8.4.1 Evaluation Metrics

The models' performance is measured using the common accuracy and F1 scores as shown in equations 8.1 and 8.2 respectively, where *TP=True Positive*, *TN= True Negative*, *FP= False Positive* and *FN= False Negative*. In addition, the Matthews Correlation Coefficient (MCC) is used as shown in equation 8.3, which takes the value between -1 and +1. It produces a reliable score in evaluating binary classifications, where a high score is produced only if the model preforms well on the majority of the negative instances and the majority of positive instances (Chicco & Jurman, 2020).

$$Accuracy = \frac{TP + TN}{TN + TP + FP + FN} \tag{8.1}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{8.2}$$

$$MCC = \frac{TP * TN – FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{8.3}$$

8.4.2 Baseline Model

Textual data must be first represented in a mathematical form for the computer to process it. Recently, text is converted to low dimensional vectors (known as embeddings) where lexical units are represented by an n-dimensional vector that captures word analogy and relationships. This is achieved by training a neural network on a corpus to generate embeddings for each word[48] in the corpus, taking into account its surrounding words. For the baseline, a character embedding model is used (Bojanowski *et al.*, 2017) since this has proven to be a major breakthrough in the field of semantic similarity, specifically for morphologically rich languages. The key idea is that words and subword information are learnt in the process of training the embedding model. This way word embedding vectors can be the aggregation of their constituent character n-grams.

To build the baseline model, the methodology proposed by Nagoudi *et al.* (2017) is adopted, which produced one of the top results at the SemEval 2017 Task1. Their approach enhances the embedding model by incorporating Inverse Document Frequency (idf) weighting and Part-of-Speech (POS) tagging to give more weight to highly descriptive words in a text. The is depicted in Figure 8.1, which illustrates the process of computing the relatedness between a given pair of texts (S^1, S^2).

To compute the idf, the Quran is used as the background corpus. The idf for each word is measured using Eq 8.4, where s is the total number of sentences in

[48] A maximum vocabulary size can be set, or a minimum frequency threshold can be set for a word to be included in the corpus.

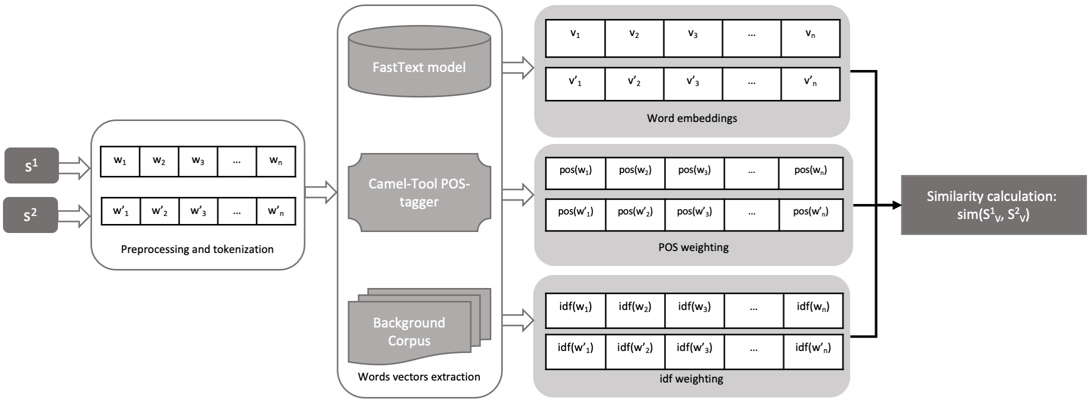

s^1
s^2
w_1 w_2 w_3 ... w_n
w'_1 w'_2 w'_3 ... w'_n
Preprocessing and tokenization
FastText model
Camel-Tool POS-tagger
Background Corpus
Words vectors extraction
v_1 v_2 v_3 ... v_n
v'_1 v'_2 v'_3 ... v'_n
Word embeddings
$pos(w_1)$ $pos(w_2)$ $pos(w_3)$ ... $pos(w_n)$
$pos(w'_1)$ $pos(w'_2)$ $pos(w'_3)$ ... $pos(w'_n)$
POS weighting
$idf(w_1)$ $idf(w_2)$ $idf(w_3)$ ... $idf(w_n)$
$idf(w'_1)$ $idf(w'_2)$ $idf(w'_3)$ ... $idf(w'_n)$
idf weighting
Similarity calculation: $sim(S^1_V, S^2_V)$

Figure 8.1: Baseline model architecture

the corpus (Quran verses), and ws is the number of sentences that contain the word w. The POS tags were obtained using CAMeL Tools (Obeid *et al.*, 2020).

$$idf(w) = log(s/ws) \tag{8.4}$$

To obtain the word embeddings, a character n-grams model $FastText(FT)$ is trained using the Genism[49] library on several corpora including Quran, Hadith, and commentaries, which consist of 15,222,814 words of which 207,347 are unique words. The model training hyperparameters were set to vector size 300, which refers to the dimensionality of the vectors. The window is set to 5, which indicates the number of words considered around the pivot word to capture context. The minimum count is set to 3 to ignore rare words.

The limitation of $FastText$ and other context-free models is that it generates a single word embedding representation for each word in the data. However, since the models were trained on a specialized corpus, I assume they will capture the semantics of the religious CA words. Once the embedding for each word is obtained, the verse (sentence) embedding S_v is calculated as shown in Eq 8.5, where w_i represents a word in the verse, $POSw_i$ represents the POS tag of the word w_i,

[49] https://radimrehurek.com/gensim/models/fasttext.html

which is used to assign the corresponding weight as proposed by Nagoudi *et al.* (2017), and v_i is the word vector obtained from the *FastText* model. Once the embedding for each verse is calculated, the cosine similarity between the pairs is measured considering those pairs with more than 0.5 as related.

$$S_v = \sum_{i=1}^{n} (idf(w_i)) \times POS\_weight(POSw_i) \times v_i \tag{8.5}$$

Another baseline model is considered where several machine learning classifiers (SVM, Random Forests, Naive Bayes) were trained with the aforementioned embeddings as the features. The best performance was obtained using the Random Forests as shown in the results in Section 8.5.

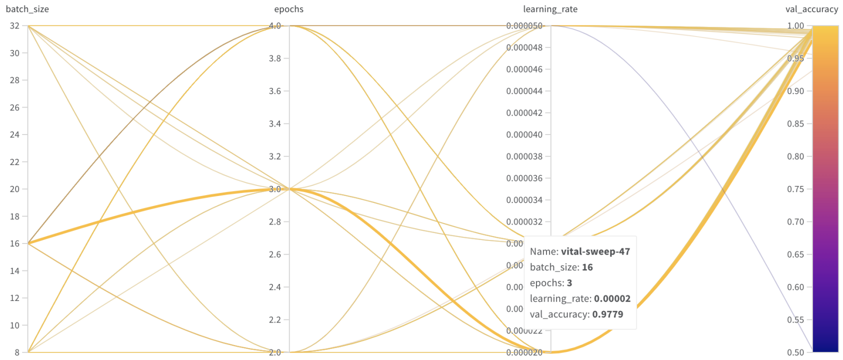
8.4.3 Transformer-based Models

Transformer-based models such as BERT are context-based and aim to address the limitation of the aforementioned context-free models that generate a single word embedding representation regardless of word contexts. Hence, polysemic words with different meanings could be captured in these transformer-based models.

Hyper-Parameters and Technical Details

The next experiments consist of fine-tuning pre-trained transformer-based models to classify pairs as related or not. The Tesla K80 GPU available on Google Colab is used in all experiments. For a given model, the experiment is run three times to ensure the results obtained are stable using the same hyper-parameter values of Adam optimizer with learning-rate 2e-5, patch-size 16. These values were determined after using Weights and Biases[50] for experiment tracking as shown in Figure 8.2

[50] https://wandb.ai/site



Name: **vital-sweep-47**
batch_size: **16**
epochs: **3**
learning_rate: **0.00002**
val_accuracy: **0.9779**

Figure 8.2: Weights and Biases is used to determine the hyper-parameters

Transformer-based models have a predefined input length requirement. To accommodate this, a maximum sequence length of 128 tokens is used. This is determined by tokenizing each sample in the training data and plotting the number of tokens in a histogram as shown in Figure 8.3. It is observed that the majority of samples have fewer tokens than 128. Furthermore, early stopping is used to avoid over-fitting since the models are observed to overfit after the second epoch as depicted in Figure 8.4.

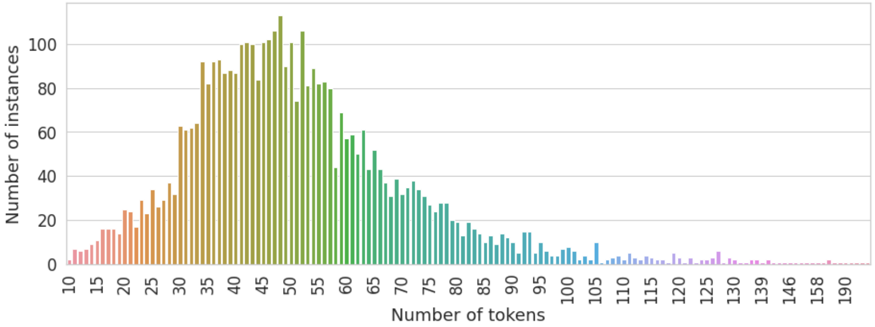


Figure 8.3: The number of tokens in each instance of the training data

Monolingual Models

In this experiment, the focus is on using pre-trained transformer-based models that were trained on Classical Arabic (CA) or Modern Standard Arabic (MSA),

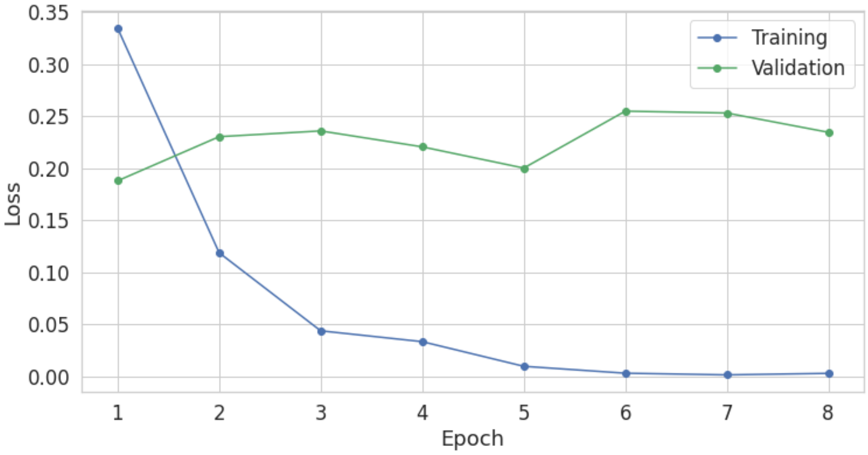


Figure 8.4: Fine-tuning CAMeLBERT-CA for 8 epochs

but not Dialect Arabic (DA). This is because a recent study found that the variant proximity of pre-training data to fine-tuning data is the most critical factor (Inoue *et al.*, 2021). One of the widely used models is AraBERT (Antoun *et al.*, 2020) which is trained on 24GB of Arabic text in the news domain. There are several versions of this model, but AraBERTv02 is used because previous studies showed its superior performance (Inoue *et al.*, 2021; Alsaleh *et al.*, 2021). A similar model incorporated into the experiments is ArabicBERT (Safaya *et al.*, 2020), which is trained on 95GB of text mainly from the Arabic portion of the OSCAR corpus. Another model used is the CAMeLBERT-CA (Inoue *et al.*, 2021), which is trained on the 6GB OpenITI Corpus consisting of CA texts (Belinkov *et al.*, 2018).

Multilingual Models

Zero-shot learning showed promising results where multilingual models are fine-tuned on one language then tested on another language. In this study, experiment on multilingual modules is conducted by fine-tuning two models on the English and the multilingual datasets. These models are the mBERT (Devlin *et al.*, 2018) which is trained on Wikipedia dump, and XLMRoberta (Conneau *et al.*, 2019) trained on filtered Common-Crawl data. In addition to this, the performance of a bilingual module named GigaBERT (Lan *et al.*, 2020) is studied. This model is designed for English-to-Arabic cross-lingual transfer tasks and trained on newswire, Wikipedia, and web crawl data.

Sentence-BERT

Sentence-BERT (SBERT) is a framework introduced to enable obtaining semantically meaningful sentence embeddings that can be compared using cosine similarity (Reimers & Gurevych, 2019). According to the authors, SBERT results in better representations than BERT for tasks involving text comparison. Therefore, their multilingual models are incorporated into these experiments as shown in Table 8.4. The threshold of 0.5 and above is used to classify pairs as related.

8.5 Results

This section shows the performance of models on the validation dataset of QQ pairs. It discusses the best model's performance on the testing dataset of QH pairs.

8.5.1 Validation

Table 8.3 shows the models' performance on the validation dataset of 1,019 Arabic QQ pairs. Moreover, it shows the fine-tuning dataset types, the number of pairs used in fine-tuning, and the fine-tuning time. The table is categorized into three section based on the model's type. The first section shows the performance of baseline models using FastText embeddings and cosine similarity as explained in Section 8.4.2. The enhanced baseline incorporates a machine learning (ML) model trained on the two Arabic datasets of 4,072 QQ pairs and 8,144 Tafsir pairs.

The second section records the performance of the monolingual models fine-tuned on the Arabic datasets. The third section presents the results of using zero-shot learning where the cross-lingual GigaBert model is fine-tuned on the English translations of the QQ pairs. The multilingual models, mBert and XLM-Roberta, are fine-tuned on the English QQ pairs in the first iteration, and on the

Model	Fine-tuning	# of pairs	Time	MCC	Acc.	F1
Baseline FastText Nagoudi *et al.* (2017)	-	-	-	-2.8	49.2	64.1
FastText + Random Forests	Ar	4,072	0:00:65	+57.6	78.4	76.5
FastText + Random Forests	Ar + Taf	12,216	0:24:32	+55.4	77.4	75.9
AraBERT Antoun *et al.* (2020)	Ar	4,072	0:01:57	+74.6	86.7	85.4
AraBERT	Ar + Taf	12,216	0:04:37	+81.0	90.2	89.6
ArabicBERT Safaya *et al.* (2020)	Ar	4,072	0:01:33	+88.4	94.1	**94.0**
ArabicBERT	Ar + Taf	12,216	0:04:37	+94.1	97.0	**97.0**
CAMeLBERT-CAInoue *et al.* (2021)	Ar	4,072	0:01:38	+82.2	90.9	**90.5**
CAMeLBERT-CA	Ar + Taf	12,216	0:04:51	+86.6	93.2	**93.1**
GigaBERT Lan *et al.* (2020)	En	20,360	0:07:51	+16.0	55.9	67.0
XLM-RoBERTa Conneau *et al.* (2019)	En	20,360	0:08:44	+20.8	57.8	68.2
XLM-RoBERTa	M, Ar, En	260,608	1:52:40	+29.0	64.3	61.6
mBERT Devlin *et al.* (2018)	En	20,360	0:08:10	-24.7	44.2	61.3
mBERT	M, Ar, En	260,608	1:52:40	+39.0	69.0	65.2

Table 8.3: Models' performance on the QQ validation dataset shown in %. Overall best models based on F1 score are highlighted in bold. Ar = Arabic Quran, En = English Quran, M = Multilingual Quran (43 languages), Taf = Arabic Tafsir(commentaries)

multilingual QQ pairs in the second iteration. Finally, the performance of SBERT models are shown separately in table 8.4 since they do not require fine-tuning.

Model	MCC	Acc.	F1
paraphrase-xlm-r-m-v1	+19.6	59.8	61.4
distiluse-base-m-cased-v2	+30.5	60.4	37.8
distiluse-base-m-cased-v1	+30.3	56.8	35.3
stsb-xlm-r-m	+16.6	58.4	60.0

Table 8.4: Performance of SBERT multilingual models on the QQ dataset

The results demonstrate that the best performing models on the Arabic QQ pairs validation dataset are the monolingual models. Contrarily, multilingual and SBERT models showed the worst performance across their different models, which highlights the impact of the pretraining data. Additionally, the baseline model does not perform as well as it does on the SemEval task. This could be attributed to the dataset of QQ pairs, since they are considered complex with embedded meanings and polysemy, which is hardly captured in non-contextual representations. For this reason, I believe there is potential in contextual models to capture the underlying meaning of Hadith and Quran.

Table 8.5 shows examples of CAMeLBERT-CA classification results. It seems that the model produced promising results even with some of the incorrectly classified pairs. For example, the first pair, although not related as a sentence, contains words that have semantic similarities like the word حميم which mean 'warm' or 'close [in relationships]' . The transliteration of this example is shown to indicate the possible shallow similarity that was identified by the model. The second example was classified as not-related, which is appropriate, because this relatedness requires knowing what the pronouns are referring to. Bear in mind that the Arabic text is different from the translations which usually explicitly state the meaning embedded in the original Arabic text.

ID	Label	Predic.	Quran Verse1	Quran Verse2
Ex1	0	1	نار حامية Transliteration: [Naarun hamiyah] It is a Fire, intensely hot. [101:11]	ولا يسأل حميم حميما Transliteration: [Wa laa yas'alu hameemun hameemaa] And no friend will ask [anything of] a friend. [70:10]
Ex2	1	0	وإذ قلنا للملائكة اسجدوا لآدم فسجدوا إلا إبليس أبى واستكبر وكان من الكافرين And [mention] when We said to the angels, "Prostrate before Adam"; so they prostrated, except for Iblees. He refused and was arrogant and became of the disbelievers. [2:34]	قال فاهبط منها فما يكون لك أن تتكبر فيها فاخرج إنك من الصاغرين [Allah] said," Descend from Paradise, for it is not for you to be arrogant therein. So get out; indeed, you are of the debased". [7:13]
Ex2	1	1	لكل امرئ منهم يومئذ شأن يغنيه For every man, that Day, will be a matter adequate for him. [80:37]	ولا يسأل حميم حميما And no friend will ask [anything of] a friend. [70:10]

Table 8.5: Examples of QQ pairs classified by CAMeLBERT-CA

Although monolingual modules produced promising results, can their performance be generalized to other CA datasets in the same domain? To answer this question, these models are evaluated on the testing dataset of Quran-Hadith (QH) pairs.

8.5.2 Testing

This section investigates how well the best-performing models generalize to the dataset of QH pairs. The results are shown in Table 8.6. It is clear that

CAMeLBERT-CA's performance is superior. However, there is a significant drop across all the models' F1 score compared to their performance on the validation dataset of QQ pairs in Table 8.3.

Model	Fine-tuning	MCC	Acc.	F1
ArabicBERT	Ar	+3.6	51.7	50.0
ArabicBERT	Ar+Taf	+32.3	66.1	65.1
CAMeLBERT-CA	Ar	+31.9	65.7	67.9
CAMeLBERT-CA	Ar+Taf	**+55.3**	**77.2**	**74.8**
AraBERT	Ar	+12.9	56.3	59.1
AraBERT	Ar+Taf	+41.4	70.6	69.6
GigaBERT	En	-8.3	46.2	56.2
XLM-RoBERTa	En	-14.0	42.7	38.0
XLM-RoBERTa	En+Ar+M	-6.8	46.6	46.4
mBERT	En	-4.2	48.5	62.7
mBERT	En+Ar+M	+7.0	53.4	46.8

Table 8.6: Performance of models on the Quran-Hadith testing dataset

8.6 Analysis and Discussion

This section analyses the results of the ∼20 points drop in F1 score across the models on the QH dataset. Several examples classified by CAMeLBERT-CA are shown in Table 8.7 to illustrate where it did well and discuss the causes of misclassification. Example 1 shows a Hadith-teaching and Quran-verse that consist of mostly different words except for one keyword that occurs in different morphological forms (الإيلاء - يؤلون). The model was able to identify the relatedness. However, many of the correctly classified pairs involve of a clear message as in Example 2. While 70% of the incorrectly classified pairs have the label "1" (related), but the model fails to identify the relatedness and predicts "0" (not-related). This could be attributed to several reasons. First, the two texts consist of different words but have the same underlying message as shown in Example 3. Second, several words in these texts are vague, hence reference to exegesis is required to understand them as shown in Example 4. Third, many Hadiths are a narration of an incident or a story that has a moral behind it, while the Quran gives the explicit guidance, as shown in Example 5.

To further analyse the model, the steps it takes to process an input and produce an output are scrutinized. Consider example 6 in Table 8.7. The input text is converted to the appropriate format required by the pre-trained model as shown in Figure 8.5. It expects the input to start with the [CLS] token and end with the [SEP] token. Since there are two texts (Quran-verse and Hadith-teaching), they are separated by the [SEP] token. Finally, a [PAD] token is used for padding, because the model expects each input to be at a fixed length of 128 tokens as discussed in Section 8.4.3.

The way the Quran-verse and the Hadith-teaching are tokenized depends on the model's tokenizer. Generally, pre-trained models are trained on specific corpora (e.g. Classical Arabic), where they learn a vocabulary set. Hence, during fine-tuning, they might not recognize a word that is not found in their vocabulary. This out-of-vocabulary (OOV) problem is solved using the WordPiece tokenizer, which breaks down such words into sub words as indicated by the ## in Figure 8.5.

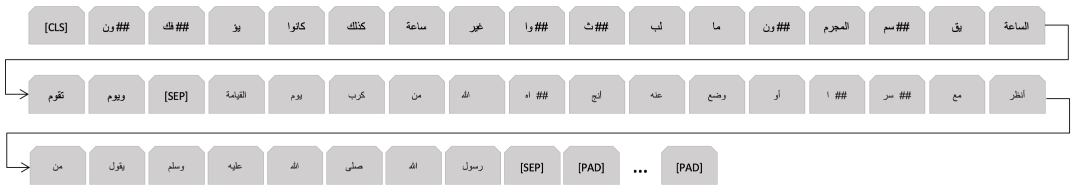


Figure 8.5: How the WordPiece tokenizes a Quran-verse and a Hadith-teaching from Example 6 in Table 8.7

The figure shows that the tokenizer identified the words سَاعَةٍ and السَّاْعَةُ in the Quran-verse and did not break them down into sub words. Although they convey the same meaning out of context, they have completely different meanings in this verse. The first means "Day of Judgment" while the second means

ID	Label	Predic.	Quran	Hadith
Ex1	1	1	للذين يؤلون من نسائهم تربص أربعة أشهر فإن فاءوا فإن الله غفور رحيم	كان يقول في الإيلاء الذي سمى الله لا يحل لأحد بعد الأجل إلا أن يمسك بالمعروف أو يعزم بالطلاق كما أمر الله عز وجل
			For those who swear not to have sexual relations with their wives is a waiting time of four months, but if they return [to normal relations] - then indeed, Allah is Forgiving and Merciful. [2:226]	If the period of Ila expires, then the husband has either to retain his wife in a handsome manner or to divorce her as Allah has ordered.
Ex2	1	1	وإني لغفار لمن تاب وآمن وعمل صالحا ثم اهتدى	قال رسول الله صلى الله عليه وسلم التائب من الذنب كمن لا ذنب له
			But indeed, I am the Perpetual Forgiver of whoever repents and believes and does righteousness and then continues in guidance. [20:82]	He who repents of a sin is like him who has committed no sin.
Ex3	1	0	يا أيها الذين آمنوا لا تأكلوا الربا أضعافا مضاعفة واتقوا الله لعلكم تفلحون	النبي صلى الله عليه وسلم قال ليأتين على الناس زمان لا يبالي المرء بما أخذ المال أمن حلال أم من حرام
			O you who have believed, do not consume usury, doubled and multiplied, but fear Allah that you may be successful. [3:130]	Certainly a time will come when people will not bother to know from where they earned the money, by lawful means or unlawful means.
Ex4	1	0	يا أيها الذين آمنوا إنما الخمر والميسر والأنصاب والأزلام رجس من عمل الشيطان فاجتنبوه لعلكم تفلحون.	أن رسول الله صلى الله عليه وسلم قال لا سبق إلا في نصل أو خف أو حافر
			O you who have believed, indeed, intoxicants, gambling, [sacrificing on] stone alters [to other than Allah], and divining arrows are but defilement from the work of Satan, so avoid it that you may be successful. [5:90]	Wagers are allowed only for shooting arrows, or racing camels or horses.
Ex5	1	0	وما أرسلنا من قبلك إلا رجالا نوحي إليهم فاسألوا أهل الذكر إن كنتم لا تعلمون	خرجنا في سفر فأصاب رجلا منا حجر فشجه في رأسه ثم احتلم فسأل أصحابه فقال هل تجدون لي رخصة في التيمم فقالوا ما نجد لك رخصة وأنت تقدر على الماء فاغتسل فمات فلما قدمنا على النبي صلى الله عليه وسلم أخبر بذلك فقال قتلوه قتلهم الله ألا سألوا إذ لم يعلموا فإنما شفاء العي السؤال إنما كان يكفيه أن يتيمم ويعصب على رأسه خرقة ثم يمسح عليها ويغسل سائر جسده
			And We sent not before you except men to whom We revealed [Our message]. So ask the people of the message if you do not know.[16:43]	We set out on a journey. One of our people was hurt by a stone, that injured his head. He then had a sexual dream. He asked his fellow travelers: Do you find a concession for me to perform tayammum? They said: We do not find any concession for you while you can use water. He took a bath and died. When we came to the Prophet (PBUH), the incident was reported to him. He said: They killed him, may Allah kill them! Could they not ask when they did not know? The cure for ignorance is inquiry. It was enough for him to perform tayammum and to pour some drops of water or bind a bandage over the wound (the narrator Musa was doubtful); then he should have wiped over it and washed the rest of his body.
Ex6	1	0	ويوم تقوم الساعة يقسم المجرمون ما لبثوا غير ساعة كذلك كانوا يؤفكون	رسول الله صلى الله عليه وسلم يقول من أنظر معسرا أو وضع عنه أنجاه الله من كرب يوم القيامة
			And the Day the Hour appears the criminals will swear they had remained but an hour. Thus they were deluded[30:55]	God's Messenger say, "He who grants a respite to one who is in straitened circumstances or who remits his debt will be saved by God from the anxieties of the day of resurrection.".

Table 8.7: Examples of QH pairs classified by CAMeLBERT-CA

“hour”. Hence, the word السَاْعَة in the Quran is semantically similar to the word الْقِيَامَةِ in the Hadith of Example 6. To check if the model captured the contextual meaning, the cosine similarity between the contextual embeddings is measured. This is done by obtaining the tensor produced by the model for this record.

The model produces [43 x 12 x 768] tensor since there are 43 tokens, 12 layers, and 768 features. There are several ways to extract the contextual word embeddings as mentioned by the creators of BERT (Devlin *et al.*, 2018), but the suggested approach is followed where the last four layers are added. Below are the first five vector values for each word.

الساعة	[-0.5590, -2.8905, -3.1850, -2.1183, -4.8249, ...]
ساعة	[-1.2987, 0.6788, -2.4608, -0.4145, -8.4386, ...]
القيامة	[-3.6930, -0.5412, 0.8372, -6.3855, -3.2655, ...]

The cosine similarity is 0.78 between الساعة and القيامة,while the cosine similarity between الساعة and ساعة is 0.83. Hence, the model did not capture that الساعة is more similar to القيامة than it is to ساعة in that context.

The issue with transformer-based models is that they are challenging to interpret. There is an ongoing field of research aimed at making models more interpretable. Therefore, it is difficult to determine the reasons for the inadequate detection of semantic similarities by these models, particularly the poorer performance of multilingual models as opposed to the baseline models on the validation data. Research by Rust *et al.* (2020) suggests that the superior performance of monolingual models compared to multilingual models may be attributed to the way data is tokenized. The tokenization process uses the WordPiece algorithm (Wu *et al.*, 2016), which breaks texts into words present in its vocabulary. If a word is not in the vocabulary (OOV), it is tokenized into its constituent syllables or letters. This is known as “tokenizer fertility”, and measures the average number of subwords produced per word. A fertility score of 1 indicates that the tokenizer's vocabulary contains every single word in the text, while high fertility scores indicate that more sub words are produced for a word.

To test this hypothesis, the Quran verses (310 sentences) and Hadith teachings (310 sentences) from the QH dataset were tokenized using two approaches. The first approach is white space tokenization, which is performed to show the normal sentence length. The second approach is using the models' tokenizers to show the number of tokens produced for each sentence by the WordPiece algorithm. Figure 8.6 shows sentence length distribution using the two approaches, with the x-axis representing the sentence length in tokens and the y-axis representing the number of instances of a certain length. Although every tokenizer is different, the bilingual and monolingual models show similar distributions, while the multilingual models show different distributions. For more analysis, the overall fertility scores of the tokenizers on the QH dataset are depicted in Figure 8.7.

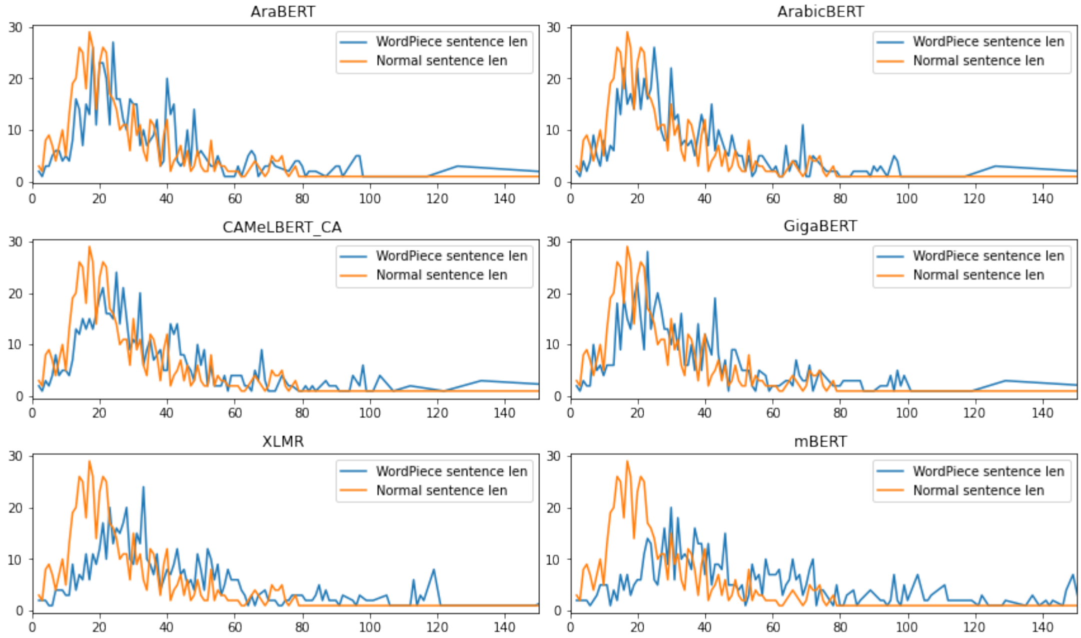


Figure 8.6: Sentence length distribution where x-axes represent the number of tokens and the y-axes reflect the number of instances

It is clear from Figures 8.7 and 8.6 that mBERT has the highest tokenizer fertility. However, as shown in Table 8.6, the mBERT's performance on identifying relatedness between Quran and Hadith is comparable to one of the monolingual models (ArabicBERT). This indicates that the training and fine-tuning data has a greater impact on the results than the tokenizer's fertility. This is because for the models to capture a specific meaning of a word, they should be presented with multiple examples that use that particular meaning. However, some words in Arabic have more prevalent meanings than that meaning found in a particular Quran-verse, such as الساعة. Hence, to improve the performance of these models, the data should be increased. However, the limitation of a relatively small training/fine-tuning data is difficult to overcome, especially for Arabic religious texts, which are static, unlike MSA or DA, where the training data can be increased by, for instance, collecting textual posts from social media platforms.

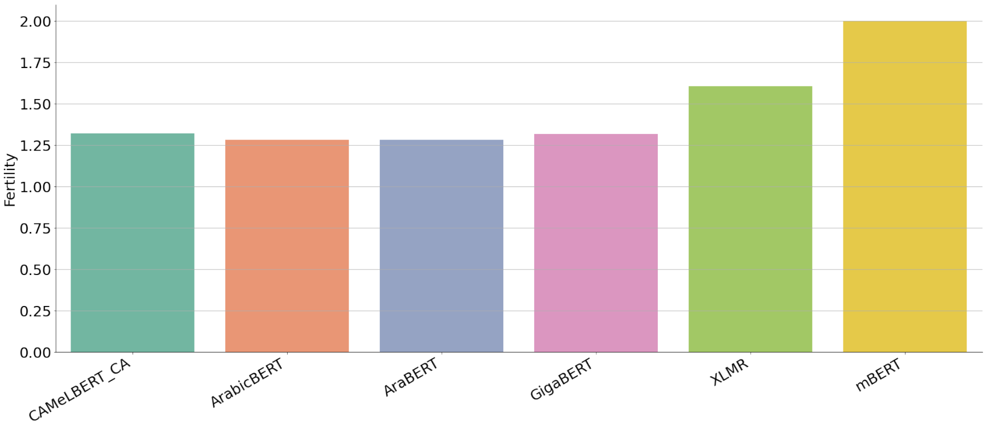


Figure 8.7: Models' tokenizers' fertility

The analysis leads to the following conclusion. It might be more possible for machines to capture the underlying meanings in the Quran and the Hadith through a knowledge-based approach. Although many ontologies have been developed for the Quran, none of them cover the larger scope of Hadith as discussed in Chapter 6. Creating such resources is expensive, and utilizing deep learning models could be the alternative. Yet, the possibilities of the current models reaching a human-level understanding of such texts remains somewhat elusive. This aligns with a recent study by Chandrasekaran & Mago (2020), who tested transformer-based models on a complex dataset of 50 English sentence pairs. Their performance decreases by more than 10 points compared to their performance on the common benchmarks (e.g. SICK). This is amplified for under-resourced languages as shown by the results in this chapter. Hence, measuring the sensitivity of the models to an increase in the complexity of the text should be considered.

8.7 Conclusion

In this chapter, various transformer-based models are evaluated on a binary classification task of identifying the relatedness between Quran-verses and Hadith-teachings. Datasets of Arabic Quran pairs, Arabic Tafsir pairs, and translations of the Quran pairs are used for fine-tuning these models. The results indicate that the monolingual model CAMeLBERT-CA achieved the highest performance, but there was a significant drop in the F1 score on the testing dataset compared to the validation dataset. This suggests that current state-of-the-art models still struggle with complex data and highlights the need for improved representation methods for such complex texts.

Chapter 9

Conclusion and Future Work

9.1 Introduction

Research in AI and NLP aims to model and computationally interpret the meaning of human languages such as Arabic, which has several variants: Modern Standard Arabic (MSA), Classical Arabic (CA) and Dialect Arabic (DA) as discussed in Chapter 2. Although Arabic NLP has received more attention than previously in the last decade (Guellil *et al.*, 2021), more work is required to reach maturity for major languages such as English. The Arabic language is more complex than English and what works for one variant does not necessary apply to another (Habash *et al.*, 2012b). This is evident from the number of attempts to create more fine-grained resources specific to each variant such as Abdulrahim *et al.* (2022); Khalifa *et al.* (2018) and El-Haj *et al.* (2022a).

Although some work has been done on CA, most is dedicated to the Quran while the Hadith does not benefit from the same level of computational focus (Bounhas, 2019). This thesis is based on five publications that make significant contributions to enriching Hadith computational research.

9.2 Overall Findings

At the outset of this work, I aimed to answer two research questions:

1. **Is annotating the Hadith components, Isnad and Matn, Feasible?**

Chapter 4 answers this question by discussing Hadith features that make its structure unique. It is clear that Isnad, the chain of narrators, follows a special pattern of scholar's names linked by transmission method such as 'said', 'heard', and 'told us', which are exploited in building the segmenter. The experiment show that the bigram representation model is the best in capturing the Isnad pattern as discussed in Section 4.5.2. Furthermore, Section 4.6 describes how Random Forests outperforms other ML model to classify bigrams as Isnad or Matn. Additionally, Arabic text preprocessing is discussed and experiments show that normalization has a positive effect on the segmenter performance as shown in Section 4.8. These findings were published in Altammami *et al.* (2019, 2020a) and facilitated building a Hadith segmenter that deconstruct Hadiths to their Isnad and Matn components with 92.5% accuracy.

Although there is still room for improvement, specifically for Hadiths with irregular structure, the proposed system produced acceptable performance for building a well-structured Hadith corpus of the six canonical Hadith books. This is because most Hadiths in these books follow a regular structure. Chapter 5 described the creation of the Leeds and KSU (LK) Hadith corpus, which incorporates Hadiths in their original Classical Arabic and their English translation aligned at the Hadith level. Moreover, each Hadith's meta-data is incorporated including book name, chapter, section, Hadith number, and grade, along with the Isnad and Matn produced by the Hadith segmenter. This is one of the main contributions of this thesis, and is published in Altammami *et al.* (2020b) and made available to the research community on GitHub [51]. To the best of my knowledge, this corpus addresses the shortage in common Hadith datasets as identified by Bounhas (2019). Additionally, the value of this corpus is demonstrated by its current usage in other research (Habash *et al.*, 2022; Tarmom *et al.*, 2021).

[51] https://github.com/ShathaTm/LK-Hadith-Corpus

2. **Is semantic similarity in Quran and Hadiths detectable?**

The second research question aims to discover whether SOTA AI tools and methods are able to model the meanings of Hadiths. This is vital for identifying similarities within the Hadith teachings and other religious texts like the Quran. Such findings facilitate better understanding of these text by computationally organizing and mining knowledge to serve those studying the Hadith and shed light on interpretations and discovery of new embedded knowledge.

To address this question, an investigation was conducted to explore various methods for computationally capturing the meaning of Hadith teachings (Matns). Two main approaches were considered: a non-deep-learning approach and a deep-learning approach. As there are several Quran ontologies available, the non-deep-learning approach was initially explored. This is because both the Quran and the Hadith are written in CA and cover the same domain of religious teachings. Therefore, the study evaluated how well existing Quran ontologies cover Hadith concepts using a corpus-based evaluation, as discussed in Chapter 6. To perform this corpus-based evaluation, Hadith keywords must be extracted. Instead of using feature extraction methods such as Latent Dirichlet Allocation (LDA) topic modelling, the Hadith section headings were utilized. This is because domain-relevant terms tend to occur in section headings as the scholar who compiled each Hadith book organized the Hadiths into chapters and sections with titles reflecting the topic in the incorporated Hadiths. The results showed that the best Quran ontology covered only 26.8% of Hadith concepts. This experiment and findings were published in my paper (Altammami *et al.*, 2021), which led to the conclusion that extending the ontology to cover Hadiths is expensive. Therefore, the second approach of using deep-learning models is investigated.

Deep-learning transformer-based models have shown promising results on various downstream tasks, but their performance on CA texts has not been widely studied. To address this gap, I evaluated monolingual, bilingual, and multilingual SOTA models for detecting relatedness between the Quran and the Hadith considering their underlying meanings, which require deep human understanding, as discussed in Section 7.2. To accomplish this, a benchmark dataset was required. Since there were no existing datasets appropriate for the task, I developed a methodology for building a dataset of related and non-related Quran-verse and Hadith-Matn pairs by consulting reputable religious experts, as discussed in Section 7.4. These datasets are made available on GitHub [52]. Chapter 8, which is based on my publication (Altammami & Atwell, 2022), discusses the different experiments conducted and the results, which indicate that the current models' ability to reach a human-level understanding of Quran and Hadith remains somewhat elusive.

9.3 Limitations and Future Work

Further work is certainly required to address limitations and improve performance. The following points are ideas worth exploring.

- Building a Hadith segmentation tool is a challenging task due to the unique structure of Hadith and the difficulties in recognizing sentence boundaries in Arabic without strict punctuation rules and capitalization. Additionally, this is a domain-specific task that can be difficult for non-specialists. Moreover, some Hadiths contain several Isnad and Matns which makes it more appropriate to deconstruct the Hadith into finer-grained segments. This can be achieved by using sequence modeling methods, such as HMM, which have shown promising results as discussed in Section 4.7. Furthermore, NER tools, such as CAMel-tool (Obeid *et al.*, 2020), can also be used to identify names of narrators in the Isnad. This tool was not available at the time of building the Hadith segmenter, but it is worth exploring to segment Hadith to fine-grained components.

[52] https://github.com/ShathaTm/Quran_Hadith_Datasets

- The LK Hadith corpus can be extended to include Hadith commentaries aligned with the Hadith at the narrative level. Moreover, the translations of Hadiths in other languages could be incorporated.

- The evaluation of Quran ontologie's "fit" for Hadith can be extended by utilizing word embeddings as an alternative to WordNet discussed in Section 6.6. By using word embeddings, it is possible to identify and link concepts in the Quran to topics in Hadith that are semantically similar, even if they do not have an explicit lexical relationship. This can provide a more comprehensive and nuanced evaluation of the ontologies.

- When I presented the paper Altammami & Atwell (2022) at LREC 2022 in Marseilles, other researchers made excellent suggestions and comments. Unfortunately, time does not allow me to incorporate these in this thesis, but they are certainly worth exploring in future work. One of the ideas is regarding the dataset created in Chapter 7. The framework proposed in Section 7.4.1 to create the Quran-Hadith pairs can be tweaked to collect larger numbers of pairs. Instead of imposing a strict rule of finding a Fatwa with only one Quran verse and one Hadith, a Fatwa can have Quran and Hadith with a one-to-many relationship, hence increasing the number of samples. This data can then be used in fine-tuning the models only, while the 310 gold standard pairs is still used for testing. Another suggestion worth mentioning is to build a simple neural network trained on the QQ dataset as one of the baseline models to compare its performance to the transformer-based models.

- After the experiments in the thesis were conducted and finalized, other language models were created. It is worth testing these models, such as ARBERT and MARBERT (Abdul-Mageed *et al.*, 2021; Elmadany *et al.*, 2022).

- Future work can be improved by collaborating with Arabic and Islamic scholars to apply the research outcomes and resources in the fields of Arabic and Islamic studies, as well as using them to educate the general public in understanding the Quran and Hadith.

9.4 Research Contributions

My major research contributions are summarized in the following points:

- The first contribution is a novel Hadith segmentation tool that deconstructs the Hadith to Isnad and Matn with 92.5% accuracy.
- The second contribution is the first well-structured parallel corpus of the six canonical Hadith collections in their original CA form and English translations aligned at the Hadith level. This Leeds and KSU (LK) Hadith Corpus is freely available for the research community[53].
- The third contribution is the enumeration, comparison, and evaluation of Quran ontologies' "fit" for Hadith. The result identifies the most comprehensive Quran ontology that can be extended to form a larger Islamic ontology to cover the Hadith. This can be done by utilizing Hadith section-headings which contain the main concepts of the incorporated Hadiths. These section-headings can be accessed from the LK Hadith corpus.
- The fourth contribution is a framework for creating a datasets of Quran-Hadith pairs by extracting relatedness information from a reliable source of an archived Fatwa from an Islamic scholar. The Quran-Hadith pairs dataset is provided on my GitHub [54].
- The fifth contribution is the development, evaluation, and analysis of SOTA models performance in capturing the meaning of the Quran and Hadith texts.

9.5 Concluding Summary

In conclusion, this thesis has made significant contributions to the field of Classical Arabic NLP, specifically in the area of computational Hadith research. By

[53] https://github.com/ShathaTm/LK-Hadith-Corpus

[54] https://github.com/ShathaTm/Quran_Hadith_Datasets

utilizing AI methods, I have demonstrated the feasibility of annotating the components of Hadiths, the Isnad and the Matn. I have also created the Leeds and KSU (LK) Hadith corpus, addressing the scarcity of existing datasets in this field. Additionally, I have presented a dataset of Quran-Hadith pairs, along with a methodology for expanding them. Lastly, this thesis has explored the potential for detecting semantic similarity between Quran and Hadith using state-of-the-art AI tools and methods. While further research is needed, this thesis provides a solid foundation for future advancements in the field of computational Hadith research.

References

Abbas, N., Luluh, A., Al-Khalifa, H., Alqassem, Z., Atwell, E., Dukes, K., Sawalha, M. & Sharaf, A.B.M. (2013). Unifying linguistic annotations and ontologies for the Arabic Quran. In *Proceedings of WACL-2 Second Workshop on Arabic Corpus Linguistics*. 103

Abbas, N.H. (2009). Quran "search for a concept" tool and website. *The University of Leeds*. 104, 105, 106, 108, 110, 112, 148

Abdul-Mageed, M., Elmadany, A. *et al.* (2021). ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7088–7105. 170

Abdul-Raof, H. (2013). *Qur'an translation: Discourse, texture and exegesis*. London and New York: Routledge. 122, 147

Abdulrahim, D., Inoue, G., Shamsan, L., Khalifa, S. & Habash, N. (2022). The Bahrain Corpus: A Multi-genre Corpus of Bahraini Arabic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2345–2352. 166

Abouenour, L., Bouzoubaa, K. & Rosso, P. (2013). On the evaluation and improvement of Arabic WordNet coverage and usability. *Language resources and evaluation*, **47**, 891–917. 113

Ahmad, O., Hyder, I., Iqbal, R., Murad, M.A.A., Mustapha, A., Sharef, N.M. & Mansoor, M. (2013). A survey of searching and information extraction on a classical text using ontology-based semantics modeling: a case of Quran. *Life Science Journal*, **10**, 1370–1377. 105

Al-Aswadi, F.N., Chan, H.Y. & Gan, K.H. (2020). Automatic Ontology Construction from Text: A Review from Shallow to Deep-learning Trend. *Artificial Intelligence Review*, **53**, 3901–3928. 32

Al Faraby, S., Jasin, E.R.R., Kusumaningrum, A. *et al.* (2018). Classification of Hadith into positive suggestion, negative suggestion, and information. In *Journal of Physics: Conference Series*, vol. 971, IOP Publishing. 32

Al-Sanasleh, H.A. & Hammo, B.H. (2017). Building domain ontology: Experiences in developing the prophetic ontology from Quran and Hadith. In *2017 International Conference on New Trends in Computing Sciences (ICTCS)*, 223–228, IEEE. 33

Al-Thubaity, A.O. (2015). A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *In Language Resources and Evaluation*, **49**, 721–751. 85

Al-Yahya, M., Al-Khalifa, H., Bahanshal, A., Al-Odah, I. & Al-Helwah, N. (2010). An ontological model for representing semantic lexicons: An application on time nouns in the Holy Quran. *Arabian Journal for Science and Engineering*, **35**, 21. 105, 108

Aldhaln, K., Zeki, A., Zeki, A. & Alreshidi, H. (2012). Improving knowledge extraction of Hadith classifier using decision-tree algorithm. In *then International Conference on Information Retrieval & Knowledge Management*, 148–152, IEEE. 31

Alhaj, Y.A., Xiang, J., Zhao, D., Al-Qaness, M.A., Abd Elaziz, M. & Dahou, A. (2019). A study of the effects of stemming strategies on Arabic document classification. *IEEE Access*, **7**, 32664–32671. 47

Alharbi, A.I. & Lee, M. (2020). Combining character and word embeddings for affect in arabic informal social media microblogs. In *Natural Language Processing and Information Systems: 25th International Conference on Applications of Natural Language to Information Systems, NLDB 2020, Saarbrücken, Germany, June 24–26, 2020, Proceedings 25*, 213–224, Springer. 38

Alhawarat, M.O., Abdeljaber, H. & Hilal, A. (2021). Effect of stemming on text similarity for Arabic language at sentence level. *PeerJ Computer Science*, **7**, e530. 47

Alkhatib, M., Monem, A.A. & Shaalan, K. (2017). A Rich Arabic WordNet Resource for Al-Hadith Al-Shareef. *Procedia Computer Science*, **117**, 101–110. 32

Alosaimy, A. & Atwell, E. (2017). Sunnah Arabic Corpus: Design and Methodology. In *Proceedings of the 5th International Conference on Islamic Applications in Computer Science and Technologies (IMAN 2017)*. 86, 87

Alqahtani, M.M. & Atwell, E. (2018). Developing Bilingual Arabic-English Ontologies of Al-Quran. In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, 96–101, IEEE. 18, 31, 103, 104

Alrabiah, M., Al-Salman, A. & Atwell, E. (2013). The design and construction of the 50 million words KSUCCA. In *Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics*, 5–8. 85, 87

Alrehaili, S.M. & Atwell, E. (2018). Discovering Qur'anic Knowledge through AQD: Arabic Qur'anic Database, a Multiple Resources Annotation-level Search. In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, 102–107, IEEE. 31, 105

Alromima, W., Moawad, I.F., Elgohary, R. & Aref, M. (2015). Ontology-based model for Arabic lexicons: An application of the Place Nouns in the Holy Quran. In *11th International Computer Engineering Conference (ICENCO)*, 137–143, IEEE. 105, 108

Alsaleh, A.N., Atwell, E. & Altahhan, A. (2021). Quranic Verses Semantic Relatedness Using AraBERT. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 185–190. 140, 148, 155

Alshammeri, M., Atwell, E. & Alsalka, M.A. (2020). Quranic Topic Modelling Using Paragraph Vectors. In *Proceedings of SAI Intelligent Systems Conference*, 218–230, Springer. 148

Altammami, S. & Atwell, E. (2022). Challenging the Transformer-based models with a Classical Arabic dataset: Quran and Hadith. In *Proceedings of the Language Resources and Evaluation Conference (LREC'22)*, 1462–1471, European Language Resources Association, Marseille, France. 146, 169, 170

Altammami, S., Atwell, E. & Alsalka, A. (2019). Text Segmentation Using N-grams to Annotate Hadith Corpus. In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, 31–39. 40, 80, 167

Altammami, S., Atwell, E. & Alsalka, A. (2020a). Constructing a Bilingual Hadith Corpus Using a Segmentation Tool. In *Proceedings of The 12th Language Resources and Evaluation Conference LREC'20*, 3390–3398. 40, 80, 167

Altammami, S., Atwell, E. & Alsalka, A. (2020b). The Arabic-English parallel corpus of authentic Hadith. *International Journal on Islamic Applications in Computer Science And Technology*, **8**, 1–10. 84, 167

Altammami, S., Atwell, E. & Alsalka, A. (2021). Towards a Joint Ontology of Quran and Hadith. *International Journal on Islamic Applications in Computer Science And Technology*, **9**, 01–12. 140, 168

Alyafeai, Z., Masoud, M., Ghaleb, M. & Al-shaibani, M.S. (2021). Masader: Metadata sourcing for Arabic text and speech data resources. *arXiv preprint arXiv:2110.06744*. 35, 131

Antoun, W., Baly, F. & Hajj, H. (2020). AraBert: Transformer-based model for Arabic language understanding. *arXiv preprint arXiv:2003.00104*. 148, 155, 157

Atwell, E. (2018). *Arabic Corpus Linguistics: Using the Web to model Modern and Quranic Arabic*. Edinburgh: University Press. 33, 85

Atwell, E., Brierley, C., Dukes, K., Sawalha, M. & Sharaf, A.B. (2011). An Artificial Intelligence approach to Arabic and Islamic content on the internet. In *Proceedings of NITS 3rd National Information Technology Symposium*, 1–8. 1

Atwell, E.S. (1983). Constituent-likelihood grammar. *International Computer Archive of Modern and Medieval English Journal*, **7**, 34–67. 50

Azmi, A. & Badia, N.B. (2010). iTree-Automating the construction of the narration tree of Hadiths (Prophetic Traditions). In *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010)*, 1–7, IEEE. 32, 43, 45

Azmi, A.M., Al-Qabbany, A.O. & Hussain, A. (2019). Computational and natural language processing based studies of Hadith literature: A survey. *Artificial Intelligence Review*, 1–46. 2, 3, 26, 30, 84

Banerjee, S., Pedersen, T. *et al.* (2003). Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *International Joint Conferences on Artificial Intelligence*, vol. 3, 805–810. 37

Baraka, R.S. & Dalloul, Y. (2014). Building Hadith ontology to support the authenticity of Isnad. *International Journal on Islamic Applications in Computer Science And Technology*, **2**. 31

Bashir, M.H., Azmi, A.M., Nawaz, H., Zaghouani, W., Diab, M., Al-Fuqaha, A. & Qadir, J. (2022). Arabic natural language processing for Qur'anic research: A systematic review. *Artificial Intelligence Review*, 1–54. 31, 119, 132

Belinkov, Y., Magidow, A., Barrón-Cedeño, A., Shmidman, A. & Romanov, M. (2018). Studying the History of the Arabic Language: Language Technology and a Large-Scale Historical Corpus. *arXiv preprint arXiv:1809.03891*. 86, 87, 155

BENDER, E.M. & KOLLER, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. 147

BILAL, K. & MOHSIN, S. (2012). Muhadith: A cloud based distributed expert system for classification of Ahadith. In *2012 10th international conference on Frontiers of Information Technology*, 73–78, IEEE. 31

BINBESHR, F., KAMSIN, A. & MOHAMMED, M. (2021). A systematic review on Hadith authentication and classification methods. *Transactions on Asian and Low-Resource Language Information Processing*, **20**, 1–17. 31

BLECHER, J. (2016). *Hadith Commentary*. Oxford: University Press. 29

BOELLA, M. (2011). Regular expressions for interpreting and cross-referencing Hadith texts. *Langues et Littératures du Monde Arabe (LLMA)*, **9**, 25–39. 43, 45, 46, 49, 50

BOJANOWSKI, P., GRAVE, E., JOULIN, A. & MIKOLOV, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146. 38, 151

BOUNHAS, I. (2019). On the usage of a classical arabic corpus as a language resource: Related research and key challenges. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, **18**, 23. 2, 30, 31, 33, 84, 132, 166, 167

BREWSTER, C., ALANI, H., DASMAHAPATRA, S. & WILKS, Y. (2004). Data Driven Ontology Evaluation. In *International Conference on Language Resources and Evaluation (LREC'04)*. 106, 113

BROWN, J.A. (2017). *Hadith: Muhammad's legacy in the medieval and modern world*. Simon and Schuster. 2, 18, 25, 26, 27, 88, 109, 129

CER, D., DIAB, M., AGIRRE, E., LOPEZ-GAZPIO, I. & SPECIA, L. (2017). Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*. 2, 132

CHANDRASEKARAN, D. & MAGO, V. (2020). Domain Specific Complex Sentence (DSCS) Semantic Similarity Dataset. *arXiv preprint arXiv:2010.12637*. 147, 165

CHANDRASEKARAN, D. & MAGO, V. (2021). Evolution of Semantic Similarity—a Survey. *ACM Computing Surveys (CSUR)*, **54**, 1–37. 4, 36, 38

CHICCO, D. & JURMAN, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, **21**, 1–13. 150

CONNEAU, A., KHANDELWAL, K., GOYAL, N., CHAUDHARY, V., WENZEK, G., GUZMÁN, F., GRAVE, E., OTT, M., ZETTLEMOYER, L. & STOYANOV, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*. 8, 155, 157

DALLOUL, Y.M. (2013). An Ontology-Based Approach to Support the Process of Judging Hadith Isnad. *International Conference on Advanced Computer Science Applications and Technologies*, 1–108. 4, 31, 33

DARWISH, K., HABASH, N., ABBAS, M., AL-KHALIFA, H., AL-NATSHEH, H.T., BOUAMOR, H., BOUZOUBAA, K., CAVALLI-SFORZA, V., EL-BELTAGY, S.R., EL-HAJJ, W. *et al.* (2021). A Panoramic Survey of Natural Language Processing in the Arab World. *Communications of the ACM*, **64**, 72–81. 85

DEVLIN, J., CHANG, M.W., LEE, K. & TOUTANOVA, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. 8, 39, 155, 157, 162

DICKINS, J., HERVEY, S. & HIGGINS, I. (2016). *Thinking Arabic Translation: A course in translation method: Arabic to English*. London and New York: Routledge. 21

DUKES, K. (2015). Statistical parsing by machine learning from a classical Arabic treebank. *arXiv preprint arXiv:1510.07193*. 104, 105, 108

Duwairi, R. & El-Orfali, M. (2014). A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. *Journal of Information Science*, **40**, 501–513. 47

El-Haj, M., de Souza, E., Khallaf, N., Rayson, P. & Habash, N. (2022a). AraSAS: The Open Source Arabic Semantic Tagger. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, 23–31. 166

El-Haj, M., Zmandar, N., Rayson, P., Litvak, M., Pittaras, N., Giannakopoulos, G., Kosmopoulos, A., Carbajo-Coronado, B., Moreno-Sandoval, A. *et al.* (2022b). The Financial Narrative Summarisation Shared Task (FNS 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @ LREC2022*, 43–52. 1

Elangovan, A., He, J. & Verspoor, K. (2021). Memorization vs. generalization: Quantifying data leakage in nlp performance evaluation. *arXiv preprint arXiv:2102.01818*. xii, 142, 143, 147

Elmadany, A., Nagoudi, E.M.B. & Abdul-Mageed, M. (2022). ORCA: A Challenging Benchmark for Arabic Language Understanding. *arXiv preprint arXiv:2212.10758*. 170

Faqeer, H. (2017). A Survey of Qur'an Translations In English [1649 - 2014]. 122

Farghaly, A. & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, **8**, 14. 14

Ferguson, C.A. (1959). Diglossia. *Word*, **15**, 325–340. 12

Fischer, W. (2013). *The Semitic languages: Classical Arabic*. London and New York: Routledge. 14, 16

Ghazizadeh, M., Zahedi, M.H., Kahani, M. & Bidgoli, B.M. (2008). Fuzzy Expert System In Determining Hadith Validity. In *advances in computer and information sciences and engineering*, 354–359, Dordrecht: Springer. 31

Goldberg, Y. (2017). *Neural Network Methods in Natural Language Processing*. USA: Morgan & Claypool Publishers. 34

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning*. MIT Press, http://www.deeplearningbook.org. 1

Gruber, T.R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge acquisition*, **5**, 199–220. 103

Guellil, I., Saâdane, H., Azouaou, F., Gueni, B. & Nouvel, D. (2019). Arabic Natural Language Processing: An overview. *Journal of King Saud University-Computer and Information Sciences*. 2, 84, 85

Guellil, I., Saâdane, H., Azouaou, F., Gueni, B. & Nouvel, D. (2021). Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, **33**, 497–507. 3, 166

Habash, N., Diab, M. & Rambow, O. (2012a). Conventional Orthography for Dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 711–718. 17

Habash, N., Eskander, R. & Hawwari, A. (2012b). A morphological analyzer for Egyptian Arabic. In *Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology*, 1–9. 166

Habash, N., AbuOdeh, M., Taji, D., Faraj, R., Gizuli, J. & Kallas, O. (2022). Camel Treebank: An open multi-genre Arabic dependency treebank. In *Proceedings of the Language Resources and Evaluation Conference (LREC'22), Marseille, France*. 7, 100, 167

Habash, N.Y. (2010). *Introduction to Arabic natural language processing*. USA: Morgan & Claypool Publishers. 2, 12, 14, 17, 128, 147

HADJ TAIEB, M.A., ZESCH, T. & BEN AOUICHA, M. (2020). A Survey of Semantic Relatedness Evaluation Datasets and Procedures. *Artificial Intelligence Review*, **53**, 4407–4448. 36, 38

HAKAK, S., KAMSIN, A., ZADA KHAN, W., ZAKARI, A., IMRAN, M., BIN AHMAD, K. & AMIN GILKAR, G. (2022). Digital Hadith authentication: Recent advances, open challenges, and future directions. *Transactions on Emerging Telecommunications Technologies*, **33**, 3977. 31

HAKKOUM, A. & RAGHAY, S. (2016). Ontological approach for semantic modeling and querying. *International Journal on Islamic Applications in Computer Science and Technology*, 37–37. 31, 105, 106, 108, 110, 112, 113, 117, 140

HAMMO, B., YAGI, S., ISMAIL, O. & ABUSHARIAH, M. (2016). Exploring and exploiting a historical corpus for Arabic. *Language Resources and Evaluation*, **50**, 839–861. 85, 87

HAMOUD, B. & ATWELL, E. (2016). Quran question and answer corpus for data mining with WEKA. In *2016 Conference of Basic Sciences and Engineering Studies (SGCAC)*, 211–216, IEEE. 31

HARRAG, F. (2014). Text mining approach for knowledge extraction in Sahih Al-Bukhari. *Computers in Human Behavior*, **30**, 558–566. 43, 45, 46

HASAN, S. (1994). *An introduction to the science of Hadith*. London: Al-Quran Society. 31

HEARST, M.A. (1997). Text Tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, **23**, 33–64. 34

HENRY, S., BUCHAN, K., FILANNINO, M., STUBBS, A. & UZUNER, O. (2020). 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, **27**, 3–12. 1

HLOMANI, H. & STACEY, D. (2014). Approaches, Methods, Metrics, Measures, and Subjectivity in Ontology Evaluation: A survey. *Semantic Web Journal*, **1**, 1–11. 106

INOUE, G., ALHAFNI, B., BAIMUKAN, N., BOUAMOR, H. & HABASH, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*. 17, 155, 157

JAAFAR, A.H. & CHE PA, N. (2017). Hadith commentary repository: An ontological approach. *In 6th International Conference on Computing Informatics (ICOCI2017)*. 33

JBARA, K. (2010). Knowledge Discovery in Al-Hadith Using Text Classification Algorithm. *Journal of American Science*, **6**. 32

JURAFSKY, D. & MARTIN, J.H. (2020). Sequence labeling for parts of speech and named entities. *Speech and Language Processing*. 71

KENNEDY, G. (2014). *An Introduction to Corpus Linguistics*. London and New York: Routledge. 35

KHALIFA, S., HABASH, N., ERYANI, F., OBEID, O., ABDULRAHIM, D. & AL KAABI, M. (2018). A morphologically annotated corpus of Emirati Arabic. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC'18)*. 166

KHAN, H.U., SAQLAIN, S.M., SHOAIB, M. & SHER, M. (2013). Ontology based semantic search in Holy Quran. *International Journal of Future Computer and Communication*, **2**, 570. 105, 108

KHAN, S.H. (1987). *Al-Hitta Fi Dhikr Al-sihah Al-sitta*. Beirut: Dar Ammar. 26, 89

KIDWAI, A.R. (1987). Translating the untranslatable: A survey of English translations of the Quran. *The Muslim World Book Review*, **7**, 66–71. 148

KIESSLING, B., MILLER, M.T., MAXIM, G., SAVANT, S.B. *et al.* (2017). Important new developments in Arabographic optical character recognition (OCR). *The Journal of Middle East Medievalists*, **25**, 1. 86

LAHBIB, W., BOUNHAS, I. & ELAYEB, B. (2014). Arabic-English domain terminology extraction from aligned corpora. In *On the Move to Meaningful Internet Systems: OTM 2014 Conferences. OTM 2014. Lecture Notes in Computer Science, vol 8841.*, 745–759, Berlin, Heidelberg: Springer. 33

LAN, W., CHEN, Y., XU, W. & RITTER, A. (2020). An empirical study of pre-trained transformers for Arabic information extraction. *arXiv preprint arXiv:2004.14519*. 156, 157

LE, Q. & MIKOLOV, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, 1188–1196. 38, 148

LEECH, G. (1993). Corpus Annotation Schemes. *Literary and linguistic computing*, **8**, 275–281. 35

LEWIS, P., STENETORP, P. & RIEDEL, S. (2020). Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint arXiv:2008.02637*. 147

LUTHFI, E.T., SURYANA, N. & BASARI, A.H. (2018). Digital hadith authentication: A literature review and analysis. *Journal of Theoretical & Applied Information Technology*, **96**. 2, 4, 31, 41, 84

MAHMOOD, A., KHAN, H.U., ALARFAJ, F.K., RAMZAN, M. & ILYAS, M. (2018). A multilingual datasets repository of the Hadith content. *International Journal of Advanced Computer Science and Applications*, **9**, 165–172. 44, 45, 46, 86, 87

MANNING, C.D. (2015). *Computational linguistics and deep learning*, vol. 41. USA: MIT Press. 16

MARAOUI, H., HADDAR, K. & ROMARY, L. (2018). Segmentation tool for Hadith corpus to generate TEI encoding. In *International Conference on Advanced Intelligent Systems and Informatics*, 252–260, Cham: Springer. 43, 45, 46, 49, 50

MARELLI, M., MENINI, S., BARONI, M., BENTIVOGLI, L., BERNARDI, R., ZAMPARELLI, R. *et al.* (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *International Conference on Language Resources and Evaluation (LREC'14)*, 216–223, European Language Resources Association, Paris, France. 146

MCDANIEL, M. & STOREY, V.C. (2019). Evaluating domain ontologies: Clarification, classification, and challenges. *ACM Computing Surveys (CSUR)*, **52**, 1–44. 106

MCENERY, T. & WILSON, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press. 34

MIHALCEA, R., CORLEY, C., STRAPPARAVA, C. *et al.* (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Aaai*, vol. 6, 775–780. 38

MIKOLOV, T., CHEN, K., CORRADO, G. & DEAN, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. 38

MILLER, G.A. (1995). WordNet: a Lexical Database for English. *Communications of the ACM*, **38**, 39–41. 37

MOZANNAR, H., MAAMARY, E., EL HAJAL, K. & HAJJ, H. (2019). Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 108–118, Association for Computational Linguistics, Florence, Italy. 131

MUAZZAM SIDDIQUI, A.B., MOSTAFA SALEH (2014). Extraction and visualization of the chain of narrators from Hadiths using named entity recognition and classification. *International Journal of Computational Linguistics Research*, **5**, 14–25. 32

MUBARAK, H., DARWISH, K., MAGDY, W., ELSAYED, T. & AL-KHALIFA, H. (2020). Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on open-source arabic corpora and* *processing tools, with a shared task on offensive language detection*, 48–52. 1, 131

Nagoudi, E.M., Ferrero, J. & Schwab, D. (2017). LIM-LIG at SemEval-2017 Task1: Enhancing the semantic similarity for Arabic sentences with vectors weighting. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 134–138. 38, 151, 153, 157

Navigli, R. & Ponzetto, S.P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artificial intelligence*, **193**, 217–250. 37

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A. & Habash, N. (2020). CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th language resources and evaluation conference*, 7022–7032. 111, 120, 152, 169

Och, F.J. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**, 19–51. 33

Pak, I. & Teh, P.L. (2018). Text Segmentation Techniques: A Critical Review. *Innovative Computing, Optimization and Its Applications*, 167–181. 33

Pennington, J., Socher, R. & Manning, C.D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543. 38

Peuriekeu, Y.M., Noyum, V.D., Feudjio, C., Goktug, A. & Fokoue, E. (2021). A Text Mining Discovery of Similarities and Dissimilarities Among Sacred Scriptures. *arXiv preprint arXiv:2102.04421* . 148

Purver, M. (2011). Topic Segmentation . In *Spoken language understanding: systems for extracting semantic information from speech*, 291–317, New Jersey: Wiley. 34

QAHL, S.H.M. (2014). *An automatic similarity detection engine between sacred texts using text mining and similarity measures*. Rochester Institute of Technology. 148

RAAD, J. & CRUZ, C. (2015). A survey on ontology evaluation methods. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. xii, 106, 107

RADA, R., MILI, H., BICKNELL, E. & BLETTNER, M. (1989). Development and Application of a Metric on Semantic Nets. *IEEE transactions on systems, man, and cybernetics*, **19**, 17–30. 37

REIMERS, N. & GUREVYCH, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*. 156

REYNAR, J.C. (1998). *Topic Segmentation: Algorithms and Applications*. Ph.D. thesis, University of Pennsylvania. 34

RUST, P., PFEIFFER, J., VULIĆ, I., RUDER, S. & GUREVYCH, I. (2020). How good is your tokenizer? on the monolingual performance of multilingual language models. *arXiv preprint arXiv:2012.15613*. 162

SAAD, S., SALIM, N., ZAINAL, H. & MUDA, Z. (2011). A process for building domain ontology: An experience in developing salat ontology. In *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, 1–5, IEEE. 33, 108

SAEED, S. (2018). *Intraquranic Hermeneutics: Theories and Methods in Tafsīr of the Qurān through the Qurān*. Ph.D. thesis, SOAS University of London. 7, 21

SAEED, S., HAIDER, S. & RAJPUT, Q. (2020). On Finding Similar Verses from the Holy Quran using Word Embeddings. In *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, 1–6, IEEE. 147

Safaya, A., Abdullatif, M. & Yuret, D. (2020). Kuisail at semeval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2054–2059. 155, 157

Safeena, R. & Kammani, A. (2013). Quranic computation: A review of research and application. In *2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*, 203–208, IEEE. 31

Saloot, M.A., Idris, N., Mahmud, R., Ja'afar, S., Thorleuchter, D. & Gani, A. (2016). Hadith data mining and classification: A comparative analysis. *Artificial Intelligence Review*, **46**, 113–128. 4, 32, 41

Sayoud, H. (2012). Author discrimination between the Holy Quran and Prophet's statements. *Literary and Linguistic Computing*, **27**, 427–444. 109, 115, 119

Seelawi, H., Mustafa, A., Al-Bataineh, H., Farhan, W. & Al-Natsheh, H.T. (2019). NSURL-2019 Task 8: Semantic question similarity in Arabic. In *Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers*, 1–8. 39, 132

Shao, Y. (2017). Hcti at semeval-2017 task 1: Use Convolutional Neural Network to Evaluate Semantic Textual Similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 130–133. 39

Sharaf, A.B. & Atwell, E. (2012a). QurAna: Corpus of the Quran annotated with Pronominal Anaphora. *International Conference on Language Resources and Evaluation (LREC'12)*, 130–137. 104, 108

Sharaf, A.B. & Atwell, E. (2012b). QurSim: A corpus for evaluation of relatedness in short texts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2295–2302. 104, 108, 118, 131, 132, 140, 147

Sherif, M.A. & Ngonga Ngomo, A.C. (2015). Semantic Quran. *Semantic Web*, **6**, 339–345. 105, 108

Siddiqui, M.A., Saleh, M. & Bagais, A.A. (2014). Extraction and visualization of the chain of narrators from hadiths using named entity recognition and classification. *International Journal of Computational Linguistics Research*, **5**, 14–25. 41, 42, 45, 46, 50, 72

Solbiati, A., Heffernan, K., Damaskinos, G., Poddar, S., Modi, S. & Cali, J. (2021). Unsupervised topic segmentation of Meetings with BERT Embeddings. *arXiv preprint arXiv:2106.12978*. 34

Tarmom, T., Atwell, E. & Alsalka, M. (2019). Non-authentic Hadith corpus: Design and methodology. In *Proceedings of IMAN 2019*. 100

Tarmom, T., Atwell, E. & Alsalka, M. (2021). Deep Learning vs Compression-Based vs Traditional Machine Learning Classifiers to Detect Hadith Authenticity. In *Information Management and Big Data: 8th Annual International Conference, SIMBig 2021, Virtual Event, December 1–3, 2021, Proceedings*, Cham: Springer. 7, 100, 167

Tashtoush, Y.M., Al-Soud, M.R., AbuJazoh, R.M. & Al-Frehat, M. (2017). The noble Quran Arabic ontology: Domain ontological model and evaluation of human and social relations. In *2017 8th International Conference on Information and Communication Systems (ICICS)*, 40–45, IEEE. 105, 108

Vajjala, S., Majumder, B., Gupta, A. & Surana, H. (2020). *Practical natural language processing: A comprehensive guide to building real-world NLP systems*. USA: O'Reilly Media. 1

Varghese, N. & Punithavalli, M. (2019). Lexical and Semantic Analysis of Sacred Texts using Machine Learning and Natural Language Processing. *International Journal of Scientific & Technology Research*, **8**, 3133–3140. 148

Verma, M. (2017). Lexical analysis of religious texts using text mining and machine learning tools. *International Journal of Computer Applications*, **168**, 39–45. 148

Watson, J.C. (2007). *The phonology and morphology of Arabic*. Oxford: Oxford University Press. 18

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. *et al.* (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*. 162

Zaghouani, W. (2017). Critical survey of the freely available Arabic corpora. *arXiv preprint arXiv:1702.07835*. 86

Zerrouki, T. & Balla, A. (2017). Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems. *Data Brief*, **11**, 147. 85, 87