Deep Learning and Distributional Semantics for the Qur'an

Menwa Hayef K Alshammeri

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Computing

August, 2022

Declaration

I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. I was the lead author and the co-authors acted in an advisory capacity, providing supervision and review. Further details of the jointly-authored publications are listed below. I confirm that appropriate credit has been given within the thesis where reference has been made to the work of others.

(Menwa Hayed K Alshammeri)

The work in **Chpater 4** of the thesis is submitted for publication to the Journal of Quran and Tafseer Studies QiST.

The work in **Chapter 6** of the thesis is based on publication as follows:

Alshammeri M., Atwell E., Alsalka M.A. (2021) Quranic Topic Modelling Using Paragraph Vectors. In: Arai K., Kapoor S., Bhatia R. (eds) Intelligent Systems and Applications. IntelliSys 2020. Advances in Intelligent Systems and Computing, vol 1251. Springer, Cham. https://doi.org/10.1007/978-3-030-55187-2_19

The work in **Chapter 7** of the thesis has appeared in proceedings of the 5th International Conference on AI in Computational Linguistics (ACLing2021), and is based on publication as follows:

Menwa Alshammeri, Eric Atwell, Mhd ammar Alsalka, Detecting Semantic-based Similarity Between Verses of The Quran with Doc2vec, Procedia Computer Science, Volume 189, 2021, Pages 351-358, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.05.104.

The work in **Chapter 8** of the thesis has appeared in proceedings of the 18th International Conference on Natural Language Processing (ICON2021), and is based on publication as follows:

Alshammeri, M, Atwell, E. and Ammar Alsalka, M. Classifying Verses of the Quran using Doc2vec. In Proceedings of the 18th International Conference on Natural Language Processing (ICON2021, pages 284–288 Silchar, India. December 16 - 19, 2021

The work in **Chapter 9** of the thesis has been presented in the 9th International Conference on Islamic Applications in Computer Sciences and Technologies (IMAN2021), and is based on publication as follows:

Alshammeri, M., Atwell, E., & Alsalka, MA. (2022). A Siamese Transformer-based Architecture for Detecting Semantic Similarity in the Quran. International Journal On Islamic Applications In Computer Science And Technology, 9(4). Retrieved from http://www.sign-ific-ance.co.uk/index.php/IJASAT/article/view/2467

Alshammeri M, Atwell E, & Alsalka MA. 2021. A Siamese Transformer-based Architecture for Detecting Semantic Similarity in the Quran. The 9th International Conference on Islamic Applications in Computer Science and Technologies (IMAN 2021).

Dedication

أهدي هذه الأطروحة لذكرى والدي الحبيب، معلمي وملهمي الأول، المرحوم بإذن الله **هايف كايد الشمري**، أسأل الله أن يجازيه عني خير الجزاء وأن يسكنه الفردوس الأعلى من الجنة.

This dissertation is dedicated to the memory of my beloved father, Hayef Alshammeri, my first teacher and my inspiration to pursue my doctoral degree, who passed away before I reap the fruits of this research. May Allah reward him with Jannah.

Acknowledgements

First and foremost, I praise Almighty Allah for helping me to complete this dissertation and leverage computational systems in the service of the Holy Qur'an, and explore intrinsic knowledge.

I would like to express my sincere gratitude to my two PhD supervisors, Eric Atwell and Mhm Ammar Alsalka, at the University of Leeds, for their guidance, support, and encouragement.

I am deeply indebted to Dr. Eric for his thorough comments and insightful recommendations on this dissertation. His immense knowledge steered me in this research and resulted in high-quality publications. I benefited from his experience in the field of computational linguistics and the Arabic NLP. I extend my appreciation to Dr. Ammar for his generous guidance in key turn-points during my research journey. His advice and support have made this an inspiring experience for me. Many thanks to Eric and Ammar, who were always there whenever I needed a lift up. They were incredibly understanding of all the challenges I have been through, and I found a moral lift in their advice and encouragement.

I am also grateful for the support from Jouf University for sponsoring my research. I would also like to extend my thanks to Dr. Mustapha Sheikh, who reviewed my annual progress reports. Thanks for all the guidance, support, and outstanding feedback.

I would also like to acknowledge Abdul-Baquee Sharaf for Qursim, his valuable resource for the evaluation of relatedness in the Qur’an. I have used it extensively to train and evaluate my models during my research. Many thanks also go to Kais Dukes for his Qur'anic Arabic Corpus (QAC) and Noorhan Abbas for Qurany' Search for a topic' tool. I benefited a lot from their works during my research.

Finally, I'd like to express my gratitude to my family for their unwavering support throughout the compilation of this dissertation. I am grateful for my mother, who always shielded me with sincere prayers. She is my role model in life, and she is the source that gives me the strength and desire to continue after the loss of my father. A special thanks to my husband, Majed Alshammari, who supported me and helped me every step of the way. Thank you for being so patient, making the past four years much more enjoyable, and keeping me sane throughout the process. I want to give my deepest appreciation to my children; you are the joy of my life, made me smile, and motivated me to keep forward. Finally, I am grateful to my sisters and brothers, who have been consistent in their support, I couldn't have done it without all of you, and I'll never forget your contributions.

Abstract

This research presents an empirical framework for examining the semantic similarity task in the Qur'an, aiming to promote the acquisition of knowledge from the sacred text.

The framework employs recent breakthroughs in feature embedding to encode the verses of the Qur'an and get their embeddings, and then apply a semantic similarity metric to score the relationship between the Qur'anic verses. The framework utilizes deep learning models based on distributional semantics to achieve state-of-the-art semantic textual similarity task results.

Embeddings can be encoded using topic models (like LDA); to encode the vectors with the topic, or learned from encoding using neural networks, such as Word2vec, Doc2vec, and BERT. This research investigates a range of machine learning approaches to modelling semantics of Qur'an verses: Qur'an topic modelling and verse clustering using Latent Dirichlet Allocation; learning Qur'an word meanings using word2vec; learning distributed representations of Qur'an verse meanings using doc2vec; detecting semantic similarity between verses using doc2vec; classifying Qur'an verses using doc2vec; and deep learning of Qur'an verse meaning similarity using BERT and Siamese Transformer architecture.

The most successful novel contribution is the Siamese Transformer model of Qur'an verse similarity. The architecture exploits both the pre-trained contextualized representations for the Arabic language and the Siamese architecture to derive semantically meaningful verse embeddings and achieve impressive results in pairwise semantic similarity detection in the Qur'an. The F1 score of 95% on the Qur'anic semantic similarity test was impressively high.

Performance results obtained by the experiments are significant contributions of this research. The document vector approach proved to be more useful to retrieve the semantically close verses for a given verse. Classifiers and neural networks were trained on top of the derived vectors for classification, regression, and semantic similarity, yielding performance results that are comparable to or better than the reported ones.

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

{ هُوَ الَّذِي أَنزَلَ عَلَيْكَ الْكِتَابَ مِنْهُ آيَاتٌ مُّحْكَمَاتٌ هُنَّ أُمُّ الْكِتَابِ وَأُخَرُ مُتَشَابِهَاتٌ ۖ فَأَمَّا الَّذِينَ فِي قُلُوبِهِمْ زَيْغٌ فَيَتَّبِعُونَ مَا تَشَابَهَ مِنْهُ ابْتِغَاءَ الْفِتْنَةِ وَابْتِغَاءَ تَأْوِيلِهِ ۗ وَمَا يَعْلَمُ تَأْوِيلَهُ إِلَّا اللَّهُ ۗ وَالرَّاسِخُونَ فِي الْعِلْمِ يَقُولُونَ آمَنَّا بِهِ كُلٌّ مِّنْ عِندِ رَبِّنَا ۗ وَمَا يَذَّكَّرُ إِلَّا أُولُو الْأَلْبَابِ}

It is He who has sent down to you, [O Muhammad], the Book; in it are verses [that are] precise - they are the foundation of the Book - and others unspecific. As for those in whose hearts is deviation [from truth], they will follow that of it which is unspecific, seeking discord and seeking an interpretation [suitable to them]. And no one knows its [true] interpretation except Allah. But those firm in knowledge say, "We believe in it. All [of it] is from our Lord." And no one will be reminded except those of understanding.

The Qur’an, verse [3: 7]

Table of Contents

List of Tables

List of Figures

Chapter1
Introduction

The Qur’an is the scripture of the followers of Islam (Cook, 2000). The Qur’an was revealed to Prophet Muhammad by God through the agency of the angel Gabriel (Cook, 2000). The Qur'an was revealed in the Arabic language given its eloquence and articulateness, as the Almighty stated:

إِنَّآ أَنزَلۡنَٰهُ قُرۡءَٰنًا عَرَبِيّٗا لَّعَلَّكُمۡ تَعۡقِلُونَ ٢

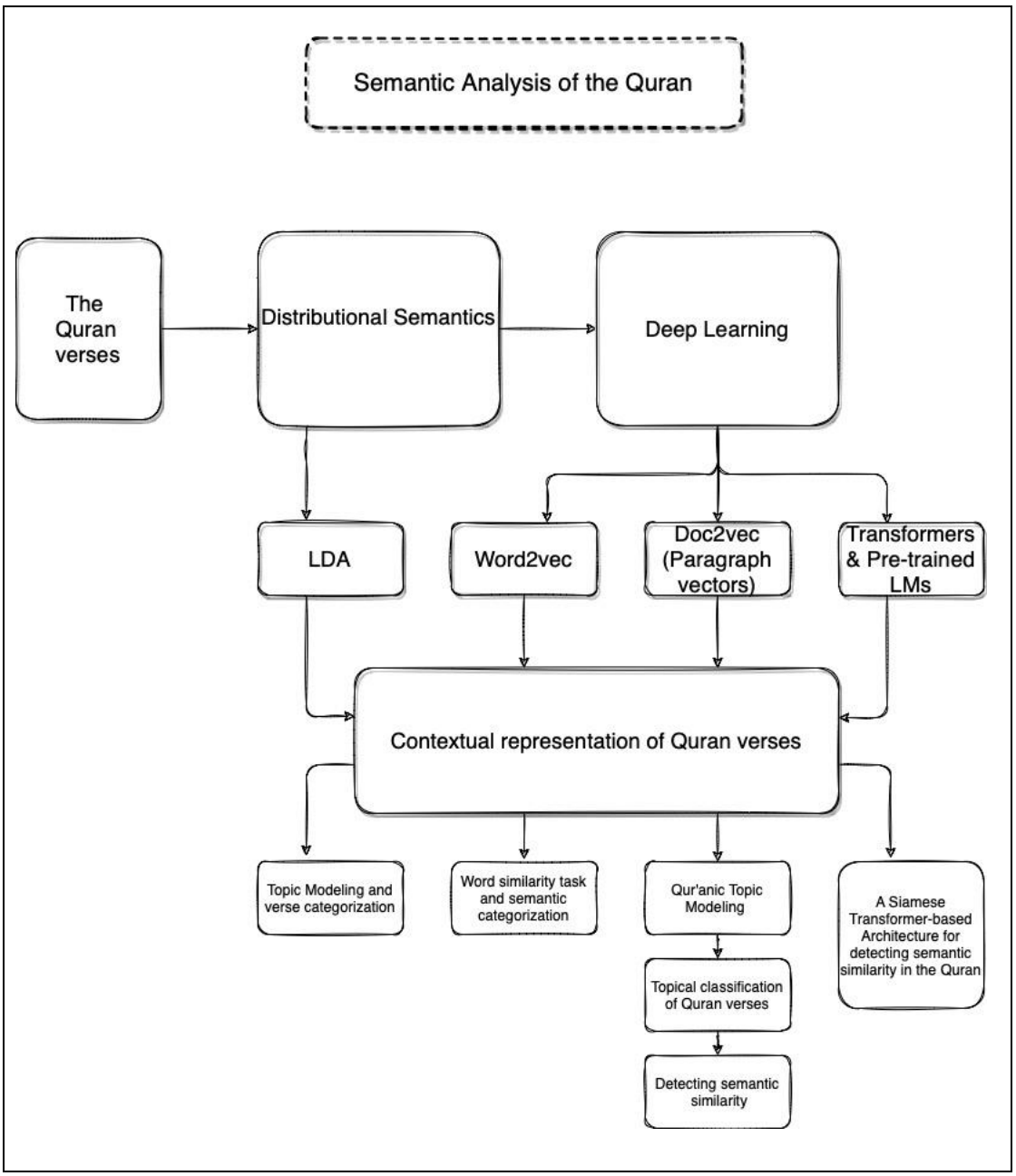
“I have sent it down as an Arabic Qur'an that you may be wise” [12:2]

The Qur'an is a significant source of knowledge and wisdom that regulates Muslims’ lives. Therefore, mining the content of the sacred text is a substantial task for scholars and researchers, particularly computer scientists. Their mission is to build intelligent systems that probe the knowledge enshrined in the Qur'anic verses and present it through computing systems to facilitate understanding and interpretation of the sacred text.

Analysing the Qur’anic text is not a trivial task due to the overlapping of meanings over its verses. Therefore, extracting the implied connections would require deep semantic analysis and domain knowledge. It also requires dedicated tools and software for the complex classical Arabic. With the tremendous advancements in the field of natural language processing (NLP), and in particular, the latest revolution of deep learning (DL) and the growing trend of transfer learning and pretrained language models, the task has become much more feasible.

This research work seeks to employ recent advancements in the field of AI and NLP to present novel computational approaches within the context of deep learning and distributional semantics for the semantic similarity task in the Qur’an to aid in mining the Qur’anic guidance and knowledge. The contributed resources and models would be extremely useful to classical Arabic research and the Arabic

NLP community as well. The diagram below is a roadmap to guide the reader throughout the dissertation.



1.1 This Dissertation

This dissertation presents an empirical framework for learning the semantic relations between the Qur'an verses. The framework employs computational methodologies based on distributional semantics and deep learning to achieve impressive results in the semantic similarity task. Figure 1 depicts the proposed framework.

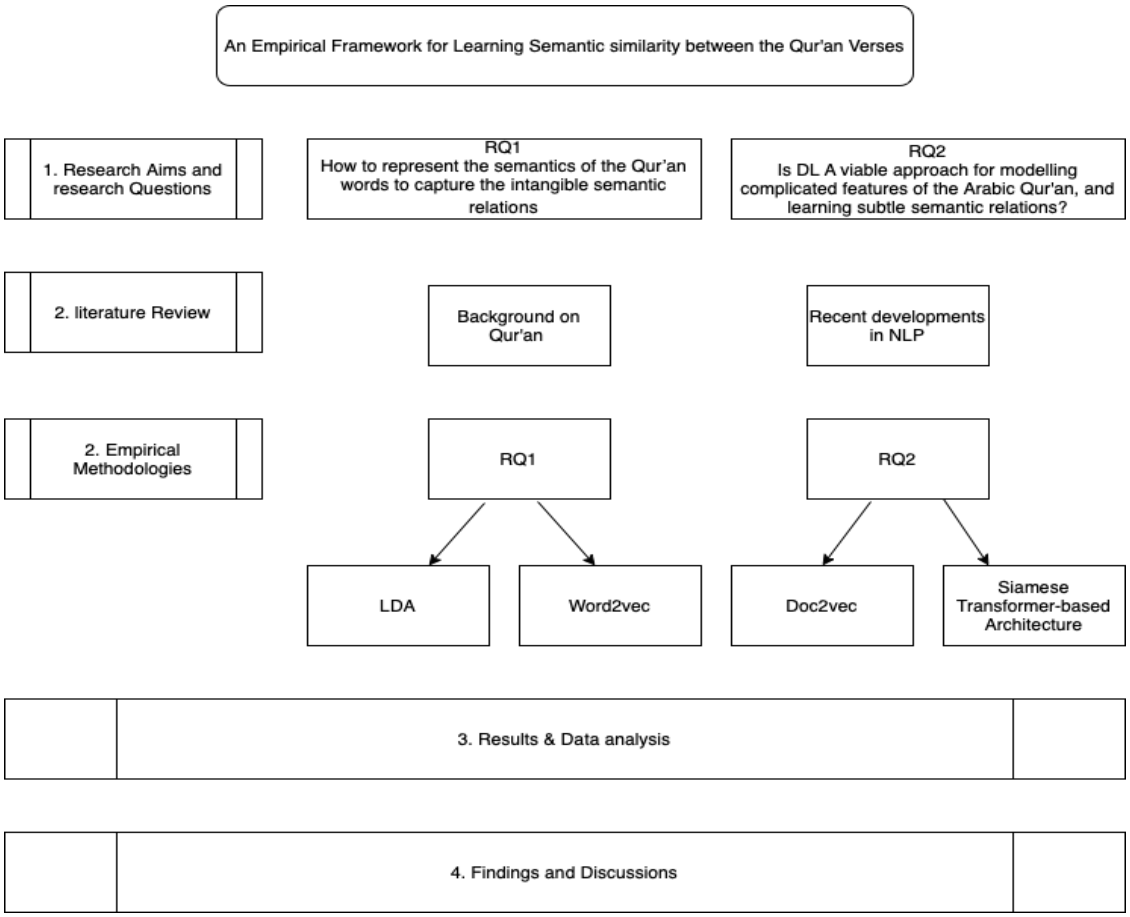


Figure 1: An Empirical Framework for learning semantic relations in the Qur'an

The dissertation presents a novel approach for detecting semantic similarity in the Qur'an, the Siamese transformer-based architecture, with impressive performance results. The dissertation also provides significant resources including machine learning tools for NLP, which leverage the power of distributional semantics using deep learning to model the complex Qur'anic properties. The outcome is a representation, of the Qur'anic verses, that is contextually relevant and semantically significant. Such representation allows for the discovery of semantic relations between the Qur'anic verses, achieving state-of-the-art results on the semantic similarity search and unsupervised clustering, compiling useful resources and innovative computational findings, expanding the index of Islamic teachings and helping exegetes and scholars who use the thematic approach to interpret and comprehend the Qur'anic text.

This research lays out the space of potential models, situates existing work in this space, and assesses which approaches appear most promising toward achieving the ultimate goal. This dissertation indeed provides corpora and software resources to a low-resource language like the classical Arabic of the Qur'an.

1.2 Motivation & Research Aims

This dissertation is driven by the goal of modelling the semantics of the Qur'an verses using powerful deep learning approaches such as sentence embeddings and transformers that are based on distributional semantics and then applying ML-based methods to achieve state-of-the-art semantic textual similarity task results. The main objective is to present an efficient approach for detecting the semantic similarity between the Qur'an verses.

Detecting semantic similarity in the Arabic language is a great challenge, and it is getting more complex in the case of classical Arabic. The Qur'an sets a unique example of classical Arabic in its utmost form. Most research into detecting semantic similarity/ relatedness has focused on English. Recent works have addressed the problem in the Arabic language, and most of them involves the Modern Standard Arabic (MSA) and Arabic Dialects (Mohamed et al., 2015; Habash et al., 2009; Al-Bataineh et al., 2019; Al-Theiabat and Al-Sadi, 2019). However, studying the semantic similarity in the Qur'an is an emerging topic and few works examined the semantic similarity in the Qur'anic text, some of them used the English translation of the Qur'an while others considered specific chapters of the text. One recent work by Alsaleh et al. (2021) used the pretrained language model AraBERT (Antoun et al., 2020) for detecting semantic similarity in the Qur'an, achieving an accuracy score of 92%.

This research emerged from a desire to further efforts and develop novel computational approaches for detecting the semantic similarity in the Qur'an and contribute to the sustained efforts to approach the Qur'an, in particular thematic and Modernist exegesis. We foresee the advantage of studying the semantic similarity in the Qur'an as a practical and useful approach for extracting the meanings and knowledge embedded within the sacred text. It has the potential of revealing the intrinsic structure behind the ancient book that the manual investigation could not have uncovered. It could also reveal some novel

computational results that aren't covered by any interpretation or exegesis, adding to the corpus of Islamic knowledge.

Thus, the main research aims are:

1. Presenting a semantic representation of the Qur'anic text that facilitates semantic analysis and detecting intangible semantic relations with high performance.
2. Presenting an empirical framework for learning semantic relations between the Qur'an verses that can act as a strong baseline for potential research.
3. Conducting experiments to evaluate my models on the semantic textual similarity task, and building suitable datasets needed for training and testing.
4. Contributing quality datasets to be used in the computational semantic analysis of the Qur'an.
5. Providing corpora and software resources to a low-resource language[1], the classical Arabic, and helping scholars and the public to comprehend the Arabic language in its utmost form like in the Qur'an.

1.3 Research Questions

1.3.1 How to represent the semantics of the Qur'an words to capture the intangible semantic relations?

Acquiring semantic knowledge and using it in language understanding and processing has always been an active area of study. Researches have resulted in various approaches and techniques related to semantics representation. Approaches to semantics representation can be classified into three main categories. They are semantic networks, feature-based models, and semantic spaces (Mitchell, 2010).

[1] In NLP, a Low-resource languages are languages for which statistical methods cannot be directly applied due to lack of data (Magueresse et al., 2020). The Arabic language is considered a low-resource language. This is primarily caused by the relative youth of research attempts to create natural language processing (NLP) tools for Arabic, the difficulty of handling its complicated morphology, and the scarcity of extensive Arabic opinion resources (Al-Sallab et al., 2017).

Probabilistic topic models (Blei et al., 2003; Griffiths et al., 2007), such as Latent Dirichlet Allocation (LDA), offer an alternative to semantic spaces based on the assumption that words observed in a corpus manifest some latent structure linked to topics. The words that have a high probability under the same topics will tend to be highly semantically related to each other (Griffiths et al., 2007). Probabilistic topic models represent each document as a mixture of latent topics in which a topic is a multinomial distribution over words. They are generative; they specify a probabilistic procedure by which documents can be generated (Blei et al., 2003; Griffiths et al., 2007).

Part of learning and using language is identifying the latent semantic structure responsible for generating a set of words. Probabilistic generative models provide solutions to this problem, making it possible to use powerful statistical learning to infer structured representations. The topic model is one instance of this approach and serves as a starting point for exploring how generative models can be used to address questions about human semantic representation.

Distributional representations, also known as distributional semantics, are studied in more traditional NLP as a more flexible way to express natural language semantics. Distributional semantics tries to use vectorial representations to characterise the meaning of words and sentences. Distributional semantics is based on the distributional hypothesis (Harris, 1954) – words have similar meaning when employed in similar settings. Words are represented by distributional vectors, which describe information about the contexts in which they appear. Sentence representations are created by merging vectors representing words. Distributional semantics is computationally capable of modelling what humans do when they make similarity judgements (Blei et al., 2003). Distributional semantics is the basis of sophisticated deep learning models.

In recent years, machine learning and in particular deep learning with neural networks has played a central role in corpus-based Natural Language Processing NLP (Goldberg, 2017). One of the most prominent developments in NLP is the use of machine learning to capture the distributional semantics of words, and in particular deep learning of word embeddings, where words are represented as

vectors in a continuous space, capturing many syntactic and semantic relations (Soliman et al., 2017). Word embeddings are significantly useful in capturing semantic relations among words, as a step towards presenting meaningful semantic structure of text. Use of deep learning word embeddings has led to outstanding advancements in semantic textual similarity tasks.

Thus, the research question that will be answered is how to present a semantic representation of the Qur'an words that capture the semantic similarity between its verses. This question will be answered by investigating different approaches to model the semantics of the Qur'an words. First, this research will examine the ability of a probabilistic approach, LDA, to model latent topics from a concise and short text like the Qur'an. Then we use a word embedding method, word2vec, to learn the semantic representation of the Qur'an words. To tackle this question, the following experiments will be conducted:

1. Qur'an topic modelling and verse clustering using Latent Dirichlet Allocation (Chapter 4)
2. Learning the semantic representation of Qur'an words using word2vec (Chapter5)

1.3.2 Is deep learning a viable approach for modelling the complicated features of the Arabic Qur'anic text, and learning subtle semantic relations?

NLP has reached a remarkable level of maturity for the English language. However, Arabic NLP still lags behind. This may be attributed to the richness and complex grammatical and syntactic patterns of the Arabic language. The most common and literary version of Arabic, Modern standard Arabic (MSA), is used for the majority of work on Arabic NLP. It is getting more challenging when it comes to the Arabic language in its utmost form like in the Qur'an (Bashir et al., 2021; Farghaly and Shaalan, 2009).

Due to the advancements in computational techniques, particularly in the field of NLP, we can employ natural language processing (NLP) techniques for Qur'anic research and create new applications that can aid those who are interested in learning and comprehending the Qur'anic message (Bashir et al., 2021). Turning

a new corner, NLP has been increasingly focused on using deep learning neural networks to solve many NLP problems (Brownlee, 2018).

Deep learning approaches toward understanding language are achieving state-of-the-art results in many tasks such as text classification, clustering, and semantic textual similarity (Goldberg, 2017; Goodfellow et al., 2016). Deep learning approaches are made up of multiple nonlinear transformations that combine to produce more abstract and eventually more useful representations. Representation learned from deep neural networks enable machine learning models to have an informative numerical representation of the input text (Goldberg, 2017).

Deep learning techniques have been incorporated into many systems and proved successful. Recently the field has witnessed a switch from linear models over sparse inputs to nonlinear neural network models over dense inputs. For years, shallow models (e.g., SVM and logistic regression) trained on high-dimensional and sparse data have been used in machine learning techniques for NLP tasks. In recent years, dense vector representation neural networks have surpassed traditional neural networks in a number of NLP tasks. Word embeddings have launched this movement (Young et al., 2018). Due to their outstanding success, word embeddings are close to replace the traditional Distributional Semantic Models. The impressive impact of these models has motivated researchers to consider vector representation to larger pieces of texts. Sentence embeddings, can be used to further integrate word semantic representations to determine the meaning of a wider portion of a text. Accordingly, Mikolov and Le (2014) have released a sentence embedding model, Doc2vec. It is another breakthrough in embeddings that maps sentences/documents to informative vector representations that preserve more semantic and syntactic information.

Moreover, the transformer architecture, in particular, has seen significant breakthrough in deep learning over the last few years. Recent pre-trained contextualized representations, such as ELMo (Peters et al., 2018) and a Bidirectional Encoder Representations from Transformers BERT (Devlin et al., 2019), have significantly increased performance across various NLP tasks, including semantic similarity detection. The Transformer is an encoder-decoder-based neural network that was first proposed in the paper Attention is all you

need (Vaswani et al.,2017). Its main characteristics include the use of so-called attention (i.e., a mechanism that determines the importance of words to other words in a sentence or which words are more likely to come together) and the absence of recurrent connections (or recurrent neural networks) to solve tasks that involve sequences. The BERT model was proposed by Devlin et al. (2019), and it is a way of learning representations of a language that makes use of a transformer, more specifically, the encoder portion of the transformer. Moreover, the NLP community accommodates a wide range of powerful components that are open-sourced and available to download and use in our models and pipelines, as well as pre-trained models that had already been pre-trained on large datasets (Devlin et al., 2018).

The primary research question that will be answered in this thesis is to determine whether or not deep learning models that are based on distributional semantics is a viable approach for modelling the semantics of the religious text in a way that aids the Qur'anic semantic analysis and achieves state-of-the-art semantic similarity accuracy. This thesis will leverage the power of recent advances in the field to present a semantic representation of the Qur'anic text capturing the semantic similarity between its passages. This question will be answered by using embeddings techniques and transformers to derive semantically meaningful and contextual representations of the Qur'an verses, enabling machine learning algorithms to achieve state-of-the-art results on topical classification, regression and semantic similarity. Thus, a range of machine learning experiments will be conducted to address this research question as the following:

1. Learning the semantic representation of Qur'an verses using Doc2vec (Chapter6), then training classifiers and neural networks on top of the derived vectors for classification (Chapter7) and semantic similarity (Chapter8).
2. Learning deep semantic similarity between Qur'an verses using a novel Siamese transformer-based architecture (Chapter9).

1.4 The Qur'an: A Spiritual and Linguistic Scripture

The Qur'an is one of the most extensively read books in the world every day. Millions of Muslims read it every day, whether in its original Arabic form or in

translated variants. The vocabulary, morphology, and syntactic patterns of the Qur'anic text are extensive. Its language is characterized by its elegance and unique style, which distinguishes it from other classical Arabic texts (Haleem, 2010). Some Qur'anic words are intended to have many contextual interpretations within the same passages in which they appear. The Qur'an uses the classical Arabic language to its maximum potential and explains concepts in the briefest yet most accurate of forms (Durakovic, 2005). In addition to the physical and spiritual unity that can be conveyed in terms of the rhythms and rhymes dominating any particular sura, the verses of the Qur'an are unified by the fact that they all serve in delivering the message of Islam to mankind (El-Awa, 2009).

Perhaps one of the essential things that draw the reader's attention to this miraculous text is the semantic similarity between its verses. The Qur'an contains several major themes, all of which revolve around the core theme of God's relationship with humans. Other prominent themes in the Qur'anic text include creation, early prophets, and life after death. It can be difficult to distinguish between the key themes because references to them emerge throughout the Qur'anic text. The Qur'an emphasises a different aspect of each of these themes in the specific phrasing of passages each time one of them is referenced (Saeed, 2008, Ali, 2017).

Also, the Qur'an depicts various relatedness between its texts. As Qur'anic verses vary in size, a pair of two large size verses might relate on a small phrase within these verses. Such verses can be divided into chunks with varying sizes gradually until an optimal size is reached so only meaningful phrases are preserved. The same strategy applies for the small verses that share a concept with adjacent verses (Sharaf and Atwell, 2012b).

1.5 Approaching the Qur'an

The Qur'anic text is not simply factual, but encodes subtle religious meanings. Much knowledge is encoded through the subtle use of words, syntax, allusions, relationships, and cross-references. Therefore, accessing the underlying knowledge, wisdom, and law needs interpretation and inference (Atwell et al., 2010).

Numerous studies have emerged on the Holy Qur'an and scholars have accumulated a far broader corpus of analyses, interpretations, and inference chains as a result. Computational analysis of the Qur’an aims to embody this knowledge, wisdom, and law in computer systems. The task for computer scientists is to construct intelligent systems that can answer any question using Qur’anic knowledge, and that can help Muslims and non-Muslims alike understand and appreciate the Qur’an (Atwell et al., 2010; Bashir et al., 2021; Atwell, 2018). Re-analysing the text content, extracting and capturing underlying knowledge in a Knowledge Representation and Reasoning formalism, and enabling automated, objective inference and querying are all potential prospects in Computer Science and Artificial Intelligence (Atwell et al., 2010).

Classical Arabic is the language of Holy Qur’an and Islamic documentary (Farghaly and Shaalan, 2009). Therefore, studies pertinent to the classical Arabic of the Qur’an have significance to promote the acquisition of knowledge from the sacred text; and provide a robust resource for religious scholars, educators, and the public to understand and learn the Qur’an. The Qur’an has been the subject of numerous studies due to its linguistic and spiritual value. Scholars have studied the Qur’an and drawn out knowledge and patterns that were the base for many applications to allow search in the holy book (Bashir et al., 2021; Atwell, 2018).

The study of semantic relationships between Qur'anic verses is a useful approach for gaining a balanced understanding of what the Qur'an says on any given topic. Studying the themes in the Qur’an would help the public, and also the scholars to understand the Muslim scripture. 'Topical presentation of the Qur'anic verses' has been proposed by a number of scholars, with the aim of providing the reader with a full and holistic comprehension of the Qur'anic Message.

In 1976, Fazlur Rahman recognised the urgent need for "an introduction to major themes of the Qur'an", and his book of the same name is a significant step forward in this direction. Moreover, Haleem (2010), presented his book entitled “Understanding the Qur’an: Themes and Style”, to further study in that direction. He explored some important themes of the Qur’an, with a focus on their literacy, figurative, and rhetorical aspects. According to Haleem (2010), to understand and interpret the Qur'anic text, Muslim scholars articulated the principle of "internal relationships." The concept is that some parts of the Qur'anic text clarify, strengthen, and amplify the meanings of other parts of the text.

Since its revelation, Muslims have been striving to understand the meanings and comprehend the divine message inherent in the sacred book. The Qur'an has been extensively studied by Muslims and non-Muslims (Rahman, 2009). The numerous Muslim commentaries on the Holy Book frequently interpret the text verse by verse. They are unable to provide insight into the Qur'an's clearly coherent outlook on the universe and life. Scholars have recently developed topical groupings of Qur'anic verses; while they can serve the scholar as a source or an index to varied degrees, they are of little use when trying to learn what the Qur'an has to say about specific concepts such as God, human, or society (Rahman, 2009). According to Rahman (2009), Major Themes of the Qur'an, the approach for synthesising themes is logical rather than chronological; except for a few key themes like religious variety, miracles' plausibility and reality, and Jihad, which all exhibit growth through the Qur'an.

Rather than focusing on individual verses, Thematic analysis is a systematic examination of the sacred text according to specific themes. Thematic exegesis emphasizes the unity of the Qur'an rather than the interpretation of individual verses. It gives significance to related verses on a particular concept throughout the Qur'an. In this approach, concepts and themes can be explored in depth by studying all aspects of the subject discussed in the Qur'an over different verses. Such an approach allows for a more objective assessment of the issues at hand. Thematic exegesis is claimed to be effective in today's life when dealing with modern issues like human rights, women's rights, and ethical dilemmas, to name a few (Saeed, 2008).

1.5.1 Semantic Similarity in the Qur'an

One interesting characteristic in the Qur'an as a text, is the scattering of a particular concept over many different verses within different chapters. The Qur'an uses the different themes to reinforce its message in various places. The different concepts in Qur'an such as: historical events, stories of prophets, attributes and qualities of Gods, and emphasis on a command, are mentioned at different times in different locations in the Qur'an; with more elaboration in meanings each time. Indeed, the overall subject is entirely covered (Sharaf & Atwell, 2012b). Moreover, the text encodes subtle religious meanings that are uncovered by direct and simple analysis. This property of the Qur'an made it the

right text for the purpose of analysing semantic relatedness between individual verses, or a group of verses of the Qur'an (Sharaf and Atwell, 2012b).

The Qur'anic text contains details and concepts that are scattered all over its passages. The text addresses several concepts in a novel manner, as one concept extends to cover more than one sentence; a verse. The same concept also may emerge in different places in the Qur'anic text. Hence, studying a subject must consider all related verses on that topic. Understanding the semantic relations between the Qur'an verses can facilitate extracting meanings and concepts and eventually presenting insightful knowledge. Hence, understanding the text and inferring the underlying meanings entail semantic similarity analysis.

Studying the semantic similarity in the Qur'an can facilitate extracting meanings and concepts and revealing the intrinsic structure behind the holy book that the manual investigation could not have uncovered. The semantic relatedness task is computationally complex and it requires deep semantic analysis and domain specific knowledge to relate the texts in a pair of verses (Sharaf and Atwell, 2012b).

Recently, the NLP community has witnessed the release of powerful developments, which set state-of-the-art results for various NLP tasks, including the semantic textual similarity. Hence, this research leverages the power of such models to examine the semantic similarity in the Qur'an, aiming to uncover overlooked places in the Qur'an by digging and scrutinizing the secrets of the semantics inherent in the sacred text. For example, Table 1 shows examples of semantically related verses derived as novel outcomes of this thesis, representing new computational results, not in any Tafsir interpretation or relevant knowledge resources.

30: 28	3: 118
{ ضرب لكم مثلا من أنفسكم هل لكم من ما ملكت أيمانكم من شركاء في ما رزقناكم فأنتم فيه سواء تخافونهم كخيفتكم أنفسكم كذلك نفصل الآيات لقوم يعقلون}	{ يا أيها الذين آمنوا لا تتخذوا بطانة من دونكم لا يألونكم خبالا ودوا ما عنتم قد بدت البغضاء من أفواههم وما تخفي صدورهم أكبر قد بينا لكم الآيات إن كنتم تعقلون}
He presents to you an example from yourselves. Do you have among those	O you who have believed, do not take as intimates those other than yourselves,

for they will not spare you [any] ruin. They wish you would have hardship. Hatred has already appeared from their mouths, and what their breasts conceal is greater. We have certainly made clear to you the signs, if you will use reason.	whom your right hands possess any partners in what we have provided for you so that you are equal therein [and] would fear them as your fear of one another [within a partnership]? Thus, do We detail the verses for a people who use reason.
8: 52	59: 15
{ كدأب آل فرعون والذين من قبلهم كفروا بآيات الله فأخذهم الله بذنوبهم إن الله قوي شديد العقاب} [Theirs is] like the custom of the people of Pharaoh and of those before them. They disbelieved in the signs of Allah, so Allah seized them for their sins. Indeed, Allah is powerful and severe in penalty.	{ كمثل الذين من قبلهم قريبا ذاقوا وبال أمرهم ولهم عذاب أليم} [Theirs is] like the example of those shortly before them: they tasted the bad consequence of their affair, and they will have a painful punishment.
80: 26	99: 2
{ ثُمَّ شَقَقْنَا الْأَرْضَ شَقًّا} Then We broke open the earth, splitting [it with sprouts].	{ وَأَخْرَجَتِ الْأَرْضُ أَثْقَالَهَا} And the earth discharges its burdens.
30: 35	26: 6
{ أَمْ أَنزَلْنَا عَلَيْهِمْ سُلْطَانًا فَهُوَ يَتَكَلَّمُ بِمَا كَانُوا بِهِ يُشْرِكُونَ} Or have We sent down to them an authority, and it speaks of what they were associating with Him?	{ فَقَدْ كَذَّبُوا فَسَيَأْتِيهِمْ أَنبَاءُ مَا كَانُوا بِهِ يَسْتَهْزِئُونَ } For they have already denied, but there will come to them the news of that which they used to ridicule.

Table 1: Examples of new findings related to semantic similarity in the Qur'an

1.6 Contributions

The novelty of this research lies in applying computational linguistics techniques to the classical Arabic of the Qur’an, one of the most distinctive and miraculous sacred texts in human history. This research uses NLP methods within the context of deep learning and distributional semantics to mine and probe the Qur’anic guidance and knowledge.

This dissertation presents a novel architecture within the context of deep learning and distributional semantics for the semantic similarity task in the Qur’an to aid in mining the Qur’anic knowledge. The result is a semantic representation of the Qur’anic verses that captures the semantic similarity with high accuracy.

This pioneering study has sparked the interest of other scholars, which is evidenced by frequent requests for cooperation from the NLP research community and inquiries pertaining to the published chapters. Indeed, this research has contributed to Qur'anic NLP in the following:

1 An empirical framework that encompasses computational methods and novel approaches for modelling semantic relations between the verses of the Qur'an, which should be easily reused and customized for relevant work. It utilizes recent advances in NLP and deep Learning to provide machine learning models with an informative numerical representation of the sacred text, and potentially allow for the discovery of knowledge-related connections between the Qur'anic verses.

2 The most successful novel contribution is the Siamese Transformer-based Architecture for detecting semantic similarity in the Qur'an, presented in chapter 9. This work was also presented at IMAN2021 and ResCompLeedsCon2022. The F1-score 95% on the Qur'anic semantic similarity test was impressively high.

3 A dataset is derived to cover subtle knowledge and deep semantic relations in Qur'anic Verses. The dataset constitutes pairs of related verses, supported with evidence and verified against acceptable sources; Tafsir books[2].

4 A collection of software resources and corpora, including the verses embedding, the developed language models, and the training datasets released to the research community.

The three main benefits of the thesis will also impact linguistics, AI and Islamic Studies, as resources for future research:

1 Annotated textual similarity datasets drawn from the Qur'an can be utilized as a linguistic resource for potential training and research.

2 A dataset of semantically related verses representing new computational results that are not in any Tafsir interpretation, thus adding to the catalogue

[2] Tafsir is the Arabic word of exegesis. The science of tafsir aims to explain the meanings in the Qur'an and is usually known as "Qur'anic interpretation" or "exegesis. Al-Tabari Tafsir and Jami' al-Bayan an Ta'wil al-Qur'an are examples of the most comprehensive work of tafsir (Ali, 2017).

of Islamic knowledge. The dataset can act as a significant resource for Islamic scholars to further analyse and explore embedded meanings and teachings.

3 The delivered knowledge can be used by exegetes and scholars who adopt the thematic approach for interpretation and understanding the Qur'anic text as they study the sacred text according to specific themes, rather than focusing on individual verses.

4 Finally, a potential future application can be developed to utilize these contributions and the derived knowledge to support understanding and interpreting the Qur'an.

1.7 Roadmap

This thesis is organized into five parts with 10 chapters.

Part I is an introduction and background, and includes chapters 1, 2, and 3.
Chapter 1 is an introductory chapter that outlines research aims, research questions and the main contributions of the dissertation
Chapter 2 is dedicated to the literature review, starting with a background introduction to the Qur'an, approaches to Qur'anic exegesis, and a detailed discussion on the unique structure of the sacred text, followed by relevant background on the Arabic language, Arabic Computational Linguistics, and Arabic NLP.
Chapter 3 reviews the text analytics methods for understanding texts. It introduces the concept of distributed representation, the basis of sophisticated deep learning models. It then provides a review of recent deep learning models and methods that have been employed for NLP tasks.

Part II examines the distributional semantics to model the semantic similarity in the Qur'an. It includes chapters 4 and 5.
Chapter 4 investigates a Qur'anic topic model based on the Latent Dirichlet Allocation algorithm, generating topics that could be semantically meaningful. In addition, LDA generates a distribution over topics for each document/ chapter in the sacred text.
Chapter 5 uses state-of-the-art word embedding methods, Word2Vec, for learning dense embeddings. The aim is to present a semantic representation that captures semantic similarity within the passages of the Qur'an.

Part III uses recent breakthroughs in feature embeddings and state-of-the-art deep learning models to generate a robust representation of the sacred text. It includes chapters 6, 7, and 8.
Chapter 6 uses a sentence embedding algorithm, Paragraph vectors, or Doc2vec, to map the Qur'anic verses to vectorized form.
Chapter 7 and **Chapter 8** describes ML experiments training classifiers and neural networks on top of the derived vectors for classification, regression, and semantic similarity.

Part IV presents the most successful novel contribution in the thesis which is the Siamese Transformer-based model for detecting semantic similarity in the Qur'an.
Chapter 9 exploits both the pre-trained contextualized representations for the Arabic language and the Siamese architecture to derive semantically meaningful verse embeddings and achieve impressive results in pairwise semantic similarity detection in the Qur'an.

Part V concludes the thesis in chapter 10.
Chapter 10 summarizes the main contributions and presents recommendations for future research. Finally, the chapter concludes with a discussion of the challenges and limitations of the work and its implications for future Qur'anic semantic similarity research and the study of exegesis in general.

Chapter 2
Literature Review: The Qur’an, Arabic language, and Computational Studies

This chapter provides a background introduction to the Qur’an as scripture, followed by a review of Qur’anic exegesis, its evolution, and trends. The chapter concludes with thoughts on the pressing need for the Qur’an interpretation.

2.1 The Qur’an

The Qur’an is the holy scripture of Islam from which Islamic ethics, law, and practise are drawn. It encompasses not only religious teachings, but also a way of life for millions of people (Saeed, 2008; Cook, 2000). The Qur'an is one of the most significant works of religious literature in world history. It is frequently likened to the Gospels and the Torah as Islam's holy scripture. The majority of Muslims believe that revelation is a divine undertaking through which God communicates His Will to humanity through chosen prophets. Muslims believe in a number of prophets, including Muhammad, who is regarded as the final recipient of divine revelation. Muslims believe Muhammad was a divinely inspired messenger, but he does not represent God's Being (Saeed, 2008; Cook, 2000).

2.1.1 Structure of the Qur’anic Text

The Qur'an is divided into 114 chapters (suras), each of which is different in length. Each chapter consists of a number of verses (ayas), each of which varies in length. Some verses are made up of various sentences, while others are made up of a single word or a short phrase. The Qur'an is normally organised according to the length of its chapters, with the exception of the first chapter, al-Fatiha (the Opening). The chapters of the Qur'an gradually get shorter, starting from the second chapter, al-Baqara (the Cow), which is the longest at 286 verses. As a result, the Qur'an's shortest chapters, 110, 108, and 103, all arrive near the finish and include only three verses apiece (Saeed, 2008; Cook, 2000; Ali, 2017).

The Qur'an is regarded by Muslims as the most perfect representation of the Arabic language; a one-of-a-kind piece of writing that, as the Qur'an itself asserts, can be matched by no human composition. This feature of the Qur'an, known as its 'inimitability' (i'jaz al-qur'an), has been the focus of major works by Muslim linguists, Qur'an interpreters, and literary critics. A number of Qur'anic verses support the idea of the Qur'an's inimitability, through which the Prophet Muhammad's opponents in Mecca were challenged to create a literary compilation equivalent to the Qur'an. These challenges were in reaction to claims made by the Prophet's opponents that the Qur'an was written by him rather than God (Abdul-Raof, 2013; Saeed, 2008; Ali, 2017). The Qur'an states that in many places:

{ قُل لَّئِنِ ٱجْتَمَعَتِ ٱلْإِنسُ وَٱلْجِنُّ عَلَىٰٓ أَن يَأْتُوا۟ بِمِثْلِ هَٰذَا ٱلْقُرْءَانِ لَا يَأْتُونَ بِمِثْلِهِۦ وَلَوْ كَانَ بَعْضُهُمْ لِبَعْضٍ ظَهِيرًا}

'Say: "Surely if men and jinns get together to produce the like of this (Qur'an), they will not be able to produce the like of it, however they might assist one another.' [17: 88][3]

{ أَمْ يَقُولُونَ ٱفْتَرَىٰهُ ۖ قُلْ فَأْتُوا۟ بِسُورَةٍ مِّثْلِهِۦ وَٱدْعُوا۟ مَنِ ٱسْتَطَعْتُم مِّن دُونِ ٱللَّهِ إِن كُنتُمْ صَٰدِقِينَ}

'Do they say (of the Prophet) that: "He has composed it?" Say to them: "Bring a Surah like this, and call anyone apart from God you can (to help you), if what you say is true."' [10:38][4]

It is claimed that the distinctive style and content of the Qur'an is considered as evidence of the Qur'an's inimitability. This argument revolves around the Qur'an's exceptional eloquence and unique language. The Qur'an's content, particularly its inclusion of historical information about earlier prophets and their communities that would have been impossible for anyone living during the Prophet's time to know, as well as the text's apparent lack of contradictions, are both seen as proof of the Qur'an's inimitability. Recently, scholars have approached the Qur'an's inimitability from scientific and mathematical perspectives, which are

[3] The translation by Ahmed Ali from the Tanzil project: https://tanzil.net/#trans/en.ahmedali/17:88

[4] https://tanzil.net/#trans/en.ahmedali/10:38

representative of some of the new ways that Muslims are striving to exhibit the 'truth' of the Qur'an today (Abdul-Raof, 2013; Ali, 2017).

2.1.2 The Context of the Qur’an

The revelation of the Qur'an took place in the large political, social, intellectual, and religious context of Arabia in the seventh century CE, specifically in the Hijaz area, which includes Mecca and Medina. Understanding the key components of this context allows us to connect the Qur'anic text to the environment in which it was written. This encompasses the region's spiritual, social, economic, political, and legal climates, as well as the norms, customs, institutions, and values that go along with them (Saeed, 2008).

Understanding the context of the Qur'an necessitates a thorough understanding of the Prophet's life events, both in Mecca and Medina. Many key events in the Prophet's life are referenced in the Qur'an, but not in detail. Therefore, an understanding of the Prophet's life and the events that occurred at the time is required to comprehend the relevance of numerous verses (Saeed, 2008).

2.1.3 Cultural and Ethical Language of the Qur’an

The cultural environment of Hijaz was a starting point for both the Qur'an and the Prophet in articulating the terms of the new religion forming in Mecca and Medina. The Hijaz people's manner of life and some aspects of their worldview were largely preserved. The Prophet's innovations were mostly in the areas of theology, spirituality, law, and ethical-morality. The Qur'an includes symbols, metaphors, terminologies, and expressions that were utilised in Hijaz, as well as its own culturally specific language appropriate to the worldview of its initial recipients (Saeed, 2008; Ali, 2017).

When we examine the Qur’anic text more closely, we typically discover that much of the Qur’an's language is primarily ethical. During the first three centuries of Islam, a focus on legal issues was necessary, as jurists sought an authoritative foundation for developing law and devising a jurisprudence system. This emphasis grew extreme, however, when plainly ethical texts were considered as solely legal, and the Qur'an's language and spirit were lost to increasingly strict legal interpretations (Saeed, 2008; Ali, 2017).

A good example is the position of women, which is addressed several times in the Qur'an. Distinctions based on gender and social class occurred in pre-Islamic and early Islamic societies. This is mirrored in the way women are referred to in certain Qur'anic texts. The Qur'an, on the other hand, did not establish gender discrimination as Islamic rule; in fact, it did the reverse. With Islam, the woman gained her right to inherit, and clear divorce guidelines were developed, giving women more rights than they had previously (Saeed, 2008).

Thus, some of the Qur'an's references to women may appear discriminatory nowadays, but they must be considered in the context of the complete Qur'an, as well as the cultural and societal values at the time of its revelation (Saeed, 2008). According to the Qur'an, the only distinction that matters to God among human beings is their piety, and women and men are judged equally in this regard, which is stated here:

> { إِنَّ ٱلْمُسْلِمِينَ وَٱلْمُسْلِمَـٰتِ وَٱلْمُؤْمِنِينَ وَٱلْمُؤْمِنَـٰتِ وَٱلْقَـٰنِتِينَ وَٱلْقَـٰنِتَـٰتِ وَٱلصَّـٰدِقِينَ وَٱلصَّـٰدِقَـٰتِ وَٱلصَّـٰبِرِينَ وَٱلصَّـٰبِرَٰتِ وَٱلْخَـٰشِعِينَ وَٱلْخَـٰشِعَـٰتِ وَٱلْمُتَصَدِّقِينَ وَٱلْمُتَصَدِّقَـٰتِ وَٱلصَّـٰٓئِمِينَ وَٱلصَّـٰٓئِمَـٰتِ وَٱلْحَـٰفِظِينَ فُرُوجَهُمْ وَٱلْحَـٰفِظَـٰتِ وَٱلذَّٰكِرِينَ ٱللَّهَ كَثِيرًا وَٱلذَّٰكِرَٰتِ أَعَدَّ ٱللَّهُ لَهُم مَّغْفِرَةً وَأَجْرًا عَظِيمًا }
>
> Verily men and women who have come to submission, men and women who are believers, men and women who are devout, truthful men and truthful women, men and women with endurance, men and women who are modest, men and women who give alms, men and women who observe fasting, men and women who guard their private parts, and those men and women who remember God a great deal, for them God has forgiveness and a great reward. [33: 35] (Translation from Tanzil project)

When individual verses are read in isolation, the Qur'an's position on women appears to be somewhat confusing. In most instances, it appears to treat both sexes equally, yet there are times when women's position appears to be lower than men. However, it is evident that the Qur'an and the Prophet's mission had the overall advantage of giving women in the Islamic era greater privileges than they had in pre-Islamic Arabia (Saeed, 2008; Ali, 2017; Abdul-Raof, 2013).

Women have always played a major role in Islamic scholarship, politics, and social life in Muslim societies. Many Muslim scholars' unfavourable attitudes and opinions regarding women are now being questioned by an increasing number of

Muslims, both men and women, who say that such attitudes do not reflect the entire Qur'anic message and must be reconsidered (Saeed, 2008).

2.2 Approaches to Qur’anic Exegesis

Since the seventh century CE, Qur'anic exegesis has been important to the intellectual development and practical implementation of Islam as a religion. Muslim scholars have created a variety of ideas and ways for approaching the Qur'an over time, all of which are intended to aid in the interpretation and understanding of its meaning. Some scholars have focused on reading the Qur'an according to the Qur'an itself, or as it was interpreted by the Prophet and the first Muslims, while others have emphasised on the use of independent reasoning and individual scholars' capability to extract meaning from the text. All such efforts have been accompanied by a long history of debate and discussion (Ali, 2017; Abdul-Raof, 2013).

Approaches for interpreting the Qur’an have evolved constantly over time. The 'textualist' and 'contextualist' approaches are two of the many diverse approaches that are often referred to, which stress the relevance of the ethical and legal content of the Qur’an to contemporary Muslim’s life (Saeed, 2008). The textualist method is still the most popular among Muslim Qur'an interpreters today, especially among Sunni Muslims. Proponents of this method typically engage in linguistic analyses of texts such as the Qur'an and Hadith in order to grasp the Qur'an's meanings, which are frequently thought to be fixed and unchanging over time. Textualists advocate for a strict adherence to the text in order to preserve the traditional interpretation of the Qur’an's ethical legal content.

On the other hand, the contextualist approach is beginning to gain traction in the current day. Contextualists argue for understanding the ethical legal content in light of the social, cultural, and political conditions at the time of revelation in order to understand the Qur'an's meanings, of which the essence is assumed to be unchangeable (Saeed, 2008).

2.2.1 A Holistic View on Qur’anic Exegesis (Tafsir/ Interpretation): its Origins, Evolution, and Trends

Without addition or modification, the Prophet faithfully delivered what had been "dictated" to him by God in the Arabic language through the angel of revelation,

typically designated as Gabriel. The Prophet's "received" revelations were passed down precisely to his followers, who in turn passed them on to following generations. This view of the revelation has been maintained throughout Islamic history, and it is the foundation of most Islamic exegetical studies (Ali, 2017; Abdul-Raof, 2013).

The Prophet's mission, according to the Qur'an, included assisting in the explanation of the Qur'an's meanings. The Prophet accomplished this through both words and actions, but primarily through his actions. According to history, he only verbally interpreted parts of the Qur'an to his followers (Saeed, 2008).

2.2.1.1 Early Exegesis

An in-depth explanation of the entire book would not have been necessary because most of the Companions[5] spoke Arabic and were familiar with the Qur'an's wider context and contents. However, there was a need for some explanation of verses that articulated new concepts or employed pre-Islamic words in novel ways, or where there were language challenges, especially for those unfamiliar with Mecca's Arabic dialect. At the Prophet's era and the Companions, there were primitive attempts to interpret the Qur'anic text. For example, the Companions relied on the text itself to find the meaning of a verse in another verse and referred to the Prophet's sayings. They also used diligence and referred to the pre-Islamic literature and customs common in the pre-Islamic era and the era of the beginning of Islam (Saeed, 2008; Abdul-Raof, 2013).

The Prophet interprets the Qur'an in Hadith and sometimes in practical exegesis. Much of the Prophet's interpretation exists in the form of 'practical exegesis'. Practical exegesis is when the Qur'an uses a term or concept that the Prophet subsequently demonstrates via his actions, rather than explaining it in the form of Hadith. The Prophet's thorough demonstration of how to perform the five daily prayers is an example of this. Fortunately, much of this 'practical exegesis' was preserved in the Companions' memories and community practise. A significant portion of this was later documented in Hadith literature. The Prophet's time should be considered the most affluent period of exegetical effort through

[5] The immediate followers of the prophet Muhammad are known as Companions, examples are Abu Bakr as-Siddiq and Umar ibn al-Khattab.

practice, based on the enormous amount of information included in the Hadith. (Saeed, 2008; Abdul-Raof, 2013).

After the death of the Prophet, the Qur’an was in its final and complete form, and it was compiled ("collected") during the caliphate of the third caliph, Uthman (Saeed, 2008; Ali, 2017). Among the Prophet's Companions, some of them contributed directly to Qur’anic exegesis. They are the first four caliphs, Abu Bakr, Umar, Uthman and, as well as Aishah, the Prophet's wife. In addition, many of the Companions made significant attempts to explain the Qur'an in their time; among them are Abd Allah ibn Masoud, Ubay ibn K’aab, Abd Allah ibn Abbas, and Zayd ibn Thabit were among them. The most famous is Abd Allah ibn Abbas, renowned as the "Interpreter of the Qur’an." In addition, Abd Allah ibn Masoud is credited with many exegetical traditions. Indeed, the abundance of sound exegetical Hadith from most Companions suggests that there was no urgent need for a large-scale "explanatory effort" during their time. (Saeed, 2008; Ali, 2017).

For understanding and interpreting the Qur’an, the Companions used different approaches to interpreting the Qur’an: they explained sections of the text using other parts of it, oral and practice knowledge from the Prophet, and their understanding of the Qur’an. They also knew the Qur’an's language, the context of the revelation, the Prophet's ways of thinking, and Arab norms, values, and practices, all of which gave them a unique foundation for understanding the Qur’anic text within the context of the community's evolving "established practice." The People of the Book (Jews and Christians, or Ahl al-Kitab) traditions were the last source, especially regarding the Qur’an's narratives about prior prophets, peoples, and events. Many Companions referred to testimonies by converts to Islam, particularly Abd Allah ibn Sallam and Kab al-Ahbr, both were Jewish because the Qur’an only referred to these narratives briefly (Ali, 2017; Abdul-Raof, 2013).

In the seventh century, the Islamic empire expanded, creating new social contexts and further inquiries. In addition, the need for interpretation increased with the second generation of Muslims, known as "successors," from different backgrounds, including children of Arab Companions, Arabic-speaking converts to Islam, and non-Arabic speaking converts to Islam from other religions. Such

diversity and the wide gap between their era and the Prophet's created the demanding need to handle exegesis issues (Saeed, 2008; Ali, 2017).

Thus, exegesis began to take shape on a greater scale, primarily in Mecca, Medina, Damascus, Yemen, and Iraq, but it was still informal. Also, material from other religions began to enter the discourse of exegesis via converts to Islam. Exegetes and storytellers then wanted to fill out details pertinent to the Qur'an's narratives about past prophets. Hence, many people began to turn to the Qur'an and its interpretation for guidance in the face of these new societal contexts (Saeed, 2008; Ali, 2017).

In the eighth century, various disciplines have appeared and provided further support to the emerging tradition of exegesis (Tafsir). Hadith, Arabic Linguistics, Literature, and Qira'at have contributed to the tradition of exegesis in different ways. In addition, a rudimentary version of Qur'anic exegesis was forming. The social and political environment of the time influenced all of these fields to differing degrees as they emerged together. Early exegesis (dating back to the Prophet and his Companions) was essentially oral and depended on oral transmission; subsequently, written exegesis emerged. They had begun to appear by the early second/ eighth century (Wielandt, 2002; Saeed, 2008).

These exegetical writings were not comprehensive Qur'anic commentaries; rather, they were the start of a process of documenting Qur'anic exegesis. These works began with brief explanatory comments on terms and expressions that were obscure, difficult, or ambiguous. They also reviewed legal and ritual issues like as how to pray, calculate zakat (alms), and conduct the pilgrimage, as well as some of the Qur'an's commands and prohibitions. These early exegetical writings aimed to provide explanations and fill in the gaps based on the Prophet's example and the experience of the earliest Muslims in issues where the Qur'an merely supplied a broad guidance, such as doing daily prayers (Ali, 2017; Saeed, 2008).

2.2.1.2 Trends of Exegesis

Within Islam, several schools of thought arose in the ninth centuries. Sunni, Shi'a, and Khariji religious-political groups had established diverse approaches to legal and theological issues, as well as Qur'anic exegesis (Saeed, 2008; Ali, 2017).

A number of other major kinds of exegesis evolved in the first three centuries of Islam, in addition to the general religio-political developments outlined above. Exegetes who worked in these fields (theological, legal, mystical, and philosophical exegesis) were typically affiliated with one of the three major religious groups (Sunni, Shi'a, or Khariji). Numerous modern forms of exegesis have evolved in the contemporary time, while all of the aforementioned forms of exegesis are still studied and employed. Many individuals, including Muslims, are seeking a balance between traditional and modern perspectives of life in reaction to global advances in areas as diverse as politics, the environment, and ethics (Saeed, 2008). Figure 2 illustrates the different trends in the modern era.

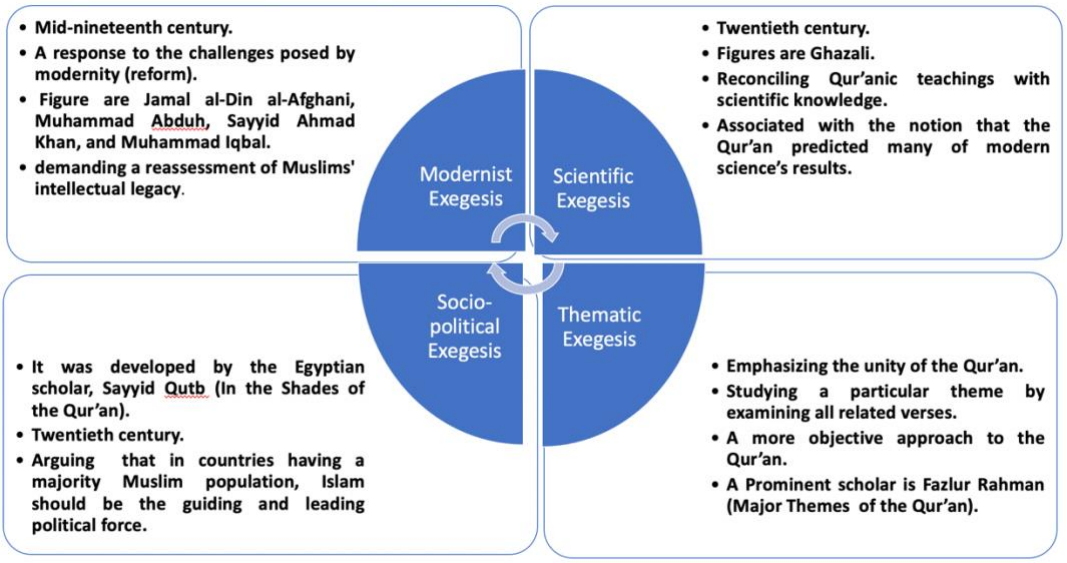


Figure 2: Exegesis Trends in the modern time

2.2.2 Modern Interpretation of the Qur'an

An alternative, contextualist approach is beginning to gain traction in the current day. Proponents of this method say that textual study must be supported by understanding of the social, cultural, and political conditions at the time of revelation in order to understand the Qur'an's meanings, of which the essence is

assumed to be unchangeable. They primarily engage in linguistic examinations of sources such as the Qur'an and Hadith in order to understand the meanings of the Qur'an (Saeed, 2008; Ali, 2017).

As a result, modern exegetes have not made a significant addition to Qur'anic exegesis in terms of a new school that takes a creative approach to exegesis that differs from the well-established ways. Modern exegetes, on the other hand, are affected by contemporary socio-political requirements and scientific discoveries, resulting in the establishment of an exegetical school that considers modern scientific, medical, social, and political changes (Abdul-Raof, 2013; Ali, 2017).

2.2.2.1 Fazlur Rahman and Modernity

Fazlur Rahman is regarded as one of the most influential reforming scholars of the twentieth century. He is best known for his significant contribution to modern Islamic reform discourse (Saeed, 2004). Fazlur Rahman is an educationist who believes that a particular type of education is required for Muslims to successfully blend the essence of their faith with modern practises and institutions. A unique and personal study of the Qur'an was his strategy for integrating intellectual thought and religious commitment (Rahman, 1989). This endeavour was to be a source of inspiration for many of his students and followers, as well as a source of innovation for him (Panjwani, 2012).

Fazlur Rahman has established himself as one of the most respected Muslim philosophers in the United States, both within and outside of academia. He wrote on a wide range of subjects, including Islamic education, Qur'anic studies, law, and historical and philosophical subjects. 'Islam and Modernity: Transformation of an Intellectual Tradition', his most important work on Islamic education, was published in 1982 (Panjwani, 2012). The book's defining characteristic is the author's proposed new methodology for reinterpreting the Qur'an. According to Fazlur Rahman, Muslims' stagnation is due to their failure to develop an effective methodology for interpreting the Qur'an (Husain, 1983).

The methodology is called a "double movement theory", which operates as follows: from the present situation to Qur'anic periods, then back to the present situation (Rahman, 1985). Rahman aims to find the underlying moral values that the distinct verses of the Qur'an were supposed to attain by analysing the seventh century socioeconomic and moral context in which the various provisions of the

Qur'an were revealed. These fundamental moral principles, according to the author, are everlasting, despite the fact that the particular verses of the Qur'an from which the moral values might be drawn are not (Husain, 1983). Then, returning to the present, he employs these value criteria to judge if a specific social or economic regulation is in accordance with the Qur'an. If a law meets these criteria, it is said to be compatible with the Qur'an, even if the literal meaning of a verse dealing with the same law may lead to a different outcome (Husain, 1983). Rahman has established himself as one of the few Islamic scholars in the world today who has a deep understanding of the Qur'an as an internally consistent, open-ended process of socio-political and moral norms, thanks to his use of the double movement methodology (Husain, 1983).

In summation, Rahman's primary contribution to the twentieth-century discussion on Islam was his statement that Muslims must move away from reductionist and formulaic methods to understanding the Qur'an that ignore the Qur'an's social, historical, and linguistic context. One of the fundamental aims of the Qur'an, Rahman felt, was to create a society based on justice (Saeed, 2004; Panjwani, 2012). In his views, Islam as a religion, and the Qur'anic teachings in particular, are "oriented toward the building of positive and effective equality among human beings." As a result, the Islamic goal cannot be achieved unless genuine human freedom is restored and independence from all forms of exploitation – social, spiritual, political, and economic – is guaranteed (Panjwani, 2012).

This research goes in line with the overall viewpoint of Rahman as to the significance of the Qur’an as a comprehensive approach to life and the necessity to gain a deep understanding of its contents. Indeed, there is an "urgent necessity" to establish a coherent Islamic worldview, which can only be accomplished by returning to the Qur’an.

2.3 Summary and Conclusion

This chapter presented a background introduction to the Qur’an as a scripture, its structure, and context. It then gave an overview of the different approaches to Qur’anic exegesis, offering a Holistic View of Qur’anic Exegesis (Tafsir/ Interpretation): its origins, evolution, and trends.

Like many ancient texts that were meant for guidance to humanity, the Qur’an can be a book of inspiration. Over one-fifth of the world's population relies on the Qur'an for guidance. It is the primary source from which Islamic ethics, law, and practises are drawn as the holy scripture of Islam. It encompasses not only religious beliefs and ideas, but also a way of life for millions of people (Saeed, 2004; 2008; Haleem, 2010). Therefore, the interpretation of the Qur’an has become one of Islam's most renowned subjects. Given that the early Muslims' lives revolved around the Qur’an from the beginning, one of their priorities was to understand the sacred text's meanings.

The Qur'an is a complicated text with many different interpretations. Scholars and lay people alike have discussed the meanings and interpretations of the Qur'an from the time of the Prophet Muhammad (d.11/632) to the current age. The discipline of Qur’an interpretation developed through time, leading to the emergence of numerous interpretation schools and trends (Ali, 2017). The literature on Qur’an interpretation in the modern era shows that there is an intense desire on Muslims, scholars, and thinkers to find the relevance of the Qur’anic text to contemporary issues without compromising the Qur’anic value system and its essential and core beliefs and practices (Saeed, 2004; 2008; Abdul-Raof, 2013).

Muslim laymen and scholars have consistently used their Arabic skills to interpret the Qur'an, both formally and informally (Ali, 2017; Abdul-Raof, 2013). They have also relied on historical knowledge of the time of the Qur'an's revelation as well as the Prophet Muhammad's person to comprehend the context in which the Qur'an was revealed and on historical accounts of the Prophet's interpretation of the Qur'an, as well as early Muslims' and other scholars' accounts throughout Islam's history. As a result, in the twenty-first century, there exist a wealth of resources to draw from while striving to understand the meanings of the Qur'an and Islam (Saeed, 2008). Given contemporary challenges and continual inclusion of Islam and Muslims in global debates, there has never been a more pressing need for individuals of all backgrounds to develop a more balanced and comprehensive understanding of Islam. A fundamental comprehension of the Qur'an becomes all the more important in this situation.

In recent years, Muslim scholars have been debating current issues and evaluating what the Qur'an has to say about them. War, marriage, and tolerance in Islam are among the persistent issues that Muslims believe have been misunderstood by many Western scholars, and they believe that the Qur'an has yet to be fully explored on these themes (Haleem, 2010). The Qur'an is a text that primarily strives to deliver a certain message to humanity. It is a text that requires every effort to explain. Any method used to gain a deeper knowledge of the text should be legal as long as it follows scientific research principles (Haleem, 2010).

Chapter3
Literature Review: Distributional Semantics and Deep learning for Understanding Text

This chapter provides the reader with background information relevant this research. The chapter first introduces important concepts such as AI, NLP, Machine learning, Deep learning, and transfer learning. It then reviews AI methodologies and techniques used for understanding text, Arabic NLP, and NLP for the Qur'an. Next, the chapter summarizes the well-established approaches to semantic representation, followed by a detailed overview of the topics of distributed and distributional representations for NLP. Finally, it summarizes recent deep learning trends that have been used for solving NLP tasks.

3.1 NLP Review on Text Corpus Analytics Methods

Text mining is a field that aims to extract meaningful information from unstructured text data by identifying and exploring interesting patterns (Feldman and Sanger, 2007). There are numerous approaches to the identification of global patterns in text, using different learning representations for texts. This can be done through various techniques such as: Natural Language Processing (NLP), Computational Linguistics (CL), and numerical algorithms such in Machine Learning (ML) (Srinivasa-Desikan, 2018).

3.1.1 Natural Language Processing

Text mining uses natural language processing (NLP) to allow machines to understand and analyse human language automatically (Deng and Liu, 2018). Natural Language Processing is a branch of computer science, particularly Artificial intelligence (AI). AI is a broad phrase that refers to machines that can mimic human intelligence; it includes systems that simulate cognitive abilities, such as learning from examples and solving problems. The branch of AI known as Natural Language Processing (NLP) explores how machines interact with human language (Deng and Liu, 2018). NLP enables computers to perform a wide range of natural language related tasks at all levels, ranging from parsing

and part-of-speech (POS) tagging, to machine translation and dialogue systems (Young et al., 2018). Natural Language Processing leverages techniques, and algorithms to process and understand unstructured natural language-based data such as text, speech, and so on (Young et al., 2018).

3.1.1.1 Tokenization

Tokenisation is the process of breaking up a text's sequence of characters by determining where one word ends and another begins (Palmer, 2000). For researchers working on English and similar languages, where word boundaries are generally coincident with space characters, this was not seen as a serious problem; however, it is a more difficult task for Chinese and some other languages, where a word can be a single character or a series of two or more characters, and there may be no spaces to separate words (Hu and Atwell, 2003).

3.1.1.2 Part-of-Speech Tagging

In POS-tagging, each word must be allocated to the correct Part-of-Speech, such as noun, verb, adjective, or adverb; additionally, most POS-taggers include grammatical elements such as singular/plural number, tense, and gender. The number of tags utilised by various systems varies significantly (Hu and Atwell, 2003). To correctly identify Parts of Speech in order to recognise entities, extract themes, and process sentiment, POS tagging is the foundation of a number of essential Natural Language Processing activities.

3.1.1.3 Parsing

An automatic parser's job is to take a formal grammar and a sentence and apply the grammar to the sentence in order to generate a parse-tree structure. The task of parsing is well-known in the field of natural language processing. Two different starting points reflect two different perspectives: one starts with the sentence's words and develops the tree from the bottom up, while the other starts with the sentence and builds the tree from the top down (Hu and Atwell, 2003).

3.1.1.4 Semantic Annotation

Data is augmented with semantic annotation to allow for automatic recognition of the underlying semantic content and structure. Labelling documents with thesaurus classes for document classification and management is a frequent approach in this regard. Semantic annotation has been used in conjunction with machine-learning software trainable on annotated corpora for word-sense

disambiguation, co-reference resolution, summarization, information extraction, measuring semantic similarity or difference between documents, and other tasks (Hu and Atwell, 2003).

3.1.2 Machine Learning for NLP and Text Analytics

In natural language processing NLP and text analytics, the role of machine learning and AI is to improve, accelerate, and automate the underlying text analytics functions and NLP features that convert unstructured text into usable data and insights. The area of machine learning (ML) is concerned with the subject of how to build computer programmes that improve themselves over time (Radovanović and Ivanović, 2008).

A machine learning model is made up of all of the knowledge it has gained from its training data. As additional knowledge is gained, the model evolves. A machine learning model, unlike algorithmic programming, can generalise and cope with novel scenarios. If a situation looks similar to something the model has seen before, it can utilise its previous "learning" to evaluate it. Machine learning for NLP and text analytics encompasses a set of statistical techniques for detecting portions of speech, entities, sentiment, and other features of text are used in (Radovanović and Ivanović, 2008).

There have been significant advances in Machine Learning approaches for automatic textual analysis, covering a range of linguistic levels, including tokenisation, Part-of-Speech tagging, Partial parsing, Semantic analysis, and Discourse annotation, in the field of Natural Language Processing research (Hu and Atwell, 2003).

The distinction between supervised and unsupervised learning methods is an essential one. Computer programmes capture structural information and derive conclusions (predictions) from previously labelled samples in supervised learning (instances, points). Without the use of labels, unsupervised learning finds groups in data (Radovanović and Ivanović, 2008). Clustering, representation learning, and density estimation are the most popular unsupervised learning problems. The aim is to learn the data's inherent structure without requiring labels that are explicitly provided. K-Means clustering, principal component analysis, and autoencoders are examples of prevalent algorithms. In most unsupervised

learning methods, there is no explicit way to compare model performance because no labels are provided (Radovanović and Ivanović, 2008).

3.1.3 Distributed and Distributional Representations for Natural Language Processing

Natural language is a discrete symbolic representation of human knowledge by its own nature (Chomsky, 1957). Recent breakthroughs in machine learning (ML) and natural language processing (NLP) appear to contradict the foregoing intuition: discrete symbols are slipping away, being replaced by distributed and distributional representations, which are vectors or tensors. Discrete symbols and distributed/distributional representations, on the other hand, are intimately connected, the first being an approximation of the second (Ferrone and Zanzotto, 2020).

Deep learning models are being pushed to achieve incredible outcomes in a variety of high-level tasks thanks to distributed representations (LeCun et al., 2015; Schmidhuber, 2015). They have applications in image generation (Goodfellow et al., 2016), image captioning (Vinyals et al., 2015b; Xu et al., 2015), machine translation (Zou et al., 2013; Bahdanau et al., 2014), syntactic parsing (Vinyals et al., 2015a; Weiss et al., 2015) and in other NLP tasks (Devlin et al., 2019).

Distributional representations, also known as distributional semantics, are studied in more traditional NLP as a more flexible way to express natural language semantics (Turney and Pantel, 2010). Words and sentences are represented as real-number vectors or tensors. Word vectors are created by looking at how certain words appear in document collections with other words. Furthermore, vectors for phrases (Clark et al., 2008; Mitchell and Lapata, 2008; Baroni and Zamparelli, 2010; Zanzotto et al., 2010; Grefenstette and Sadrzadeh, 2011) and sentences (Socher et al., 2011, 2012; Kalchbrenner and Blunsom, 2013) are obtained by composing vectors for words, just as they are in traditional compositional representations.

The success of distributed and distributional representations over symbolic approaches can be attributed to the introduction of new parallel paradigms that drove neural networks (Rosenblatt, 1958; Werbos, 1974) toward deep learning (LeCun et al., 2015; Schmidhuber, 2015).

3.1.3.1 Distributed Representations

In a distributed representation, the informative content is distributed (thus the name) among numerous units (Plate, 1995; Hinton et al., 1986), and each unit can contribute to the representation of multiple components at the same time. In comparison to a distributed local representation (One-hot encoding; will be explained in next section), distributed representation has two obvious advantages: it is more efficient, and it does not treat each element as equally different from the others (Ferrone and Zanzotto, 2020).

3.1.3.2 Distributional Representations and Distributional Semantics

Distributional semantics is a branch of natural language processing that tries to use vectorial representations to characterise the meaning of words and sentences (Turney and Pantel, 2010). Distributional representations are the name for these types of representations. Distributional semantics is based on the distributional hypothesis (Harris, 1954) – words have similar meaning when employed in similar settings. The underlying idea is that "a word is characterized by the company it keeps" as popularized by Firth (1957).

Distributional representations are clearly a subset of distributed representations, and the difference in terminology is merely a reflection of the context in which these techniques were developed. Words are represented by distributional vectors, which describe information about the contexts in which they appear. Sentence representations are created by merging vectors representing words.

As a result, distributional semantics is a subset of distributed representations with a limitation on what can be utilised as features in vector spaces: features represent a piece of contextual information. Once this is decided, massive matrices representing words in context are gathered, and dimensionality reduction techniques are used to create vectors that are more manageable and discriminative (Ferrone and Zanzotto, 2020).

3.1.3.3 Distributional Semantics Models (DSM)

Distributional semantic models are motivated by the distributional hypothesis and are typically implemented through vector space models (Lenci, 2008; Sahlgren, 2008). They are based on the assumption that the linguistic environment determines word meanings, and words with similar meanings occur in similar contexts. Word meaning is represented by taking large amounts of text as input and producing a distributional model, with semantic representations in the form of vectors that determine points in a multi-dimensional space — through an abstraction mechanism. The collection of points forms a vector space or semantic space in a distributional model, in which semantic relations can be expressed as geometric relations. The similarity between the semantic representations of words is usually measured using cosine similarity between the respective vectors (Fabre and Lenci, 2015).

3.1.3.4 Semantic Similarity and Relatedness

One of the potentialities of distributional models is to model semantic similarity. The distributional hypothesis is based on a claim about semantic similarity, making distributional models an excellent baseline for distributional representations. The primary result of DSMs is a continuous semantic space defined by linguistic units' mutual closeness relations.

The most popular and basic method of evaluating the effectiveness of distributional models is semantic similarity. DSMs were evaluated for accuracy and correlation with human similarity ratings, using different datasets and models. DSMs has achieved perfect accuracy on the TOEFL test (Bullinaria and Levy, 2012) and a Spearman correlation of 0.8 or better with similarity ratings. Thus, the performance of DSMs is often better than that obtained with measures based on manually designed lexical resources like WordNet (Agirre et al., 2009; Lofi, 2015). Indeed, DSMs usually obtain excellent results in tasks that involve semantic similarity, such as categorizing nouns (Baroni and Lenci, 2010; Riordan and Jones, 2011), modelling semantic priming (Jones et al., 2006; Mandera et al., 2017), and predicting patterns of functional magnetic resonance imaging

(fMRI) activation (Mitchell et al., 2008; Anderson et al., 2017). However, one significant limitation is that DSMs produce a network of word associations rather than a semantically structured space.

3.1.4 An Overview for Text Representations in NLP

The success of machine learning algorithms is largely determined by data representation. Since ML models are only capable of processing numerical values, texts are transformed into different input formats. This section will explore these input representations, starting with the simplest format, and going up to the most complex representations, in the same order they will be investigated in this thesis.

3.1.4.1 One-Hot Encodings

One-hot encodings is s straightforward technique for text representation in NLP. It works by replacing each category with a vector of zeros, except for the position of its associated index value, which has a value of 1. Using this approach, a vector with length equal to the number of categories in the corpus—each of which is represented by a single unique word—is created.

Tokens are substituted by their one-hot vectors when one-hot encoding is used on a text document, and a sentence is then turned into a 2D matrix with the shape of (n, m), where n is the number of tokens in the sentence and m is the vocabulary size. One-hot encoding is very simple form of representation with very easy implementation. However, it has some drawbacks. For example, processing and storing such vectors requires a lot of Memory. Along with these vectors' sparse nature. Also, no notion of similarity is captured which means no semantic information is getting expressed with this representation technique (Babić et al., 2020).

3.1.4.2 Topic Models

Probabilistic topic models (Blei et al., 2003; Griffiths et al., 2007) offer an alternative to semantic spaces (Mitchell and Lapata, 2010). Probabilistic topic models are based on the premise that words seen in a corpus have some latent structure associated to topics. Words are represented as a probability distribution over a range of topics, rather than as points in a high-dimensional space. Each topic is a probability distribution over words, and the words to which it gives high

probability reflect the content of the topic. Topic models are generative in that they define a probabilistic technique for generating documents. To create a new document, one must first select a topic distribution. Then, for each word in the document, a topic is chosen at random from this distribution and a word from that topic is selected (Mitchell and Lapata, 2010).

Latent Dirichlet allocation (LDA) is a generative probabilistic model for a collection of documents (text corpora.) LDA is a three-level hierarchical Bayesian model, in which each item (a document) of a collection is modelled as a finite mixture over an underlying set of topics, and each topic is modelled as an infinite mixture over an underlying set of topic probabilities (Blei et al., 2003). The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words (Blei, 2003). LDA is a popular algorithm for topic modelling that is used to extract the hidden topics from large volumes of text. LDA is a topic modelling unsupervised machine learning method that helps us discover hidden semantic structures in a text (Mitchell and Lapata, 2010).

LDA is a classic technique for document embedding. The dimensions of a document embedding space created by topic modelling techniques like LDA can be thought of as latent semantic structures that are hidden in the data because they are intended to model and explain word distribution in the corpus. LDA, as a quite different method for distributional semantics, is relevant in the context of this thesis and will be explored in more details in Chapter 4.

3.1.4.3 Word Embeddings

When deeper understanding of the context is required, on-hot encodings will not be sufficient. Complex NLP tasks such as Classification, Question-answering and semantic similarity would require a deep representation of the text. One of the most distinguished recent developments in NLP is the use of word embeddings, where words are represented as vectors in a continuous space, capturing many syntactic and semantic relations among them (Mikolov et al., 2013a). With embeddings, each word is represented by a dense vector of fixed size, with values corresponding to a set of features (Babić et al., 2020).

A neural network is used to learn embeddings on a supervised task. The network's parameters, or weights, are formed by the embeddings and are modified to minimize task loss. The generated embedded vectors depict categories where similar categories are closer to one another in relation to the task. A number of frameworks were developed for learning vector representations of words using neural networks (Bengio et al., 2003); (Collobert et al., 2008); (Mikolov et al., 2010) & (Mikolov et al., 2013a). Word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b) is a word embeddings technique that has two main architectures: Continuous Bag of Words (CBOW) and Skip-gram. Doc2vec extends words2vec to work with larger units of texts. The paragraph vectors method developed by Le and Mikolov (2014) has been tested on a number of text classification and sentiment analysis tasks, while Dai et al. (2015) examined it in the context of document similarity tasks. Word2vc and Doc2vec are state-of-the-art embeddings methods that will be examined in this thesis, in chapters 5 and 6.

3.1.4.4 Contextualized Embeddings

Polysemy refers to words that, despite having the same form, can mean entirely different things depending on the context. An issue that word embeddings fail to handle. Each word embedding must consider the context in which the word is found and adjust its values as necessary to solve the problem. To address the problem, a more complex structure is used to process the inputs to obtain contextualized embeddings, which in turn is used to address the issue of polysemy. ELMo, OpenAI GPT, and BERT (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019) are examples of the latest breakthroughs in NLP to obtain contextualized embeddings and achieves state-of-the-art-results in many NLP tasks (Babić et al., 2020). By utilising Siamese and triplet network architectures to create semantically significant sentence embeddings that can be compared using cosine-similarity, SBERT intends to adapt the BERT architecture (Reimers and Gurevych, 2019). These models will be discussed in details in subsequent sections.

3.2 The Arabic Language, Arabic Linguistics, Arabic Computational Linguistics, and Arabic Natural Language Processing ANLP

3.2.1 The Arabic Language

Arabic is the fifth most spoken language in the world and the main language of all Islamic sources (Al-Mahmoud and Al-Razgan, 2015). Arabic is known for its complexity, as it has three variations used in different situations. They are the Modern standard Arabic, Modern Arabic dialects, and the Classical Arabic of the Qur'an. Arabic is the native language of more than 400 million[6] people. In addition, more than 2 billion Muslims[7] use it in their daily prayers. Classical Arabic is the language of Holy Qur'an and Islamic documentary (Farghaly and Shaalan, 2009).

The Arabic language is made up of several varieties, one of which has a unique status as the formal written standard for the media, culture, and education in the Arab world. The other varieties are informal spoken dialects that are used in everyday life as a means of communication. Of course, both historically and geographically, language is part of a natural continuum (Habash, 2010).

The Arab World's official language is Modern Standard Arabic (MSA). The media and education use MSA as their primary language. MSA is based on Classical Arabic, the language of the Qur'an (Islam's Holy Book), in terms of syntax, morphology, and phonology. MSA, on the other hand, has a considerably more modern lexicon. MSA is primarily a written language, not a spoken one. In contrast, Arabic dialects are authentic native language variants. In most cases, they are only used for casual daily communication. Although there is a rich popular dialect culture of folktales, music, movies, and TV series, they are not taught in schools or even standardised. Dialects are predominantly spoken rather than written languages. This is changing, though, as more Arabs get access to

---

[6] https://worldpopulationreview.com/country-rankings/arabic-speaking-countries

[7] https://worldpopulationreview.com/country-rankings/muslim-population-by-country

electronic media platforms like email and newsgroups. Arabic dialects are vaguely connected to Classical Arabic (Habash, 2010).

3.2.2 Arabic Linguistics

The linguistic analysis of Arabic by early Arab grammarians was primarily motivated by a proper understanding and interpretation of the Qur'an and the Prophet's sayings (Farghaly, 2010). To date, the only complete Arabic grammar is that written by early Arab grammarians. At least to some extent, Arabic computational linguistics is based on the analyses of those leading Arab grammarians (Farghaly, 2010).

The methodology of traditional Arabic Linguistics is corpus-based (Farghaly, 2010). To determine the structural regularities underlying a corpus, Arab grammarians concentrated on a fixed corpus. The Qur'an, pre-Islamic poetry, and Prophet Muhammad's sayings made up their corpus. Despite their goal of establishing rules for proper Arabic use, their method was descriptive rather than prescriptive. They never imposed their opinions on proper usage without first obtaining proof from the corpus. They did, however, frequently consult fluent Arabic speakers in order to elicit their intuitions about the grammaticality of a sentence or phrase (Farghaly, 2010).

3.2.3 Arabic Computational Linguistics

Most of computational and corpus linguistics studies have been conducted using the English language, and other languages such as French, German, and Spanish. However, researches in Arabic NLP and Arabic corpus linguistics are still immature and a lot of work yet to be done (Al-Mahmoud and Al-Razgan, 2015).

Arabic computational linguistics is a subset of computational linguistics in general. In terms of current paradigms, computational linguistics, like any other science, undergoes periodic change (Oepen et al., 2000). New technologies arose in the late 1970s and 1980s to make language engineering more efficient. Two areas were targeted by the new technologies: processing efficiency and language descriptions. In the field of computer processing, more efficient techniques such as chart parsing (Kay, 1973), definite clause grammars (Pereira and Warren, 1980), and unification (Shieber, 1986) were created. However,

during the 1990s, the field has shifted dramatically from language engineering to statistical methodologies (to mention some: Manning et al., 1999). Many computational linguists were more concerned with maintaining strict adherence to linguistic theory than with designing systems that might be used in real-world situations. They were more concerned with theoretical correctness. In addition, the lack of fluency in the target language plagued rule-based machine translation systems, and translated texts were loaded with grammatical phrases and sentences (Farghaly, 2010).

The achievement of successful voice recognition systems based on probability theory and machine learning cleared the way for a new empirical approach (Manning et al., 1999), demonstrating its cost effectiveness, validity, and resilience. Therefore, since the 1990s, statistical approaches to natural language processing have dominated the area. The use of training and testing data characterises the new paradigm. The better the outcomes, the closer the testing data is to the training set. Also, both supervised and unsupervised learning approaches are used (Farghaly, 2010). Such approach uses probability theory to view language and cognition as probabilistic and anticipates the next word based on the preceding one. It assumed that linguistic information exists in linguistic data, and that machine learning algorithms may derive this knowledge through training and retraining cycles until the language is "learned." Moreover, it was stressed that language modelling is used to give the output fluency that is lacking in symbolic systems (Farghaly, 2010).

Statistical approaches to NLP, on the other hand, have flaws. First, when the input differs from the training data, performance quickly degrades. Symbolic systems, on the other hand, are frequently consistent in their performance. While, in ML, the algorithm learns rules as it establishes correlations between input and output, in symbolic reasoning, rules are created through human intervention. Second, it can be difficult to forecast the type of data a system will get in some instances. As a result, it is difficult to train such a system. Third, the language used in emails, chat rooms, and blogs is noisy, which degrades system performance. Fourth, there may come a point where adding more training data will cause the system to get confused; this is referred to as the threshold problem. Fifth, there is inadequate evidence to "learn" from when a phenomenon occurs seldomly; that is known as the "sparse data problem" (Farghaly, 2010).

On the other hand, renowned computational linguists (Zaenen, 2006; Reiter, 2007; Jones, 2007) have recently highlighted concerns about relying exclusively on machine learning methodologies while ignoring symbolic language processing's contributions[8]. As a result, the same philosophical trend that inspired natural language processing paradigms has influenced research in Arabic computational linguistics (Farghaly, 2010). Bender (2009) believes that generalisations from language typology, which specifies how languages differ from one another, are essential even when designing language-independent NLP systems. It's becoming increasingly clear that adding language information into statistical NLP systems improves their performance. In recent years, there has been a growth in interest in the Arabic language and culture, resulting in the development of various cutting-edge computational tools based on machine learning and engineering knowledge (Farghaly, 2010).

3.2.4 Arabic Natural Language Processing (ANLP)

Arabic, as a natural language, has a lot in common with other languages, such as English. It is, nonetheless, unique in terms of its history, diglossia[9] nature, internal structure, and complicated link with Islam, as well as Arabic culture and identity. Due to the unique characteristics of the Arabic language, NLP tools designed for use with western languages are difficult to adapt to Arabic. Realizing the importance of developing tools for Arabic is substantial for ANLP progress (Farghaly and Shaalan, 2009).

NLP has reached a decent level of maturity after more than fifty years of development. For obvious reasons, this is especially true of the English language. When compared to English, Arabic NLP falls behind by at least a decade. This

---

[8] Symbolic AI was the dominant paradigm of AI research since 1950. Computational linguistics uses Symbolic AI (rule-based reasoning systems) to teach the machine how to understand languages using humans-like rules. The approach is deterministic and transparent as everything is visible and explainable, unlike the black box created by ML.

[9] Diglossia is a phenomenon in which two or more varieties of the same language coexist in the same speech community (Ferguson 1959; Ferguson 1996). Each has a particular use and is applied in a particular circumstance. Arabic displays a true diglossic situation where at least three different varieties of the same language are utilized within a speech community and in specific contexts. (Farghaly and Shaalan, 2005).

might be attributed to the Arabic language's richness and complex grammatical and syntactic structures (Bashir et al., 2021). First, developing NLP systems in a diglossic situation like Arabic is difficult. Processing data from all the Arabic variations is exceedingly challenging and nearly impossible for a single ANLP application. Despite sharing some traits, each variant has its distinct grammar, vocabulary, and morphology. So, the variant has to be stated upfront for an ANLP application to process. In a nutshell, building a system that can handle all the Arabic varieties simultaneously is challenging, thus the solution is to build resources for the various varieties of Arabic (Salloum et al., 2018; Farghaly and Shaalan, 2009).

Next, the Arabic script has special features that place another challenge for NLP systems. There are no special letters to represent short vowels, the form of the letter varies depending on where it appears in the word, there is no capitalization, and there is little punctuation. Therefore, an NLP application must have knowledge of the Arabic language's structure and syntax in order to spot patterns in the absence of these rules (Shaalan and Raza 2008; Shaalan and Raza 2009).

The problem of normalization is another obstacle that computational linguists working on Arabic face. The issue is brought about by the inconsistent use of diacritical markings and some letters in modern Arabic writings. For example, a dot, a hamza, or a madda added above or below the letter can separate some Arabic letters that have the same shape from one another[10]. To distinguish between letters that appear to be similar, one must be able to recognize these marks above or below a letter (Farghaly and Shaalan, 2009).

Finally, the existence of the many levels of ambiguity poses a grand challenge to researchers developing NLP tools for the Arabic language (Attia, 2008). All the above-mentioned features contribute to the ambiguity of Arabic in addition to the lack of short vowels. Two different forms of linguistic information are lost when

[10] Three different letters depending on whether it has a hamza above as in (‘أ’) or a hamza below as in (‘إ’) or a madda above as in (‘آ’). Some variants of Arabic do not incorporate voweling, nor do they include such letters and marks, which complicate the role of Arabic normalizer.

short vowels are absent. The majority of case markers[11] that specify the grammatical role of Arabic nouns and adjectives fall within the first category. The relatively open word order in Arabic and the fact that it is a pro-drop language[12] cause numerous difficulties regarding the absence of case markers and, consequently, the grammatical function of a word. Lexical and part of speech information[13] is the second sort of information that is lost as a result of the Arabic script's design. As a result, without contextual cues, it might be difficult to identify the part of speech (POS) when internal voweling is absent (Farghaly and Shaalan, 2009).

According to Salloum et al. (2018), computational text mining in Arabic literature can yield a lot of information, but the effort in this subject is inadequate. This is attributed to the difficult nature of conducting Arabic NLP research. Over the last decade, Arabic and its dialects have made strides in the field of Natural Language Processing research (NLP). Many projects focused on various aspects of how this language and its dialects are processed, such as morphological analysis, resource development, machine translation, and so on (Guellil et al., 2021). Several works surveyed Arabic language and its dialects with a thorough analysis to its features (Habash, 2010; Farghaly and Shaalan, 2009; Shoufan and Alameri, 2015). A recent survey by Guellil et al. (2021) provided and categorised the most recent works on Arabic. The majority of them came out between 2015 and 2018. Their survey presented the work that has been done on the three variations of the Arabic language: CA, MSA, and DA.

Recently, with the trend of deep learning and transfer learning, several tools and models were developed to serve many Arabic NLP tasks. For example,

---

[11] For instance, a fatHa, a low front vowel in the last position of a common noun, marks the accusative case, and a kasra, a high front vowel, marks the genitive case. Damma, a high back rounded vowel, at the end of a common noun or adjective, marks the nominative case.

[12] Subject pronouns may be freely dropped in Arabic (Farghaly, 1982) as long as the Recoverability of Deletion Condition is met (Chomsky, 1965). The ability to omit the subject pronoun and use "subjectless sentences" is not unique to the Arabic language; other languages that do this include Italian, Spanish, and Korean.

[13] An Arabic token like (‘كتب’) *ktb* without internal voweling could be a plural noun 'books', an active past tense verb 'wrote,' a passive past tense verb 'was written' or a causative past tense verb 'he made him write'(Farghaly and Shaalan, 2009).

HuggingFace's Transformers (Wolf et al., 2019) to fine-tune multilingual BERT (mBERT) (Devlin et al., 2018) and AraBERT (Antoun et al., 2020) on the tasks of Arabic Sentiment Analysis (SA), and named entity recognition (NER). Also, a collection of software suites was generated that provide multiple capabilities in the form of a unified toolkit such as MADAMIRA (Pasha et al., 2014) and Farasa (Abdelali et al., 2016; Darwish and Mubarak, 2016) or multi-lingual, such as Stanford CoreNLP (Manning et al., 2014).

Moreover, recently, Obeid et al. (2020) presented CAMeL Tools, an open-source set of tools for Arabic NLP providing utilities for pre-processing, morphological modelling, dialect identification, named entity recognition, and sentiment analysis. The suite is still under development.

3.3 AI Review for Understanding the Qur'an

The Qur'an, as a significant religious text written in the Classical Arabic, has been the. subject of numerous studies due to its linguistic and spiritual value and it is argued that Artificial Intelligence modelling of the Qur'an has a significant potential impact (Atwell, 2018). Scholars have studied the Qur'an and drew out knowledge and patterns that were the base for many applications on the holy book, for recent surveys see (Alrehaili and Atwell 2014; Guellil et al., 2021; Bashir et al., 2021).

3.3.1 Computational and Corpus Linguistics Resources for the Qur'an

Several researches have contributed to the development of Arabic corpus linguistics resources. The majority of contributions were devoted to Modern Standard Arabic and modern Arabic dialects. However, few researches have focused on the classical Arabic of the Qur'an (Atwell, 2018). The Leeds University researchers have developed a wide range of Arabic-language resources and corpora. Such researches have employed Artificial Intelligence with Corpus linguistics to provide substantial resources for religious knowledge, and a robust base for potential research and development (Atwell, 2018).

One significant Classical Arabic work is the Qur'anic Arabic Corpus (Dukes et al. 2013). QAC is an impressive open-source project that is developed by Artificial

Intelligence research group at University of Leeds. The corpus provides layers of annotation including POS tagging, morphological annotation, a syntactic treebank and a semantic ontology of Qur'anic concepts.

QurAna is a large corpus created from the original Qur'anic text annotated with antecedent references of pronouns (Sharaf and Atwell, 2012a). Another significant resource on the Arabic Qur'anic text is Qursim, that contains over 7000 pairs of related verses collected from scholarly resources (Sharaf and Atwell, 2012b). The dataset is incorporated into a website where users can visualize, for a given verse, a network of all directly and indirectly related verses. The dataset can serve as an evaluation resource for potential works on textual similarity and relatedness in short texts. Also, Sharaf and Atwell (2012b) presented a manually annotated large corpus (QurSim), created from the original Qur'anic text, where semantically similar or related verses are linked together. In the years followed, the QurSim corpus has been used as a resource for evaluation in a number of studies (Sharaf and Atwell, 2012b) that investigated semantic relatedness in the Qur'an, including this research.

Another resource for searching concepts in the Qur'an is Qurany. Qurany is a web-based search tool for concepts in the Qur'an (Abbas, 2009). Qurany Explorer is a comprehensive tool that covers all the themes and concepts mentioned in the Qur'an. The Qur'an corpus is augmented with an ontology or index of nearly 1200 key concepts, taken from a recognized expert source. Expert knowledge used in annotating the Qur'an corpus is obtained from 'Mushaf Al Tajweed'. On Qurany, a user can navigate for a specific concept through a sequence of concepts, with a list of verses which allude to this concept, along with different eight alternatives English translations.

In 2016, Asda et al. proposed using Mel-Frequency Cepstral Coefficient (MFCC) feature extraction and Artificial Neural Networks to construct a Qur'an reciter detection and identification system. Characteristics from utterances will be retrieved from each speech using neural network models. There are two parts to

the proposed system. The feature extraction process is the first element, and the neural network identification process is the second (Asda et al., 2016).

At another level, Al-Kabi et al. (2013) and Adeleke et al. (2017) concentrated on the Qur'an's classification. The first project intended to categorise Qur'anic verses based on their topics, while the second offered a feature selection approach for automatically labelling Qur'anic verses.

3.3.2 NLP for the Qur'an

The development of AI-based NLP tools for the Arabic language has sparked a lot of interest recently. However, Qur'anic NLP study is less developed than Arabic NLP research, which is itself comparable to a low-resource language with fewer tools and data compared to works focusing on the English language (Bashir et al., 2021; Guellil et al., 2021; Atwell et al., 2010).

In addition to the challenges associated with the Arabic NLP (highlighted previously in 3.2.4), there are unique challenges that limit the Qur'anic NLP researches. The Qur'an is a concise text, and it is accompanied by a massive body of work that spans tens of thousands of pages, including commentaries, exegesis, and other works. Understanding the text requires deep semantic analysis and domain knowledge. The richness in subtle meanings and unique orthography are essential features of the Qur'an. Along with all of the aforementioned obstacles, one of the most challenging aspects when dealing with the Qur'an is that it is a divine scripture, which necessitates extra caution in order to preserve the semantics or information received. Current NLP efforts are aimed at overcoming all of these obstacles with a variety of techniques. Performing NLP tasks on a language with all of these challenges is a difficult effort that necessitates both technical and linguistic assistance (Bashir et al., 2021).

According to Atwell et al. (2010), "Understanding the Qur'an" can be viewed as a big AI challenge because many tasks such as reasoning, knowledge representation, and knowledge extraction based on Qur'anic text, among others, must be solved using the latest NLP approaches. In 2011, Atwell et al. assessed some of Leeds University's Arabic and Qur'an AI and Corpus Linguistics research, which has resulted in a variety of software and corpus datasets for research on Modern Standard Arabic and, more recently, Qur'anic Arabic. Based

on that they proposed the Qur'anic Knowledge Map, a machine-readable organised collection of linguistic and semantic information, as well as a very effective teaching website, to further their research.

The Qur'an Researchers are attempting to address the challenges associated with all aspects of NLP, but more work is required. Many of these have been aided by recent breakthroughs in computational and natural language processing (NLP) methods. These methods aid in the development of techniques that enable both Muslims and non-Muslims to simply obtain teachings and knowledge from the Qur'an (Bashir et al., 2021). In a recent survey, Bashir et al. (2021) provided a comprehensive survey of Qur'anic Arabic focused NLP techniques, tools, and applications. Figure 3 illustrates the timeline for milestones related to Qur'anic NLP per their survey.

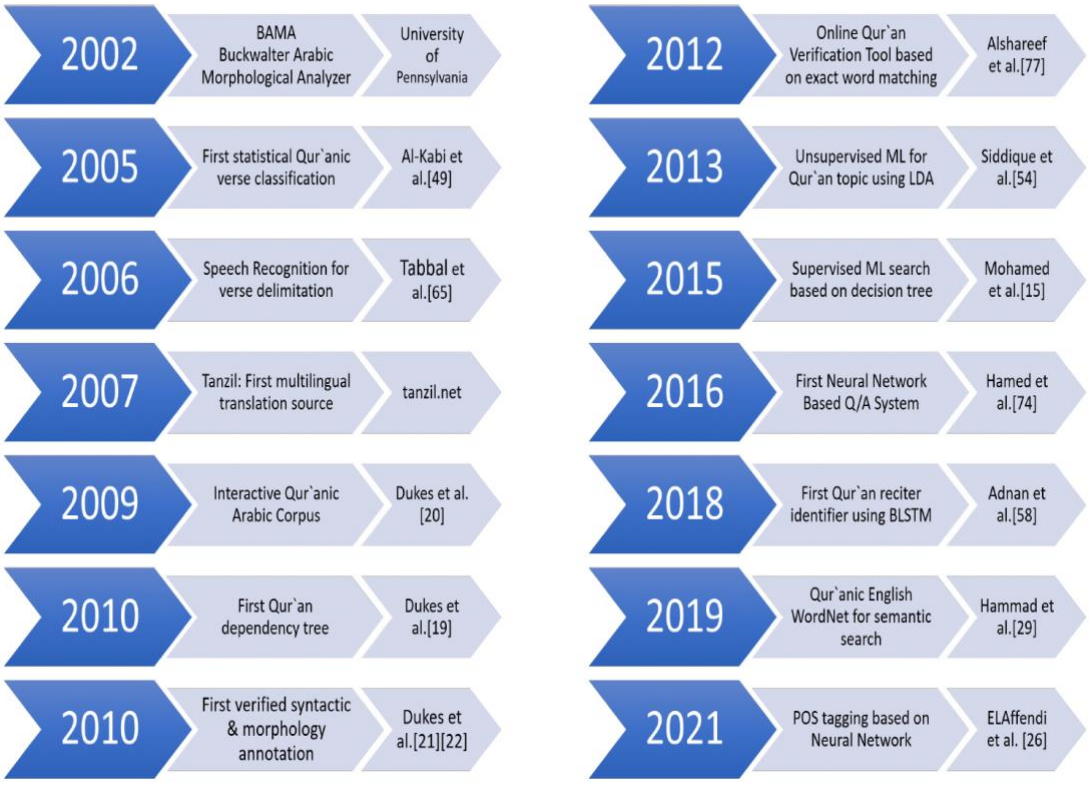


Figure 3: The timeline for milestones related to Qur'anic NLP (Bashir et al., 2021)

Moreover, Bashir et al. (2021) added that deep learning techniques have not been widely used for Qur'anic NLP. In various Qur'anic NLP tasks, recurrent

neural networks (RNN) in the form of LSTM and BiLSTM, CNN, and AraBERT can be used.

With the advent of the Transformer architecture by Vaswani et al. (2017) in recent years, Natural Language Processing (NLP) has advanced. BERT, which is based on the transformer layer, has demonstrated and generated state-of-the-art accuracy in a variety of NLP applications, including text categorization and machine translation. As a recent effort in this direction, Alsaleh et al. (2021) presented an experiment using a transformer-based language model, AraBERT, to classify pairs of verses to be semantically related or not. His experiment showed promising results with a lot of potential for improvement.

Recently, researchers at Leeds University have examined the transformer-based model with the Question-answering (QA) task (Alsaleh et al., 2022). The experiments have been conducted using three Arabic language models AraBERT, CAMeL-BERT, and Ara- bicBERT. They applied the simple transformers model to the Shared Task Question-Answering over the Holy Qur'an. It focused primarily on customizing the Simple Transformers model to extract the necessary information from Qur'anic passages and enhance the accuracy of findings using three pre-trained Arabic language models that are based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018).

3.4 Deep Learning for NLP

3.4.1 Deep Learning

Deep learning is a subfield of machine learning, while both fall under the broad category of AI. Deep learning is an area of machine learning that deals with artificial neural networks, which are algorithms inspired by the structure and function of the brain. It is ML and functions in a similar way but with different capabilities. Deep learning is what powers the most human-like AI. Deep learning algorithms are a sophisticated and mathematically complex progression of machine learning algorithms. The field has gotten great attention recently. This is fuelled by the consistent growth of computational power, including GPUs and the

massive amount of data we can feed to these algorithms (Goodfellow et al., 2016; Goldberg, 2016).

3.4.2 Deep Learning and Machine Learning

Machine learning uses algorithms to parse data, learn from that data, and make informed decisions based on what it has learned. Machine learning models become progressively better at their function, but they still need some guidance. For example, if an ML algorithm makes an incorrect prediction, we have to make adjustments. On the other hand, with deep learning, the algorithm can determine whether a prediction is accurate or not. DL structures algorithms in layers to create an artificial neural network that can learn and make intelligent decisions independently. DL models are designed to continually analyse data with a logic structure similar to how humans draw conclusions (Goldberg, 2016; Goodfellow et al., 2016; Young et al., 2018). Figure 4 illustrates a comparison between ML and DL in processing data.

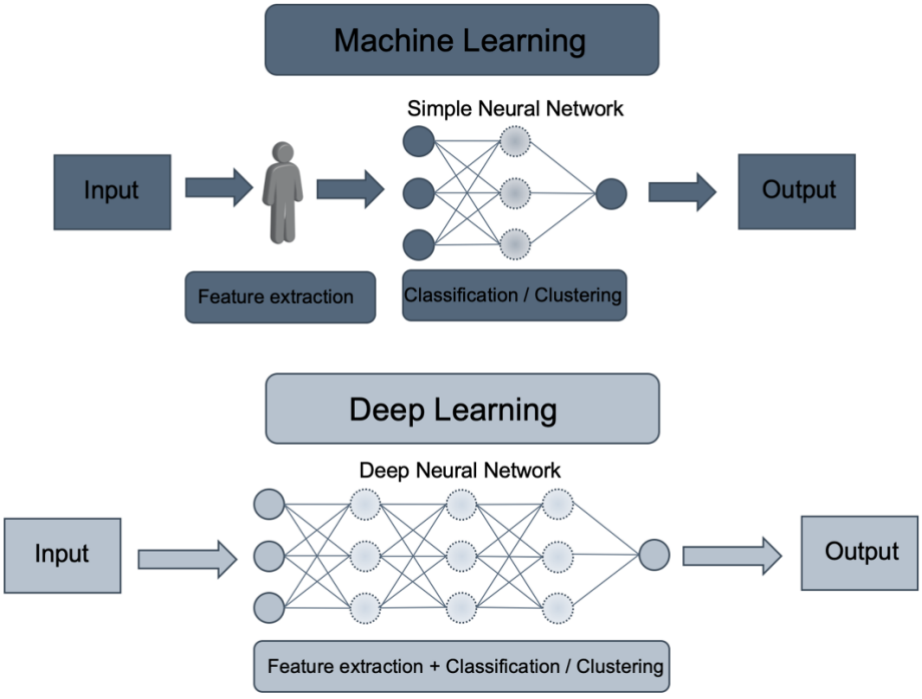


Figure 4: Machine Learning vs Deep Learning (Xantaro et al., 2018)

Machine Learning algorithms are frequently used for NLP problems; hence ML and NLP have some overlap (Cambria and White, 2014). Figure 5 illustrates how NLP is related to ML and Deep Learning.

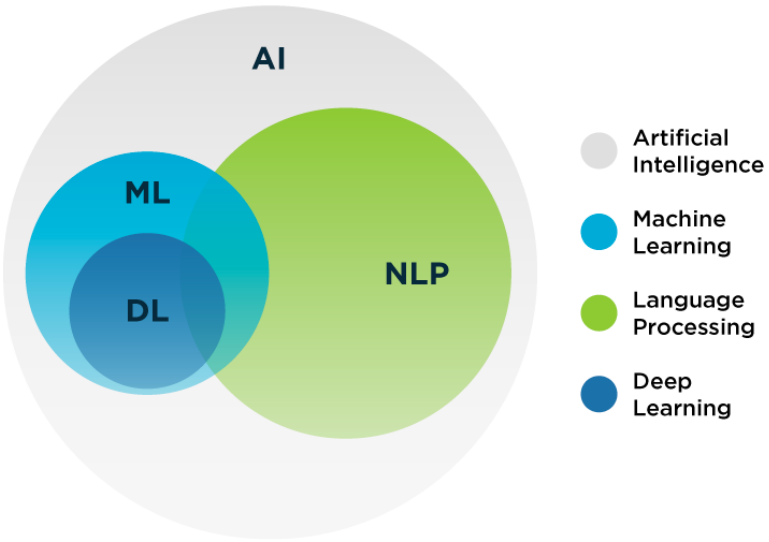


Figure 5: Relationship between AI, ML, DL, and NLP (7wData, 2021)

3.4.3 Deep Learning and NLP

New deep learning methods are rapidly being used in contemporary NLP research. Shallow models (e.g., SVM and logistic regression) trained on very high dimensional and sparse features have been the foundation of machine learning approaches to NLP problems for decades. In recent years, neural networks based on dense vector representations have outperformed traditional neural networks on a variety of NLP tasks. The success of word embeddings (Mikolov et al., 2010; Mikolov et al., 2013) and deep learning approaches (Socher et al., 2013) has fuelled this movement. Multi-level automatic feature representation learning is possible with deep learning. Traditional machine learning-based NLP systems, on the other hand, rely primarily on hand-crafted features. Hand-crafted features take time and are frequently incomplete (Young et al., 2018).

In numerous NLP tasks, such as named-entity recognition (NER), semantic role labelling (SRL), and POS tagging, Collobert et al. (2011) have shown that a simple deep learning framework surpasses most state-of-the-art techniques. Since then, various complicated deep learning-based algorithms for solving

challenging NLP tasks have been presented. In the following sections, we review recent deep learning related models applied to NLP tasks.

3.5 Language Models

The context problem still exists even when neural networks are able to solve the sparsity problem. In order to tackle the context problem more effectively, language models were first developed, allowing an increasing number of context words to have an impact on the probability distribution (Goldberg, 2017). Many essential natural language processing tasks rely on language modelling. Recent research has shown that neural-network-based language models outperform traditional methods in both isolated and multi-task natural language processing tasks (Brownlee, 2017).

3.5.1 Language Model in NLP

A language model is a probability distribution over sequences on an alphabet of tokens in natural language processing. Learning a language model from examples, such as a model of English sentences from a training set of phrases, is a core problem in language modelling (Goldberg, 2017). There are two primary types of language models, they are: statistical language models and neural language models.

3.5.2 Statistical Language models

Statistical Language Modelling, often known as Language Modelling or LM, is the creation of probabilistic models that can predict the next word in a sequence based on the words that came before it (Brownlee, 2017). According to Goldberg (2017), A language model uses machine learning to create a probability distribution over words that is used to predict the most likely word that will come after the current entry in a sentence. Language models can be used to generate original text and predict the next word in a text since they learn from text.

To learn the probability distribution of words, these models employ standard statistical techniques such as N-grams, Hidden Markov Models (HMM), and Conditional random fields (CRFs) (Nadkarni et al., 2011). Language modelling

has many real-world applications such as speech recognition, optical character recognition, handwriting recognition, machine translation, and spelling correction.

3.5.3 Neural Language Models

The use of neural networks in the development of language models has recently gained a lot of momentum, to the point that it might now be the dominant method. Neural Language Modelling, or NLM for short, is the use of neural networks in language modelling (Goldberg, 2017). They surpassed statistical models as they yield state-of-the-art accuracy with minimal human engineering using a recent breakthrough: deep neural networks. They also exploit the power of distributed representation: word embeddings (Brownlee, 2017).

The sparsity issue is made easier by neural network-based language models due to the way they encode inputs. Word embedding layers produce a vector of each word of any size that includes semantic links. The probability distribution of the subsequent word gains the crucial granularity from these continuous vectors. As part of the training process, the parameters are learned. Word embeddings created with NLMs have the feature that semantically similar words are also semantically similar in the induced vector space (Kim et al., 2016; Bengio et al., 2003).

3.5.3.1 Classic neural language model

Word embeddings are common to all word-level neural language models. The classic neural language model proposed by Bengio et al. (2003) consists of a one-hidden layer feed-forward neural network that predicts the next word in a sequence as shown in Figure 6. The main features of their model can still be found in today’s neural language and word embedding models. They are:

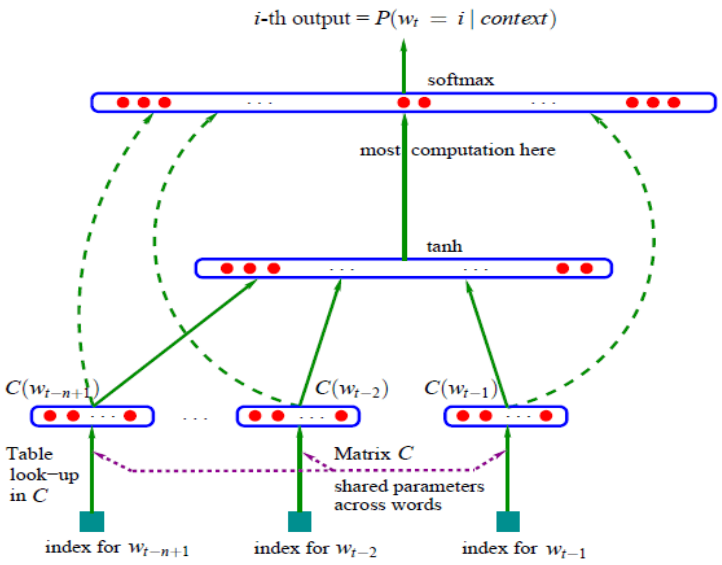


Figure 6: Neural Language Model proposed by (Bengio et al., 2003)

1 Embedding Layer: this layer generates word embeddings by multiplying an index vector with a word embedding matrix;
2 Intermediate Layer(s): one or more layers that produce an intermediate representation of the input, e.g. a fully-connected layer that applies a non-linearity to the concatenation of word embeddings of (n) previous words;
3 Softmax Layer: the final layer that produces a probability distribution over words in (V).

An activation function gives the neural network some form of nonlinear property. The main activation functions used deep learning are Sigmoid function, Tanh, Rectified Linear unit (ReLU), and Softmax function. Tanh or sigmoid are inappropriate for the output layer of a neural network if it needs to predict values larger than 1, and ReLU must be used in their place. On the other hand, if the output values are expected to be in the range [0,1] or [-1, 1], then ReLU is not a good choice for the output layer sigmoid or tanh must be used. The softmax activation function should be employed in the last layer if the classification task requires the neural network to predict a probability distribution over the class labels that are mutually exclusive (Bengio et al., 2003).

Sigmoid is used for binary classification methods where we only have 2 classes, while Softmax applies to multiclass problems. Sigmoid receives just one input and only outputs a single number that represents the probability of belonging to

class1 or class 2. On the other hand, SoftMax is vectorized, which means that it takes a vector with the same number of entries as the classes we have and produces a second vector, the components of which each reflect the chance that a given class will be selected. Softmax normalizes the results by forcing the values of the output neurons to take values between 0 and 1, so that they can represent probability values in the interval [0, 1], thus building a probability distribution over all the predicted classes (Bengio et al., 2003).

3.5.3.2 Word2vec

Word2Vec is one of the most popular techniques to learn word embeddings using a shallow neural network. It was developed by Tomas Mikolov in 2013 (Mikolov et al., 2013a). Word2vec has two architectures: the continuous bag-of-words model (CBOW) and the Skip-gram model (SKIP-G). In the first model, CBOW predicts a pivot word according to the context by using a window of context words. The second model SKIP-G predicts surrounding words of the current pivot word. Both have their own advantages and disadvantages. According to Mikolov (2013a), Skip Gram works well with small amount of data and is found to represent rare words well. On the other hand, CBOW is faster and has better representations for more frequent words.

3.5.3.3 Glove

Pennington et al. (2014) proposed a Global Vectors (Glove) to build a words representation model. Glove uses the global statistics of word-word co-occurrence to build a co-occurrence matrix. Then, this matrix is used to calculate the probability of a word to appear in the context of another word. This probability represents the relationship between words. As co-occurrence counts can be directly encoded in a word-context co-occurrence matrix, GloVe takes such a matrix rather than the entire corpus as input.

3.5.3.4 Recurrent Neural Networks (RNN)

Regarding sparsity, recurrent neural networks (RNNs) represent an advancement. RNNs (Socher et al., 2011; Socher et al., 2013; Tai et al., 2015) consider all previous words while selecting the next one since they can be either long short-term memory (LSTM) or gated recurrent unit (GRU) cell-based

networks (Tai et al., 2015). By implementing a bidirectional LSTM, AllenNLP's ELMo (Peters et al., 2018) expands on this idea by taking into account the context both before and after the word counts. On the other hand, RNN-based architectures are sequential in nature which is their main flaw. Due to the lack of parallelization, training times for long sequences increase dramatically. The transformer architecture offers a solution to this issue.

3.5.3.5 Transformers

Vaswani et al. (2017) presented the transformer, an encoder-decoder architecture based on attention layers. One key distinction is that the input sequence can be passed simultaneously, maximizing the utilization of the GPU and speeding up training. It also relies on the multi-headed attention layer; therefore, the vanishing gradient problem is easily resolved. The transformer is applied to an NMT (Neural Machine Translator) in their paper. The architecture is shown in Figure 7.

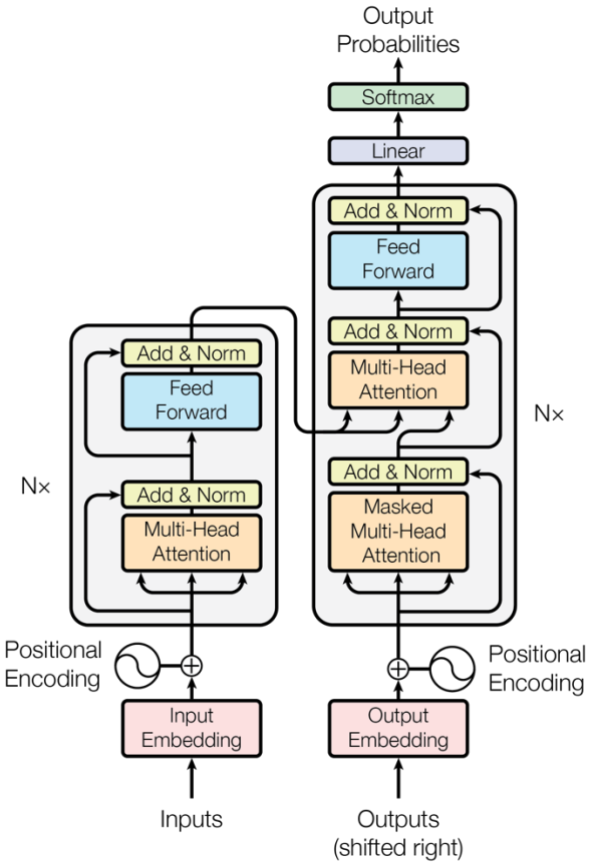


Figure 7: The Transformer architecture (Vaswani et al., 2017)

By handling these parts of learning in a completely different architecture, the Transformer architecture gets rid of the time-dependent aspect of the RNN architecture. As a result, the transformer contains as many linear layers as there are words in the longest phrase, but unlike RNNs, these levels are largely prime and time-independent. As a result, it is very parallel and simple to calculate.

Transformers laid the road for more contemporary generative models as well as more advanced language models. The transformer design is used by Google's BERT (Devlin et al., 2018) and OpenAI's GPT (Radford et al., 2018) models. These models include a "Attention" mechanism, which enables the model to learn which inputs deserve more attention than others in certain cases. Moreover, OpenAI released Whisper, a general-purpose speech recognition model (Radford et al., 2022). It is a multi-task model that can do multilingual speech recognition, speech translation, and language identification and was trained on a large dataset of varied sounds.

3.5.4 Transfer Learning

Transfer learning is the process of improving learning in a new task by transferring knowledge from a previously learned related task (Olivas et al., 2009; Torrey and Shavlik, 2010). The emergence of transfer learning, or the use of pre-trained models, is one of the most significant advances in the field of deep learning. The rationale for this is that transfer learning can be thought of as a cure for the enormous training datasets that are required for ANNs to yield meaningful results. Transfer learning enables algorithms to learn a new task by using pre-trained models (Olivas et al., 2009; Brownlee, 2019).

Pre-training of language models has been proven to significantly improve a variety of language comprehension tasks (Peters et al., 2018; Radford et al., 2018; Phang et al., 2018; Devlin et al., 2018). The main concept is to train a large generative model on huge corpora and then apply the generated representations to tasks where labelled data is scarce (Edunov et al., 2019).

Conventional machine learning and deep learning algorithms, so far, have been traditionally designed to work in isolation. These algorithms are trained to solve

specific tasks. The models have to be rebuilt from scratch once the feature-space distribution changes. Transfer learning is the idea of overcoming the isolated learning paradigm and utilizing knowledge acquired for one task to solve related ones (Goergen, 2022). Figure 8 shows the difference between traditional ML and transfer learning.

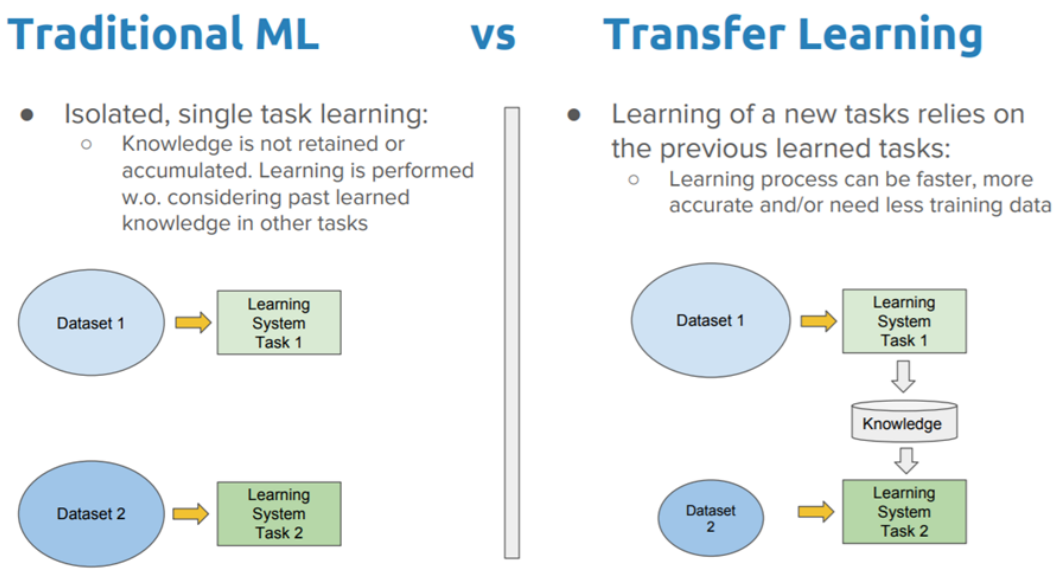


Figure 8: Traditional Machine Learning vs. Transfer Learning (Sarkar, 2018)

A machine learning algorithm works in isolation in classical learning. Having a large enough dataset, it learns how to perform a specific task. When faced with a challenge, however, it is unable to use any previously acquired knowledge. A traditional algorithm, on the other hand, requires a second dataset to begin a new learning phase. In transfer learning, the learning of new tasks is based on previously learned tasks. The algorithm has the ability to store and retrieve information. Instead of being detailed, the model is broad (Goergen, 2022).

3.5.5 Pre-trained Language Models

Transfer learning is commonly used with natural language processing tasks that involve text as input or output. For these types of problems, a word embedding is utilised, which is a mapping of words to a high-dimensional continuous vector

space with similar vector representations for distinct words with similar meanings (Goldberg, 2017; Weiss et al., 2016).

There are efficient techniques for learning these distributed word representations, and it is usual for research organisations to offer pre-trained models under a permissive licence that have been trained on very large corpora of texts (Edunov et al., 2019). These distributed word representation models can be downloaded and included into deep learning language models for either word interpretation as input or word generation as output. These pre-trained networks/models form the basis of transfer learning in the context of deep learning (Goldberg, 2017).

Pre-training of language models has been proven to significantly improve a variety of language comprehension tasks (Peters et al., 2018; Radford et al., 2018; Phang et al., 2018; Devlin et al., 2019). Pre-trained models are usually shared in the form of the millions of parameters/weights the model achieved while being trained to a stable state (Edunov et al., 2019).

The presence of models that perform well on source tasks is one of the most basic conditions for transfer learning. Many of the most cutting-edge deep learning architectures have been openly released by their creators. NLP is one popular domain for deep learning applications (Goldberg, 2017). Examples include word embedding models such as Word2vec, Glove, and FastText. There have also been some remarkable breakthroughs in NLP transfer learning that hold a lot of promise, one example is Universal Sentence Encoder by Google (Cer et al., 2018).

Moreover, in many areas of NLP, transformers have boosted the state of the art (Vaswani et al., 2017), by delivering greater parallelization and better modelling of long-range dependencies. BERT (Devlin et al., 2019) is the most well-known Transformer-based model; it achieved state-of-the-art results in numerous benchmarks and is still a must-have baseline (Rogers et al., 2022). Transfer

learning and pre-trained language models, indeed, have many potentials in relation to this thesis, and will be leveraged to answer its research questions.

3.6 Summary and Conclusions

This chapter reviewed previous work in four areas: Arabic NLP, Qur'anic NLP, distributional semantics, and recent deep learning methods that have been employed for NLP tasks. Reviewing the literature provided a holistic view of opportunities the field can offer to help address the research questions. This section summarizes the implications of the reviewed work concerning thesis research questions.

Several works have contributed to the development of Arabic NLP resources. The majority of contributions were devoted to Modern Standard Arabic and modern Arabic dialects. However, few studies have focused on the classical Arabic of the Qur'an (Atwell, 2018). The review determined the status of Qur'anic NLP research. Qur'anic NLP research is still evolving and deep learning-based techniques have not been much utilized. Thus, the trending deep learning methods and pre-trained language models have many potentials with Qur'anic NLP tasks, in particular detecting semantic similarity and topic modelling.

Deep learning approaches toward understanding language are achieving state-of-the-art results in many tasks such as text classification, clustering, and semantic textual similarity. However, because these tasks are essentially built on Language Modelling, there has been a tremendous research effort with excellent results to use Neural Networks for Language Modelling. Neural language models use distributed representations for input words.

From the literature on distributed representations, and word embeddings, a key theme is that adopting learned distributed representations has the potential to pick up on beneficial semantic relatedness properties of words. Recently, in ML applied to NLP, distributed representations of texts push deep learning models toward achieving impressive results in solving complex NLP tasks. Word embeddings are effective in capturing context similarities and analogies, as well as being fast and efficient in performing key NLP tasks due to their reduced dimensionality (Young et al., 2018).

Work reviewed for the different embedding techniques, in addition to the availability of powerful pre-trained models, showed that they have the potential to

capture semantic relations between the verses of the Qur'an and potentially reveal embedded meanings within the sacred text. They proved their success in solving challenging NLP tasks. The implication of this work is that deep learning approaches such as sentence embeddings and transformers that are based on distributional semantics may be an appropriate approach for achieving impressive results in the semantic textual similarity task, a research question that will be addressed in Chapter 6, 7, 8, and 9.

For classical Arabic, the language of the Qur'an, having an informative representation of the text ensures the success of machine learning. A thesis research question asks, how to represent the semantics of the Qur'an words to capture the intangible semantic relations. The question will be answered in chapters 4 and 5. This thesis will argue that complicated features of the sacred text can be captured in dense vector representations in the embedding space, where words with similar meanings are locally clustered. Embeddings can be encoded using topic models (like LDA); to encode the vectors with a topic, or learned from encoding using neural networks, such as Word2vec. Reviewing the literature has demonstrated the feasibility of LDA and Word2vec for unsupervised clustering and semantic similarity.

In a nutshell, NLP is developing so fast. We foresee the significance of this research as to present a baseline for research and potential improvement concerning empirical methods and data. Today's available language models are trained using English-language datasets as demanded. Accuracy for other languages such as Arabic is acceptable, but there is undoubtedly insufficient training data, and the quality of their output is consequently inferior. Furthermore, some models support the multilingual feature, however, they are inadequate and exclude classical Arabic.

With the rapid advancements, as we learn how to take advantage of emerging techniques, we will be able to train massively multilingual models that can exploit transfer learning between multiple languages. Also, developing monolingual systems for classical Arabic, and in particular, the Qur'an, is a grand opportunity the research community can take, by creating training data with supplementary knowledge resources including Hadith and ontologies.

Chapter4
Modelling topics from the Qur’an using Machine Learning and NLP

4.1 Introduction

Topic modelling is a technique for analysing documents to learn relevant word patterns (Dieng et al., 2020). Latent Dirichlet Allocation (LDA) is a powerful model and it is widely used to model latent topics in concise and short texts. LDA model represents each document as a mixture of latent topics in which a topic is a multinomial distribution over words. Every document then has its own mixing proportion of topics, and each topic has its own word distribution. In LDA, a topic is a probability distribution function over a set of words (Blei et al., 2003; Blei, 2012).

This chapter addresses the first research question and presents a semantic representation of the Qur’an words that capture the semantic similarity between its verses. It presents a computational method for extracting topics from the Holy Qur’an as a significant religious resource using a probabilistic approach. It is based on an unsupervised machine learning technique for topic modelling which is Latent Dirichlet Allocation (LDA), in order to extract the latent topics in the sacred text. The input are the 114 chapters of the Qur’an, in classical Arabic; each chapter represents a document. LDA, indeed, examines the documents of the Qur’an to learn what words tend to be used in the same documents. The topics are probabilistic word clusters with semantic relationships. The chapter then interprets and evaluates the LDA output by manually checking the derived clusters and examines if the words in each topic represents it. For a robust conclusion, we use human evaluator and existing knowledge resources.

4.2 Background

4.2.1 A Review on Clustering and Topic modelling Approaches

Topic models are statistical tools for discovering the hidden semantic structure in a collection of documents (Blei et al., 2003; Blei, 2012). Text mining and clustering methods, as well as latent topic models and neural embedding approaches, have all been developed over the years.

For representing the content of documents in large document collections, probabilistic latent topic models, such as probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003), as well as their numerous variants, have received extensive research. They offer a solid, unsupervised framework for carrying out shallow latent semantic analysis of the themes (or topics) mentioned in texts. The premise underlying all of these probabilistic latent topic model families is that there are latent variables, or topics, which influence how words in documents are generated.

Probabilistic topic models were initially introduced for text data, but they have also naturally found many uses in NLP. To detect main themes discussed in texts and to provide summaries for large text collections, previously discovered distributions of words over topics and distributions of topics over documents can be directly applied (see, e.g., Hofmann, 1999; Blei et al., 2003; Griffiths & Steyvers, 2004; Griffiths et al., 2007). Lu et al. (2011) examined task performance of pLSA and LDA as representative monolingual topic models in the context of common tasks like document clustering, text categorization, and ad hoc information retrieval.

Numerous natural language processing (NLP) tasks have made use of probabilistic topic models, including inferring captions for images (Blei and Jordan, 2003), sentiment analysis (e.g., Mei et al., 2007; Titov and McDonald, 2008), language modelling in information retrieval (e.g., Wei and Croft, 2006; Yi and Allan, 2009), document classification (e.g., Blei et al., 2003; Lacoste-Julien et al., 2008), word sense disambiguation (e.g., Boyd-Graber et al., 2007), modelling distributional similarity of terms (e.g., Ritter et al., 2010; Dinu and Lapata, 2010).

4.2.2 Latent Dirichlet Allocation for Topic Modelling

LDA is a powerful model and it is widely used to model latent topics from concise and short texts. LDA (Latent Dirichlet Allocation) is a generative probabilistic model for text corpora. LDA represents each document as a mixture of latent topics in which a topic is a multinomial distribution over words. LDA is a three-level hierarchical Bayesian model in which each collection item is represented as a finite mixture over an underlying set of topics. Each topic is thus represented as an infinite mixture over a collection of topic probabilities. The topic probabilities give an explicit representation of a document in the context of text modelling (Blei et al., 2003).

LDA uses Dirichlet distributions to model a topic per document and a word per topic model. It is a model that is applied to documents (a collection of words) and includes a latent variable for topic assignment for every word. The likelihood of a given word being used in combination with a specific topic represent the output values. As a result, words can be assigned to various clusters (Blei et al., 2003).

LDA is a powerful and extensively used model. However, it has a widespread technical flaw: it fails to cope with big vocabularies. In order to construct strong topic models, practitioners must dramatically reduce their vocabulary by eliminating the most and least frequently used words. This trimming may remove essential terms from huge collections, limiting the scope of the models (Dieng et al., 2020). Thus, LDA can be a good approach when the corpus is small like the Qur'an, and it allows exploring the thematic structure of the sacred text, instead of parsing through every detail in each chapter.

4.3 Methodology

In order to implement our system, we use Genism; an open-source Python library for natural language processing. Gensim[14] supports LDA implementation as one

[14] Gensim is a well-optimized library for topic modelling and document similarity analysis among the available Python NLP libraries

topic modelling algorithm. We build the model and set up its parameters. We then train the LDA model using the 114 chapters of the Qur'anic text in the Classical Arabic.

4.3.1 Latent Dirichlet Allocation LDA

LDA assumes that a document is made up of a mix of topics. The probability distribution of those topics is then used to generate words. Given a collection of documents, LDA tries to determine out what topics led to the creation of those documents.

LDA is a matrix factorization technique. The input corpus (collection of documents) can be represented as a document-term matrix in vector space as in Table 2, where the corpus is composed of N documents $D1$, $D2$, $D3$ … Dn and vocabulary size of M words $W1$, $W2$ … Wn. In this experiment, the corpus is the Qur'anic text consisting of 114 documents[15] and vocabulary size of 78,418 words, including Bismillah at the beginning of each chapter except chapter number 9 (At-Tawba). The frequency count of word Wj in Document Di is given by the value of the i,j cell. For example, $W2$ occurs 2 times in $D1$, 4 times in $D2$, and 2 times in D3.

	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	0	3	0

Table 2: Document-term Matrix as represented by LDA

[15] The Qur'an is organized in 114 chapters. Each chapter represents a document.

The Document-Term Matrix is subsequently converted via LDA into two lower-dimensional matrices, M1 and M2. M1 is a document-topics matrix as shown in Table 3, and M2 is a topic-terms matrix as shown in
Table 4, with dimensions (n, K) and (K, m), respectively, where n represents the number of documents, K represents the number of topics, and m represents the vocabulary size.

	T1	T2	T3	TK
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
Dn	1	0	1	0

Table 3: Document-topics matrix M1

	W1	W2	W3	Wm
T1	0	1	1	1
T2	1	1	1	0
T3	1	0	0	1
TK	1	1	0	0

Table 4: Topic-terms matrix M2

Although the two matrices already provide topic-word and document-topic distributions, the main goal of LDA is to improve these distributions. In order to improve these matrices, LDA employs sampling techniques. It goes over each word for each document and tries to replace the old topic – word assignment with a new one. With a probability P that is the product of two probabilities p1 and p2, a new topic k is assigned to the word w. Two probabilities, p1 and p2, are determined for each topic as the following:

The proportion of words in document d that are currently attributed to topic t	P1 = p (topic t / document d)
The proportion of assignments to topic t out of all documents originating from word w.	P2 = p (word w / topic t)

Then, the probability, product of p1 and p2, updates the current topic- word assignment. The model assumes that all existing word – topic allocations are correct except for the current word in this stage. Thus, changing the current word's topic to reflect the new likelihood. After a number of iterations (it is configured in model parameters), LDA reaches its point of convergence where the document-topic and topic-term distributions are reasonably good.

4.3.2 Model Parameters

We tune the different parameters of the model, and Table 5 below shows the parameters used settings. The parameters are:

- Alpha and Beta Hyperparameters – The density of document-topic is represented by alpha, and the density of topic-word is represented by beta. Documents with a higher alpha value have more topics, whereas those with a lower alpha value have fewer topics. On the other hand, topics with a high beta value have a large number of words in the corpus, while those with a low beta value have a small number of words.
- Number of Topics – Number of topics to be extracted from the corpus. The experiment used values ranging from 2 up to 20.
- Number of Iterations / passes – Maximum number of iterations needed for the LDA algorithm to converge. The model continues iterating until the topic/word assignments settle down and hardly change at all. The topic distributions are updated after each pass. Iterations must be high enough to guarantee that a sufficient number of documents converge before going on. When topics still don't make sense, we try increasing the parameters. Each round will iterate each document’s probability distribution assignments for a maximum of 50 times, moving to the next document before 50 times if it already reached convergence. We started by 10 and went up to 50, and number of passes is 40. Most of the documents should have converged by the time the passes are finished. If not, we make more iterations and passes.

Alpha	Beta	#Topics K	# Iterations	# Passes
1.0	1.0	8	50	40

Table 5: The parameters settings for training LDA model

4.4 Corpus

We import the Qur'anic text in the classical Arabic from the Tanzil project website, which is a genuine and certified source for the holy Qur'an text. The Qur'an's whole text is included in the downloaded file, with no diacritics. The file is divided into 114 documents, each of which is a chapter of the Qur'an. Each document in a separate text file, will be used to train the LDA model. We first pre-process the input documents using tokenization, removing stop- words[16].

4.5 Results

The LDA model was built with 8 topics using the Gensim implementation, where each topic is a combination of keywords and each keyword contributes a certain weightage to the topic. We chose an average value to be the number of topics. the number of topics was determined to be close to predefined topics' classification as in Qurany corpus[17] (Abbas, 2009). Later on, we compared the model with different versions with varying range of the number of topics to justify the results. Table 6 shows the keywords for each topic (8 topics) and the weightage (importance) of each keyword.

Topic NO.	Keywords	Topic No.	Keywords
1	('0.065256834,'آمنوا_وعملوا_الصالحات'), ('0.06195262 ,'أحسن'), ('0.052242465 ,'الإنسان'), ('0.023092434 ,'خلق'), ('0.015145395 ,'يعلم'), ('0.013572931 ,'علم'), ('0.013308047 ,'أمر'),	2	('0.030179208 ,'خلق'), ('0.022892356 ,'عذاب'), ('0.018320555 ,'السماء'), ('0.017156547 ,'كفروا'), ('0.015140297 ,'ربهم'), ('0.011654903 ,'الملائكة'), ('0.011439755 ,'الإنسان'),

[16] Qur'an stop-words are taken from the Qur'an analysis project by Karim Ouda (2015). The list comprises 809 words, verbs, and derivations.

[17] http://quranytopics.appspot.com/Frames/AboutQurany.html

	('0.011465097', 'كفروا'), ('0.010535497', 'يوم_القيامة'), ('0.010143053', 'بالحق')		('0.011136561', 'يعلمون'), ('0.011107912', 'ربه'), ('0.010948545', 'الناس')
3	('0.07799424', 'آمنوا'), ('0.04333792', 'والله'), ('0.042738605', 'الناس'), ('0.03898799', 'أنزل'), ('0.035447124', 'كفروا'), ('0.022805687', 'بالله'), ('0.022619778', 'سبيل'), ('0.022221725', 'كثيرا'), ('0.019030198', 'الصلاة'), ('0.018171648', 'الرسول')	4	('0.041670576', 'العالمين'), ('0.04155643', 'رسول'), ('0.032981228', 'فرعون'), ('0.027974773', 'السماء'), ('0.027031964', 'ربهم'), ('0.026133852', 'اليوم'), ('0.022494521', 'الكافرين'), ('0.022019353', 'أحد'), ('0.02069598', 'بالحق'), ('0.019315574', 'القوم')
5	('0.08075203', 'الناس'), ('0.03385899', 'موسى'), ('0.030999897', 'علم'), ('0.025651563', 'يوم_القيامة'), ('0.024483034', 'حق'), ('0.024377543', 'الحق'), ('0.024284752', 'ربكم'), ('0.020491023', 'عذاب'), ('0.020012695', 'خير'), ('0.019549813', 'هدى')	6	('0.04919948', 'كفروا'), ('0.024695322', 'النار'), ('0.023871372', 'ربنا'), ('0.020756548', 'السماء'), ('0.020222254', 'عذاب'), ('0.018623084', 'آياته'), ('0.017977882', 'الحق'), ('0.014977201', 'العذاب'), ('0.0148941735', 'آمنوا'), ('0.014613527', 'يعملون')
7	('0.041767485', 'الكتاب'), ('0.0325054', 'والله'), ('0.030614592', 'خير'), ('0.028221443', 'ربي'), ('0.02592292', 'ويوم'), ('0.025729772', 'آمنوا'), ('0.024477309', 'إبراهيم'),	8	('0.044974405', 'ربنا'), ('0.034364205', 'الناس'), ('0.031965658', 'كفروا'), ('0.026441472', 'ربهم'), ('0.02483638', 'الحق'), ('0.02195795', 'موسى'), ('0.021803726', 'آمنوا'),

	('0.01813483 ,الشيطان'), ('0.0174927 ,كفروا'), ('0.017327826 ,ورسوله')		('0.021721635 ,النار'), ('0.021400757 ,الظالمين'), (' ,خلق_السماوات_والأرض' 0.019344022)

Table 6: Keywords for each topic with their weightage

To interpret that, for example, the top 10 keywords that represent topic 1 are:

{' بالحق','، يوم القيامة' ,' كفروا','، أمر' ,' علم','يعلم' ,'خلق' ,'أحسن' , ' آمنوا_وعملوا_الصالحات'}, with the first word contributing the most to the topic. By looking up these words, we observed that they tend to occur in similar verses representing one theme in Qurany corpus, Faith. However, some words represent another theme as well. For example, the words 'آمنوا_وعملوا_الصالحات' occurs in another theme, that is Man and moral relations.

4.6 Evaluation

4.6.1 Manual Technique

To evaluate the model, we examined the generated clusters manually. We examined by eye the interpretability of the LDA model, and the extent to which keywords are related and represent each topic. Table 7 displays the keywords representing each topic (0,1, ... ,9); numbering starts from 0, which mean topic #1. Other than keywords, we looked at the document that a specific topic has contributed to the most; for each topic, there is representative document/chapter that contributed the most to that topic. The table lists, for each topic, the keywords and the top representative document. The Perc Contribution field represents the topic's percentage contribution to the document. For example, chapter 96 (العلق /Al-Alaq) has contributed the most to topic1, chapter 92 (الليل/ Al-Lail) has contributed the most to topic 2, chapter 4 (النساء/ An-Nisa) has contributed the most to topic 3, chapter 18 (الكهف/ Al-Kahf) has contributed the most to topic 4, chapter 22 (الحج/ Al-Hajj) has contributed the most to topic 5, chapter 41 (فصلت/ Fussilat) has contributed the most to topic 6, chapter 19 (مريم/Maryam) has contributed the most to topic 7, and chapter 14 (ابراهيم/ Ibrahim) has contributed the most to topic 8.

	Topic_Num	Topic_Perc_Contrib	Keywords	Representative Text
0	0.0	0.9340	أمنوا_وعملوا_الصالحات, أحسن, الإنسان, خلق, يعلم, علم, أمر, كفروا, يوم_القيامة, بالحق	[بسم, الله, الرحمن_الرحيم, اقرأ, باسم_ربك, خلق, خلق, الإنسان, اقرأ, وربك, الأكرم, علم, بالقلم, علم, الإنسان, يعلم, الإنسان, ليطغى, رآه, استغنى, ربك, الرجعى, أرأيت, ينهى, عبدا, صلى, أرأيت,] [الهدى, أمر, بالتقوى, أرأيت, كذب_وتولى, يعلم, الله, يرى, بينه, لنسفعا, بالناصية, ناصية, كاذبة, خاطئة, فليدع, ناديه, سندع, الزبانية, تطعه, واسجد, واقترب
1	1.0	0.7411	خلق, عذاب, السماء, كفروا, ربهم, الملائكة, الإنسان, يعلمون, ربه, الناس	[بسم, الله, الرحمن_الرحيم, والليل, يغشى, والنهار, تجلى, خلق, الذكر_والأنثى, سعيكم, لشتى, أعطى, واتقى, وصدق, بالحسنى, فسنيسره, لليسرى, بخل, واستغنى, وكذب, بالحسنى, فسنيسره,] [للعسرى, يغني, ماله, تردى, للهدى, للآخرة, والأولى, فأنذرتكم, نارا, تلظى, يصلاها, الأشقى, كذب_وتولى, وسيجنبها, الأتقى, يؤتي, ماله, يتزكى, لأحد, نعمة, تجزى, ابتغاء_وجه, ربه, الأعلى, يرضى
2	2.0	0.9977	أمنوا, والله, الناس, أنزل, كفروا, بالله, سبيل, كثيرا, الصلاة, الرسول	[بسم, الله, الرحمن_الرحيم, الناس, اتقوا, ربكم, خلقكم, واحدة, وخلق, زوجها, وبث, رجالا, كثيرا, ونساء, واتقوا, الله, تساءلون, والأرحام, الله, رقيبا, وآتوا, اليتامى, أموالهم, تتبدلوا, الخبيث, بالطيب,] [تأكلوا, أموالهم, أموالكم, حوبا, كبيرا, خفتم, تقسطوا, اليتامى, فانكحوا, طاب, النساء, مثنى, وثلاث, ورباع, خفتم, تعدلوا, فواحدة, ملكت, أيمانكم, أدنى, تعولوا, وآتوا, النساء, صدقاتهن, نحلة, طبن, ...نفسا, فكلوه
3	3.0	0.9929	العالمين, رسول, فرعون, السماء, ربهم, اليوم, الكافرين, أحد, بالحق, القوم	[بسم, الله, الرحمن_الرحيم, طسم, آيات, الكتاب, المبين, باخع, يكونوا, مؤمنين, نشأ, ننزل, السماء, آية, فظلت, أعناقهم, خاضعين, يأتيهم, ذكر, الرحمن, محدث, معرضين, كذبوا, فسيأتيهم, أنباء,] [يستهزئون, يروا, الأرض, أنبتنا, زوج, كريم, لآية, مؤمنين, ربك, العزيز_الرحيم, نادى, ربك, موسى, ائت, القوم_الظالمين, قوم, فرعون, يتقون, أخاف, يكذبون, ويضيق, صدري, ينطلق, لساني, فأرسل, ... ,هارون, ذنب, فأخاف, يقتلون, فاذهبا
4	4.0	0.9610	الناس, موسى, علم, يوم_القيامة, حق, الحق, ربكم, عذاب, خير, هدى	[بسم, الله, الرحمن_الرحيم, الناس, اتقوا, ربكم, زلزلة, الساعة, عظيم, يوم, ترونها, تذهل, مرضعة, أرضعت, وتضع, حمل, حملها, وترى, الناس, سكارى, بسكارى, عذاب, الله, شديد, الناس, يجادل,] [الله, علم, ويتبع, شيطان, مريد, كتب, تولاه, يضله, ويهديه, عذاب, السعير, الناس, ريب, البعث, خلقناكم, تراب_نطفة, علقة, مضغة, مخلقة, لنبين, ونقر, الأرحام, نشاء, مسمى, نخرجكم, طفلا, ...,لتبلغوا, أشدكم, يتوفى, يرد, أرذل
5	5.0	0.9917	كفروا, النار, ربنا, السماء, عذاب, آياتنا, الحق, العذاب, أمنوا, يعملون	[بسم, الله, الرحمن_الرحيم, تنزيل, الرحمن_الرحيم, كتاب, فصلت, آياته, قرآنا_عربيا, لقوم, يعلمون, بشيرا_ونذيرا, فأعرض, يسمعون, قلوبنا, أكنة, تدعونا, آذاننا, وقر, حجاب, فاعمل, عاملون, بشر,] [يوحى_أنما, إلهكم_إله, واحد, فاستقيموا, واستغفروه, ويل, للمشركين, يؤتون, الزكاة, بالآخرة_كافرون, أمنوا_وعملوا_الصالحات, أجر_ممنون, أئنكم, لتكفرون, خلق, الأرض, يومين, وتجعلون, أندادا, ...العالمين, رواسي, وبارك, وقدر
6	6.0	0.9908	الكتاب, والله, خير, ربي, ويوم, أمنوا, إبراهيم, الشيطان, كفروا, ورسوله	[بسم, الله, الرحمن_الرحيم, كهيعص, ذكر, رحمت, ربك, عبده, زكريا, نادى_ربه, نداء, خفيا, العظم, واشتعل, الرأس, شيبا, أكن, بدعائك, شقيا, خفت, الموالي, ورائي, وكانت, امرأتي, عاقرا, وليا,] [يرثني, ويرث, آل, يعقوب, واجعله, رضيا, زكريا, نبشرك, بغلام, اسمه, يحيى, نجعل, سميا, يكون_غلام, وكانت, امرأتي, عاقرا, بلغت, الكبر, عتيا, ربك, هين, خلقتك, ك, اجعل, آية, آيتك, تكلم, الناس, ...,ثلاث, ليال, سويا, فخرج, قومه
7	7.0	0.9923	ربنا, الناس, كفروا, ربهم, الحق, موسى, أمنوا, النار, الظالمين, خلق_السماوات_والأرض	[بسم, الله, الرحمن_الرحيم, الر, كتاب_أنزلناه, لتخرج, الناس, الظلمات_النور, بإذن, ربهم, صراط, العزيز, الحميد, الله, السماوات_الأرض, وويل, للكافرين, عذاب_شديد, يستحبون, الحياة_الدنيا,] [الآخرة, ويصدون_سبيل, الله, ويبغونها_عوجا, ضلال_بعيد, أرسلنا, رسول, بلسان, قومه, ليبين, فيضل, الله, يشاء_ويهدي, يشاء, العزيز_الحكيم, أرسلنا, موسى, بآياتنا, أخرج, قومك, الظلمات_النور, ...,وذكرهم, بأيام, الله, لآيات_صبار

Table 7: Top representative documents for each topic

Here we determine the topic number that has the highest percentage contribution in a given document/ chapter. Table 8 shows the first 12 document/chapters in the Qur'an, and the dominant topic in each. It also shows the keywords in a document/ chapter along with their contributions to the dominant topic. The reader here can observe that the same word can contribute to different topics.

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	5.0	0.9416	كفروا, النار, ربنا, السماء, عذاب, آياتنا, الحق, العذاب, آمنوا, يعملون	[بسم, الله, الرحمن_الرحيم, كفروا, أهل_الكتاب, والمشركين, منفكين, تأتيهم, البينة, رسول, الله, يتلو, صحفا, مطهرة, كتب, قيمة, تفرق, أوتوا_الكتاب, جاءتهم, البينة, أمروا,] ...ليعبدوا, الله, مخلصين_الدين, حنفاء, ويقيموا, الصلاة, ويؤتوا, الزكاة, دين, القيمة, كفروا, أهل_الكتاب, والمشركين, نار_جهنم, خالدين, شر, البرية, ...آمنوا_وعملوا_الصالحات, خير, البرية, جزاؤهم, ربهم, جنات_عدن, تجري_الأنهار, خالدين_رضي, ال
1	1	3.0	0.4303	العالمين, رسول, فرعون, السماء, ربهم, اليوم, الكافرين, أحد, بالحق, القوم	[بسم, الله, الرحمن_الرحيم, والكتاب, المبين, جعلناه, قرآنا_عربيا, تعقلون, الكتاب, حكيم, أفنضرب, الذكر, صفحا, قوما, مسرفين, أرسلنا, نبي, الأولين, يأتيهم, نبي, يستهزئون, فأهلكنا, أشد, بطشا, ومضى, الأولين, سألتهم_خلق, السماوات_والأرض, ليقولن, خلقهن, العزيز, العليم, الأرض, مهدا, سبلا, تهتدون, نزل, السماء_ماء, بقدر, ...فأنشرنا, بلدة_ميتا, تخرجون, خلق, الأزواج, الفلك, والأنعام, تركبون, لتستووا, ظهوره, تذ
2	2	2.0	0.8580	آمنوا, والله, الناس, أنزل, كفروا, بالله, سبيل, كثيرا, الصلاة, الرسول	[بسم, الله, الرحمن_الرحيم, سبح_لله, السماوات_الأرض, العزيز_الحكيم, آمنوا, تقولون, تفعلون, كبر, مقتا, الله, تقولوا, تفعلون, الله, يحب, يقاتلون, سبيله, صفا, بنيان,] ...مرصوص, موسى_لقومه, قوم, تؤذونني, تعلمون, رسول, الله, زاغوا, أزاغ, الله, قلوبهم, والله, يهدي_القوم, الفاسقين, عيسى_مريم, بني_إسرائيل, رسول, الله, مصدقا, ...يدي, التوراة, ومبشرا, برسول, يأتي, اسمه, أحمد, جاءهم, بالبينات, سحر_مبين, أظلم_ممن
3	3	5.0	0.7072	كفروا, النار, ربنا, السماء, عذاب, آياتنا, الحق, العذاب, آمنوا, يعملون	[بسم, الله, الرحمن_الرحيم, تنزيل, الكتاب, الله, العزيز_الحكيم, السماوات_والأرض, لآيات, للمؤمنين, خلقكم, بث, دابة, آيات, لقوم_يوقنون, واختلاف_الليل, والنهار, أنزل, الله, السماء, رزق, فأحيا, الأرض_موتها, وتصريف, الرياح, آيات, لقوم_يعقلون, آيات, الله, نتلوها_عليك_بالحق, حديث, الله, وآياته, يؤمنون, ويل, أفاك, أثيم, يسمع, آيات, الله, ...تتلى, يصر, مستكبرا, يسمعها, فبشره, بعذاب_أليم, علم, آياتنا, اتخذها, ه
4	4	7.0	0.7478	ربنا, الناس, كفروا, ربهم, الحق, موسى, آمنوا, النار, الظالمين, خلق_السماوات_والأرض	[بسم, الله, الرحمن_الرحيم, اقتربت, الساعة, وانشق, القمر, يروا, آية, يعرضوا, ويقولوا, سحر, مستمر, وكذبوا, واتبعوا_أهواءهم, أمر, مستقر, جاءهم, الأنباء, مزدجر,] حكمة, بالغة, تغن, النذر, فتول, يوم, يدع, الداع, نكر, خشعا, أبصارهم, يخرجون, الأجداث, جراد, منتشر, مهطعين, الداع, الكافرون, يوم, عسر, كذبت_قبلهم, قوم_نوح, ... فكذبوا, عبدنا, مجنون, وازدجر, فدعا, ربه, مغلوب, فانتصر, ففتحنا, أبواب, السماء, بماء
5	5	5.0	0.9071	كفروا, النار, ربنا, السماء, عذاب, آياتنا, الحق, العذاب, آمنوا, يعملون	[بسم, الله, الرحمن_الرحيم, والمرسلات, عرفا, فالعاصفات, عصفا, والناشرات, نشرا, فالفارقات, فرقا, فالملقيات, ذكرا, عذرا, نذرا, توعدون, لواقع, النجوم, طمست, السماء,] ...ماء, مهين, فجعلناه, قرار, مكين, قدر, معلوم, فقدرنا, القادرون, ويل, للمكذبين
6	6	3.0	0.3744	العالمين, رسول, فرعون, السماء, ربهم, اليوم, الكافرين, أحد, بالحق, القوم	[بسم, الله, الرحمن_الرحيم, تنزيل, الكتاب, الله, العزيز, العليم, غافر, الذنب, وقابل, التوب, شديد_العقاب, الطول, إله, المصير, يجادل, آيات, الله, كفروا, يغررك, تقلبهم,] البلاد, كذبت_قبلهم, قوم_نوح, والأحزاب, وهمت, أمة, برسولهم, ليأخذوه, وجادلوا, بالباطل, ليدحضوا, الحق, فأخذتهم, عقاب, حقت_كلمت, ربك, كفروا, أصحاب_النار, ...يحملون, العرش, يسبحون_بحمد, ربهم, ويؤمنون, ويستغفرون, آمنوا, ربنا, وسعت, رحمة, و
7	7	4.0	0.4892	الناس, موسى, علم, يوم_القيامة, حق, الحق, ربكم, عذاب, خير, هدى	[بسم, الله, الرحمن_الرحيم, والسماء, البروج, واليوم, الموعود, وشاهد, ومشهود, قتل, أصحاب, الأخدود, النار, الوقود, قعود, يفعلون, بالمؤمنين, شهود, نقموا, يؤمنوا, بالله,] العزيز, الحميد, ملك, السماوات_والأرض, والله, شهيد, فتنوا, المؤمنين_والمؤمنات, يتوبوا, عذاب, جهنم, عذاب_الحريق, آمنوا_وعملوا_الصالحات, جنات, تجري_الأنهار, ...الفوز, الكبير, بطش, ربك, لشديد, يبدئ, ويعيد, الغفور, الودود, العرش, المجيد, فعا
8	8	5.0	0.9190	كفروا, النار, ربنا, السماء, عذاب, آياتنا, الحق, العذاب, آمنوا, يعملون	[بسم, الله, الرحمن_الرحيم, سبح, اسم, ربك, الأعلى, خلق, قدر, فهدى, أخرج, المرعى, فجعله, غثاء, أحوى, سنقرئك, تنسى, شاء, الله, يعلم, الجهر, يخفى, ونيسرك,] لليسرى, فذكر, نفعت, الذكرى, سيذكر, يخشى, ويتجنبها, الأشقى, يصلى, النار, الكبرى, يموت, يحيى, أفلح, تزكى, وذكر, اسم, ربه, فصلى, تؤثرون, الحياة_الدنيا, [والآخرة, خير_وأبقى, الصحف, الأولى, صحف, إبراهيم, وموسى
9	9	5.0	0.1820	كفروا, النار, ربنا, السماء, عذاب, آياتنا, الحق, العذاب, آمنوا, يعملون	[بسم, الله, الرحمن_الرحيم, أرأيت, يكذب, بالدين, يدع, اليتيم, يحض, طعام_المسكين, فويل, للمصلين, صلاتهم, ساهون, يراءون, ويمنعون, الماعون]
10	10	4.0	0.7161	الناس, موسى, علم, يوم_القيامة, حق, الحق, ربكم, عذاب, خير, هدى	[بسم, الله, الرحمن_الرحيم, طه, أنزلنا, القرآن, لتشقى, تذكرة, يخشى, تنزيلا, خلق, الأرض, والسماوات, العلى, الرحمن, العرش, استوى, السماوات_الأرض, الثرى,] تجهر, بالقول, يعلم, السر, وأخفى, الله, إله, الأسماء_الحسنى, أتاك_حديث, موسى, رأى, نارا, لأهله, امكثوا, آنست_نارا, آتيكم, بقبس, أجد, النار, هدى, أتاها, نودي, ...موسى, ربك, فاخلع, نعليك, بالواد, المقدس, طوى, اخترتك, فاستمع, يوحى, إنني, الله, إله, فاعب
11	11	7.0	0.8375	ربنا, الناس, كفروا, ربهم, الحق, موسى, آمنوا, النار, الظالمين, خلق_السماوات_والأرض	[بسم, الله, الرحمن_الرحيم, أنى, الإنسان, الذكرى, مذكورا, خلقنا_الإنسان, نطفة, أمشاج, نبتليه, فجعلناه, سميعا_بصيرا, هديناه, السبيل, شاكرا, كفورا, أعتدنا,] للكافرين, سلاسل, وأغلالا, وسعيرا, الأبرار, يشربون, كأس, مزاجها, كافورا, عينا, يشرب, عباد, الله, يفجرونها, تفجيرا, يوفون, بالنذر, ويخافون, يوما, شره, مستطيرا, ...ويطعمون, الطعام, حبه, مسكينا, ويتيما, وأسيرا, نطعمكم, لوجه, الله, نريد, جزاء, شكورا, ن
12	12	2.0	0.4692	آمنوا, والله, الناس, أنزل, كفروا, بالله, سبيل, كثيرا, الصلاة, الرسول	[بسم, الله, الرحمن_الرحيم, سورة, أنزلناها, وفرضناها, وأنزلنا, آيات_بينات, تذكرون, الزانية, والزاني, فاجلدوا, واحد, مائة, جلدة, تأخذكم, رأفة, دين, الله, تؤمنون_بالله, واليوم_الآخر, وليشهد, عذابهما, طائفة, المؤمنين, الزاني, ينكح, زانية, مشركة, والزانية, ينكحها, زان, مشرك, وحرم, المؤمنين, يرمون, المحصنات, يأتوا, بأربعة, شهداء, ...فاجلدوهم, ثمانين, جلدة, تقبلوا, شهادة, الفاسقون, تابوا_وأصلحوا, الله

Table 8: Dominant topic for each document

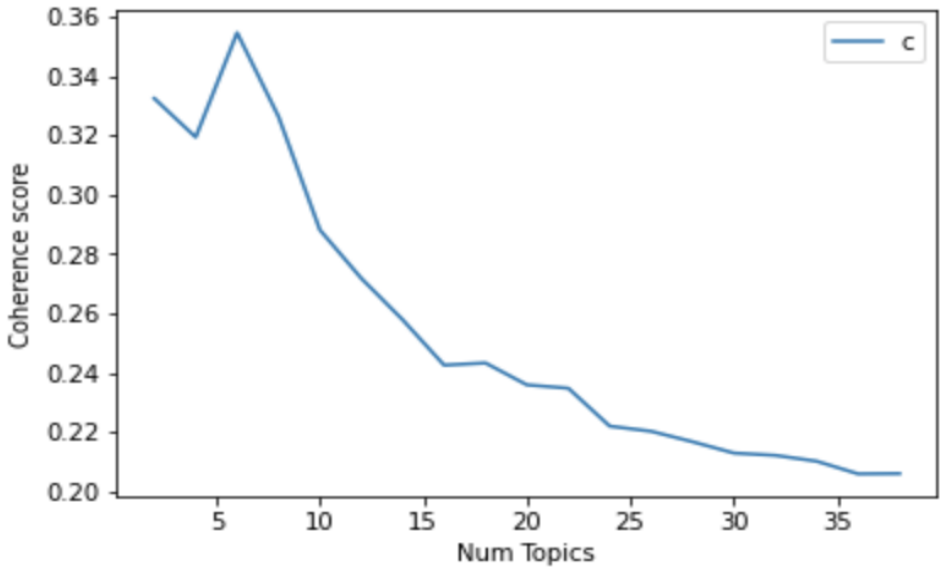
The topic distribution among documents is another way to help interpreting the LDA model. We determine how widely topics were discussed by assessing the volume and distribution of topics. Table 9 provides information such as number of documents contributing to each topic and their percentage for each topic. The table displays the information for the first 8 chapters/documents in the Qur'an. For example, the dominant topic in chapter 3 is topic 3, and the number of documents/chapters contributing to the topic are 26 chapters (out of 114 chapters) with a percentage of about 10%.

	Dominant_Topic	Topic_Keywords	Num_Documents	Perc_Documents
0.0	5.0	كفروا, النار, ربنا, السماء, عذاب, آياته, الحق, العذاب, آمنوا, يعملون	8.0	0.0702
1.0	3.0	العالمين, رسول, فرعون, السماء, ربهم, اليوم, الكافرين, أحد, بالحق, القوم	3.0	0.0263
2.0	2.0	آمنوا, والله, الناس, أنزل, كفروا, بالله, سبيل, كثيرا, الصلاة, الرسول	26.0	0.2281
3.0	5.0	كفروا, النار, ربنا, السماء, عذاب, آياته, الحق, العذاب, آمنوا, يعملون	21.0	0.1842
4.0	7.0	ربنا, الناس, كفروا, ربهم, الحق, موسى, آمنوا, النار, الظالمين, خلق_السماوات_والأرض	8.0	0.0702
5.0	5.0	كفروا, النار, ربنا, السماء, عذاب, آياته, الحق, العذاب, آمنوا, يعملون	32.0	0.2807
6.0	3.0	العالمين, رسول, فرعون, السماء, ربهم, اليوم, الكافرين, أحد, بالحق, القوم	7.0	0.0614
7.0	4.0	الناس, موسى, علم, يوم_القيامة, حق, الحق, ربكم, عذاب, خير, هدى	9.0	0.0789

Table 9: Topic distribution across documents

4.6.2 Topic Coherence

Using eight as the number of topics, we generated a LDA model with coherence score of 0.3262. Choosing bigger values of K (number of topics) leads to the same keyword being repeated in multiple topics, which explains the ambiguity of the Arabic language as same words can be used in different contexts. To determine the optimal number of topics, we created several LDA models with varying numbers of topics and chose the one with the highest coherence value. Figure 9 shows coherence values corresponding to the LDA model with respective number of topics. The figure illustrates that the best coherence score was achieved with LDA model composed of 2,4,6,8, and 10 as the number of topics.




```
Num Topics = 2  has Coherence Value of 0.3324
Num Topics = 4  has Coherence Value of 0.3193
Num Topics = 6  has Coherence Value of 0.3545
Num Topics = 8  has Coherence Value of 0.3262
Num Topics = 10  has Coherence Value of 0.2881
Num Topics = 12  has Coherence Value of 0.2718
Num Topics = 14  has Coherence Value of 0.2578
Num Topics = 16  has Coherence Value of 0.2425
Num Topics = 18  has Coherence Value of 0.2433
Num Topics = 20  has Coherence Value of 0.2359
Num Topics = 22  has Coherence Value of 0.2348
Num Topics = 24  has Coherence Value of 0.222
Num Topics = 26  has Coherence Value of 0.2203
Num Topics = 28  has Coherence Value of 0.2167
Num Topics = 30  has Coherence Value of 0.2129
Num Topics = 32  has Coherence Value of 0.2121
Num Topics = 34  has Coherence Value of 0.2102
Num Topics = 36  has Coherence Value of 0.2059
Num Topics = 38  has Coherence Value of 0.206
```

Figure 9: Coherence values corresponding to the LDA model with respective number of topics (produced by the output of Matplotlib[18])

The best coherence occurs when number of topics is 2 and 6. The verses of the Qur'an can be grouped in two general themes, they are Makki and Mdani (Cook, 2000). Each theme is oriented around specific concepts that have common key words. Usually Makki verses discusses faith, pillar of Islam and the basics of Aqeedah, while Maddani verses discusses rules, and regulations that govern people life. Choosing a higher value yield more specific sub-topics, when K=6.

[18] Matplotlib is a comprehensive library for creating visualizations in Python.

4.6.3 Visualization using pyLDAvis

We use the pyLDAvis[19] package's interactive chart to visualize the topics-keywords to get an insightful view on the derived clusters. The pyLDAvis is a Python library for visualising interactive topic models (Sievert and Shirley, 2014). The package uses data from a fitted LDA topic model to create a web-based interactive representation.

Here we examine the produced topics and the associated keywords. Figure 10 illustrates the output of pyLDAvis. Each topic is represented by a bubble on the left-hand plot. The greater the bubble, the more prominent the topic. Instead of being grouped in one region, a strong topic model will have reasonably large, non-overlapping bubbles dispersed around the chart. When selecting a bubble (a topic), the words and bars on the right-hand side will be updated to show the main keywords that make up the selected topic. Figure 11 shows topic 1 along with the top words and their contribution. Appendix B contains the figures representing the rest of the seven topics.

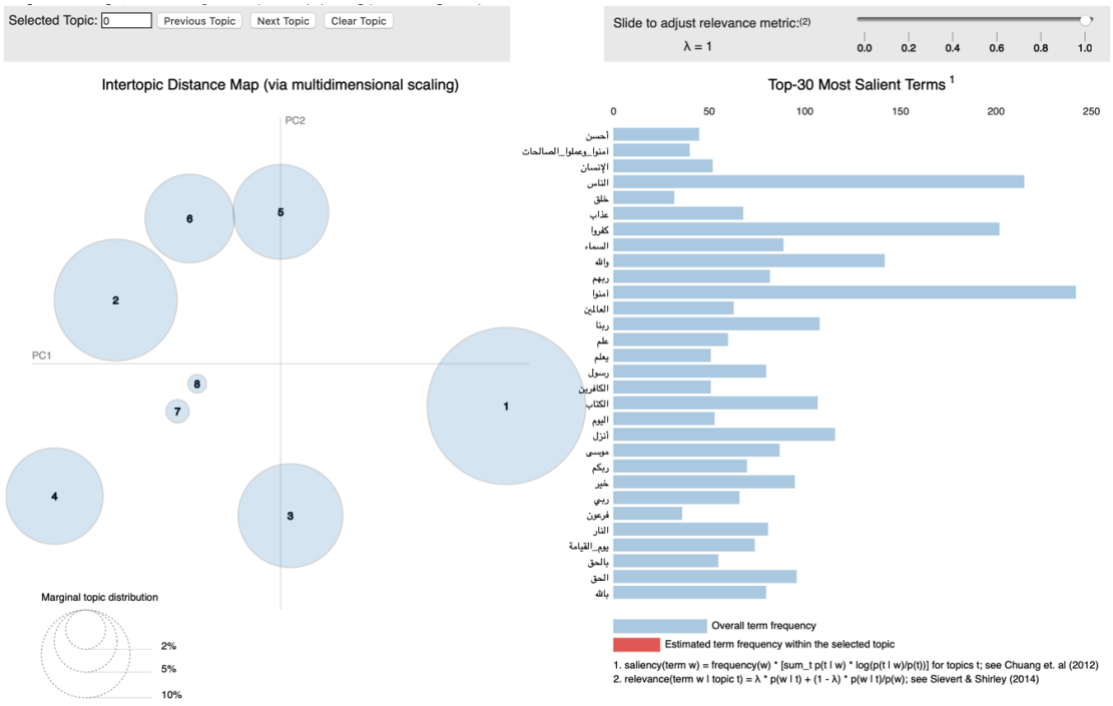


Figure 10: Output of pyLDAvis

[19] https://pyldavis.readthedocs.io/en/latest/readme.html

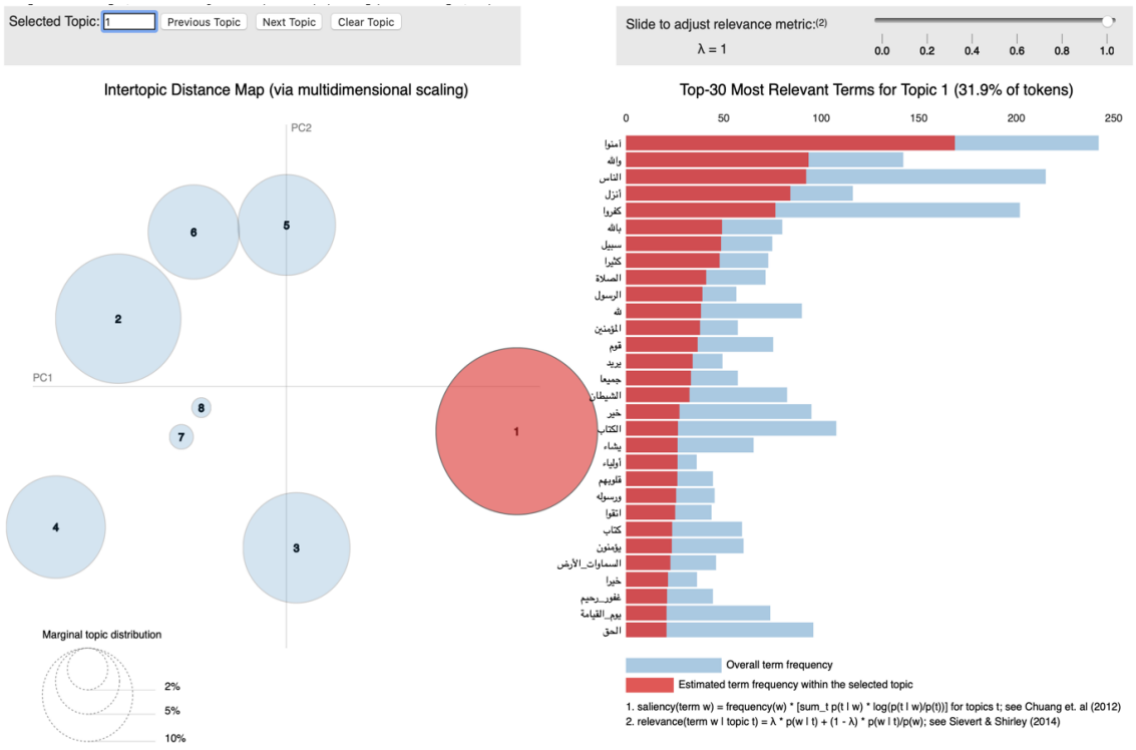


Figure 11: Topic 1 represented by pyLDAvis

4.7 Discussion

The LDA algorithm is a well-known and easy-to-implement algorithm to organize a group of documents into themes or topics. Each topic is a collection of keywords in a particular proportion. Here, we examined the ability of a probabilistic approach, LDA, to model latent topics from a concise and short text like the Qur'an. The derived keywords represent prevalent themes in the Qur'an, including the book, Creation, Believers, Disbelievers, Afterlife, and stories of Past Prophets. However, it is challenging to model these topics due to them overlapping, as the same keywords appear in multiple clusters describing multiple topics. Furthermore, because references to the main themes appear throughout the Qur'anic text, it might be difficult to distinguish between them. Each time one of these topics is mentioned, the Qur'an emphasises a distinct aspect of it in the exact phrasing of sections.

Hence, the challenge is determining how to extract high-quality, unique, and relevant topics. It depends on the text preparation quality and the approach for selecting the most appropriate number of topics. Both issues were observed

during this experiment. For example, after removing the stop words, we had to filter out frequent words with all their derivatives to get to the point where results are interpretable and meaningful. Such words are frequent and tend to appear without distinguishing the meaning or the topic. Here is a list of these words:

{'قال', 'قالوا', 'قل', 'كنا', 'كانوا', 'قيل', 'كنتم', 'وقالوا',
'تكن',
'وكانوا','فقالوا','وقال','كانت','يقول','يا','منا','فقال','قالت','من',
'ا','شيئا','جاء','شيء','تر','كنت','قبله','يكن','يقولوا','قليلا','يقول'
ون'}

Another challenge is choosing the number of topics that best describes topics distribution in the Qur'an. we had to train multiple LDA models aiming to reach a point where no word overlapping occurs between clusters as much as possible. Despite all that, some of the derived clusters still have the same keywords repeated in other clusters. Maintaining high coherence is challenging as no stopping point is predefined, with no gold standard outcomes. Intuition and domain knowledge were mainly used.

Moreover, another concern is the robustness of the generated clusters and how to evaluate them. To evaluate the quality of word clusters generated, manual techniques can be employed using an approach that has been suggested in (Chang et al., 2009)[20]. For example, the procedure is to take the top 10 words in a topic and replace one/two of them with a word that occurs very low in the sequence for the topic. Then, a human evaluator[21] has to spot the 'odd-one-out.' The process was repeated multiple times using the different clusters. The technique was applied to English text (Chang et al., 2009). However, when trying it here, it didn't help. We may attribute that to the ambiguity associated with the Arabic language and the complex dependencies that appeared when describing topics in the Qur'an. For example, words like the book, Moses, and creation can

[20] A technique for evaluating the interpretability of a topic model was presented by Chang et al. (2009). To explicitly assess the quality of the topics the model infers as well as how well the algorithm allocates subjects to documents, we developed two human evaluation tasks. The first, word incursion, examines how semantically "cohesive" the topics a model infers are and determines whether or not topics match up with human natural groupings. The second, topic incursion, evaluates how closely a topic model's breakdown of a document into a collection of topics accords with what people often associate a topic with in a document.

[21] We used an Arabian student who is a good reader of the Qur'an.

fit different topics. Therefore, it is challenging to apply such a test when dealing with the Qur'anic text.

4.8 Summary and Conclusion

This chapter describes an experiment to build a topic model using Gensim's LDA and visualize the clusters using pyLDAvis. The experiment examined the ability of a probabilistic approach, LDA, to model latent topics from a concise and short text like the Qur'an. The derived clusters may provide scholars with directions to look at the chapters that contributed to a specific topic and explore patterns that may support the teachings. We then combined and presented the results to provide meaningful insights to help judge how well the LDA algorithm models the topics in the Qur'an. We also calculated the volume and percentage contribution of each topic to determine its significance. Next, the model was evaluated using the coherence metric, and it was used to find the optimal number of topics to generate meaningful and interpretable topics.

Latent Dirichlet Allocation (LDA), as a statistical method, was mainly adopted in most of the works related to Qur'anic topic modelling (Siddiqui et al., 2013; Alhawarat, 2015; Panju, 2014; Putra et al., 2017; Rolliawati et al., 2020). However, they were limited to a unigram model and examined specific chapters and documents of the text. Moreover, most research projects focused on the translation of the Qur'an into different languages instead of the original text. This chapter examined LDA using the whole chapters of the book in Arabic. By looking at the key terms in each topic, It is observed that it is hard to identify one unique topic for each document. Interpreting the topic is challenging when multiple topics share same keywords in different contexts. Moreover, LDA is tough to tune, and results are troublesome to evaluate.

As a clustering problem, we will look at another algorithm, K-Means, in which given verses are clustered into a known number of groups, depending on the distance of words in verses from centroids of each group. K-Means algorithm is simpler in implementation as compared to LDA. While, with LDA, clustering documents is based on word usage, Word embeddings can generate better features for clustering. Thus, we examine the distributional representation of text, word embeddings, in the next chapter to capture the distributional semantics of words.

Chapter5
Understanding the Qur’an with Deep Neural Networks – Experiment with Word2vec

5.1 Introduction

Deep learning models built on dense vector representations are increasingly being used in recent NLP research. The success of distributed representation of words as embeddings has driven this trend (Mikolov et al., 2010, 2013a). Deep learning-based NLP models use these embeddings to represent their words, phrases, and even sentences, which is a significant distinction between conventional word count models and deep learning models.

Word2vec is a novel distributional semantic technique based on deep learning. Word2vec (Mikolov et al., 2013a) learns word vectors using contextual information. These vectors try to capture the characteristics of the neighbours of a word. The main advantage of distributional vectors is that they capture similarity between words. There is no explicitly computed co-occurrence matrix in this technique, nor is there an explicit association feature between pairs of words; instead, the network learns the regularities and distribution of the words implicitly (Ferrone and Zanzotto, 2020).

Hence, the research question that will be answered in this chapter is how to present a semantic representation of the Qur’an words that capture the semantic similarity between its verses. We investigate word-level neural networks' capability to learn the semantic representation of words in a small corpus. In particular, we examine the optimality of deep learning methods and word embeddings to achieve insightful semantic representations from a small specific-domain corpus, the Qur'an. Here, we experiment using the two architectures, CBOW and Skip-gram models, with different settings of hyperparameters to learn semantic representations of Qur’an words. We then evaluate the effectiveness of derived representation by performing the semantic categorization test.

5.2 Background

5.2.1 Word Embeddings

Word embeddings are a modern approach for representing text where individual words are represented as real-valued vectors in a predefined vector space. Each word is mapped to one vector and the vector values are learned in a way that resembles a neural network (Goldberg, 2017). The vector space representation of the words provides a projection where words with similar meanings are locally clustered within the space (Firth, 1957).

The term, word embeddings, was originally coined by Bengio et al. (2003) who trained them in a neural language model together with the model's parameters. However, Collobert and Weston demonstrated the power of pre-trained word embeddings in their paper (Collobert and Weston, 2008). They established word embeddings as a highly effective tool when used in downstream tasks, while also presenting a neural network architecture that has become the basis for later models. The authors built a model based on the architecture of Collobert and Weston (Mikolov et al. 2013a). The work created word2vec, an efficient tool that uses the surrounding words to represent the target words with a Neural Network whose hidden layer encodes the word representation.

5.2.2 DL for specific-domain corpora

Word embedding models were mainly tested on large texts with a general domain. Such studies involve general domain texts, and their results do not apply to a specific domain like sociological and biomedical literature. Fewer studies have used deep learning methods to represent words from a small corpus. A study (Chiu et al., 2016) found that more extensive corpora do not necessarily produce better word embeddings in the biomedical domain. This result is confirmed in other articles (Muneeb et al., 2015), with different tests and corpus. Another study (Altszyler et al., 2016) compared the capabilities of Skip-gram and LSA to learn accurate word embeddings in small text corpora. They investigated the ability of both models to identify word associations in dream reports, which could bring new insights into this area of psychology.

5.2.3 Word2vec

Word2vec is a shallow, two-layer artificial neural network that processes text by converting them into numerical "vectorized" words. It outputs a vector space, usually hundreds of dimensions, with each unique word in the corpus represented with that vector space generated. Word vectors are positioned in the vector space so that the words that share common contexts are located close to one another in that multidimensional space. Word2vec captures both syntactic and semantic similarities between the words. Word2vec generates an embedding with two different deep learning architectures: Skip Gram and Continuous Bag of Words (CBOW). The CBOW takes the word's context as input and tries to predict the word according to the context. Skip-gram model actually works in a reverse pattern than the CBOW model; it tries to predict context words based on given target words. Figure 12 illustrates the two architectures (Mikolov et al., 2013a).

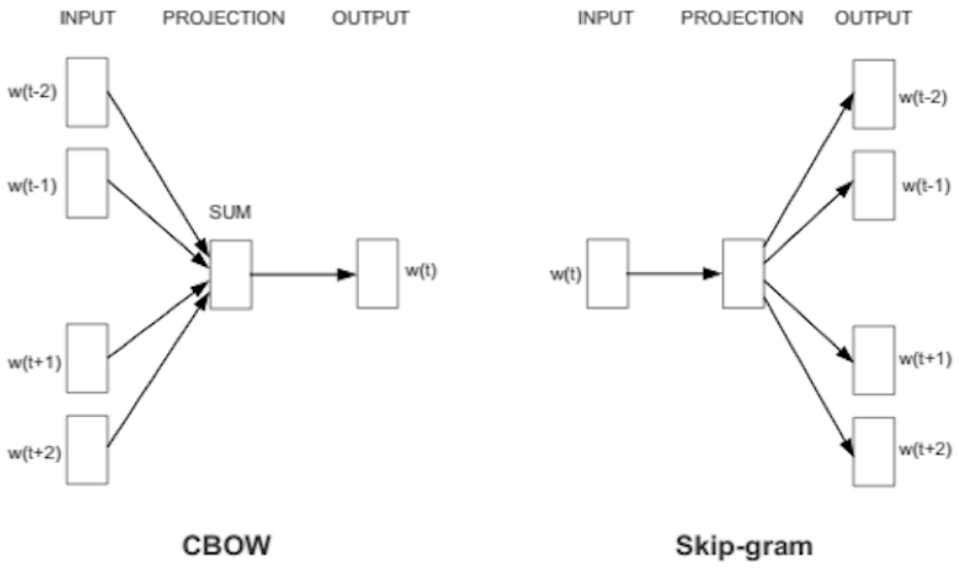


Figure 12: Architecture of word2vec models: CBOW and Skip-Gram (Mikolove et al., 2013a)

5.3 Experiment

The aim is to investigate the ability of word embedding for semantic representations of a small domain-specific corpus, the Qur'an, in classical Arabic. In this experiment, we map the Qur'an's words to numerical vectors using Word2vec. We then examine the model capabilities of capturing semantic and syntactic similarities in the Qur'an words, generating a document embedding

space that models and explains word distribution in the Holy Qur'an. The model is trained on the Arabic text of the Qur'an, imported from the Tanzil project. Different configurations of the hyper-parameters are used to produce the optimal model with high coherence and accuracy scores. Three main factors affect the quality of the neural word embedding. They are the model architecture, hyper-parameters, and corpus (Chiu et al., 2016).

5.3.1 Model Architecture

The two architectures of Word2vec, CBOW and Skip-gram, will be used to learn distributed representations of the Qur'an words. We compare the results of the two models to determine the better-performing model architecture with the optimal vales of hyperparameters. The process will be repeated multiple times for every parameter. Some parameters need tuning multiple times to achieve the best model performance such as vector size and number of iterations. They are explained in the following section.

5.3.2 Hyper-parameters

We trained the model with different settings of hyperparameters. Table 10 describes the values of each parameter explored in the experiment. We underline the values of the model's parameters that achieved the best representations of Qur'an words; identified qualitatively and quantitatively, see results and evaluation in subsequent sections. We used recommended default values, however, the experiment showed that using different values for some hyperparameters, namely minimum word count, context window size, vector dimension and number of iterations, leads to significantly better performances on semantic similarity task.

Hyperparameter	Value
Negative sample size	5
Minimum word count	2, 5
Learning rate	0.025
Context window size	5, 10, 15
Vector dimension	10, 15, 25, 50, 100
Number of iterations	10, 20, 25, 30, 40, 50
Architecture (CBOW / Skip gram)	0, 1

Table 10: Different settings of hyperparameters for training word2vec model

The model hyper-parameters are explained as the following:

1 Negative sample size (neg): The representation of a word is learned by maximizing its predicted probability to co-occur with its context words, while minimizing the probability for others. A good number of negative samples, according to the original paper (Mikolov et al., 2003), is 5-20. It also claims that if you have a large enough dataset, 2-5 seems to suffice. The default for Gensim is five negative samples.

2 Minimum-count (min-count): The minimum-count defines the minimum number of occurrences required for a word to be included in the word vectors. This parameter allows control over the size of the vocabulary and, consequently, the resulting word embedding matrix. The default value is 2.

3 Learning Rate: Neural networks are trained by gradually updating weight vectors along a gradient to minimize an objective function. The magnitude of these updates is controlled by the learning rate. The default value is 0.025.

4 Vector dimension (dim): The vector dimension is the size of the learned word vector. While a higher dimension tends to capture better word representations, their training is more computationally costly and produces a larger word embedding matrix.

5 Context window size (win): The size of the context window defines the range of words to be included as the context of a target word. For instance, a window size of 5 takes five words before and after a target word as its context for training. Larger window sizes (15-50) result in embeddings in which similarity is a better indicator of word relatedness. Smaller window sizes (2-15) result in embeddings with high similarity scores between them, indicating that the words are interchangeable.

6 Number of iterations: Learning depends on the number of iterations. It means that while a high number of iterations can cause to overfitting problem, having few repetitions will not lead to proper learning.

5.3.3 Corpus

We trained the model on the Qur'anic corpus that contains 6236 verses. We imported the Arabic text from the Tanzil project. We saved the verses in a CSV file, where each row represents a verse. We imported the data and pre-process them by tokenizing and removing stop-words. The preliminary results showed that training with stemming does not improve the quality of the final vectors; the stemming step generates poor vectors as some words were missing or reduced to their stems. In addition, because the holy Qur'an is God's word, stemming algorithms are not appropriate for it, and mistakes are not tolerated. Therefore, we excluded the stemming step. For a more discussion on this see section 5.6. Next, we used the Qur'an stop-words from the Qur'an analysis project by Karim Ouda (2015). The list comprises 809 words, verbs, and derivations. We then convert the input into a convenient format for training the model. Finally, we derived a vocabulary of 5447 tokens.

5.4 Results

After training the word2vec model, it generated an embedding space that contains all the Qur'an words as in Figure 13.

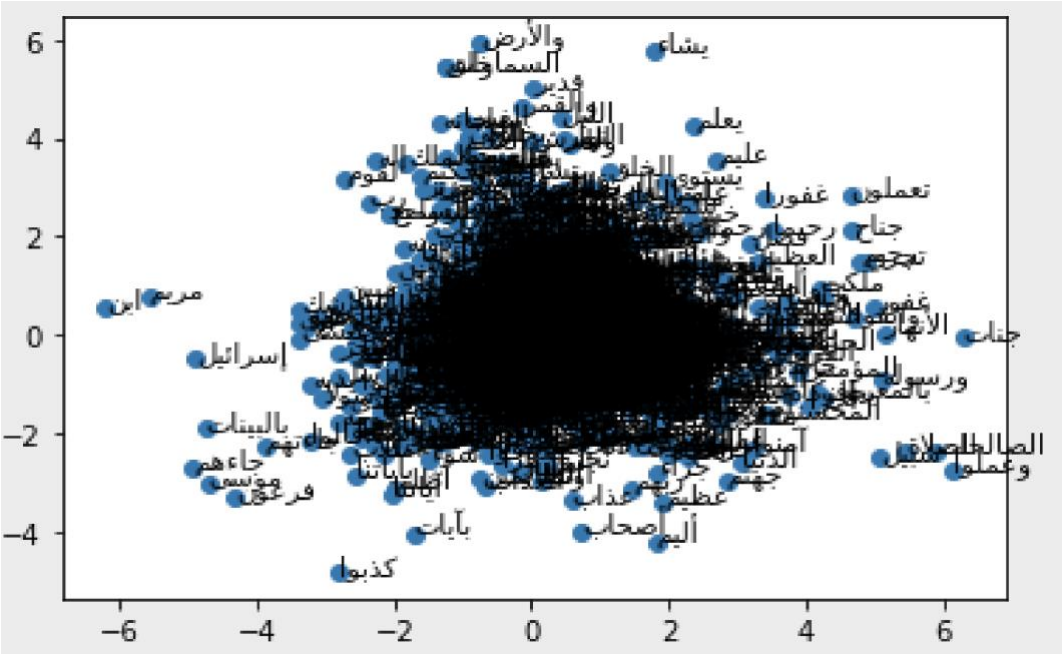


Figure 13: A visualization of a 2-Dimensional space containing the Qur'an words' vectors generated by the word2vec model[22]

For a better insight, we need to lower the dimensions to be able to visualise the vector representation of all the input words and see which word is closest in terms of co-occurrence and context to other words. Therefore, we used the two dimensionality reduction methods: PCA and t-SNE[23] for visualizing the semantic groups of words in the embedding space; examples are given in Figure 14.

[22] PCA converts the 100-dimensional vectors of words into a 2-dimensional vector that are represented by the X Y values.

[23] PCA and t-SNE are used to reduce dimensional space keeping relative pairwise distance between points.

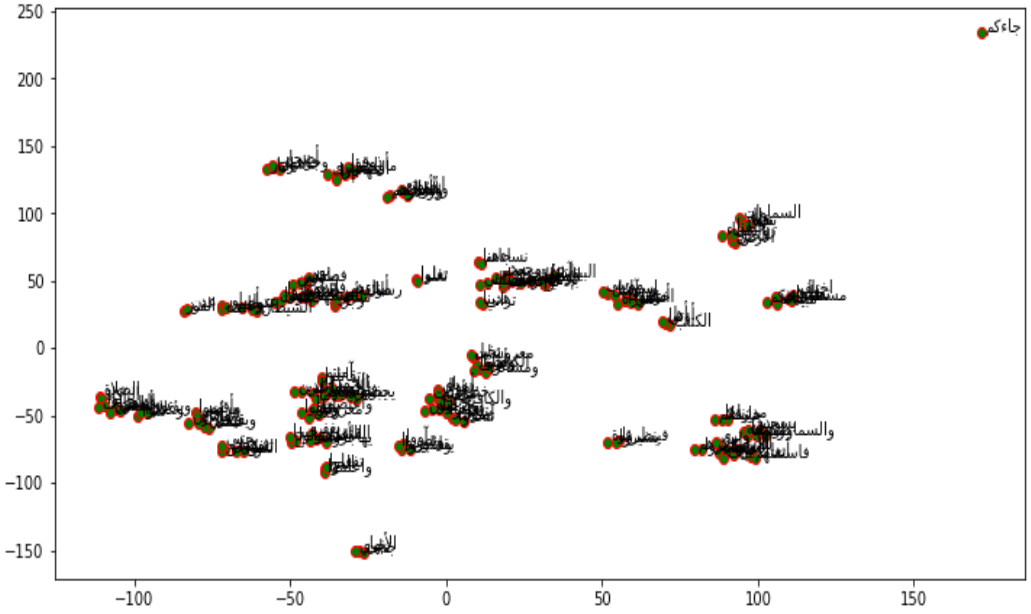


Figure 14: A visualization of semantic groups of words in the embedding space using t-SNE

To measure the quality of the words' vectors, we used groups of words and their most similar words, and were able to understand them intuitively. Table 11 shows groups of similar words derived by the trained model.

Word		Similar words			
الحياة	Worldly life	المقربين	Those brought near to Allah	الآخرة	Hereafter
آيات	Verses	بينات	Clear proofs	النور	The light
القرآن	The Qur'an	يسرنا	Made the Qur'an easy for remembrance	مدكر	Who will remember
الثواب	Reward	عيون	Springs	ظلال	In shade
جزاء	Reward	الإحسان	The good	سعيكم	Efforts

Table 11: Examples of words pairs with semantic relationship

5.4.1 Qualitative Results

To compare the quality of different versions of word vectors, we examined the semantic similarity between words. we computed the similarity between any two words in the vocabulary by measuring the cosine of the angle between their corresponding vectors projected in a multi-dimensional space. The cosine

similarity is used to measure the similarity between two vectors. It captures the angle of the word vectors and not the magnitude.

Moreover, to assess whether the model has clustered semantically similar words, we tried to get a set of the most similar words for a given word based on cosine similarity, by setting different numbers of nearest words. Table 12 and Table 13 show the results. The task here is to find related/similar words to the words, prayer and fire; respectively 'الصلاة' and 'النار' [24].

Word: الصلاة/ prayer		
Similar words		**Cosine similarity**
الزكاة	**Zakat**	**0.954**
وآتوا	**give**	**0.946**
وأقاموا	**Establish**	**0.904**
رزقناهم	**We have provided them**	**0.882**
سرا	**secretly**	**0.835**
وعلانية	**publicly**	**0.820**
وإقام	**performing**	**0.815**
المنكر	**wrongdoing**	**0.676**
عاهدوا	**they promised**	**0.662**
الفحشاء	**Immorality**	**0.659**

Table 12: The top 10 similar words to the term prayer/ 'الصلاة'

The derived words are considered similar/ related as they occur in similar context and share a common subject. For example, words similar to prayer are usually used in verses that discuss pillars of Islam such as prayer. Here are examples of verses that are related using words from Table 12 . They share a common theme discussing Prayer and Zakat[25], their rewards, and their effects on human life.

[24] The two terms represent key themes in the Qur'an.

[25] Relations between verses are defined based on Qurany corpus (Abbas, 2009).

إِنَّ الَّذِينَ يَتْلُونَ كِتَابَ اللَّهِ وَأَقَامُوا الصَّلَاةَ وَأَنفَقُوا مِمَّا رَزَقْنَاهُمْ سِرًّا وَعَلَانِيَةً يَرْجُونَ تِجَارَةً لَّن تَبُورَ **Surely those who read the Book of God, are firm in devotion, and spend of what We have given them in secret or openly, can hope for a commerce that will not decline. [35, 29]**
وَنُفَصِّلُ الْآيَاتِ لِقَوْمٍ يَعْلَمُونَ ۝ فَإِن تَابُوا وَأَقَامُوا الصَّلَاةَ وَآتَوُا الزَّكَاةَ فَإِخْوَانُكُمْ فِي الدِّينِ **But if they repent and are firm in devotion and pay the zakat, then they are your brothers in faith. We explain Our commands distinctly for those who understand. [9, 11]**
اتْلُ مَا أُوحِيَ إِلَيْكَ مِنَ الْكِتَابِ وَأَقِمِ الصَّلَاةَ ۖ إِنَّ الصَّلَاةَ تَنْهَىٰ عَنِ الْفَحْشَاءِ وَالْمُنكَرِ ۗ وَلَذِكْرُ اللَّهِ أَكْبَرُ ۗ وَاللَّهُ يَعْلَمُ مَا تَصْنَعُونَ **Recite what has been revealed to you of this Book, and be constant in devotion. Surely prayer keeps you away from the obscene and detestable, but the remembrance of God is greater far; and God knows what you do. [29, 45]**

Word: النار/ Hell		
Similar words		Cosine similarity
وجوههم	their faces	0.828
أولياؤهم	their allies	0.826
يعرض	does not respond	0.826
خالدون	eternally therein	0.822
الخلد	eternity	0.625
ذوقوا	Taste punishment	0.619
تمسنا	touch us	0.604
تجزون	awarded	0.585
الحريق	Burning fire	0.560
ذلة	humiliation	0.535

Table 13: The Top 10 similar words to the term Hell/ ‘النار’

Also, the t-SNE in Figure 15 plotted similar words in the embedding space representing the word ‘the Qur’an’/ ‘القرآن’. The space contains words that are close or similar (e.g., Rivers, Eden, Reward, Garden, Eternally therein).

Figure 15: A visualization of similar words to the word “The Qur’an”/ "القرآن " using t-SNE

Here are examples of verses that discuss the Qur’an using the same terms in Figure 15.

وَلَقَدْ صَرَّفْنَا فِي هَٰذَا الْقُرْآنِ لِيَذَّكَّرُوا وَمَا يَزِيدُهُمْ إِلَّا نُفُورًا **We have explained (the truth) in various ways in this Qur'an, that they may be warned; but it only increased their refractoriness. [17, 41]**
وَكَانَ الْإِنسَانُ أَكْثَرَ شَيْءٍ جَدَلًا ۚ وَلَقَدْ صَرَّفْنَا فِي هَٰذَا الْقُرْآنِ لِلنَّاسِ مِن كُلِّ مَثَلٍ **We have explained in various ways all things to men in this Qur'an; but of all things man is most contentious. [18, 54]**
وَلَقَدْ يَسَّرْنَا الْقُرْآنَ لِلذِّكْرِ فَهَلْ مِن مُّدَّكِرٍ **Easy have We made the Qur'an to understand: So is there any one who will be warned? [54, 40]**
أَوْ زِدْ عَلَيْهِ وَرَتِّلِ الْقُرْآنَ تَرْتِيلًا **Or a little more, and recite the Qur'an slowly and distinctly. [73, 4]**
وَإِذَا قُرِئَ الْقُرْآنُ فَاسْتَمِعُوا لَهُ وَأَنصِتُوا لَعَلَّكُمْ تُرْحَمُونَ **When the Qur'an is recited listen to it in silence. You may perhaps be blessed. [7, 204]**
أفلا يتدبرون القرآن ولو كان من عند غير الله لوجدوا فيه اختلافا كثير **Do they not ponder over the Qur'an? Had it been the word of any other but God they would surely have found a good deal of variation in it. [4, 82]**

Having a domain knowledge, we randomly picked some examples that demonstrate existing semantic relations between words in the Qur’an. Here, we

list some of these examples in Table 14 along with similarity score as detected by the trained model.

1st Word		2nd Word		Cosine Similarity
جنات	gardens	وعيون	springs	0.89783615
جنات	gardens	عدن	Eden	0.950336
السماوات	heavens	مددناها	split	0.79673016
الأرض	earth	السحاب	clouds	0.6336701
الشيطان	Satan	عدو	enemy	0.69264305
الصلاة	prayer	وأقاموا	performed	0.94589204
الكتاب	The book/ register	مسطورا	inscribed	0.8377156
النار	Hell /fire	خالدون	Abide therein eternally	0.8080162

Table 14: Pairs of related words that were detected by the model

To get better insights on the words' distribution and their importance, we visualized them using the word cloud technique[26]. For each input word, the trained model learns a group of 30 nearby words that were frequently mentioned in similar contexts. For example, when passing the word heavens /'جنات' to the model, the word cloud displays the top n words occur in the same context as in Figure 16. Examples are: 'عدن'/Eden, 'الأنهار'/rivers, 'وعيون'/springs, and 'الثواب'/reward.

[26] https://github.com/amueller/word_cloud

Also, the word cloud in Figure 17 and Figure 18 contain the Qur’anic words that are most likely to appear around the words 'الشيطان' / Satan and 'الكتاب'/ the book, respectively.

Figure 16: Similar words to word heavens 'جنات'

Figure 17: Similar words to the word 'الشيطان' / Satan

Figure 18: Similar words to the word 'الكتاب'/ the book

By refereeing to Qurany corpus, we identified a group of related verses that include words from Figure 18.

وَمِنْهُمْ أُمِّيُّونَ لَا يَعْلَمُونَ الْكِتَابَ إِلَّا أَمَانِيَّ وَإِنْ هُمْ إِلَّا يَظُنُّونَ **Among them are heathens who know nothing of the Book but only what they wish to believe, and are only lost in fantasies. [2, 78]**
وَمَن يَكْفُرْ بِهِ فَأُولَٰئِكَ هُمُ الْخَاسِرُونَ ۞ الَّذِينَ آتَيْنَاهُمُ الْكِتَابَ يَتْلُونَهُ حَقَّ تِلَاوَتِهِ أُولَٰئِكَ يُؤْمِنُونَ بِهِ **Those to whom We have sent down the Book, and who read it as it should be read, believe in it truly; but those who deny it will be losers. [2,121]**
وَآتَيْنَا مُوسَى الْكِتَابَ وَجَعَلْنَاهُ هُدًى لِّبَنِي إِسْرَائِيلَ أَلَّا تَتَّخِذُوا مِن دُونِي وَكِيلًا **We gave Moses the Book, and made it a guidance for the children of Israel that they should not take another protector apart from Me. [17, 2]**
وَلَقَدْ أَنزَلْنَا إِلَيْكُمْ آيَاتٍ مُّبَيِّنَاتٍ وَمَثَلًا مِّنَ الَّذِينَ خَلَوْا مِن قَبْلِكُمْ وَمَوْعِظَةً لِّلْمُتَّقِينَ **We have sent down clear instructions to you, and illustrations from (the accounts) of those who have gone before you, and a warning for those who take heed for themselves. [24, 34]**
وَيُعَلِّمُهُ الْكِتَابَ وَالْحِكْمَةَ وَالتَّوْرَاةَ وَالْإِنْجِيلَ **He will teach him the Law and the judgement, and the Torah and the Gospel, [3, 48]**
وما أنزلنا عليك الكتاب إلا لتبين لهم الذي اختلفوا فيه وهدى ورحمة لقوم يؤمنون **We have sent down this Book to you that you may explain to them what it is that they are differing about, and as guidance and a grace for those who believe. [16, 64]**

The verses are related and use terms from the ones generated by our model, which means that word2vec succeeded to capture the semantic similarity to some extent. However, polysemy is not handled properly, where words with multiple meanings are combined in a single representation (a single vector).

5.5 Evaluation

The semantic vectors should be tested across semantic tasks like semantic categorization and distance comparison (Bullinaria and Levy, 2012). To evaluate the derived model and examine its ability of semantic representation of Qur'an words, we perform a semantic categorization test and distance comparison. In order to do that, we created two datasets of Qur'anic words to act as an evaluation resource.

5.5.1 Distance comparison

The test is intended to evaluate the semantic space's large-scale structure using words that are widely scattered in the corpus (Bullinaria and Levy, 2007).

We create a dataset of 150[27] semantically similar pairs of words from the Qur'an. Examples are shown in Table 15 . The test includes 150 target words and the comparison is between one semantically related word and other ten words randomly chosen from the 150 pairs. For example, related words are Satan and deviators, we choose ten control words from other pairs such as: paradise, sky, truth, religion, forgiveness, the book, Gospel, prayer, believers, and reward. The performance is the percentage of control words that are further than the related word from the target word. By computing the cosine distance between each target word's semantic vector and that of its related word and each of ten others randomly chosen control words from the 150 pairings.

[27] We started by creating 50 pairs at an initial stage. Then we extended the dataset, aiming to cover more instances adding robustness to our evaluation. Potentially, we aim to enlarge the dataset. Codes and produced dataset will be accessible at Github repository: Mhalshammeri

Word		Related word		Word		Related word	
الشيطان	Satan	الغاوين	deviators	الشيطان	Satan	يوسوس	Whispers[evil]
الكتاب	The book	الإنجيل	Gospel	الدين	Religion	الحق	The Truth
يعلمون	They Know	يخشون	Fear their lord	الأرض	The Earth	السماء	The Sky
الصلاة	Prayer	المؤمنون	Believers	الرسل	Messengers	مبشرين	Bringers of good tidings
الصالحات	Righteous deeds	مغفرة	Forgiveness	الصدقات	Charities	الثواب	Reward

Table 15: Pairs of related words from the Qur'an

First, we examine the cosine distance between a target word and a related one, and other ten control words chosen randomly from other pairs in the dataset. We apply the cosine measure on the derived vectors from the trained w2v model. We then compute the percentage of control words that are further from the target than its related word. Table 16 shows an instance of the results, the target word is 'الشيطان/ Satan', related word is ' الغاوين/ deviators', with other ten control words. The distances show that the target word is closer to its related one that control words.

Target word		**Related word**		**Cosine distance**			
الشيطان/Satan		**الغاوين/ deviators**		**0.268**			
Control words/ cosine distance from target word							
المكذبين	**0.873**	**حافظون**	**0.769**	**الفقراء**	**0.895**	**الناصحين**	**0.706**
تنصرون	**0.785**	**البنون**	**0.972**	**الأرحام**	**0.951**	**يخشى**	**0.858**
الصادقون	**0.828**	**ترتيلا**	**0.933**	**Performance[28] = 100%**			

Table 16: The cosine distance between a target word, a related word, and other 10 control words from other pairs.

[28] The performance is the percentage of control words that are further than the related word from the target word. All the ten control words are further from the target word 'Satan', so performance is 100%.

It is observed that the proportion is always 100%, as the cosine distance between a target word and its related one is always smaller compared with any words from other pairs. The control words were chosen randomly from the dataset, however, when multiple pairs share the same topic, we excluded them to give more validity to the test. For example: the two pairs <Garden, Eden> and <Garden, Firdous>, both address the paradise theme, so only one was kept during the comparison. The aim here is to show how distinct are control words from related ones, in terms of cosine similarity, which provides an indication of the ability of the embedding model (word2vec) in capturing the semantic structure in the Qur'anic text. the results of the test are shown in Appendix C.

5.5.2 Semantic Categorization Test

We measure the capabilities of the embeddings model to represent the known semantic categories using semantic categorization test (Patel et al., 1997). This test is designed to explore the extent to which semantic categories are represented in the vector space. It measures how often individual word vectors are closer to their own semantic category centre rather than one of the other category centres (Patel et al., 1997). We use ten[29] words from each of 50 semantic categories (e.g., Paradise, Hell, Satan, Prayer, The Qur'an, the Book). The percentage of the 500 words that fall closer to their own category centre rather than another was computed.

To perform this test, we create a dataset of 50 semantic categories of Qur'an words[30], as an initial stage. Each category is composed of collection of words that represent that category; ten words that tend to occur in the same context. The new dataset is derived from available ontologies produced in previous scholarly work (Dukes, 2013; Alzain, 1995) to guarantee the quality of the data. Table 17 shows examples of the semantic categorization of the Qur'an words. For more details on the dataset, refer to section 5.6.

---

[29] We chose 10 as in the paper, ten words were taken from each of 53 semantic categories based on human category norms (Battig & Montague, 1969).

[30] Codes and produced dataset will be accessible at Github repository: Mhalshammeri

Word	Category/ tag	Word	Category/ tag	Word	Category/ tag	النار	Category/ tag	Word	Category/ tag	Word	Category/ tag
الكتاب	The Book	الصلاة	Prayer	جنات	Paradise	النار	Hell	القرآن	The Quran	الشيطان	Satan
العلم	1	الزكاة	2	الخلد	3	ذوقوا	4	وهدى	5	يعدهم	6
التوراة	1	وأقاموا	2	الأنهار	3	خالدون	4	يتدبرون	5	خطوات	6
البينة	1	وآتوا	2	أعناب	3	تجزون	4	صرفنا	5	ينزغنك	6
والإنجيل	1	المنكر	2	تجري	3	بسيماهم	4	مدكر	5	قرينا	6
والنبوة	1	وأقام	2	عدن	3	الحريق	4	قرئ	5	الغاوين	6
ونور	1	والصابرين	2	الثواب	3	عذاب	4	يسرنا	5	فوسوس	6
الأميين	1	ينفقون	2	رضي	3	أولياؤهم	4	تصديق	5	شيعته	6
ليحكم	1	وعلانية	2	الفوز	3	المشأمة	4	ترتيلا	5	الرجيم	6
أوتوا	1	فأقيموا	2	وعيون	3	تمسنا	4	مدكر	5	عدو	6

Table 17: Examples of 6 semantic categories of the Qur'an words

Table 18 and Table 19 shows the results of the test for two semantic categories, the Qur'an and paradise. We calculate the proportion of target word vectors with a smaller cosine distance to their own semantic category centre than one of the other category centres. The category centres are just the averages of the vectors that correspond to each category's words (excluding the target word). For each category, we use the ten words from that category and compute 10 different category's centres. We then calculate the cosine distance between each target word' vector and its category centre, and the other nine centres for the same category. Finally, we compute an aggregate percentage for each semantic category, that determines the number of the target word vectors that are closer to its own category centre than one of the other centres for the same category.

Thus, we compute ten category's centres; we compute ten different centres for the same category where the target word belongs; it is the Qur'an. Each centre is calculated as the mean value of nine vectors of the other nine words, excluding the target word's vector. In the first example category here, the category centres

for the two the target words ‘ترتيلا/ reciting’ and ‘قـرئ/read’ was computed as the following.

```
center4 = (vector1 + vector2 + vector3 + vector4 + vector5 + vector7  + vector8 + vector9 + vector10)/ 9
Dis = spatial.distance.cosine(center4, vector6)
print(Dis)
```

```
قـرئ
0.17824822664260864
```

```
center2 = (vector1 + vector2 + vector3 + vector4 + vector5 + vector6  + vector7 + vector8 + vector10)/ 9
Dis = spatial.distance.cosine(center2, vector9)
print(Dis)
```

```
ترتيلا
0.22428381443023682
```

Then for each word in the category, we calculate the cosine distance between each target word’ vector and its category centre, and the other nine centres for the same category. The ten category centres for the word ‘يسرنا/ we made it easy’ is computed as the following.

```
#Cosine distance between target word's 1 vector(7) and the other 9 category centres

print(list1[6])

Dis = spatial.distance.cosine(center2, vector7)
print(Dis)

Dis = spatial.distance.cosine(center3, vector7)
print(Dis)

Dis = spatial.distance.cosine(center4, vector7)
print(Dis)

Dis = spatial.distance.cosine(center5, vector7)
print(Dis)

Dis = spatial.distance.cosine(center6, vector7)
print(Dis)

Dis = spatial.distance.cosine(center7, vector7)
print(Dis)

Dis = spatial.distance.cosine(center8, vector7)
print(Dis)

Dis = spatial.distance.cosine(center9, vector7)
print(Dis)

Dis = spatial.distance.cosine(center10, vector7)
print(Dis)
```

The resulting vectors are here.

يـسرنـا

0.09840524196624756

0.10944139957427979

0.09910303354263306

0.10615068674087524

0.08438968658447266

0.10137766599655151

0.05448567867279053

0.08196276426315308

0.10318011045455933

There is only one instance where the first target word vector 'يسرنا' has a smaller cosine distance (0.108) to its centre (center1) than one of the other centres (centre 3 with distance 0.109), therefore aggregate percentage here is 10%. The percentage measures how much the sematic category is represented using the embeddings. The percentage is relatively small which means that the semantic category is not represented very well in the vector space generated by word2vec.

Semantic Category	The Qur'an/ القرآن						
Target word	Cosine distance from category centre	Cosine distance between target word vector and the other 9 category centres						
يسرنا	0.108	2	3	4	5	6	7	8
		0.098	0.109	0.099	0.106	0.084	0.101	0.054
		9	10					
		0.081	0.103					

ترتيلا	0.224	1	3	4	5	6	7	8
		0.215	0.218	0.211	0.213	0.192	0.211	0.164
		9	10					
		0.197	0.215					
للذكر	0.167	1	2	4	5	6	7	8
		0.153	0.145	0.146	0.151	0.131	0.145	0.111
		9	10					
		0.114	0.147					
قرئ	0.178	1	2	3	5	6	7	8
		0.173	0.168	0.175	0.166	0.151	0.173	0.111
		9	10					
		0.165	0.172					
مدكر	0.130	1	2	3	4	6	7	8
		0.122	0.114	0.122	0.112	0.103	0.115	0.082
		9	10					
		0.105	0.115					
تصديق	0.299	1	2	3	4	5	7	8
		0.252	0.253	0.250	0.254	0.254	0.257	0.324
		9	10					
		0.250	0.254					

يتدبرون	0.201	1	2	3	4	5	6	8
		0.197	0.188	0.194	0.194	0.192	0.177	0.142
		9	10					
		0.163	0.189					
وهدى	0.402	1	2	3	4	5	6	7
		0.221	0.227	0.222	0.225	0.222	0.247	0.226
		9	10					
		0.204	0.216					
القرآن	0.237	1	2	3	4	5	6	7
		0.122	0.125	0.118	0.127	0.124	0.124	0.123
		8	10					
		0.109	0.127					
صرفنا	0.230	1	2	3	4	5	6	7
		0.197	0.194	0.197	0.194	0.194	0.184	0.192
		8	10					
		0.143	0.193					

Table 18: Results of Semantic Categorization test on Semantic Category: The Qur'an

Moreover, in the second example category, paradise, there exist two instances where two target words vectors (الثواب/reward & عيون/springs) have a smaller cosine

distance to their centres (center7 and center10) than one of the other centres, see Table 19. Therefore, the aggregate percentage for the semantic category ‘Paradise’ is 20%. It is the proportion of word vectors with a smaller cosine distance to their own semantic category centre than one of the other category centres. The percentage is higher than the aggregate percentage for the semantic category ‘the Qur’an’. It could be attributed to the uniqueness of terms used to describe that category, ‘Paradise’. We can read that as the terms used in the Qur’an category are used to describe other topics or semantic category. Thus, word2vec is not the optimal choice for capturing the semantic representations of Qur’an words.

Semantic Category	Paradise / جنة						
Target word	Cosine distance from category centre	Cosine distance between target word vector and the other 9 category centres						
أعناب	0.224	2	3	4	5	6	7	8
		0.199	0.205	0.199	0.206	0.208	0.207	0.206
		9	10					
		0.198	0.209					
الخلد	0.729	1	3	4	5	6	7	8
		0.668	0.671	0.649	0.665	0.661	0.676	0.671
		9	10					
		0.668	0.670					
الأنهار	0.054	1	2	4	5	6	7	8
		0.033	0.034	0.036	0.052	0.024	0.031	0.031

		9	10					
		0.026	0.029					
جنات	0.034	1	2	3	5	6	7	8
		0.019	0.016	0.023	0.021	0.021	0.019	0.020
		9	10					
		0.020	0.020					
تجري	0.068	1	2	3	4	6	7	8
		0.042	0.042	0.061	0.042	0.032	0.040	0.039
		9	10					
		0.037	0.038					
عدن	0.193	1	2	3	4	5	7	8
		0.159	0.153	0.132	0.159	0.133	0.161	0.159
		9	10					
		0.168	0.167					
الثواب	0.139	1	2	3	4	5	6	8
		0.134	0.138	0.115	0.126	0.114	0.144	0.132
		9	10					
		0.133	0.141					
رضي	0.205	1	2	3	4	5	6	7
		0.182	0.181	0.173	0.181	0.168	0.185	0.182

		9	10					
		0.185	0.185					
الفوز	0.253	1	2	3	4	5	6	7
		0.208	0.210	0.190	0.209	0.195	0.223	0.212
		8	10					
		0.213	0.210					
عيون	0.173	1	2	3	4	5	6	7
		0.160	0.154	0.136	0.156	0.138	0.177	0.162
		8	10					
		0.161	0.155					

Table 19: Results of Semantic Categorization test on Semantic Category: Paradise

5.6 Datasets

We created a semantic classification of the Qur'an words to evaluate my model based on existing knowledge resources. Hence, we built a dataset of the Qur'an words linking them to semantic categories; represented by topics/ meanings in the Qur'an. To identify the semantic group of each word in the vocabulary, we used Qur'anic Arabic Corpus QAS (Dukes, 2013) and "المعجم المفهرس لمعاني القرآن العظيم" the indexed dictionary of the Qur'an meanings (Alzain, 1995); a copy is provided in Appendix A. The two resources are essential references for building a semantic categorization of Qur'an words. Both resources identify key concepts and meanings in the Qur'an and show the relationships between them through listing related verses. In both resources, we first locate multiple keywords representing a topic and the associated verses. we then create semantic groups, each containing associated keywords.

Eventually, we compiled 50 semantic groups, with ten words in each group, as an initial step towards compiling a resource for evaluation of semantic categorization in the Qur'an. For example ('القرآن'/the Qur'an, 'الذكر'/reminder, 'الكتاب'/the book, 'أنزل'/revealed, 'الإنجيل'/Gospel, 'التوراة'/Torah, 'مصدقا'/in truth and confirmation, 'تنزيل'/revelation, 'استمعوا'/you listen, 'الفرقان'/discernment of falsehood and truth) based on both resources, can be in the same semantic category representing the topic 'القرآن'/ the Qur'an. Table 20 shows an example from the dataset[31], a semantic category (Day of Resurrection), with ten keywords and verses addressing the topic using each keyword.

Category	**Day of Resurrection (يوم القيامة)**		
Key word	**Chapter**	**Verse**	**Verse Text**
الآخرة	40	39	يَا قَوْمِ إِنَّمَا هَٰذِهِ الْحَيَاةُ الدُّنْيَا مَتَاعٌ وَإِنَّ الْآخِرَةَ هِيَ دَارُ الْقَرَارِ O people, the life of this world is ephemeral; but enduring is the abode of the Hereafter.
	23	74	وَإِنَّ الَّذِينَ لَا يُؤْمِنُونَ بِالْآخِرَةِ عَنِ الصِّرَاطِ لَنَاكِبُونَ But those who believe not in the Hereafter turn away from the straight path.
	31	4	الَّذِينَ يُقِيمُونَ الصَّلَاةَ وَيُؤْتُونَ الزَّكَاةَ وَهُم بِالْآخِرَةِ هُمْ يُوقِنُونَ Who are constant in devotion, pay the zakat, and are certain of the Hereafter.
الساعة	40	59	إِنَّ السَّاعَةَ لَآتِيَةٌ لَّا رَيْبَ فِيهَا وَلَٰكِنَّ أَكْثَرَ النَّاسِ لَا يُؤْمِنُونَ The Hour will certainly come; there is no mystery about it; but most men do not believe.
	15	85	وَمَا خَلَقْنَا السَّمَاوَاتِ وَالْأَرْضَ وَمَا بَيْنَهُمَا إِلَّا بِالْحَقِّ ۗ وَإِنَّ السَّاعَةَ لَآتِيَةٌ ۖ فَاصْفَحِ الصَّفْحَ الْجَمِيلَ We have not created but with reason the heavens and the earth and all that lies within them. The Hour (of the great change) is certain to come. So turn away (from them) with a grace.

[31] All produced dataset will be accessible at Github repository: Mhalshammeri

وَلِلَّهِ غَيْبُ السَّمَاوَاتِ وَالْأَرْضِ وَمَا أَمْرُ السَّاعَةِ إِلَّا كَلَمْحِ الْبَصَرِ أَوْ هُوَ أَقْرَبُ ۚ إِنَّ اللَّهَ عَلَىٰ كُلِّ شَيْءٍ قَدِيرٌ To God belong the secrets of the heavens and the earth, and the Hour of Doom is a matter of the winking of an eye, even less, for God has certainly power over all things.	77	16	
قُلِ اللَّهُ يُحْيِيكُمْ ثُمَّ يُمِيتُكُمْ ثُمَّ يَجْمَعُكُمْ إِلَىٰ يَوْمِ الْقِيَامَةِ لَا رَيْبَ فِيهِ وَلَٰكِنَّ أَكْثَرَ النَّاسِ لَا يَعْلَمُونَ Say: "God, who gives you life and makes you die, will (raise the dead) then gather you (and your ancestors) together on the Day of Resurrection of which there is no doubt." And yet most men do not understand.	26	45	**القيامة**
ثُمَّ إِنَّكُمْ يَوْمَ الْقِيَامَةِ تُبْعَثُونَ Then will be raised up on the Day of Resurrection.	16	23	
وَلَيَحْمِلُنَّ أَثْقَالَهُمْ وَأَثْقَالًا مَّعَ أَثْقَالِهِمْ ۖ وَلَيُسْأَلُنَّ يَوْمَ الْقِيَامَةِ عَمَّا كَانُوا يَفْتَرُونَ They will carry their own loads and other loads besides their own; and will surely be questioned on the Day of Resurrection about what they contrived.	13	29	
يَوْمَ يَبْعَثُهُمُ اللَّهُ جَمِيعًا فَيُنَبِّئُهُم بِمَا عَمِلُوا ۤ أَحْصَاهُ اللَّهُ وَنَسُوهُ ۚ وَاللَّهُ عَلَىٰ كُلِّ شَيْءٍ شَهِيدٌ On the day when God will raise them up together, He will tell them what they did. God takes account of it although they forget, for all things are evident to God.	6	58	**البعث**
أَلَا يَظُنُّ أُولَٰئِكَ أَنَّهُم مَّبْعُوثُونَ Do they not think they will be raised (to life) again.	4	83	
وَقَالَ الَّذِينَ أُوتُوا الْعِلْمَ وَالْإِيمَانَ لَقَدْ لَبِثْتُمْ فِي كِتَابِ اللَّهِ إِلَىٰ يَوْمِ الْبَعْثِ ۖ فَهَٰذَا يَوْمُ الْبَعْثِ وَلَٰكِنَّكُمْ كُنتُمْ لَا تَعْلَمُونَ But those who were given the knowledge and belief will say: "You have tarried, according to	56	30	

			the Book of God, as long as the Day of Resurrection, and this is the Day of Resurrection, but 'you do not know."
الحشر	10	45	وَيَوْمَ يَحْشُرُهُمْ كَأَن لَّمْ يَلْبَثُوا إِلَّا سَاعَةً مِّنَ النَّهَارِ يَتَعَارَفُونَ بَيْنَهُمْ قَدْ خَسِرَ الَّذِينَ كَذَّبُوا بِلِقَاءِ اللَّهِ وَمَا كَانُوا مُهْتَدِينَ The day He will gather them together it will appear to them that they had lived (in the world) but an hour of a day to make each other's acquaintance. Verily those who deny the meeting with God will be lost, and not find the way.
	6	22	وَيَوْمَ نَحْشُرُهُمْ جَمِيعًا ثُمَّ نَقُولُ لِلَّذِينَ أَشْرَكُوا أَيْنَ شُرَكَاؤُكُمُ الَّذِينَ كُنتُمْ تَزْعُمُونَ The day We shall gather all of them together and say to those who ascribe (partners to God): "Where are the compeers who you claimed (were equal to God)?"
	3	158	وَلَئِن مُّتُّمْ أَوْ قُتِلْتُمْ لَإِلَى اللَّهِ تُحْشَرُونَ And if you die or are killed, even so it is to God that you will return.

Table 20: Example of a semantic category for ‘Day of Resurrection’, as one topic discussed in the Qur’an.

Table 21 shows examples of semantic categories derived from the Qur’an; each category represents a concept that has been addressed over many verses. Also, we underline the keywords describing each category as they appear in each verse. Indeed, the dataset would contain all words in the vocabulary, where each word is annotated with a label representing its semantic category. Eventually, we would compile a corpus for evaluation of semantic categorization in the Qur’an, and the corpus can be used when processing the text at both words and sentences level. The compilation is done manually using existing knowledge resources to develop semantic categories in the Qur’an and list the relevant keywords and verses under each category. In the future, the dataset will be

extended to include more topics and improved to address the embedded limitations. For more details, see 10.6.2 that discusses future work.

Category	Chapter	Verse	Verse Text
الأصنام **Idol**	6	74	وَإِذْ قَالَ إِبْرَاهِيمُ لِأَبِيهِ آزَرَ أَتَتَّخِذُ أَصْنَامًا آلِهَةً ۖ إِنِّي أَرَاكَ وَقَوْمَكَ فِي ضَلَالٍ مُّبِينٍ And [mention, O Muhammad], when Abraham said to his father Azar, "Do you take idols as deities? Indeed, I see you and your people to be in manifest error."
	7	138	وَجَاوَزْنَا بِبَنِي إِسْرَائِيلَ الْبَحْرَ فَأَتَوْا عَلَىٰ قَوْمٍ يَعْكُفُونَ عَلَىٰ أَصْنَامٍ لَّهُمْ ۚ قَالُوا يَا مُوسَى اجْعَل لَّنَا إِلَٰهًا كَمَا لَهُمْ آلِهَةٌ ۚ قَالَ إِنَّكُمْ قَوْمٌ تَجْهَلُونَ And We took the Children of Israel across the sea; then they came upon a people intent in devotion to [some] idols of theirs. They said, "O Moses, make for us a god just as they have gods." He said, "Indeed, you are a people behaving ignorantly.
	21	57	وَتَاللَّهِ لَأَكِيدَنَّ أَصْنَامَكُم بَعْدَ أَن تُوَلُّوا مُدْبِرِينَ And [I swear] by Allah, I will surely plan against your idols after you have turned and gone away."
القيامة **Day of resurrection**	45	26	قُلِ اللَّهُ يُحْيِيكُمْ ثُمَّ يُمِيتُكُمْ ثُمَّ يَجْمَعُكُمْ إِلَىٰ يَوْمِ الْقِيَامَةِ لَا رَيْبَ فِيهِ وَلَٰكِنَّ أَكْثَرَ النَّاسِ لَا يَعْلَمُونَ Say, "Allah causes you to live, then causes you to die; then He will assemble you for the Day of Resurrection, about which there is no doubt, but most of the people do not know."
	23	16	ثُمَّ إِنَّكُمْ يَوْمَ الْقِيَامَةِ تُبْعَثُونَ Then indeed you, on the Day of Resurrection, will be resurrected.
	3	185	كُلُّ نَفْسٍ ذَائِقَةُ الْمَوْتِ ۗ وَإِنَّمَا تُوَفَّوْنَ أُجُورَكُمْ يَوْمَ الْقِيَامَةِ ۖ فَمَن زُحْزِحَ عَنِ النَّارِ وَأُدْخِلَ الْجَنَّةَ فَقَدْ فَازَ ۗ وَمَا الْحَيَاةُ الدُّنْيَا إِلَّا مَتَاعُ الْغُرُورِ

			Every soul will taste death, and you will only be given your [full] compensation on the Day of Resurrection. So he who is drawn away from the Fire and admitted to Paradise has attained [his desire]. And what is the life of this world except the enjoyment of delusion.

Table 21: Examples of semantic categories from the Qur'an; each category is addressed over a group of verses of relevant words.

5.7 Silhouette Coefficients and Silhouette Score

Using the trained word embedding model (word2vec), we used the unsupervised learning to cluster the documents by applying the K-Means algorithm (Xu and Tian, 2015). We then calculated the Silhouette Coefficients[32] for the words' clusters[33], and the Silhouette score for the model. We compute Silhouette score[34] for a different number of clusters and we got the optimal number of clusters is fifteen. Silhouette score takes into consideration the intra-cluster distance between the sample and other data points within the same cluster, and inter-cluster distance between the sample and the next nearest cluster (Rousseeuw, 1987). Silhouette score for the clustering algorithms is calculated using the following equation:

Silhouette Score = (b-a)/max (a, b), where

a= average intra-cluster distance; the average distance between each point within a cluster.

b= average inter-cluster distance; the average distance between all clusters.

The results show positive values of Silhouette coefficient, indicating well defined clusters, however, the value is relatively small which means the clusters are

[32] Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. the clusters are considered well apart from each other as the silhouette score is closer to 1.

[33] Number of clusters were decided based on a number of trials (2 -20) to get the best Silhouette Score.

[34] codes will be published at Github repository: Mhalshammeri

overlapping. Incorrect clusters, on the other hand, will result in negative numbers. Next, we used the new annotated dataset as the target dataset to identify the semantic category of each word. We developed a search algorithm that scan the dataset using keywords from each cluster. Each category in the dataset contains a set of keywords representing that category. We kept track of matches along with the category. Table 22 shows the average silhouette coefficients for each cluster and the Silhouette score for the clustering algorithm when number of clusters is fifteen. It is considered to be the optimal number of clusters as its score is greater than that of other number of clusters (2-20).

Cluster#	**Size**[35]	**Average Silhouette coefficient**	**Most representative terms per cluster**[36]
Cluster1	**442**	**-0.04**	اختلفوا فهدى يعرفونه القدس ليحكم ملتهم فبعث مصدقا الكتاب والحكم الباطل بوكيل أنزلنا ونور ترضى يتولى اتبعت والنبوة جاءك وأيدناه
Cluster2	**182**	**0.21**	الرحيم الرحمن الواقعة وقعت ينصركم والنجم عبده كاشف الحكيم يمسسك عبس للمكذبين القدوس العزيز يشركون لقادرون سبح المجيد يجادلون شيطانا
Cluster3	**222**	**0.03**	قالوا أجئتنا عاكفين نعبد واشهد أحلام أضغاث لضالون المسحرين خرجوا بمؤمنين الكبرياء بالكفر أرجه تدعونا المصلين وجدنا كتابيه فأتنا موتوا
Cluster4	**213**	**0.06**	خلق السماوات يسجد والأرض تروا ويأت ستة وسخر مددناها عمد بخلق وربي خلقوا يخلق ترونها سألتهم لتعلموا استوى فأحسن صوركم
Cluster5	**73**	**0.22**	ربنا نخاف ذنوبنا قرينه أطعنا آتنا أعطى توكلنا أقدامنا ونكون وتوفنا آياتك واجعلنا قالا القواعد فنتبع وثبت خطايانا منقلبون وانصرنا
Cluster6	**2032**	**0.26**	وجمع مشركون فسنيسره لليسرى الوقت المعلوم مشرقين حافظون لفروجهم قولي ضحاها سندخلهم خبرا يؤفك فاتبعوا الأشقى يريدوا ممنون كأسا تنطقون
Cluster7	**379**	**-0.04**	الميمنة المشأمة خالدون أصحاب بخارجين لأصحاب وجوههم أعيدوا تعرف فشاربون الضالون يئسوا يسحبون ترهقهم الفردوس ليكونوا أسوأ نضرة لعنة يضحكون

[35] Size represents the number of documents/ verses in each cluster

[36] Most frequent words in each cluster

Cluster8	**133**	**0.15**	الدنيا الحياة الآخرة حياتنا ونحيا نموت والأولى متاع المقربين بمبعوثين والآخرة يعذبهم والفضة الذهب وأبقى يؤخذ وتزهق مصفرا تردن الأخسرون
Cluster9	**874**	**0.00**	الفقراء الكفر يرضى ويكفر يضلوك ماتوا قلوبكم ويولج تقدروا يؤتكم وليعلم يعلموا أصابكم وتتقوا معروف يعفو رضوا الكاذبون عليم لقيتم
Cluster10	**137**	**0.03**	خالدين جنات وأعد الأنهار الثواب ومقاما تجري وعيون المهاجرين الفوز رضي غرف حزب عدن والسابقون مرفوعة خالدا مكرمون يدخلونها أوفى
Cluster11	**423**	**0.01**	يؤتكم والصالحين فوزا أموالكم وآمنوا ويكفر عظيما يطع وجاهدوا أجورهم وأطيعوا أعد المجاهدين الفائزون مثقلون بالبخل الصادقون القاعدين والصادقين أصابكم
Cluster12	**139**	**0.07**	رب العالمين أتأتون فضلتكم كذبون لغني برب انصرني لرب فأنظرني غلام أجري عمران سوءة فتنا شريك سبقكم أكون أنعمت وأكن
Cluster13	**334**	**-0.04**	فتولى الناصحين ونصحت طاغون أبلغتكم شعيبا سيهدين آلآن مقامي وآتاني كذبتم سفاهة مسرفون المكيال تعبد سقيم المهتدين عصيت ينصرني فأغرقناهم
Cluster14	**428**	**-0.03**	فأنزلنا بلد تميد تشقق سيروا مهدا بقدر اللؤلؤ يريكم والمرجان سحابا وآثارا باليمين تخرجون يريدوا اعلموا بيض رواسي وأنبتنا الخالقون
Cluster15	**225**	**-0.06**	مبصرا لتسكنوا يوقنون عابدين قم والأعناب ألسنتكم لآيات ساهون طويلا شجر سبحا الأليم سباتا النهى لقوم كذابا بنيانهم يفلحون فسبحه
Silhouette score of the model			**0.10**

Table 22: Silhouette coefficients for the Qur'an words and the Silhouette score Results

The results show that the model succeeded, sometimes, to locate the word in its semantic category (e.g., cases 1, 2, 4, 7). However, it is challenging in some cases like 3, 5, and 6. The reason could be the uniqueness and complexity that features the sacred text, and the complicated and ambiguous classical Arabic language. Hence, one recommendation is to develop specialized tools and algorithms to handle the pre-processing stage of the Qur'anic classical Arabic. Such tools have to give significance to words that are representative of the semantic category; not necessarily nouns, and also remove irrelevant words without affecting the meaning. They must consider that words occur in different

contexts have different meanings. Another suggestion is to extend the model to go beyond word level to achieve sentence-level representations, what will be studies in next chapter.

5.8 Discussion

One significant issue to raise here is the lack of stemmer algorithms for the classical Arabic of the Qur'an, or the deficiency of existing ones. It is a grand challenge to develop a stemmer considering the unique linguistics features and complexity of the Qur'anic text. A stemming algorithm usually normalizes the words by replacing different shapes of the word with its normal form. For example, the variations of the word 'رب' or God, usually refers to the same entity; 'رب' means the God, such in the two examples here.

{اتَّبِعُوا مَا أُنزِلَ إِلَيْكُم مِّن رَّبِّكُمْ وَلَا تَتَّبِعُوا مِن دُونِهِ أَوْلِيَاءَ ۗ قَلِيلًا مَّا تَذَكَّرُونَ}
Follow, [O mankind], what has been revealed to you from your Lord and do not follow other than Him any allies. Little do you remember. [7: 3]
{وَإِنَّ رَبَّكَ هُوَ يَحْشُرُهُمْ ۚ إِنَّهُ حَكِيمٌ عَلِيمٌ}
And indeed, your Lord will gather them; indeed, He is Wise and Knowing. [15: 25]

However, the same word refers to a different entity in verse 42 of chapter 12 (Yusuf), where 'رب' means the king or master:

{وَقَالَ لِلَّذِي ظَنَّ أَنَّهُ نَاجٍ مِّنْهُمَا اذْكُرْنِي عِندَ رَبِّكَ فَأَنسَاهُ الشَّيْطَانُ ذِكْرَ رَبِّهِ فَلَبِثَ فِي السِّجْنِ بِضْعَ سِنِينَ}
And he said to the one whom he knew would go free, "Mention me before your master." But Satan made him forget the mention [to] his master, and Joseph remained in prison several years. [12: 42]

The different variations of the same word, despite the same stem, have different meanings. Applying the existing Arabic stemmers would produce the same roots or stem to the different variation of the word, ignoring the different contexts within which the word was used. As a result, applying such algorithms to the holy Qur'an is inappropriate and faults are not permitted. Hence, although Arabic language stemming algorithms exist, their accuracy still needs to be improved. The

algorithms should consider contexts of the words to allow for learning meaningful representations of the words and sentences in the corpus.

5.9 Summary and Conclusion

The chapter studied the quality of vector representations of words derived by word2vec trained on the Qur'an. It described the experiment conducted using the two architectures of word2vec using different configurations to train the model. The quality of derived representations is measured in a word similarity task and evaluated using the distance comparison and semantic categorization tests. In the end, the chapter explained the dataset developed for the evaluation of semantic categorization in the Qur'an. The corpus benefits from existing knowledge resources; they are QAC, Qurany, and Indexed dictionary of the meanings in the Qur'an.

While LDA is based on word-strings in the texts, distributional models like word2vec generate embedding representations of word-meanings and subsequently similarity and classifiers work on embeddings rather than character-strings to represent words. Both approaches were used to cluster the text into groups of topics. In LDA, each topic is represented by several words with a different distribution. However, some topics have no clear interpretation. On the other hand, word2vec generates an embedding space capturing semantic similarity between words. Both approaches learns gradually and the topics become more specific with more training. Other text representations will be investigated for word sequences such as doc2vec and Transformers to exploit the semantic relations between the verses of Qur'an.

Chapter6
Learning Distributed Representations for the Qur'an Verses using Doc2vec

6.1 Introduction

Numerical representation of text documents is a challenging task in machine learning, although it can be powerful in text understanding and solving many NLP tasks. Word embedding models allow for learning vector representations of words using neural networks (Bengio et al., 2003; Collobert et al., 2008; Mikolov et al., 2013a). They have achieved remarkable performance and become pervasive in many NLP tasks. Furthermore, the impressive impact of these models has motivated researchers to consider vector representation to larger pieces of texts. Accordingly, Mikolov & Le have released a sentence embedding model (Le and Mikolov, 2014). It is another breakthrough in embeddings that maps sentences/documents to informative vector representations that preserve more semantic and syntactic information. They called it Paragraph Vectors, and by paragraph, they mean any variable-length text ranging from a sentence to documents.

This chapter extends the work in the previous chapter and go beyond word level to achieve sentence level. The research question that will be answered is to determine whether or not deep learning models that are based on distributional semantics is a viable approach for modelling the semantics of the Qur'an and learning subtle semantic relations between its verses.

To answer this question, this chapter exploits a recent trend in machine intelligence, which is the distributed representation of text, to learn an informative representation of the passages of the Qur'an, aiming to generate representations that are useful for the semantic-based similarity task and eventually capture meanings and concepts in the Qur'an. We use Paragraph Vectors to present a new vector representation of the Qur'anic verses at the paragraph level. These vectors can be used as features and leveraged for clustering and topic analysis.

6.2 Research Survey on Sentence Embedding Techniques

Following the outstanding success of words embeddings, researchers were motivated to extend these models to achieve phrase-level or sentence-level representations (Mitchell and Lapata, 2010; Zanzotto et al., 2010; Grefenstette et al., 2013; Mikolov et al., 2013).

6.2.1 Compositional Models for Sentence Representations

Mitchell and Lapata (2010) developed a framework for representing the meaning of word combinations in vector space. It is based on vector composition, which performs additive and multiplicative functions. A wide range of composition models were introduced and have been evaluated via correlation analysis. Pagliardini et al. (2017) proposed Sent2Vec1, a simple unsupervised model allowing the composition of sentence embeddings using word vectors along with n-gram embeddings, training the composition and the embedding vectors themselves simultaneously.

Neural networks offer a powerful learning resource compelling in natural language problems. In the neural network community, Collobert and Weston (2008) proposed learning word embeddings using a feed-forward neural network, by predicting a word based on the two words on the left and two words on the right. More recently, Le and Mikolov (2014) proposed simple log-bilinear models to learn continuous representations of words on very large corpora efficiently. In Neural networks frameworks, every word is mapped to a unique vector, represented by a column in a matrix. The column is indexed by position of the word in the vocabulary. The concatenation or sum of the vectors is then used as features (Le and Mikolov, 2014) for prediction of the next word in a sentence.

In recent years, various models have been proposed for modelling sentences; mainly those that are based on neural networks (Huang, 2013). These range from basic neural bag-of words or bag-of-n-grams models (Mitchell and Lapata, 2008;

Mitchell, 2010; Yu and Dredze, 2015) to the more structured recursive neural networks (Socher et al., 2012; Socher et al., 2013), and to time-delay neural networks based on convolutional operations (Collobert et al., 2008; Kalchbrenner et al., 2014; Kim, 2014; Hu et al., 2014) and recurrent neural networks (Tai et al., 2015).

The paragraph vector model (Le and Mikolov, 2014) incorporates a global context vector into the log-linear neural language model (Mikolov et al., 2013a) to learn the sentence representation; however, at prediction time, one needs to perform gradient descent to compute a new vector. The sequence autoencoder (Dai and Le, 2015) describes an encoder-decoder model to reconstruct the input sentence, while the skip-thought model (Kiros et al., 2015) extends the encoder-decoder model to reconstruct the surrounding sentences of an input sentence. Both the encoder and decoder of the methods above are modelled as RNNs. On the other side, CNNs have achieved excellent results in various natural language applications as the sentence encoder (Kim, 2014; Kalchbrenner et al., 2014; Hu et al., 2014). Gan et al., (2016) proposed a new class of CNN-LSTM encoder-decoder models that is able to leverage the vast quantity of unlabelled text for learning generic sentence representations.

6.2.2 Task-specific and General-purpose Sentence Embeddings

According to their purposes, sentence embeddings generally fall into two categories: task-specific sentence embeddings and general-purpose sentence embeddings. The first consists of sentence embeddings trained specifically for a certain task. They are usually combined with downstream applications and trained by supervised learning. Researchers have proposed many models, and they typically use recursive neural networks (Socher et al., 2012; Socher et al., 2013), convolutional neural networks (Kalchbrenner et al., 2014; Kim, 2014), or recurrent neural networks with long short-term memory (LSTM) (Tan et al., 2016; Lin et al., 2017).

The other category consists of universal sentence embeddings, which are usually trained by unsupervised or semi-supervised learning and can be used across domains, and can serve as features for many other NLP tasks such as text classification and semantic textual similarity. This includes recursive auto-encoders (Socher et al., 2011), Paragraph-Vectors (Le and Mikolov, 2014), Skip-Thought vectors (Kiros et al., 2015). Hill et al. (2016) also proposed a sentence-level log-linear bag-of-words (BOW) model, where a BOW representation of an input sentence is used to predict adjacent sentences that are also represented as BOW. Most recently, Wieting et al. (2015); Sent2Vec (Pagliardini, 2017); GRAN (Wieting and Gimpel, 2017) developed methods in which sentences are represented as a weighted average of fixed (pre-trained) word vectors.

6.2.3 Deep Neural Models for Sentence Representations

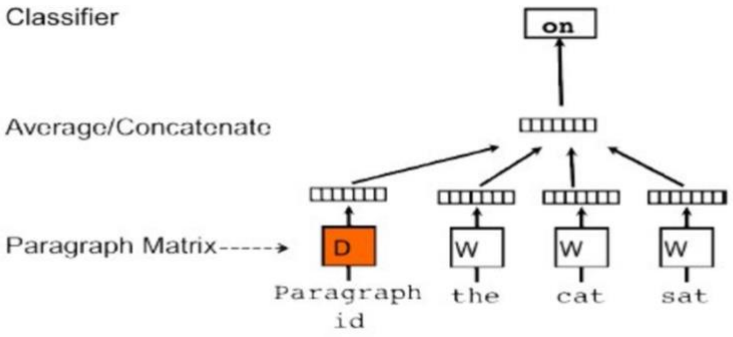
Several approaches have been proposed for learning sentence embeddings. The sequence autoencoder of Dai and Le (2015) describes an encoder-decoder model to reconstruct the input sentence, while the skip-thought model of Kiros et al. (2015) extends the encoder-decoder model to reconstruct the surrounding sentences of an input sentence. Both models use recurrent neural networks (RNNs). However, Convolutional Neural Networks (CNNs) have recently achieved tremendous results in various task dependent natural language applications as the sentence encoder (Kalchbrenner et al., 2014; Kim, 2014; Hu et al., 2014). Moreover, Gan et al. (2016) proposed a new encoder-decoder approach to learn distributed sentence representations. The model is learned by using a convolutional neural network as an encoder to map an input sentence into a continuous vector. A very recent work by Jiao et al. (2018) proposed a simple CNN model for creating general-purpose sentence embeddings that can transfer easily across domains. The model Utilizes both features of words and n-grams to encode sentences.

6.2.4 Paragraph Vector Model (Doc2vec)

In general, bag-of-words models have two main shortcomings. They lose the ordering of the words and they also ignore semantics of the words. To overcome the weaknesses of bag-of-words models, Le and Mikolov (2014) proposed Paragraph Vector, an unsupervised algorithm that learns fixed-length feature

representations from variable-length texts, such as sentences, paragraphs, and documents. Paragraph Vector is capable of constructing representations of input sequences of variable length. Unlike some of the previous approaches, it is general and applicable to texts of any length. In addition, Paragraph Vector does not require task-specific tuning of the word weighting function nor does it rely on the parse trees. The paragraph vector model incorporates a global context vector into the log-linear neural language model (Mikolov et al., 2013a) to learn the sentence representation; however, at prediction time, one needs to perform gradient descent to compute a new vector. Paragraph Vector (also known as Doc2Vec) is supposed to be an extension to Word2vec such that Word2vec learns to project words into a latent n-dimensional space whereas Doc2vec aims at learning how to project a document into a latent n-dimensional space. Paragraph Vector can detect the relationships among words and understands the semantics of the text.

There are two approaches within doc2vec: a distributed bag of words model and a distributed memory model of Paragraph Vector. The distributed bag of words model is a simpler model and ignores word order, while the distributed memory model is a more complex model with more parameters. The two techniques are illustrated in Figure 19 and Figure 20. The idea behind the distributed memory model is that word vectors contribute to a prediction task about the next word in the sentence. The model inserts a memory vector to the standard language model, which aims at capturing topics of the document. The Paragraph Vector is concatenated or averaged with local context word vectors to predict the next



word.

Figure 19: Paragraph Vector: A distributed memory model (PV-D) (Mikolov et al., 2014)

The paragraph vector can be further simplified when ignoring the context words in the input but forcing the model to predict words randomly sampled from the paragraph in the output. At inference time, the parameters of the classifier and the word vectors are not needed, and back-propagation is used to tune the paragraph vectors. That is the distributed bag of words version of the paragraph vector. The distributed bag of words model works in the same way using skip-grams (Mikolov et al., 2014), except that a special token representing the document replaces the input.

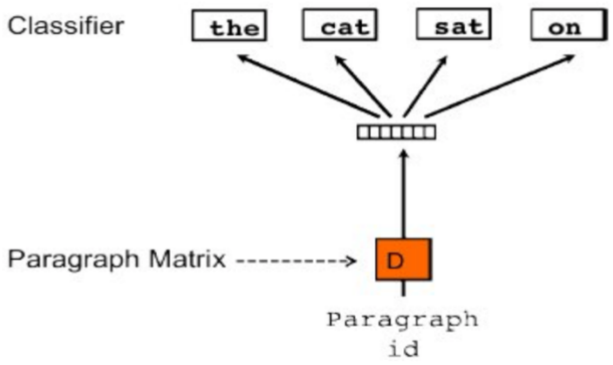


Figure 20: Paragraph Vector: A distributed bag-of-words model (PV-DBOW) (Mikolov et al., 2014)

6.3 Qur'anic Topic Modelling using Paragraph Vectors

This chapter presents a new vector representation of the Qur'anic verses at the paragraph level/verse level. Here, we evaluate the derived vectors in the ability to capture syntactic and semantic features in the text. Therefore, we apply unsupervised learning to mine insights and perceptions from the input data, Qur'an verses/ documents. We cluster the verses of the Qur'an using the trained model (doc2vec) and apply the K-Means clustering algorithm to find the best possible groupings of the documents. We use this approach to model topics from the Qur'an and reveal significant patterns that would aid in inferring coherent topics.

6.3.1 Methodology

The goal is to generate a document embedding space that models and explains the verses distribution in the Holy Qur'an. The dimensions in the space represent the semantic structure in the data and ultimately help to identify main topics and concepts in the text. Therefore, we use an unsupervised document embedding technique: Paragraph Vectors, to learn a vector representation of the verses of the Qur'an. We aim to learn embeddings of the original 6,236 verses of the Qur'an. These vectors can then be used as features and leveraged for the clustering and topic analysis. The diagram in Figure 21 shows an overview of the ML pipeline.

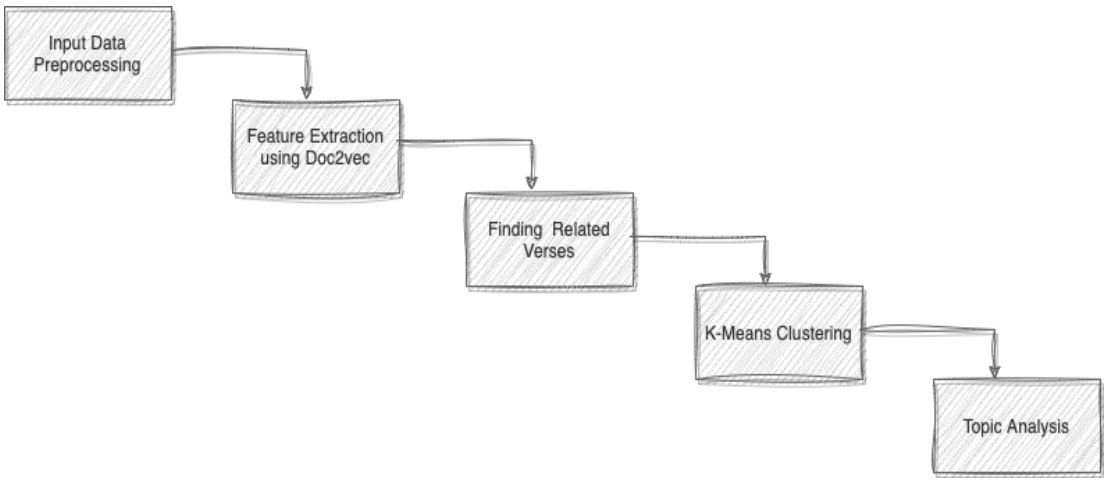


Figure 21: ML Pipeline for modelling semantic relations between the verses of the Qur'an and the topic analysis

6.3.2 Input Data Pre-processing

We used Doc2vec implemented in Gensim to learn vector representation of the Qur'anic verses. We trained the paragraph vectors on the 6,236 verses/passages of the Qur'an using the classical Arabic text from Tanzil project. First, we read the verses from a digitized version of the Qur'an as a data frame. We pre-processed and cleaned the text using the NLTK library such that the document is ready for training. We removed punctuation, Harakat, and stop-words. Figure 22 shows a snapshot of the data before it is been processed to be ready for training.

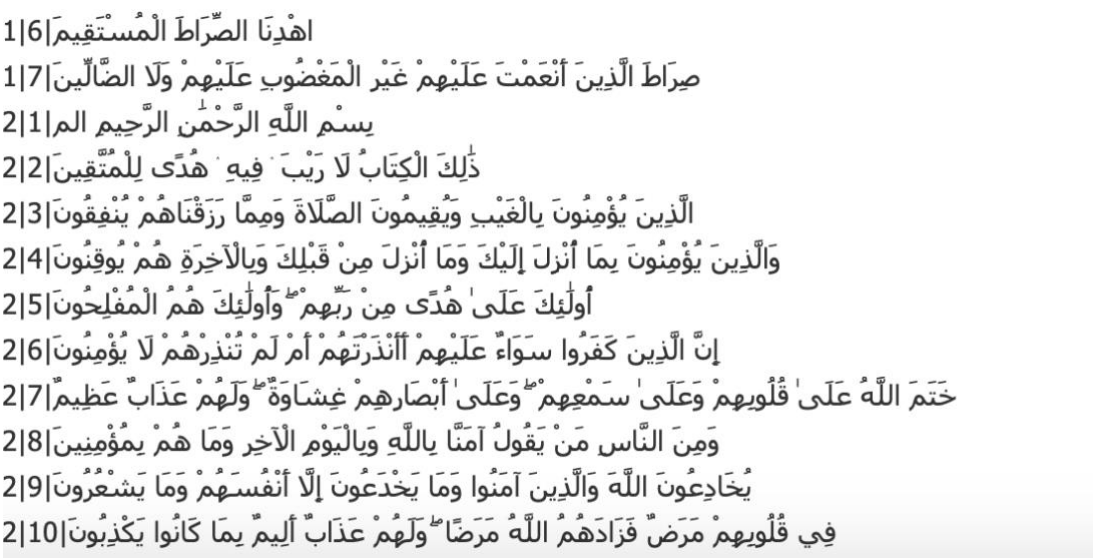


Figure 22: A sample of the input data

6.3.3 Feature Extraction

The text documents need to be encoded as numerical vectors for use as input to a machine learning algorithm, which is called feature extraction (or vectorization). To generate the verses embeddings, we use an unsupervised document embedding technique: Paragraph Vectors. This method is presented by Le and Mikolov (2014) and referred to as Doc2vec[37]. We use the python implementation of doc2vec as part of the Gensim package. We train the Paragraph Vectors model on the 80% of Qur'an verses/passages in their cleaned-format, 4961 out of the total of 6,236 verses. We train the Doc2vec model with different configurations for the hyper-parameters: vector size, number of iterations, and minimum word count. The data has undergone multiple runs to tune the hyperparameters. We have drawn on our domain knowledge to poke the model in the right direction, in addition to using the available Qur'anic knowledge resources such as: The Quranic Arabic Corpus (Dukes et al., 2013), and Qurany (Abbas,2009).

[37] Doc2vec paragraph embedding was popularised by Gensim - a widely-used implementation of Paragraph Vectors: https://radimrehurek.com/gensim/

6.4 Qualitative Results

6.4.1 Assessing the model

To assess the model, we first infer new vectors for each document of the training corpus (4961 documents /verses), compare the inferred vectors with the training corpus, and then return the rank of the document based on self-similarity. Essentially, it is pretending new unseen data in the training corpus and comparing it to the trained model. Additionally, we keep track of the second ranks for a comparison of less similar documents.

We checked the inferred vector against a training vector as a sanity check as to whether the model is behaving in a usefully consistent manner, though not a real ‘accuracy’ value. It is just to give an indication that the model is behaving correctly. If the model was able to capture such similarity, then it is likely to capture the semantic similarity between documents.

First, we trained the model with a vector size with 50. When iterating over the training corpus 40 times, removing stop words, 3617 verses were similar to itself, which is 72% of the inferred documents and about 28% of the time it is mistakenly most similar to another document. It is observed that there is a slight difference when using different techniques in pre-processing input data, by interchangeably removing and keeping the stop words.

Epochs	Vector size	Removing Stop Words	# Similar documents to itself	Correct detection (%)	Wrong detection (%)
40	50	Y	3617	72%	28%
		N	4015	80%	20%
40	100	Y	3554	71%	29%
		N	3986	79%	17%
60	50	Y	3800	76%	24%
		N	4453	89%	11%
60	100	Y	3776	75%	25%
		N	4416	88%	12%
100	50	Y	3774	75%	25%
		N	**4614**	92%	8%
100	100	Y	3735	74%	26%
		N	4578	91.50%	8.50%

Table 23: The Model similarity-detection rate with different settings of hyperparameters (vs, epochs)

Vector size = 50, epochs = 100
Counter({0: 4614, 1: 140, 2: 26, 3: 18, 8: 11, 4: 8, 5: 7, 7: 7, 10: 6, 6: 5, 11: 4, 18: 4, 16: 4, 9: 4, 17: 3, 27: 3, 37: 3, 14: 3, 15: 3, 30: 2, 28: 2, 38: 2, 34: 2, 21: 2, 26: 2, 36: 2, 49: 2, 24: 2, 22: 2, 129: 1, 65: 1, 20: 1, 72: 1, 45: 1, 111: 1, 57: 1, 4176: 1, 4596: 1, 127: 1, 202: 1, 3067: 1, 52: 1, 78: 1, 86: 1, 33: 1, 63: 1, 101: 1, 983: 1, 29: 1, 615: 1, 157: 1, 3687: 1, 47: 1, 208: 1, 3217: 1, 1069: 1, 4997: 1, 23: 1, 41: 1, 54: 1, 107: 1, 2532: 1, 123: 1, 53: 1, 630: 1, 2514: 1, 4806: 1, 862: 1, 12: 1, 529: 1, 93: 1, 572: 1, 227: 1, 3745: 1, 218: 1, 663: 1, 901: 1, 4999: 1, 122: 1, 32: 1, 193: 1, 3555: 1, 55: 1, 1307: 1, 71: 1, 62: 1, 4219: 1, 31: 1, 232: 1, 568: 1, 308: 1, 68: 1, 19: 1, 3635: 1, 3493: 1, 3585: 1, 324: 1, 330: 1, 272: 1, 484: 1, 237: 1, 255: 1, 209: 1, 141: 1, 119: 1, 4136: 1, 3672: 1, 4843: 1, 51: 1, 933: 1, 1800: 1, 3565: 1, 143: 1, 4554: 1, 2633: 1, 4771: 1, 79: 1, 3307: 1, 3469: 1, 164: 1, 56: 1, 2772: 1, 178: 1, 1649: 1, 104: 1, 883: 1, 4346: 1, 4935: 1, 490: 1, 4890: 1, 97: 1, 2459: 1, 1341: 1, 2127: 1, 374: 1, 3337: 1, 3759: 1, 4867: 1, 3274: 1})

Table 24: 92% of the inferred documents were found to be most similar to itself and about 8% of the time it is mistakenly most similar to another document

Table 23 shows how many documents were found to be similar to itself with different settings of the model parameters vector-size and epochs. When increasing the number of epochs to be 100, we got the best similarity-detection results, of 4614 verses that are similar to itself, which is 92% as shown in Table 24.

The most similar document (usually the same text) has a similarity score approaching 1.0. However, the similarity score for the second-ranked documents should be significantly lower. Table 25 and Table 26 show some examples for verses along with the most-similar and the second-most similar ones based on the similarity score. The tables show instances from the 8% where most-similar verse shares the same subject as the one in question but not similar to itself.

Vector size = 50		
Verse 13: 40		وَإِن مَّا نُرِيَنَّكَ بَعْضَ الَّذِي نَعِدُهُمْ أَوْ نَتَوَفَّيَنَّكَ فَإِنَّمَا عَلَيْكَ الْبَلَاغُ وَعَلَيْنَا الْحِسَابُ And whether We show you part of what We promise them or take you in death, upon you is only the [duty of] notification, and upon Us is the account.
Similarity score		
Most similar verse **39: 72**	0.94	قِيلَ ادْخُلُوا أَبْوَابَ جَهَنَّمَ خَالِدِينَ فِيهَا ۖ فَبِئْسَ مَثْوَى الْمُتَكَبِّرِينَ [To them] it will be said, "Enter the gates of Hell to abide eternally therein, and wretched is the residence of the arrogant."
Second-Most similar verse **11: 84**	0.58	وَإِلَىٰ مَدْيَنَ أَخَاهُمْ شُعَيْبًا ۚ قَالَ يَا قَوْمِ اعْبُدُوا اللَّهَ مَا لَكُم مِّنْ إِلَٰهٍ غَيْرُهُ ۖ وَلَا تَنقُصُوا الْمِكْيَالَ وَالْمِيزَانَ ۚ إِنِّي أَرَاكُم بِخَيْرٍ وَإِنِّي أَخَافُ عَلَيْكُمْ عَذَابَ يَوْمٍ مُّحِيطٍ And to Madyan [We sent] their brother Shu'ayb. He said, "O my people, worship Allah; you have no deity other than Him. And do not decrease from the measure and the scale. Indeed, I see you in prosperity, but indeed, I fear for you the punishment of an all-encompassing Day.

Table 25: Finding most-similar and second-most similar to a verse using the trained model (VS = 50)

Vector size = 100		
Verse **41: 28**	ذلك جزاء أعداء الله النار لهم فيها دار الخلد جزاء بما كانوا بآياتنا يجحدون That is the recompense of the enemies of Allah - the Fire. For them therein is the home of eternity as recompense for what they, of Our verses, were rejecting.	
Similarity score		
Most similar verse **76: 11**	0.92	فوقاهم الله شر ذلك اليوم ولقاهم نضرة وسرورا So, Allah will protect them from the evil of that Day and give them radiance and happiness.
Second-Most similar verse **7: 23**	0.63	قالا ربنا ظلمنا أنفسنا وإن لم تغفر لنا وترحمنا لنكونن من الخاسرين They said, "Our Lord, we have wronged ourselves, and if You do not forgive us and have mercy upon us, we will surely be among the losers."

Table 26: Finding most-similar and second-most similar to a verse using the trained model (VS = 100)

We also derived a random sample of other target-document comparisons. By picking a random document from the corpus, we inferred a vector from the model, compared and printed the second-most-similar document. When examining the model for capturing similarity, we usually consider the second-most similar verse rather than the most similar one as it is useless to know that each document is similar to itself. Table27 and Table28 show examples of the results.

Vector size = 50		
Training Document (Verse)	Similar Document (Verse)	Similarity Score
إِن تُبْدُوا الصَّدَقَاتِ فَنِعِمَّا هِيَ ۖ وَإِن تُخْفُوهَا وَتُؤْتُوهَا الْفُقَرَاءَ فَهُوَ خَيْرٌ لَّكُمْ ۖ وَيُكَفِّرُ عَنكُم مِّن سَيِّئَاتِكُمْ ۗ وَاللَّهُ بِمَا تَعْمَلُونَ خَبِيرٌ If you disclose your charitable expenditures, they are good; but if you conceal them and give them to	قَوْلٌ مَّعْرُوفٌ وَمَغْفِرَةٌ خَيْرٌ مِّن صَدَقَةٍ يَتْبَعُهَا أَذًى ۗ وَاللَّهُ غَنِيٌّ حَلِيمٌ Kind speech and forgiveness are better than charity followed by injury. And Allah is Free of need and forbearing. [2: 263]	0.78

the poor, it is better for you, and He will remove from you some of your misdeeds [thereby]. And Allah, with what you do, is [fully] Acquainted. [2: 271]		
وَلَنُذِيقَنَّهُم مِّنَ الْعَذَابِ الْأَدْنَىٰ دُونَ الْعَذَابِ الْأَكْبَرِ لَعَلَّهُمْ يَرْجِعُونَ And we will surely let them taste the nearer punishment short of the greater punishment that perhaps they will repent. [32: 21]	يُضَاعَفْ لَهُ الْعَذَابُ يَوْمَ الْقِيَامَةِ وَيَخْلُدْ فِيهِ مُهَانًا Multiplied for him is the punishment on the Day of Resurrection, and he will abide therein humiliated. [25: 69]	0.75
إِنَّا جَعَلْنَا فِي أَعْنَاقِهِمْ أَغْلَالًا فَهِيَ إِلَى الْأَذْقَانِ فَهُم مُّقْمَحُونَ Indeed, we have put shackles on their necks, and they are to their chins, so they are with heads [kept] aloft. [36: 8]	وَيَوْمَ يُحْشَرُ أَعْدَاءُ اللَّهِ إِلَى النَّارِ فَهُمْ يُوزَعُونَ And [mention, O Muhammad], the Day when the enemies of Allah will be gathered to the Fire while they are [driven] assembled in rows, [41: 19]	0.74
فَأَمَّا الَّذِينَ كَفَرُوا فَأُعَذِّبُهُمْ عَذَابًا شَدِيدًا فِي الدُّنْيَا وَالْآخِرَةِ وَمَا لَهُم مِّن نَّاصِرِينَ And as for those who disbelieved, I will punish them with a severe punishment in this world and the Hereafter, and they will have no helpers." [3: 56]	مَتَاعٌ قَلِيلٌ وَلَهُمْ عَذَابٌ أَلِيمٌ [It is but] a brief enjoyment, and they will have a painful punishment. [16: 117]	0.85
أَوَلَمْ يَرَوْا كَيْفَ يُبْدِئُ اللَّهُ الْخَلْقَ ثُمَّ يُعِيدُهُ ۚ إِنَّ ذَٰلِكَ عَلَى اللَّهِ يَسِيرٌ Have they not considered how Allah begins creation and then repeats it? Indeed that, for Allah, is easy. [29: 19]	فَانظُرْ إِلَىٰ آثَارِ رَحْمَتِ اللَّهِ كَيْفَ يُحْيِي الْأَرْضَ بَعْدَ مَوْتِهَا ۚ إِنَّ ذَٰلِكَ لَمُحْيِي الْمَوْتَىٰ ۖ وَهُوَ عَلَىٰ كُلِّ شَيْءٍ قَدِيرٌ So, observe the effects of the mercy of Allah - how He gives life to the earth after its lifelessness. Indeed, that [same one] will give life to the dead, and He is over all things competent. [30: 50]	0.69
وَيْلٌ لِّلَّذِينَ يَكْتُبُونَ الْكِتَابَ بِأَيْدِيهِمْ ثُمَّ يَقُولُونَ هَٰذَا مِنْ عِندِ اللَّهِ لِيَشْتَرُوا بِهِ ثَمَنًا قَلِيلًا ۖ فَوَيْلٌ لَّهُم مِّمَّا كَتَبَتْ أَيْدِيهِمْ وَوَيْلٌ لَّهُم مِّمَّا يَكْسِبُونَ So, woe to those who write the "scripture" with their own hands, then say, "This is from Allah," in order to exchange it for a small	إِنَّ الَّذِينَ يَكْتُمُونَ مَا أَنزَلَ اللَّهُ مِنَ الْكِتَابِ وَيَشْتَرُونَ بِهِ ثَمَنًا قَلِيلًا ۙ أُولَٰئِكَ مَا يَأْكُلُونَ فِي بُطُونِهِمْ إِلَّا النَّارَ وَلَا يُكَلِّمُهُمُ اللَّهُ يَوْمَ الْقِيَامَةِ وَلَا يُزَكِّيهِمْ وَلَهُمْ عَذَابٌ أَلِيمٌ Indeed, they who conceal what Allah has sent down of the Book and exchange it for a small price - those consume not into their bellies	0.74

price. Woe to them for what their hands have written and woe to them for what they earn. [2: 79]	except the Fire. And Allah will not speak to them on the Day of Resurrection, nor will He purify them. And they will have a painful punishment. [2: 174]	
وَالَّذِينَ صَبَرُوا ابْتِغَاءَ وَجْهِ رَبِّهِمْ وَأَقَامُوا الصَّلَاةَ وَأَنفَقُوا مِمَّا رَزَقْنَاهُمْ سِرًّا وَعَلَانِيَةً وَيَدْرَءُونَ بِالْحَسَنَةِ السَّيِّئَةَ أُولَٰئِكَ لَهُمْ عُقْبَى الدَّارِ And those who are patient, seeking the countenance of their Lord, and establish prayer and spend from what We have provided for them secretly and publicly and prevent evil with good - those will have the good consequence of [this] home – [13: 22]	أُولَٰئِكَ هُمُ الْمُؤْمِنُونَ حَقًّا لَّهُمْ دَرَجَاتٌ عِندَ رَبِّهِمْ وَمَغْفِرَةٌ وَرِزْقٌ كَرِيمٌ Those are the believers, truly. For them are degrees [of high position] with their Lord and forgiveness and noble provision. [8: 4]	0.83
ثُمَّ شَقَقْنَا الْأَرْضَ شَقًّا Then we broke open the earth, splitting [it with sprouts], [80: 26]	وَأَخْرَجَتِ الْأَرْضُ أَثْقَالَهَا And the earth discharges its burdens [99: 2]	0.83
أَمْ أَنزَلْنَا عَلَيْهِمْ سُلْطَانًا فَهُوَ يَتَكَلَّمُ بِمَا كَانُوا بِهِ يُشْرِكُونَ Or have We sent down to them an authority, and it speaks of what they were associating with Him? [30: 35]	فَقَدْ كَذَّبُوا فَسَيَأْتِيهِمْ أَنبَاءُ مَا كَانُوا بِهِ يَسْتَهْزِئُونَ For they have already denied, but there will come to them the news of that which they used to ridicule. [26: 6]	0.60

Table27 : Comparisons on different target-documents from the trained model (VS= 50)

The examples confirm that there is a relation between each pair. In each pair, the verses share a common concept that can be recognised by reading through them. For example, the two verses (80: 26) and (99:2) use different terms that explain the same meaning as to represent how the earth is broken open and discharges its burdens. The two terms are ‘ شققنا ‘ and ‘ أخرجت أثقالها ‘ . Another example is the relation between the two verses (2: 271) and (2, 263). Both verses discuss the virtue of charity.

Vector size = 100		
Training Document (Verse)	Similar Document (Verse)	Similarity Score
وَإِذَا قِيلَ لَهُمْ آمِنُوا بِمَا أَنزَلَ اللَّهُ قَالُوا نُؤْمِنُ بِمَا أُنزِلَ عَلَيْنَا وَيَكْفُرُونَ بِمَا وَرَاءَهُ وَهُوَ الْحَقُّ مُصَدِّقًا لِّمَا مَعَهُمْ ۗ قُلْ فَلِمَ تَقْتُلُونَ أَنبِيَاءَ اللَّهِ مِن قَبْلُ إِن كُنتُم مُّؤْمِنِي **And when it is said to them, "Believe in what Allah has revealed," they say, "We believe [only] in what was revealed to us." And they disbelieve in what came after it, while it is the truth confirming that which is with them. Say, "Then why did you kill the prophets of Allah before, if you are [indeed] believers?" [2: 91]**	أَمْ يَقُولُونَ افْتَرَاهُ ۖ قُلْ فَأْتُوا بِسُورَةٍ مِّثْلِهِ وَادْعُوا مَنِ اسْتَطَعْتُم مِّن دُونِ اللَّهِ إِن كُنتُمْ صَادِقِينَ **Or do they say [about the Prophet], "He invented it?" Say, "Then bring forth a surah like it and call upon [for assistance] whomever you can besides Allah, if you should be truthful." [10: 38]**	0.62
وَهُوَ الَّذِي يَتَوَفَّاكُم بِاللَّيْلِ وَيَعْلَمُ مَا جَرَحْتُم بِالنَّهَارِ ثُمَّ يَبْعَثُكُمْ فِيهِ لِيُقْضَىٰ أَجَلٌ مُّسَمًّى ۖ ثُمَّ إِلَيْهِ مَرْجِعُكُمْ ثُمَّ يُنَبِّئُكُم بِمَا كُنتُمْ تَعْمَلُونَ **And it is He who takes your souls by night and knows what you have committed by day. Then He revives you therein that a specified term may be fulfilled. Then to Him will be your return; then He will inform you about what you used to do. [6: 60]**	اللَّهُ يَبْدَأُ الْخَلْقَ ثُمَّ يُعِيدُهُ ثُمَّ إِلَيْهِ تُرْجَعُونَ **Allah begins creation; then He will repeat it; then to Him you will be returned. [30: 11]**	0.93
وَلَقَدْ أَرْسَلْنَا إِلَىٰ أُمَمٍ مِّن قَبْلِكَ فَأَخَذْنَاهُم بِالْبَأْسَاءِ وَالضَّرَّاءِ لَعَلَّهُمْ يَتَضَرَّعُونَ **And We have already sent [messengers] to nations before you, [O Muhammad]; then We seized them with poverty and hardship that perhaps they might humble themselves [to Us]. [6: 42]**	وَمَا أَرْسَلْنَا فِي قَرْيَةٍ مِّن نَّبِيٍّ إِلَّا أَخَذْنَا أَهْلَهَا بِالْبَأْسَاءِ وَالضَّرَّاءِ لَعَلَّهُمْ يَضَّرَّعُونَ **And We sent to no city a prophet [who was denied] except that We seized its people with poverty and hardship that they might humble themselves [to Allah]. [7: 94]**	0.67
وَاتَّقُوا فِتْنَةً لَّا تُصِيبَنَّ الَّذِينَ ظَلَمُوا مِنكُمْ خَاصَّةً ۖ وَاعْلَمُوا أَنَّ اللَّهَ شَدِيدُ الْعِقَابِ **And fear a trial which will not strike those who have wronged among**	فَهَلْ يَنتَظِرُونَ إِلَّا مِثْلَ أَيَّامِ الَّذِينَ خَلَوْا مِن قَبْلِهِمْ ۚ قُلْ فَانتَظِرُوا إِنِّي مَعَكُم مِّنَ الْمُنتَظِرِينَ **So, do they wait except for like [what occurred in] the days of those who passed on before them? Say,**	0.79

you exclusively, and know that Allah is severe in penalty. [8: 25]	"Then wait; indeed, I am with you among those who wait."[10: 102]	
كَلَّا ۖ بَلْ ۜ رَانَ عَلَىٰ قُلُوبِهِم مَّا كَانُوا يَكْسِبُونَ No! Rather, the stain has covered their hearts of that which they were earning. [83: 14]	كَانُوا لَا يَتَنَاهَوْنَ عَن مُّنكَرٍ فَعَلُوهُ ۚ لَبِئْسَ مَا كَانُوا يَفْعَلُونَ They used not to prevent one another from wrongdoing that they did. How wretched was that which they were doing. [5: 79]	0.80
وَلَوْ أَنَّ لِلَّذِينَ ظَلَمُوا مَا فِي الْأَرْضِ جَمِيعًا وَمِثْلَهُ مَعَهُ لَافْتَدَوْا بِهِ مِن سُوءِ الْعَذَابِ يَوْمَ الْقِيَامَةِ ۚ وَبَدَا لَهُم مِّنَ اللَّهِ مَا لَمْ يَكُونُوا يَحْتَسِبُونَ And if those who did wrong had all that is in the earth entirely and the like of it with it, they would [attempt to] ransom themselves thereby from the worst of the punishment on the Day of Resurrection. And there will appear to them from Allah that which they had not taken into account. [39: 47]	إِنَّ الَّذِينَ كَفَرُوا لَوْ أَنَّ لَهُم مَّا فِي الْأَرْضِ جَمِيعًا وَمِثْلَهُ مَعَهُ لِيَفْتَدُوا بِهِ مِنْ عَذَابِ يَوْمِ الْقِيَامَةِ مَا تُقُبِّلَ مِنْهُمْ ۖ وَلَهُمْ عَذَابٌ أَلِيمٌ Indeed, those who disbelieve - if they should have all that is in the earth and the like of it with it by which to ransom themselves from the punishment of the Day of Resurrection, it will not be accepted from them, and for them is a painful punishment. [5: 36]	0.75
وَإِذْ نَتَقْنَا الْجَبَلَ فَوْقَهُمْ كَأَنَّهُ ظُلَّةٌ وَظَنُّوا أَنَّهُ وَاقِعٌ بِهِمْ خُذُوا مَا آتَيْنَاكُم بِقُوَّةٍ وَاذْكُرُوا مَا فِيهِ لَعَلَّكُمْ تَتَّقُونَ And [mention] when We raised the mountain above them as if it was a dark cloud and they were certain that it would fall upon them, [and Allah said], "Take what We have given you with determination and remember what is in it that you might fear Allah." [7: 171]	وَإِذْ أَخَذْنَا مِيثَاقَكُمْ وَرَفَعْنَا فَوْقَكُمُ الطُّورَ خُذُوا مَا آتَيْنَاكُم بِقُوَّةٍ وَاذْكُرُوا مَا فِيهِ لَعَلَّكُمْ تَتَّقُونَ And [recall] when We took your covenant, [O Children of Israel, to abide by the Torah] and We raised over you the mount, [saying], "Take what We have given you with determination and remember what is in it that perhaps you may become righteous." [2: 63]	0.79
ذَٰلِكَ بِأَنَّ اللَّهَ يُولِجُ اللَّيْلَ فِي النَّهَارِ وَيُولِجُ النَّهَارَ فِي اللَّيْلِ وَأَنَّ اللَّهَ سَمِيعٌ بَصِيرٌ That is because Allah causes the night to enter the day and causes the day to enter the night and because Allah is Hearing and Seeing. [22: 61]	يُقَلِّبُ اللَّهُ اللَّيْلَ وَالنَّهَارَ ۚ إِنَّ فِي ذَٰلِكَ لَعِبْرَةً لِّأُولِي الْأَبْصَارِ Allah alternates the night and the day. Indeed, in that is a lesson for those who have vision. [24: 44]	0.78
هُوَ الَّذِي جَعَلَ لَكُمُ اللَّيْلَ لِتَسْكُنُوا فِيهِ وَالنَّهَارَ مُبْصِرًا ۚ إِنَّ فِي ذَٰلِكَ لَآيَاتٍ لِّقَوْمٍ يَسْمَعُونَ	إِنَّ فِي خَلْقِ السَّمَاوَاتِ وَالْأَرْضِ وَاخْتِلَافِ اللَّيْلِ وَالنَّهَارِ لَآيَاتٍ لِّأُولِي الْأَلْبَابِ	0.91

It is He who made for you the night to rest therein and the day, giving sight. Indeed in that are signs for a people who listen. [10: 67]	**Indeed, in the creation of the heavens and the earth and the alternation of the night and the day are signs for those of understanding. [3: 190]**	
وَلِبُيُوتِهِمْ أَبْوَابًا وَسُرُرًا عَلَيْهَا يَتَّكِئُونَ **And for their houses - doors and couches [of silver] upon which to recline. [43: 34]**	مُتَّكِئِينَ عَلَيْهَا مُتَقَابِلِينَ **Reclining on them, facing each other. [56: 16]**	0.95
وَلَقَدْ آتَيْنَا مُوسَى الْكِتَابَ لَعَلَّهُمْ يَهْتَدُونَ **And We certainly gave Moses the Scripture that perhaps they would be guided. [23: 49]**	فَإِنَّمَا يَسَّرْنَاهُ بِلِسَانِكَ لَعَلَّهُمْ يَتَذَكَّرُونَ **And indeed, we have eased the Qur'an in your tongue that they might be reminded. [44: 58]**	0.71
كُلُّ نَفْسٍ ذَائِقَةُ الْمَوْتِ ثُمَّ إِلَيْنَا تُرْجَعُونَ **Every soul will taste death. Then to Us will you be returned. [29: 57]**	وَهُوَ الَّذِي يَتَوَفَّاكُم بِاللَّيْلِ وَيَعْلَمُ مَا جَرَحْتُم بِالنَّهَارِ ثُمَّ يَبْعَثُكُمْ فِيهِ لِيُقْضَىٰ أَجَلٌ مُّسَمًّى ثُمَّ إِلَيْهِ مَرْجِعُكُمْ ثُمَّ يُنَبِّئُكُم بِمَا كُنتُمْ تَعْمَلُونَ **And it is He who takes your souls by night and knows what you have committed by day. Then He revives you therein that a specified term may be fulfilled. Then to Him will be your return; then He will inform you about what you used to do. [6: 60]**	0.85

Table28 : More comparisons on different target-documents from the trained model (VS = 100)

The verses (2: 91) and (10: 38) have a common theme as well. The two verses talk about disbelievers as they disbelieve in what was revealed to the prophet Muhammad. Moreover, verse (6: 6) and (29: 57) have a similarity of 85% as they both confirm that people will all return to Allah. It is observed that changing the vector dimension captured many relations between the documents.

6.4.2 Testing the model

Here, we used the same approach with documents (verses) from the test dataset. We inferred the vector for a randomly chosen test document (verse), and compared the document to our model by eye. Samples of the results are shown in Table 29. The examples show obvious correlations between each pair of verses. The relation was spotted in each pair, and confirmed by the similarity score per our model predication. Thus, a collection of pairs was compiled to act as a significant contribution of this research[38]. The dataset constitutes pairs of related verses, verified against acceptable sources; Al-Tabari and Ibn-Kathir commentaries. The dataset, indeed, can serve as an evaluation resource for the semantic similarity in the Qur'an.

Vector size = 50		
Training Document (Verse)	Similar Document (Verse)	Similarity Score
يَوْمَئِذٍ تُحَدِّثُ أَخْبَارَهَا **That Day, it will report its news [99: 4]**	وُجُوهٌ يَوْمَئِذٍ مُّسْفِرَةٌ **[Some] faces, that Day, will be bright [80: 38]**	0.98
وَلَوْ أَنَّ لِكُلِّ نَفْسٍ ظَلَمَتْ مَا فِي الْأَرْضِ لَافْتَدَتْ بِهِ ۗ وَأَسَرُّوا النَّدَامَةَ لَمَّا رَأَوُا الْعَذَابَ ۖ وَقُضِيَ بَيْنَهُم بِالْقِسْطِ وَهُمْ لَا يُظْلَمُونَ **And if each soul that wronged had everything on earth, it would offer it in ransom. And they will confide regret when they see the punishment; and they will be judged in justice, and they will not be wronged. [10: 54]**	فَكَيْفَ إِذَا جَمَعْنَاهُمْ لِيَوْمٍ لَّا رَيْبَ فِيهِ وَوُفِّيَتْ كُلُّ نَفْسٍ مَّا كَسَبَتْ وَهُمْ لَا يُظْلَمُونَ **So how will it be when We assemble them for a Day about which there is no doubt? And each soul will be compensated [in full for] what it earned, and they will not be wronged. [3: 25]**	0.63
وَمَا ظَلَمْنَاهُمْ وَلَٰكِن كَانُوا هُمُ الظَّالِمِينَ **And We did not wrong them, but it was they who were the wrongdoers. [43: 76]**	{وَكَمْ أَهْلَكْنَا مِن قَرْيَةٍ بَطِرَتْ مَعِيشَتَهَا ۖ فَتِلْكَ مَسَاكِنُهُمْ لَمْ تُسْكَن مِّن بَعْدِهِمْ إِلَّا قَلِيلًا ۖ وَكُنَّا نَحْنُ الْوَارِثِينَ **And how many a city have We destroyed that was insolent in its [way of] living, and those are their dwellings which have not been inhabited after them except briefly.**	0.46

[38] All produced dataset will be accessible at Github repository: Mhalshammeri

	And it is We who were the inheritors. [28: 58]	
قَالُوا مَا هِيَ إِلَّا حَيَاتُنَا الدُّنْيَا نَمُوتُ وَنَحْيَا وَمَا يُهْلِكُنَا إِلَّا الدَّهْرُ ۚ وَمَا لَهُم بِذَٰلِكَ مِنْ عِلْمٍ ۖ إِنْ هُمْ إِلَّا يَظُنُّونَ And they say, "There is not but our worldly life; we die and live, and nothing destroys us except time." And they have of that no knowledge; they are only assuming. [45: 24]	إِنْ هِيَ إِلَّا حَيَاتُنَا الدُّنْيَا نَمُوتُ وَنَحْيَا وَمَا نَحْنُ بِمَبْعُوثِينَ Life is not but our worldly life - we die and live, but we will not be resurrected. [23: 37]	0.77
فَتَقَطَّعُوا أَمْرَهُم بَيْنَهُمْ زُبُرًا ۖ كُلُّ حِزْبٍ بِمَا لَدَيْهِمْ فَرِحُونَ But the people divided their religion among them into sects - each faction, in what it has, rejoicing. [23: 53]	مِنَ الَّذِينَ فَرَّقُوا دِينَهُمْ وَكَانُوا شِيَعًا ۖ كُلُّ حِزْبٍ بِمَا لَدَيْهِمْ فَرِحُونَ [Or] of those who have divided their religion and become sects, every faction rejoicing in what it has. [30: 32]	0.81
قَالَ لَا تَخْتَصِمُوا لَدَيَّ وَقَدْ قَدَّمْتُ إِلَيْكُم بِالْوَعِيدِ [Allah] will say, "Do not dispute before Me, while I had already presented to you the warning. [50: 28]	فَتَوَلَّىٰ عَنْهُمْ وَقَالَ يَا قَوْمِ لَقَدْ أَبْلَغْتُكُمْ رِسَالَةَ رَبِّي وَنَصَحْتُ لَكُمْ وَلَٰكِن لَّا تُحِبُّونَ النَّاصِحِينَ And he turned away from them and said, "O my people, I had certainly conveyed to you the message of my Lord and advised you, but you do not like advisors." [7: 79]	0.79
قُلْ مَن يَرْزُقُكُم مِّنَ السَّمَاءِ وَالْأَرْضِ أَمَّن يَمْلِكُ السَّمْعَ وَالْأَبْصَارَ وَمَن يُخْرِجُ الْحَيَّ مِنَ الْمَيِّتِ وَيُخْرِجُ الْمَيِّتَ مِنَ الْحَيِّ وَمَن يُدَبِّرُ الْأَمْرَ ۚ فَسَيَقُولُونَ اللَّهُ ۚ فَقُلْ أَفَلَا تَتَّقُونَ Say, "Who provides for you from the heaven and the earth? Or who controls hearing and sight and who brings the living out of the dead and brings the dead out of the living and who arranges [every] matter?" They will say, "Allah," so say, "Then will you not fear Him?" [10: 31]	خَلَقَكُم مِّن نَّفْسٍ وَاحِدَةٍ ثُمَّ جَعَلَ مِنْهَا زَوْجَهَا وَأَنزَلَ لَكُم مِّنَ الْأَنْعَامِ ثَمَانِيَةَ أَزْوَاجٍ ۚ يَخْلُقُكُمْ فِي بُطُونِ أُمَّهَاتِكُمْ خَلْقًا مِّن بَعْدِ خَلْقٍ فِي ظُلُمَاتٍ ثَلَاثٍ ۚ ذَٰلِكُمُ اللَّهُ رَبُّكُمْ لَهُ الْمُلْكُ ۖ لَا إِلَٰهَ إِلَّا هُوَ ۖ فَأَنَّىٰ تُصْرَفُونَ He created you from one soul. Then He made from it its mate, and He produced for you from the grazing livestock eight mates. He creates you in the wombs of your mothers, creation after creation, within three darkness. That is Allah, your Lord; to Him belongs dominion. There is no deity except Him, so how are you averted? [39: 6]	0.84
يَا أَيُّهَا الَّذِينَ آمَنُوا لَا تَتَّخِذُوا بِطَانَةً مِّن دُونِكُمْ لَا يَأْلُونَكُمْ خَبَالًا وَدُّوا مَا عَنِتُّمْ قَدْ بَدَتِ الْبَغْضَاءُ مِنْ	ضَرَبَ لَكُم مَّثَلًا مِّنْ أَنفُسِكُمْ ۖ هَل لَّكُم مِّن مَّا مَلَكَتْ أَيْمَانُكُم مِّن شُرَكَاءَ فِي مَا رَزَقْنَاكُمْ فَأَنتُمْ فِيهِ سَوَاءٌ	0.86

أَفْوَاهِهِمْ وَمَا تُخْفِي صُدُورُهُمْ أَكْبَرُ ۚ قَدْ بَيَّنَّا لَكُمُ الْآيَاتِ ۖ إِن كُنتُمْ تَعْقِلُونَ **O you who have believed, do not take as intimates those other than yourselves, for they will not spare you [any] ruin. They wish you would have hardship. Hatred has already appeared from their mouths, and what their breasts conceal is greater. We have certainly made clear to you the signs, if you will use reason. [3: 118]**	تَخَافُونَهُمْ كَخِيفَتِكُمْ أَنفُسَكُمْ ۗ كَذَٰلِكَ نُفَصِّلُ الْآيَاتِ لِقَوْمٍ يَعْقِلُونَ **He presents to you an example from yourselves. Do you have among those whom your right hands possess any partners in what We have provided for you so that you are equal therein [and] would fear them as your fear of one another [within a partnership]? Thus, do We detail the verses for a people who use reason. [30: 28]**	

Table 29: More comparisons using verses from the test data

6.5 Evaluation

6.5.1 Finding similar/ related documents

We evaluated the vectors on the task of finding similar verses to examine their effectiveness in capturing the semantics of the verses/passages of the Qur'an. We inferred the vector for a randomly chosen test document/verse and compared the document to our model.

6.5.1.1 Results

Using intuitive self-evaluation, we were able to locate semantically similar verses and eventually created a dataset of pairs of related verses along with their similarity score. We decide on 50 as the vector size that produced best results in terms of the similarity between the verses in each pair. We used the Qurany ontology browser to identify how verses in each derived pair are related. Figure 23 and Figure 24 are examples of the resultant pairs.

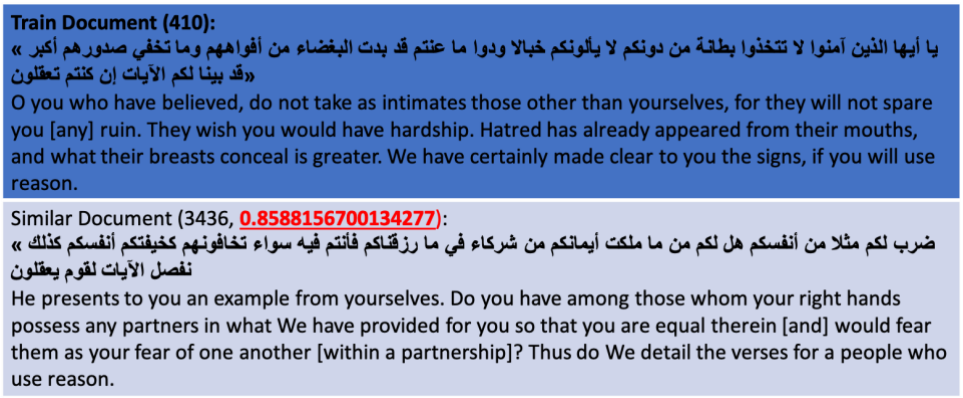


Figure 23: A pair of related verses generated by the trained model with similarity score of 0.85, example 1

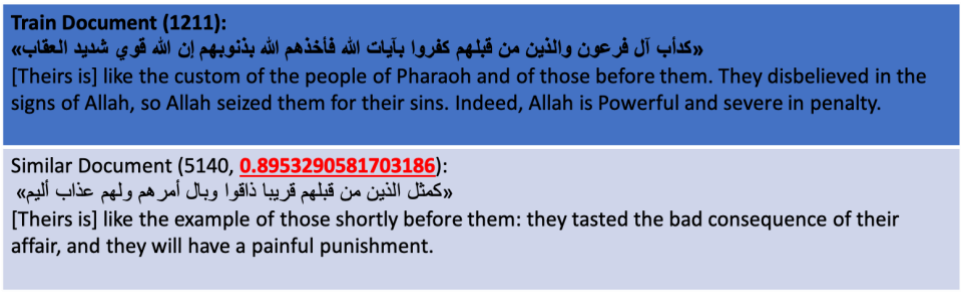


Figure 24: A pair of related verses generated by the trained model with similarity score of 0.89, example 2

However, we recognized that some instances have deep relations, which is hard to discover manually. However, the model has detected links which is reflected in high similarity scores as in **Error! Reference source not found.**. By thoroughly analysing the two verses with the aid of commentary books, we can confirm the embedded relation between the two verses. Both verses indeed entail an implicit message to believers and affirmed the messenger's keenness on their guidance and attainment in worldly life and hereafter[39]. This is an excellent example of instances where our model reveals the intrinsic knowledge embedded within the verses of the Qur'an, and with further analysis we may be able to identify patterns and hidden details that explain such connections.

Verse1 [59: 6]	Verse2 [9: 128]	Similarity score
أفاء الله على رسوله منهم فما أوجفتم عليه من خيل ولا ركاب ولكن الله يسلط رسله على من يشاء والله على كل شيء قدير	لقد جاءكم رسول من أنفسكم عزيز عليه ما عنتم حريص عليكم بالمؤمنين رءوف رحيم	0.984

39 https://quran.ksu.edu.sa/tafseer/

You did not charge with horse or camel for whatever (spoils) God gave His Apostle from them. In any case, God gives authority to His Apostle over whomsoever He please. God has power over everything.	To you has come an Apostle from among you. Any sorrow that befalls you weighs upon him; He is eager for your happiness, full of concern for the faithful, compassionate and kind.	

Table 30: A Pairs of verses with deep relation detected by proposed model

6.5.2 K-Means clustering

We cluster the verses of the Qur'an by applying a clustering algorithm on the derived vectors. The objective is to infer patterns in the data that can inform a decision to guide the learning process. With clustering, we seek to capture in some way the topics or semantic relations in our corpus.

The goal of clustering is grouping unlabelled texts in such a way that texts in the same group/cluster are more similar to each other than to those in other clusters. K-means is one of the most used clustering algorithms due to its simplicity (Xu and Tian, 2015). K-Means puts the observations into K clusters in which each observation belongs to a cluster with the nearest mean. The main idea is to define K centroids, one for each cluster. Indeed, we identify verses that are similar to each other in terms of meanings/ topics.

We implemented our K-Means clustering algorithm on our vectorized documents. We did a couple of trial/errors to find the best number of clusters. We tried different values ranging from 5 to 20. One approach we considered was to set the number of clusters to be 15 (the number of main topics from our evaluation resource: Qurany), assuming that we have a general sense of the right number of clusters. We implemented the algorithm in Python with the help of the SciKit Learn library[40].

[40] https://scikit-learn.org/stable/modules/clustering.html

We present a range of qualitative and quantitative results. We compare our results against the Qurany corpus to verify the relationships between the verses of the Qur'an, identify how they are related, and address the concepts covered in each cluster.

6.5.2.1 Qualitative Analysis

The qualitative technique demands domain knowledge to sense-check the clustering algorithm's results. To analyse the data, we find the most representative tokens and documents by looking for vectors that are closest to the clusters' centroids. As a result, in each cluster, we look for the most representative keywords/terms and verses. Our findings confirmed that the Paragraph Vectors representations offered a useful input representation that promoted the clustering performance. The list in Table 31 browses verses from cluster 2.

Chapter	Ayah	Verse
2	30	وإذ قال ربك للملائكة إني جاعل في الأرض خليفة قالوا أتجعل فيها من يفسد فيها ويسفك الدماء ونحن نسبح بحمدك ونقدس لك قال إني أعلم ما لا تعلمون Remember, when your Lord said to the angels: "I have to place a trustee on the earth," they said: "Will You place one there who would create disorder and shed blood, while we intone Your litanies and sanctify Your name?" And God said: "I know what you do not know."
2	33	قال يا آدم أنبئهم بأسمائهم فلما أنبأهم بأسمائهم قال ألم أقل لكم إني أعلم غيب السماوات والأرض وأعلم ما تبدون وما كنتم تكتمون Then He said to Adam: "Convey to them their names." And when he had told them, God said: "Did I not tell you that I know the unknown of the heavens and the earth, and I know what you disclose and know what you hide?"
2	61	وإذ قلتم يا موسى لن نصبر على طعام واحد فادع لنا ربك يخرج لنا مما تنبت الأرض من بقلها وقثائها وفومها وعدسها وبصلها قال أتستبدلون الذي هو أدنى بالذي هو خير اهبطوا مصرا فإن لكم ما سألتم وضربت عليهم الذلة والمسكنة وباءوا بغضب من الله ذلك بأنهم كانوا يكفرون بآيات الله ويقتلون النبيين بغير الحق ذلك بما عصوا وكانوا يعتدون Remember, when you said: "O Moses, we are tired of eating the same food (day after day), ask your Lord to give us fruits of the earth, herbs and cucumbers, grains and lentils and onions;" he said: "Would you rather exchange what is good with what is bad? Go then to the city, you shall have what you ask." So they were disgraced and became indigent, earning the anger of God, for they disbelieved the word of God, and slayed the prophets unjustly, for they transgressed and rebelled.

Table 31: A list of verses located in cluster 2

We qualitatively looked at the documents in each cluster and identified the semantic relation or the common subject intuitively. For example, as in Figure 25, cluster 2 contains a group of verses that are oriented around the creation concept.

```
Cluster: 0
495
   Cluster                                              Verse
0        0  ...وإذا لقوا الذين آمنوا قالوا آمنا وإذا خلوا إلى
1        0  ... والذين كفروا وكذبوا بآياتنا أولئك أصحاب النار
2        0  ...وآمنوا بما أنزلت مصدقا لما معكم ولا تكونوا أول
3        0  ...وظللنا عليكم الغمام وأنزلنا عليكم المن والسلوى
4        0  ...ولقد علمتم الذين اعتدوا منكم في السبت فقلنا له
Cluster: 1
608
   Cluster                                              Verse
0        1  ...وإذ قال ربك للملائكة إني جاعل في الأرض خليفة ق
1        1  ...وعلم آدم الأسماء كلها ثم عرضهم على الملائكة فق
2        1  ... قالوا سبحانك لا علم لنا إلا ما علمتنا إنك أنت
3        1  ...وإذ فرقنا بكم البحر فأنجيناكم وأغرقنا آل فرعون
4        1  ...وإذ قال موسى لقومه إن الله يأمركم أن تذبحوا بق
Cluster: 2
358
   Cluster                                              Verse
0        2  ...أو كصيب من السماء فيه ظلمات ورعد وبرق يجعلون أ
1        2  ...الذي جعل لكم الأرض فراشا والسماء بناء وأنزل من
2        2  ...هو الذي خلق لكم ما في الأرض جميعا ثم استوى إلى
3        2  ...وإذ استسقى موسى لقومه فقلنا اضرب بعصاك الحجر ف
4        2  ...ثم قست قلوبكم من بعد ذلك فهي كالحجارة أو أشد ق
Cluster: 3
298
   Cluster                                              Verse
0        3  ...قال يا آدم أنبئهم بأسمائهم فلما أنبأهم بأسمائه
1        3  ...وإذ قلنا للملائكة اسجدوا لآدم فسجدوا إلا إبليس
2        3  ...وقلنا يا آدم اسكن أنت وزوجك الجنة وكلا منها رغ
```

Figure 25: The derived cluster using K-Means and Doc2vec

6.5.2.2 Performance of K-Means Clustering

For the quantitative evaluation of the number of clusters, we can use metrics like inertia and Silhouette Score[41] . First, we consider the inertia metric، which is the within cluster sum of squares of distances to the cluster centre. The algorithm aims to choose centroids that minimize the inertia, which can indicate how internally coherent clusters are. The other metric is Silhouette Score which can be used to determine the degree of separation between clusters. Silhouette Score is calculated using the mean intra-cluster distance and the mean nearest-cluster distance, and goes from -1 to 1. As the coefficient approaches 1, it indicates good clustering.

So, we aim at a decreasing value of inertia as number of clusters increases, and a Silhouette score approaching 1, to indicate that we are producing a good clustering. After we calculated the inertia and silhouette scores, we plotted them and evaluated the performance of the clustering algorithm. Figure 26 shows the result of the two metrics. The inertia score always drops when we increase the number of clusters. From the Silhouette curve, when the number of clusters=14,

[41] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

this has the best average silhouette score with all clusters being above the average showing that it is actually a good choice. Joining the elbow curve with the Silhouette score curve provides valuable insight into the performance of K-Means.

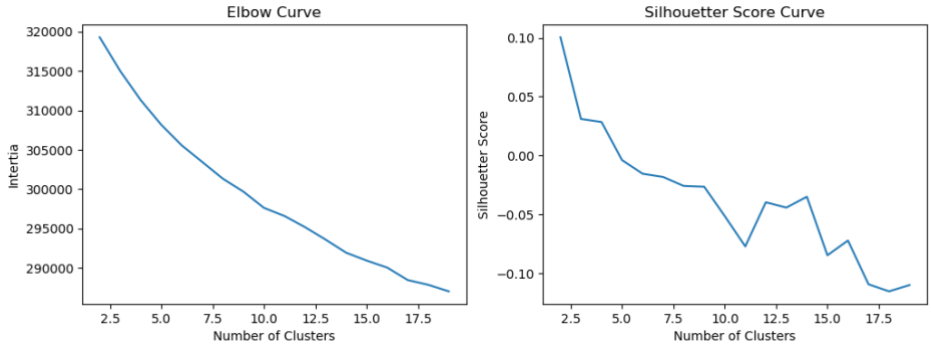


Figure 26: Performance of clustering using the metrics: Inertia & Silhouette score

6.5.2.3 Data Analysis using Word Clouds

A Word Cloud is a way to help visually interpret the text and gain insight into the most noticeable items in a text by visualizing the word frequency in the text as a weighted list. We generated the word clouds per cluster using Word Cloud Python plugin[42]

Here, we explore the text within each cluster to identify the main themes in each cluster and potentially interpret them. To do so, we created a word-cloud that represents the frequency or the importance of words in each cluster. Thus, we generated 15 word-clouds associated with the 15 clusters we had. The size of the word identifies its importance. Significant words describe the theme of each cloud. Figure 27 and Figure 28 illustrate examples of the clusters' word clouds, along with a partial list of associated verses in a cluster as in Table 32 and Table 33.

[42] https://github.com/amueller/word_cloud

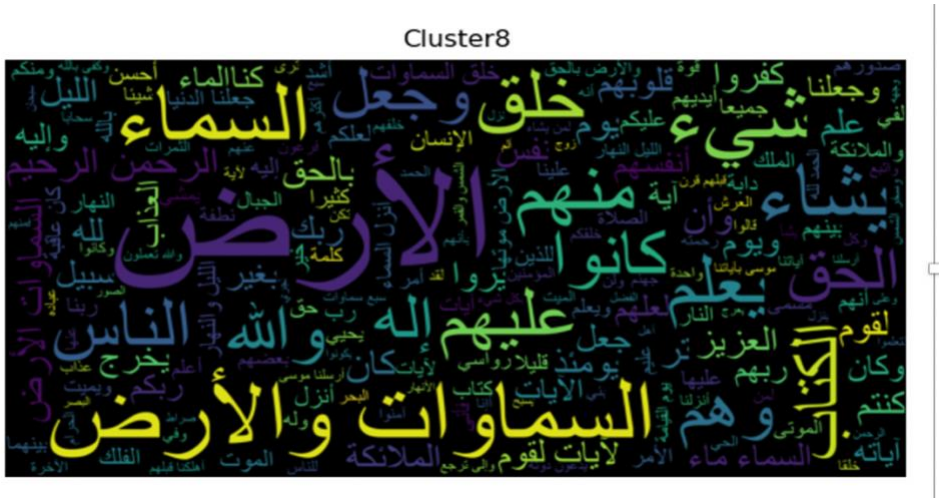


Figure 27: A word cloud visualization of cluster 8

Verses in Cluster 8
إن في خلق السماوات والأرض واختلاف الليل والنهار والفلك التي تجري في البحر بما ينفع الناس وما أنزل الله من السماء من ماء فأحيا به الأرض بعد موتها وبث فيها من كل دابة وتصريف الرياح والسحاب المسخر بين السماء والأرض لآيات لقوم يعقلون
ذلك بأن الله نزل الكتاب بالحق وإن الذين اختلفوا في الكتاب لفي شقاق بعيد
ولله ما في السماوات وما في الأرض وإلى الله ترجع الأمور
ولله ما في السماوات وما في الأرض يغفر لمن يشاء ويعذب من يشاء والله غفور رحيم
نزل عليك الكتاب بالحق مصدقا لما بين يديه وأنزل التوراة والإنجيل
أفغير دين الله يبغون وله أسلم من في السماوات والأرض طوعا وكرها وإليه يرجعون
بسم الله الرحمن الرحيم الحمد لله الذي خلق السماوات والأرض وجعل الظلمات والنور ثم الذين كفروا بربهم يعدلون

Table 32: A list of verses in cluster 8

Translation of Verses in cluster 8
Creation of the heavens and the earth, alternation of night and day, and sailing of ships across the ocean with what is useful to man, and the rain that God sends from the sky enlivening the earth that was dead, and the scattering of beasts of all kinds upon it, and the changing of the winds, and the clouds which remain obedient between earth and sky, are surely signs for the wise. [2: 164]
That is because God has revealed the Book containing the truth; but those who are at variance about it have gone astray in their contrariness. [2: 176]
For to God belongs all that is in the heavens and the earth, and to God do all things return. [3: 109]
To God belongs all that is in the heavens and the earth: He may pardon whom He please and punish whom He will. Yet God is forgiving and kind. [3: 129]
He has verily revealed to you this Book, in truth and confirmation of the Books revealed before, as indeed He had revealed the Torah and the Gospel [3: 3]
Do they seek another way than God's? But whosoever is in the heavens and the earth is submissive to God and obedient (to Him), by choice or constraint, and will be returned to Him. [3: 83]
ALL PRAISE BE to God who created the heavens and the earth, and ordained darkness and light. Yet the unbelievers make the others equal of their Lord. [6: 1]

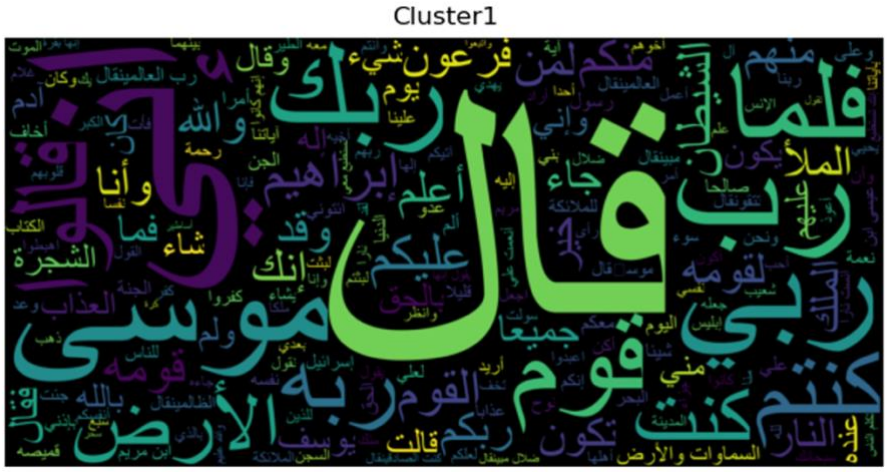


Figure 28: A word cloud visualization of cluster 1

Verses in Cluster 1
وإذ قال إبراهيم رب أرني كيف تحيي الموتى قال أولم تؤمن قال بلى ولكن ليطمئن قلبي قال فخذ أربعة من الطير فصرهن إليك ثم اجعل على كل جبل منهن جزءا ثم ادعهن يأتينك سعيا واعلم أن الله عزيز حكيم
وتلك حجتنا آتيناها إبراهيم على قومه نرفع درجات من نشاء إن ربك حكيم عليم
قالوا ادع لنا ربك يبين لنا ما هي قال إنه يقول إنها بقرة لا فارض ولا بكر عوان بين ذلك فافعلوا ما تؤمرون
قال رب اجعل لي آية قال آيتك ألا تكلم الناس ثلاثة أيام إلا رمزا واذكر ربك كثيرا وسبح بالعشي والإبكار
ألم تر إلى الملإ من بني إسرائيل من بعد موسى إذ قالوا لنبي لهم ابعث لنا ملكا نقاتل في سبيل الله قال هل عسيتم إن كتب عليكم القتال ألا تقاتلوا قالوا وما لنا ألا نقاتل في سبيل الله وقد أخرجنا من ديارنا وأبنائنا فلما كتب عليهم القتال تولوا إلا قليلا منهم والله عليم بالظالمين
قالوا يا موسى إن فيها قوما جبارين وإنا لن ندخلها حتى يخرجوا منها فإن يخرجوا منها فإنا داخلون
وإذ قال موسى لقومه يا قوم اذكروا نعمة الله عليكم إذ جعل فيكم أنبياء وجعلكم ملوكا وآتاكم ما لم يؤت أحدا من العالمين

Table 33: A list of verses in cluster 1

Translation of Verses in cluster 1
Remember, when Abraham said: "O Lord, show me how you raise the dead," He said: "What! Do you not believe?" "I do," answered Abraham. "I only ask for my heart's assurance." (The Lord) said: "Trap four birds and tame them, then put each of them on a (separate) hill, and call them, and they will come flying to you. Know that God is all-powerful and all-wise." [2: 260]
This is the argument We gave to Abraham against his people. We exalt whosoever We please in rank by degrees. Your Lord is wise and all-knowing. [6: 83]
"Call on your Lord for us," they said, "that He might inform us what kind she should be." "Neither old nor young, says God, but of age in between," answered Moses. "So do as you are bid." [2: 68]
And Zachariah said: "Give me a token, O Lord." "The token will be," was the reply, "that you will speak to no man for three days except by signs; and remember your Lord much, and pray at evening and sunrise." [3: 41]
Have you thought of the elders of Israel after Moses, and how they said to their apostle: "Set up a king for us, then we shall fight in the way of God?" He replied: "This too is possible that when commanded to

fight you may not fight at all." They said: "How is it we should not fight in the way of God when we have been driven from our homes and deprived of our Sons?" But when they were ordered to fight they turned away, except for a few; yet God knows the sinners. [2: 246]
They said: "O Moses, in that land live a people who are formidable; we shall never go there until they leave. We shall enter when they go away." [5: 22]
Remember when Moses said to his people: "O my people, remember the favours that God bestowed on you when He appointed apostles from among you, and made you kings and gave you what had never been given to any one in the world. [5: 20]

After studying each cloud, we were able to get a sense of the core themes the model was picking up. By drawing on our domain knowledge and referring to our evaluation resource Qurany, we concluded a list of keywords that we knew were related and relevant to the interpreted topic as in clusters one and eight.

In cluster1, most of the words are names of prophets such us: Abraham and Moses. In addition, we observed a recurrent use of the word "قال " ،which means "said". The word is commonly used in narration accompanying stories and history. This is considered an indication that most of the verses in this cluster are labelled by the topic: Stories and History. In cluster 8, Significant words are: "السموات"/sky, "الأرض"/earth, "خلق"/creation, and "الكتاب"/ the book. Simultaneity in the appearance of these words in one cluster can be contributed to having a common topic. By referring to the Qurany explorer, we determined the associated topic to be Faith. However, in some clusters, by referring to the associated word cloud, we observed that the theme/ topic is not obvious like in cluster 9 shown in Figure 29.

It could be that word cloud failed to capture complex theme as it is not clear form the key words how they are related. Some themes are not captured as single words. For example, the two verses here share the underlined terms from cluster 9, however, they do not describe the same theme. Therefore, word cloud failed to capture the them and produced poor results accordingly.

فَلَمَّا جَاءَهُمُ الْحَقُّ مِنْ عِندِنَا قَالُوا إِنَّ هَٰذَا لَسِحْرٌ مُّبِينٌ

Thus, when the truth had come to them from Us, they said: "Surely this is nothing but pure magic." [10: 76]

وَلَا تَلْبِسُوا الْحَقَّ بِالْبَاطِلِ وَتَكْتُمُوا الْحَقَّ وَأَنتُمْ تَعْلَمُونَ

Do not confuse truth with falsehood, nor conceal the truth knowingly. [2: 42]

Figure 29: An example of a poor cluster: contains key words that are not related to each other

6.6 Discussion

The proposed model created different representations from the same dataset, from different runs, because of random initialization, however, the performance of various representations scored consistently. It is evident in the results as shown in Table27 and Table28 . It is also confirmed when considering different values of vector size. For example, when using the values 50 and 100 to be the size of vector dimensions, it was noticed that there is a slight difference in terms of the model similarity detection rate (refer to Table 23). However, it showed that roughly 92% of the inferred documents were found to be most similar to itself, which gives an indication of how reliable the model is to predict similarity.

On the other hand, we assessed the relation between the verse's pairs qualitatively, resorting to scholarly works and hence one would assume they were accurate and verified (e.g. a similarity/ relation was detected by the model between the verses 39: 47, 13:18 and 5: 36, which occurred in different instances, as illustrated in Table 34). The three verses pointed out the inevitable punishment for those who did not respond to the God, even if they had all that is in the earth

entirely and the like of it. The different instances describe the topic uniquely and distinctively embracing embedded messages and teachings.

39: 47	ولو أن للذين ظلموا ما فى الأرض جميعا ومثله معه لافتدوا به من سوء العذاب يوم القيامة وبدا لهم من الله ما لم يكونوا يحتسبون And if those who did wrong had all that is in the earth entirely and the like of it with it, they would [attempt to] ransom themselves thereby from the worst of the punishment on the Day of Resurrection. And there will appear to them from Allah that which they had not taken into account.	
[13: 18]	للذين استجابوا لربهم الحسنى والذين لم يستجيبوا له لو أن لهم ما في الأرض جميعا ومثله معه لافتدوا به أولئك لهم سوء الحساب ومأواهم جهنم وبئس المهاد For those who have responded to their Lord is the best [reward], but those who did not respond to Him - if they had all that is in the earth entirely and the like of it with it, they would [attempt to] ransom themselves thereby. Those will have the worst account, and their refuge is Hell, and wretched is the resting place.	0.84
[5: 36]	إن الذين كفروا لو أن لهم ما في الأرض جميعا ومثله معه ليفتدوا به من عذاب يوم القيامة ما تقبل منهم ولهم عذاب أليم Indeed, those who disbelieve - if they should have all that is in the earth and the like of it with it by which to ransom themselves from the punishment of the Day of Resurrection, it will not be accepted from them, and for them is a painful punishment.	0.88

Table 34: A list of verses similar/ related to verse 39: 47

We summarize our findings into three main points:

1. The model detected similarity based on words overlapping and syntactic similarity, the verses use same wording to convey a meaning, as underlined.	
ذَٰلِكَ بِأَنَّ اللَّهَ يُولِجُ اللَّيْلَ فِي النَّهَارِ وَيُولِجُ النَّهَارَ فِي اللَّيْلِ وَأَنَّ اللَّهَ سَمِيعٌ بَصِيرٌ **That is because Allah causes the night to enter the day and causes the day to enter the night and because Allah is Hearing and Seeing. [22: 61]**	**يُقَلِّبُ اللَّهُ اللَّيْلَ وَالنَّهَارَ ۚ إِنَّ فِي ذَٰلِكَ لَعِبْرَةً لِّأُولِي الْأَبْصَارِ** **Allah alternates the night and the day. Indeed in that is a lesson for those who have vision. [24: 44]**
قَالُوا مَا هِيَ إِلَّا حَيَاتُنَا الدُّنْيَا نَمُوتُ وَنَحْيَا وَمَا يُهْلِكُنَا إِلَّا الدَّهْرُ ۚ وَمَا لَهُم بِذَٰلِكَ مِنْ عِلْمٍ ۖ إِنْ هُمْ إِلَّا يَظُنُّونَ **And they say, "There is not but our worldly life; we die and live, and nothing destroys**	**إِنْ هِيَ إِلَّا حَيَاتُنَا الدُّنْيَا نَمُوتُ وَنَحْيَا وَمَا نَحْنُ بِمَبْعُوثِينَ** **Life is not but our worldly life - we die and live, but we will not be resurrected. [23: 37]**

us except time." And they have of that no knowledge; they are only assuming. [45: 24]	
كِتَابٌ أُنزِلَ إِلَيْكَ فَلَا يَكُن فِي صَدْرِكَ حَرَجٌ مِّنْهُ لِتُنذِرَ بِهِ وَذِكْرَىٰ لِلْمُؤْمِنِينَ **[This is] a Book revealed to you, [O Muhammad] - so let there not be in your breast distress therefrom - that you may warn thereby and as a reminder to the believers. [7: 2]**	**كِتَابٌ أَنزَلْنَاهُ إِلَيْكَ مُبَارَكٌ لِّيَدَّبَّرُوا آيَاتِهِ وَلِيَتَذَكَّرَ أُولُو الْأَلْبَابِ** **[This is] a blessed Book which We have revealed to you, [O Muhammad], that they might reflect upon its verses and that those of understanding would be reminded. [38: 29]**
2. The model was able to detect a deep relation in some instances, the same meaning is delivered using different terms and distinct wording.	
يَا أَيُّهَا الَّذِينَ آمَنُوا لَا تَتَّخِذُوا بِطَانَةً مِّن دُونِكُمْ لَا يَأْلُونَكُمْ خَبَالًا وَدُّوا مَا عَنِتُّمْ قَدْ بَدَتِ الْبَغْضَاءُ مِنْ أَفْوَاهِهِمْ وَمَا تُخْفِي صُدُورُهُمْ أَكْبَرُ ۚ قَدْ بَيَّنَّا لَكُمُ الْآيَاتِ ۖ إِن كُنتُمْ تَعْقِلُونَ **O you who have believed, do not take as intimates those other than yourselves, for they will not spare you [any] ruin. They wish you would have hardship. Hatred has already appeared from their mouths, and what their breasts conceal is greater. We have certainly made clear to you the signs, if you will use reason. [3: 118]**	**ضَرَبَ لَكُم مَّثَلًا مِّنْ أَنفُسِكُمْ ۖ هَل لَّكُم مِّن مَّا مَلَكَتْ أَيْمَانُكُم مِّن شُرَكَاءَ فِي مَا رَزَقْنَاكُمْ فَأَنتُمْ فِيهِ سَوَاءٌ تَخَافُونَهُمْ كَخِيفَتِكُمْ أَنفُسَكُمْ ۚ كَذَٰلِكَ نُفَصِّلُ الْآيَاتِ لِقَوْمٍ يَعْقِلُونَ** **He presents to you an example from yourselves. Do you have among those whom your right hands possess any partners in what We have provided for you so that you are equal therein [and] would fear them as your fear of one another [within a partnership]? Thus, do We detail the verses for a people who use reason. [30: 28]**
فَأَعْرَضُوا فَأَرْسَلْنَا عَلَيْهِمْ سَيْلَ الْعَرِمِ وَبَدَّلْنَاهُم بِجَنَّتَيْهِمْ جَنَّتَيْنِ ذَوَاتَيْ أُكُلٍ خَمْطٍ وَأَثْلٍ وَشَيْءٍ مِّن سِدْرٍ قَلِيلٍ **But they turned away [refusing], so We sent upon them the flood of the dam, and We replaced their two [fields of] gardens with gardens of bitter fruit, tamarisks and something of sparse lote trees. [34: 16]**	**وَأَمْطَرْنَا عَلَيْهِم مَّطَرًا ۖ فَانظُرْ كَيْفَ كَانَ عَاقِبَةُ الْمُجْرِمِينَ** **And We rained upon them a rain [of stones]. Then see how was the end of the criminals. [7: 84]**
بَدِيعُ السَّمَاوَاتِ وَالْأَرْضِ ۖ أَنَّىٰ يَكُونُ لَهُ وَلَدٌ وَلَمْ تَكُن لَّهُ صَاحِبَةٌ ۖ وَخَلَقَ كُلَّ شَيْءٍ ۖ وَهُوَ بِكُلِّ شَيْءٍ عَلِيمٌ **[He is] Originator of the heavens and the earth. How could He have a son when He does not have a companion and He created all things? And He is, of all things, Knowing. [6: 101]**	**وَلَمْ يَكُن لَّهُ كُفُوًا أَحَدٌ** **Nor is there to Him any equivalent. [112: 4]**

3. The model sometimes made wrong or ambiguous detection; non-related verses are predicted to be related with high similarity score.	
ولا تقف ما ليس لك به علم إن السمع والبصر والفؤاد كل أولئك كان عنه مسئولا **And do not pursue that of which you have no knowledge. Indeed, the hearing, the sight and the heart - about all those [one] will be questioned. [17: 36]**	إن كانت إلا صيحة واحدة فإذا هم خامدون **It was not but one shout, and immediately they were extinguished.** **[36 :29]**
لو يجدون ملجأ أو مغارات أو مدخلا لولوا إليه وهم يجمحون **If they could find a refuge or some caves or any place to enter [and hide], they would turn to it while they run heedlessly. [9: 57]**	ولو أنا كتبنا عليهم أن اقتلوا أنفسكم أو اخرجوا من دياركم ما فعلوه إلا قليل منهم ولو أنهم فعلوا ما يوعظون به لكان خيرا لهم وأشد تثبيتا **And if We had decreed upon them, "Kill yourselves" or "Leave your homes," they would not have done it, except for a few of them. But if they had done what they were instructed, it would have been better for them and a firmer position [for them in faith].[4: 66]**

The above observations provide Islamic scholars and researchers with a rich subject that deserves a thorough investigation. They also suggest that Qur'anic scholars must manually validate the derived dataset. Indeed, studying the relations detected by the computational model can aid the thematic study of the sacred text.

6.7 Summary and Conclusion

This chapter presented a new vector representation of the Qur'anic verses at the paragraph level. These vectors can be used as features and leveraged for the clustering and topic analysis. We then examined the capabilities of paragraph vectors on finding related verses/passages. We were able to locate semantically related verses, and created a dataset of pairs of related verses; as novel outcomes. We used the Qurany ontology browser to verify our results. The Qurany corpus is augmented with an ontology, taken from a recognized expert source, and authenticated by experts with domain knowledge. Next, we fed the features to the clustering algorithm K-Means. The derived clusters suggested groups of related verses that share a common central concept.

Chapter7
Detecting Semantic-based Similarity between Verses of The Qur'an with Doc2vec

7.1 Introduction

One of the most important problems in NLP is document similarity. It has numerous applications in many natural language processing tasks. It can be achieved using lexical similarity or semantic similarity. The semantic similarity task is computationally complex, as identifying relatedness between texts does not depend only on the conventional lexical matching methods, it goes beyond that. Such a task requires in-depth semantic analysis and domain-specific knowledge (Akour et al., 2014; Oahl, 2014; Majumder et al., 2016).

If two documents are semantically similar and describe the same concept, we can call them similar/ related. To determine the similarity between documents we need to define a way to measure the similarity mathematically. We also need to represent text from documents in some form of numeric representation so that we can perform similarity calculations on top of it. Hence, to measure how similar the Qur'anic verses are to each other semantically, we convert the documents/ verses into a mathematical object and define a similarity measure.

Our knowledge of how to represent words and sentences in a way that captures underlying meanings and relationships is rapidly expanding. Distributional semantics is computationally capable of modelling what humans do when they make similarity judgements (Blei et al., 2003; Bengio et al., 2003). More recently, neural network-based sentence representation models have shown promising results in learning sentence embeddings (Wang et al., 2016). In addition, the AI community has provided a range of extremely powerful models that achieve state-of-the-art results in solving challenging NLP tasks. Paragraph Vectors (Le and Mikolov, 2014), or Doc2vec, is one of the most recent developments that is based on distributed representation for texts. Doc2Vec computes a feature vector

for every document in the corpus. The vectors generated by doc2vec can be used for tasks such as finding similarity between sentences/ paragraphs/ documents.

This chapter builds on the experiment conducted in the previous chapter. The chapter examines the use of a natural language processing method for detecting semantic-based similarity between the verses of the Qur'an. We use Doc2vec embeddings and cosine distance for similarity detection. Using Doc2vec, we scored a precision of 79% and accuracy of 76%. We scored higher than the baseline accuracy of 67%. The research question that will be addressed here is:

Is deep learning a viable approach for modelling the complicated features of the Arabic Qur'anic text, and learning subtle semantic relations?

The chapter is organized as follows: Section 2 explains the semantic similarity in the Qur'an; Section 3 provides a survey of previous work in semantic text similarity for the Qur'an and Arabic text; Section 4 describes experimental design for predicting semantic similarity between verses in the original Qur'anic text; Section 5: reports the evaluation result; finally, Section 6 concludes and provides future directions.

7.2 Semantic similarity in the Qur'an

The Qur'an is a significant religious book followed by Muslims, and considered the main resource of Islamic regulations. The text encodes subtle religious meanings that are uncovered by direct and simple analysis (Alqahtani & Atwell, 2014). This property of the Qur'an makes it an excellent text for the purpose of analysing semantic similarity/ relatedness between individual verses, or a group of verses of the Qur'an (Sharaf and Atwell, 2012b). The Qur'anic text contains details and concepts that are scattered all over its passages. The text addresses several concepts in a novel manner, as one concept extends to cover more than one sentence/verse. The same concept also may emerge in different places in the Qur'anic text. Consider the two verses given below:

وَقُرْءَانًا فَرَقْنَٰهُ لِتَقْرَأَهُۥ عَلَى ٱلنَّاسِ عَلَىٰ مُكْثٍ وَنَزَّلْنَٰهُ تَنزِيلًا ﴿١٠٦﴾

English Translation: And [it is] a Qur'an which We have separated [by intervals] that you might recite it to the people over a prolonged period. And We have sent it down progressively.

ٱللَّهُ نَزَّلَ أَحْسَنَ ٱلْحَدِيثِ كِتَٰبًا مُّتَشَٰبِهًا مَّثَانِىَ تَقْشَعِرُّ مِنْهُ جُلُودُ ٱلَّذِينَ يَخْشَوْنَ رَبَّهُمْ ثُمَّ تَلِينُ جُلُودُهُمْ وَقُلُوبُهُمْ إِلَىٰ ذِكْرِ ٱللَّهِ ۚ ذَٰلِكَ هُدَى ٱللَّهِ يَهْدِى بِهِۦ مَن يَشَآءُ ۚ وَمَن يُضْلِلِ ٱللَّهُ فَمَا لَهُۥ مِنْ هَادٍ ﴿٢٣﴾

English Translation: Allah has sent down the best statement: a consistent Book wherein is reiteration. The skins shiver therefrom of those who fear their Lord; then their skins and their hearts relax at the remembrance of Allah. That is the guidance of Allah by which He guides whom He wills. And one whom Allah leaves astray for him there is no guide.

Verses[43] 17:106 and 39:23 are semantically related (Sharaf and Atwell, 2012b); both verses discuss how the Qur’an was revealed. Therefore, studying a subject must consider all related verses on that topic. Understanding the semantic relations between the Qur’anic verses, can facilitate extracting meanings and concepts, and eventually present an insightful knowledge that helps both Muslim and non-Muslim, to understand and appreciate the Qur’an.

[43] We give the Arabic text with English word-by-word translation available at http://corpus.quran.com followed by Sahih International translation available at http://quran.com

7.3 Related Work

7.3.1 A Review on Computational Approaches for Semantic Similarity

Semantic similarity measures have gained attention in recent years due to their significant role in computational linguistics and their use in numerous applications in Natural Language Processing (Majumder et al., 2016). Semantic textual similarity (STS) is used in numerous tasks in natural language processing (NLP), including information retrieval (Kim et al., 2017), text summarization (Mohamed and Oussalah, 2019), text classification (Chen, 2015), machine translation (Zou et al., 2013), question answering (Bordes et al., 2014, Lopez-Gazpio et al., 2017).

In earlier studeis, two text samples were once regarded similar if they shared the same words/characters. To aid in the estimation of semantic similarity, approaches such as Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) were utilised to represent text as real value vectors. These strategies, on the other hand, did not take into account the fact that words have different meanings and that different words might be used to convey the same concept. These methods were straightforward to construct and captured the lexical features of the text; unfortunately, they neglected the semantic and syntactic qualities of the text. Various semantic similarity strategies have been presented during the last three decades to overcome the shortcomings of lexical metrics (Chandrasekaran and Mago, 2021).

The measure of semantic equivalency between two blocks of text is called semantic textual similarity (STS). Instead of a binary choice of similar or not similar, semantic similarity algorithms usually give a ranking or percentage of similarity between texts. Semantic relatedness and semantic similarity are sometimes used interchangeably. Semantic relatedness, on the other hand, not only accounts for semantic similarity between texts, but also examines the shared semantic qualities of two terms from a broader viewpoint. As a result, semantic similarity can be thought of as one facet of semantic relatedness. The semantic

distance, which is inversely proportional to the relationship, is used to assess the semantic relationship, which includes similarity (Hadj Taieb et al., 2020).

7.3.2 Semantic Similarity Approaches

Several surveys reviewed semantic similarity (Camacho-Colladas et al., 2018; Hadj Taieb et al., 2020; Altnel et al., 2018; Lastra-Daz et al., 2019). In a recent survey (Akila & Jayakumar, 2014), the authors presented a detailed review of semantic similarity, classifies and describes various semantic similarity methods and metrics with their advantages and limitations. Chandrasekaran and Mago (2021) provided a comprehensive overview of various semantic similarity algorithms, including the most recent advances based on deep neural networks.

Because of the growing significance of having an STS metric and as a result of the SemEval workshops[44], researchers have developed a range of STS techniques (Ranasinghe et al., 2019). Most of the early algorithms were based on conventional machine learning and mainly relied on feature engineering (Bechara et al., 2015). In task 1 of SemEval 2014, the top system (Zhao et al., 2014) used seven types of features, including text difference measurements, common text similarity measures, and so on. The data was then fed into a variety of learning algorithms, including the support vector machine regressor, Random Forest, Gradient boosting, and others.

Many of the top teams in the STS problem at Semeval 2017 used a neural network design that used word embeddings due to recent advances in word embeddings and the popularity of neural networks in other fields (Shao, 2017). In recent years, the majority of models developed have relied on neural architectures (Wang et al., 2017; Tan et al., 2018; Shao, 2017). Both supervised and unsupervised neural representation models are used.

---

[44] STS has been introduced in SemEval- 2012 (Agirre et al., 2012). Since then, shared tasks have been held annually, as part of the SemEval/*SEM family of workshops, to enable the evaluation of techniques from a diverse set of domains against a shared performance criterion.

Unsupervised techniques use pre-trained word/sentence embeddings without training a neural network model on them to do the similarity task. Cosine similarity has been employed in such models as in sent2vec (Pagliardini et al., 2017), InferSent (Conneau et al., 2017), and Doc2Vec (Le and Mikolov, 2014). Supervised techniques, on the other hand, use neural networks to project word embeddings to fixed dimensional vectors that have been trained to capture the semantic meaning of the sentence. The Kiros et al. (2015) skip-thoughts model and Tai et al. (2015) Tree LSTMS (Long short-term memory) model are two examples of models whose output can be used for sentence similarity tasks.

There is a need for an architecture that can handle two sentences at the same time in order to have an objective function that focuses just on similarity. As a result, the Siamese neural network architecture can help with this issue. In STS tasks, Siamese recurrent neural networks have lately been utilised. Using Manhattan distance as the similarity function between two sub networks, the MALSTM architecture (Mueller & Thyagarajan, 2016) attempted to project zero padded word embeddings of a sentence to fixed sized 50-dimensional vectors. They claimed that their architecture outperforms other neural network models such as Tree-LSTM (Tai et al., 2015).

7.3.3 Transformers and Pre-trained Language Models

Vaswani et al. (2017) released the breakthrough Transformer model, which constituted a significant milestone in the field of NLP. Since then, the NLP community has provided many incredibly efficient components that are publicly available for download and use in various models and pipelines. One of the most recent accomplishments in this development is the release of BERT. BERT (Bidirectional Encoder Representations from Transformers), a pre-trained language model constructed by Delvin et al. (2019), set state-of-the-art records for numerous NLP tasks, including the Semantic Textual Similarity task (STS). As a result, with just one additional output layer, the pre-trained BERT model may be fine-tuned to provide cutting-edge models for a range of tasks.

BERT is based on a number of recent NLP breakthroughs, notably ELMo (Peters et al., 2018) and the OpenAI transformer (Radford et al., 2018). Unlike Word2vec and Glove, ELMo creates a word embedding based on the context in which it's used, capturing both the word's meaning and other contextual information. To produce the embeddings, ELMo employs a bi-directional LSTM that has been trained on a specific task. ELMo was a huge step forward in terms of NLP pre-training. The paradigm has now been applied to other languages, and these pre-trained language models are valuable tools.

7.3.4 Arabic Semantic Similarity Approaches

In this section, we review work research done on semantic text similarity for Arabic text, and in particular for the Qur'an. The Qur'an has recently been regarded as a significant research subject in corpus linguistics, text analysis, and natural language processing (Dukes, 2009; Atwell et al., 2013). In the field of semantic similarity, many efforts study semantic similarity between Arabic texts (Alian & Awajan, 2018; Mahmoud et al., 2015; Schwab, 2017) focusing on Modern Standard Arabic, while others have concentrated on translation of the Qur'an.

One significant work conducted on the Arabic text of the Qur'an is Qursim (Sharaf & Atwell, 2012b). The work presents a broad dataset where semantically similar or related verses are linked together. Qursim is a large corpus of 7,600 pairs of related verses from the Qur'an. They used lexical similarity-based approaches like Term Frequency-Inverse Document Frequency (TF-IDF) to improve their results. Their experiments showed only 869 of the 7,679 pairs shared common words.

Another research published in (Akour et al., 2014) used a lexical similarity-based technique to compute text similarities in the Arabic text of the Qur'an. Using the TF-IDF technique and normalization, it aimed to produce verses from the Qur'an that are identical or relevant to a user's given query verse. It also used an N-gram and a machine learning algorithm to classify chapters (Surahs) as Makki or Madni

(LibSVM classifier in Weka3). Only common main words are used to compute similarity in this study. The lack of semantic-based similarity search is a major flaw in this study. Efforts have also been made on the Qur'an's translated version.

Hamed et al. (2016) established a Question Answering System based on a single chapter from the Qur'an, the Cow Chapter (Surah Baqarah). The authors classified the output to minimize the number of insignificant results returned. Fasting and Pilgrimage verses from Surah Baqarah were classified using neural networks.

Oahl (2014) conducted a thorough investigation into the similarities between sacred texts. This research was conducted on sacred Bible and Qur'an texts. To extract features and compute similarity between the documents, various statistical methods were used. A variety of distance scales were used, including Euclidean, Hillinger, Manhattan, and Cosine. The study looked at overall similarities between two documents based on their topics. It does not go any further in terms of comparing sentences from different texts. In the present work we utilize a natural language processing algorithm to capture semantic-based similarity between all the verses of the Qur'an in the original classical Arabic.

7.4 Experiment: Detecting Semantic-based Similarity

The objective of this experiment is to use Doc2vec method to predict if pairs of verses are related; that is share the same meaning. To measure how similar the Qur'anic verses are to each other semantically, we convert the documents/verses into a mathematical object and define a similarity measure. We build a Doc2vec model and train it on the original Qur'an corpus. We test our model for predicating similarity using test dataset created from the Qursim corpus. The experiment is composed of multiple stages: preparing the data, model training and generating embeddings, computing verses similarity and results. Each of these stages is discussed in the following sub-sections.

7.4.1 Training and Test Data

For the purpose of training and testing our model, we created annotated datasets using existing scholarly resources[45]. The new dataset is a CSV file that contains 9315 pairs of related and nonrelated verses. The dataset contains 3079 pairs of verses that are related, imported from the Qursim dataset. We picked pairs that are related with a strong relationship with a degree of relevance of 2. The dataset also contains 6236 pairs of verses that are nonrelated, randomly generated to be not in Qursim and have a degree of relevance of 0. The file contains nine columns; five of them are imported from the original Qursim dataset. In Qursim, each pair of verses <ss:sv, ts:tv> are related with a degree of relevance 0, 1, or 2. The other four columns are created for the sake of the experiment, and they are named Verse1, Verse2, vid1, and vid2. The Verses text are imported from the Arabic Original Qur'an dataset. We use the Verse text without Diacritics[46] to facilitate the training process. The dataset columns are described in Table 35 along with examples of the data in Table 36.

Column #	Column name	Description
1	ss	Surah(chapter) Id of the first item in the pair
2	sv	Verse Id of the first item in the pair within chapter ss
3	ts	Surah (chapter) Id of the second item in the pair
4	tv	Verse Id of the second item in the pair within chapter ts
5	relevance	Degree of relevance which could be 0 or 1
6	Verse1	Verse 1 text
7	Verse2	Verse 2 text
8	Vid1	Id for each verse in column 7
9	Vid2	Id for each verse in column 8

Table 35: A description of the dataset features

vid1	ss	sv	Verse1	vid2	ts	tv	Verse2	relevance
183	2	2	ذلك الكتاب لا ريب فيه هدى للمتقين	184	10	57	يا أيها الناس قد جاءتكم موعظة من ربكم وشفاء لما في الصدور وهدى ورحمة للمؤمنين	2
309	2	11	وإذا قيل لهم لا تفسدوا في الأرض قالوا إنما نحن مصلحون	310	8	73	والذين كفروا بعضهم أولياء بعض إلا تفعلوه تكن فتنة في الأرض وفساد كبير	2

Table 36: Examples of the test data

[45] All produced dataset will be accessible at Github repository: Mhalshammeri

[46] Diacritics are the symbols used beneath/above Qur'anic verses for reading purposes of the Qur'an.

7.4.2 Generating Vectors using Doc2vec

We need to transform our text documents (verses from the dataset) into a numerical vectorized form, which can later be used to calculate the cosine distance between two different verses to determine how semantically similar they are, or by the clustering algorithm to group similar documents together. We use doc2vec to generate the verses embeddings.

Le and Mikolov (2014) have proposed paragraph vectors, or Doc2vec, an unsupervised method for learning distributed representation for pieces of texts. They show that their method captures many document semantics in dense vectors and can be used for different downstream tasks (Dai, 2015). Doc2vec generates vector representations of variable-length pieces of text, such as sentences, paragraphs, or documents, using a neural network approach. These vector representations have the advantage of capturing the meanings of the input texts, their context. This means that texts with similar meanings or contexts would be closer in vector space than texts with different meanings or contexts. We used Gensim to train a Doc2vec model on our corpus and created vector representations of the Qur'anic verses. We built the model and trained it on the Arabic text of the Qur'an using the original Arabic text from Tanzil project.

7.4.3 Model Training

In order to train a doc2vec model, the training documents need to be in the form "TaggedDocument", which basically means each document/verse receives a unique id. Only the documents that are used for training purposes should be tagged. Before feeding the verses to the model, we need to pre-process them. We separated each verse into different words (tokenization) and formed lists of words for each of them along with the tagging. Because of ambiguity created when applying stemming and removing stop words, the presentation of documents was affected negatively. Therefore, we decided against removing any stop words and stemming. We trained the model and fine-tuned hyper-parameters. We experimented with different models using different settings of hyperparameters to find the optimal values for these parameters.

7.4.4 Comparing Individual Documents using Cosine similarity

To inspect relationships between documents numerically, we calculate the cosine distances between their inferred vectors. Cosine Distance/Similarity is the cosine of the angle between two vectors, which gives us the angular distance between the vectors, one of the most common and effective ways of calculating similarities. Therefore, we developed a function that takes as its parameters the doc2vec model we just trained and the two documents to be compared. As a measure of the documents' similarity, the function then returns a value between 0 and 1, where the larger the value, the more similar the documents. We iterated through each of the verses pair in the dataset and found out what is the cosine similarity for each pair. The complexity of the algorithm is (n^2), where n is the number of pairs in the dataset.

7.5 Results and Evaluation

We evaluated our model by inferring new vectors for unseen documents from the Qur'an. For each verse, we got the most similar verse from the trained model along with similarity score. Figure 30 shows a few examples.

Example 1:
Train Document (410):
«يا أيها الذين آمنوا لا تتخذوا بطانة من دونكم لا يألونكم خبالا ودوا ما عنتم قد بدت البغضاء من أفواههم وما تخفي صدورهم أكبر قد بينا لكم الآيات إن كنتم تعقلون»
O you who have believed, do not take as intimates those other than yourselves, for they will not spare you [any] ruin. They wish you would have hardship. Hatred has already appeared from their mouths, and what their breasts conceal is greater. We have certainly made clear to you the signs, if you will use reason.

Similar Document (3436, 0.8588156700134277):

«ضرب لكم مثلا من أنفسكم هل لكم من ما ملكت أيمانكم من شركاء في ما رزقناكم فأنتم فيه سواء تخافونهم كخيفتكم أنفسكم كذلك نفصل الآيات لقوم يعقلون»
He presents to you an example from yourselves. Do you have among those whom your right hands possess any partners in what We have provided for you so that you are equal therein [and] would fear them as your fear of one another [within a partnership]? Thus do We detail the verses for a people who use reason.

Example2:
Train Document (3621):
«فأعرضوا فأرسلنا عليهم سيل العرم وبدلناهم بجنتيهم جنتين ذواتي أكل خمط وأثل وشيء من سدر قليل»

But they turned away [refusing], so We sent upon them the flood of the dam, and We replaced their two [fields of] gardens with gardens of bitter fruit, tamarisks and something of sparse lote trees.

Similar Document (1037, 0.9169043898582458):

«وأمطرنا عليهم مطرا فانظر كيف كان عاقبة المجرمين»
And We rained upon them a rain [of stones]. Then see how was the end of the criminals.

Figure 30: Examples of similar verses generated by the model

7.5.1 Predicting similarity

We tested the model capability to predict if two verses are related or not based on their cosine similarity. We used the new dataset described in 7.4.1. Our dataset contains 9315 pairs of verses that are either labelled with 1 if related, and 0 if non-related. We computed the cosine similarity for each pair in the dataset by applying the cosine similarity on the associated vectors. Using a threshold of 0.60 (the same threshold throughout the thesis) for the cosine similarity, we consider the pairs with similarity equal or above the threshold to be related (1), otherwise non-related (0). The value (0-1) is the cosine similarity score to determine if a pair of verses are similar/ related or not. We then compared the actual results (1 or 0 per the annotation in the dataset) with the predicated ones (1 or 0 per the similarity score). To evaluate our model's performance, we use the Accuracy[47] metric, the proportion of prediction the model classified correctly. Accuracy can be good to establish some sort of a baseline. In this case, 67%[48] will be our baseline for accuracy. Using Doc2vec, we scored higher than the baseline. Our model scored 76% accuracy, 79% precision, and F1-score of 51%. We produced a confusion matrix report and related statistics with the results as shown in Table 37.

[47] Accuracy = (#True positives (TP)+ # True negatives (TN)) / total number of pairs

[48] We compute the baseline accuracy as: number of actual nonrelated pairs (TN) / total number of pairs; TP =0

	Precision	Recall	F1 Score	Support
0	0.75	0.95	0.84	6236
1	0.79	0.37	0.51	3079
Accuracy			**0.76**	9315

		Actual pairs
Predicted	TP	FP
	5937	299
	TN	FN
	1151	1928

Label	Actual	Predicated
0	6236	7865
1	3079	1450

Table 37: Confusion matrix and classification report on the model performance

The graphs in Figure 31 and Figure 32 below show the distribution of verses based on their similarity. It is evident that Doc2vec computes around 47% of the pairs of verses to be similar (1450), which is around half the real distribution of the similar pairs (3097).

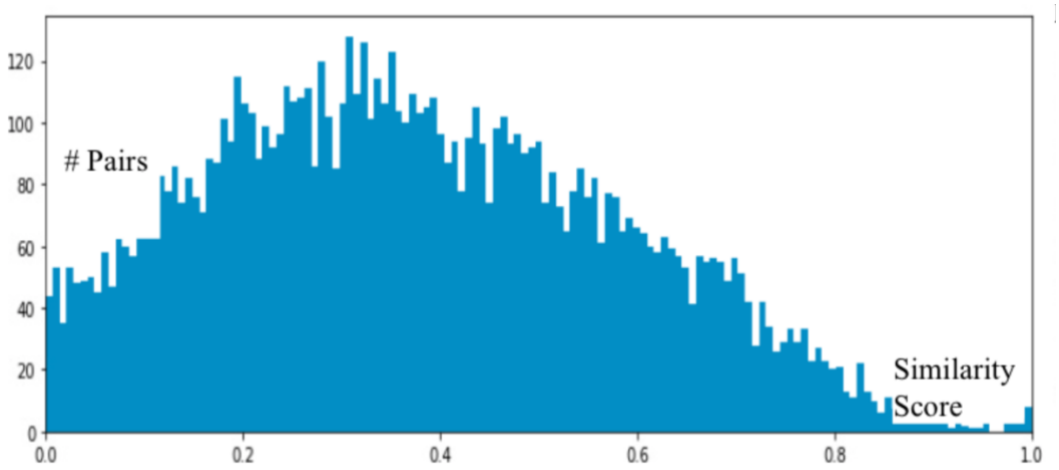


Figure 31: The distribution of verses pairs based on cosine similarity

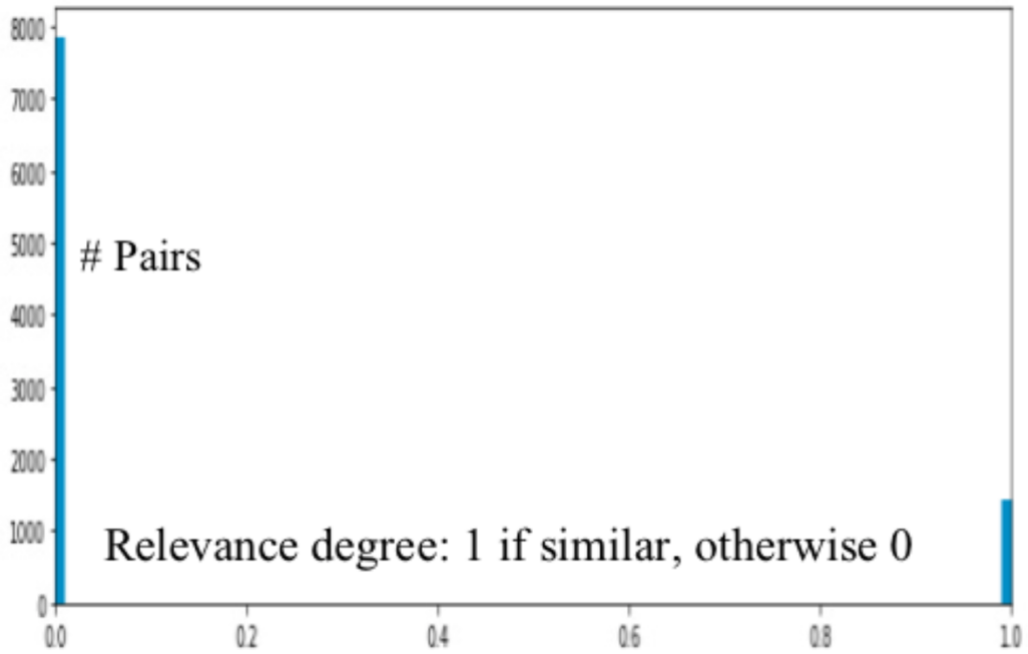


Figure 32: The distribution of verses pairs based on predicted similarity; (0: non-related, 1: related)

7.6 Discussion

In machine learning, each problem is unique in terms of techniques employed and data attributes. Modelling the semantics in the Qur'an is challenging, and detecting pairwise similarity between its passages is a unique task. We have created our dataset, and the task itself is unlikely tested before being given our data. Therefore, the first step was to look for algorithms that best model our prediction problem. Also, we need a basis for the comparison of results.

When trying different algorithms, a baseline result can inform us whether a change is bringing value. Once we have established a baseline, we can add or change the data attributes, the algorithms we are testing, or the settings of the algorithms. Eventually, we know whether we have improved our approach to the problem.

In our case, using state-of-the-art results is not possible, and we need to calculate a baseline result. It is a straightforward prediction that could be a random result or, in some cases, the most prevalent prediction. There are many ways to calculate a baseline result for a prediction problem. For example, for a

classification problem, we can select the class with the most observations and use that class as the result of all predictions. Here, we have a total of 9315 pairs in the test set. There are 3079 related pairs and 6236 non-related pairs. So the baseline would be computed as (6236/ 9315), which is 0.669 (67%).

Our baseline seems poor; however, it could imply that the task is very complex or that our algorithm has a lot of space for improvement. We, therefore, have much room for improvement. For example, we can test different settings of the algorithm, try various features of the data, or even try other algorithms.

7.7 Summary

This chapter presented a natural language processing model that can be used to predict semantic-based similarity between the verses of the Qur'an in the original Arabic text. Using Doc2vec, we mapped the Arabic Qur'anic verses to numerical vectors that encode the semantic properties of the text. We then measured similarity among those vectors. The performance of our model was judged through cosine similarity between assigned semantic similarity scores and annotated textual similarity datasets. The model scored 76% accuracy, and 51% F1-score

Chapter 8
Classifying Verses of The Qur'an using Doc2vec

8.1 Introduction

The richness of the Qur'an and the deep layers of its meaning offer immense potential for further study and experiments. The knowledge in the Qur'an was presented using different approaches, mainly using the tree-structure hierarchy (Ta'a et al., 2014). As a result, determining a concept's true meaning in the Qur'an is difficult. We want to classify the Qur'an verses based on topics or meanings to assist users in identifying the religious knowledge explained in the Qur'an. There has been previous work on classifying textual documents and sentences in English and Arabic (Al-Kabi et al., 2013). However, only a few studies in the literature attempt to classify the verses of the Holy Qur'an (Al-Kabi et al., 2013; Al-Kabi et al., 2005; Ta'a et al., 2014; Akour et al., 2014). Therefore, using NLP combined with ML, this chapter presents an approach to classify the Qur'an based on topics and meanings. To do so, we need to compute the similarity in meaning between its passages.

We focus on sentence /paragraph levels. Therefore, we represent the verses of the Qur'an as vectors of features and compare them by measuring the distance between these features. We use Doc2vec to compute features that capture the semantics of the Qur'anic verses. We then train a logistic regression classifier in a supervised way to learn the underlying meanings and classify the verses of the Qur'an into fifteen classes or categories, based on Qurany corpus (Abbas, 2009). We then use the cosine similarity measure on the vectors to examine how semantically similar the verses are in each class.

We compute two metrics: average similarity and mean similarity difference to inspect the relation between the verses in the same class and other classes. This information indicates how more similar same-category documents are to each other than to documents from different categories. A higher average similarity indicates how similar the documents are in each category. A higher mean

difference implies that the model will identify that document in one class are more distinct from those in other classes. Since we are interested in a topical classification, we use the Qurany corpus to train and evaluate our model. The Qurany corpus is a Qur'anic topic ontology obtained from 'Mushaf Al Tajweed' (Abbas, 2009). It classifies the Qur'anic verses into fifteen main themes.

The rest of the chapter is organized as follows: Section 2 presents studies related to the classification of the verses of the Qur'an. Section 3 describes our approach to classifying the Qur'anic verses and our experiments. Section 4 presents our evaluation and results. Finally, section 5 states our conclusions and future research directions.

8.2 Related Work

This section briefly reviews previous work conducted on the topical classification of holy Qur'an verses. Hamed and Ab Aziz (2018) proposed a Qur'an classification using the Neural Network classifier based on the predefined topics. The study used the English translation of the Qur'an. They applied the classification to Al-Baqarah chapter as it contains many commands and topics. They classified the verses of Al-Baqara into two classes, Fasting, and Pilgrimage.

The thesis of Al-Kabi et al. (2013) is restricted in the topical classification of only two Qur'an chapters: Fatiha (7 verses) and Yaseen (83 verses). Another study (Al-Kabi et al., 2005), evaluated the effectiveness of four well-known classification algorithms: Decision Tree, K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), and Naïve Bayes (NB), to classify Qur'an verses according to their topics. They used the manual topical classification of Qur'anic verses by (Abu Al-Khair and Kabbani, 2013) to train and evaluate the four classifiers. Three selected topics (classes) were used, and 1,227 verses were used in this study out of total 6236 verses in the whole Qur'an. Another classification has been presented by Qurany (Abbas, 2009). This project annotates the verses of the Qur'an with a comprehensive index of nearly 1100 topics, it classifies the Qur'an into fifteen main themes and subdivides the main themes into sub themes. It presents an ontology browser to identify a precise concept and a list of verses that mention this concept.

In this work, we exploit the distributed representation of text to capture the semantic properties of the 6236 Arabic verses of the Qur'an. We classify the 6236 verses of the Qur'an into topical classes using doc2vec and logistic regression. We use the Qurany corpus to train and evaluate our model.

8.3 Experiment

In this experiment, we transform the verses of the Qur'an into a numerical form using doc2vec, which can be used as input to ML methods to examine the semantic similarity between the Qur'anic verses and classify them into topical classes. The objective of this experiment is to evaluate our model in capturing the semantic properties of verses of the Qur'an. Therefore, we examine our model on the following tasks:

1. Classify the verses of the Qur'an into fifteen pre-defined categories or classes using Doc2Vec and Logistic Regression.

2. Measure how similar the verses within the same category are to each other semantically, and use this information to evaluate our model.

8.3.1 The Data

For the purpose of training and testing our model, we create a dataset[49] that contains the 6236 verses of the Qur'an categorized into 15 main topical themes, based on Qurany corpus. Table 38 shows the high-level concepts from the Qurany corpus. Each verse is annotated with a sequence of concepts besides the main concept/theme. The verses are split into training (80%) and test sets (20%). Table 39 shows an example of the dataset.

[49] All produced dataset will be accessible at Github repository: Mhalshammeri

Main Concept	English translation
أركان الإسلام	Pillars of Islam
الإيمان	Faith
القصص والتاريخ	The Stories and The History
القرآن الكريم	The Holy Quran
العمل	Work
الإنسان والعلاقات الأخلاقية	Man, and The Moral Relations
الإنسان والعلاقات الاجتماعية	Man, and The Social Relations
الجهاد	Jihad
العلوم والفنون	Science and Art
الديانات	Religions
تنظيم العلاقات المالية	Organizing Financial Relationships
الدعوة إلى الله	The Call for Allah
العلاقات القضائية	Judicial Relationships
التجارة والزراعة والصناعة والصيد	Trade, Agriculture, Industry and Hunting
العلاقات السياسية والعامة	General and Political Relationships

Table 38: The high-level concepts from the Qurany corpus

Chapter	**Ayah**	**Verse Text**	
2	3	الذين يؤمنون بالغيب ويقيمون الصلاة ومما رزقناهم ينفقون	
Translation		**Class**	**Translation**
Who believe in the Unseen, and establish worship, and spend of that We have bestowed upon them;		أركان الإسلام	Pillars of Islam

Table 39: Example of the verse annotation from the dataset

8.3.2 Classifying the Qur'an Verses using Logistic Regression

We map the verses of the Qur'an to numerical vectors using Doc2vec. We use an ML model based on our vectorized verses to classify the Qur'an verses into the

associated concepts or classes. We set up the train/ test documents, pre-process them to be ready for training and classification. We train the Doc2vec model using 80% of the data set and distributed bag of words architecture. We generate the vectors for the classifier. We infer vectors for the documents in the test set using the trained model. Then we train the logistic regression classifier.

8.3.3 Testing Category-Wise & Cross-Category Verses Similarity

Documents belonging to the same category would seem to be more similar than documents belonging to different categories, intuitively. And that's how we judge our model: a good model should generate higher similarity values for verses in the same category than for verses from different categories.

8.4 Results and Evaluation

8.4.1 Classification Results

We trained a Doc2vec model on the dataset derived from Qurany to generate the verses embeddings. We tried different configurations for the hyperparameters of the model. We then trained the classifier with different versions of derived embeddings. Using 80% of the data set to train the classifier, we achieved 68% accuracy, and 56% F1-score, using the distributed bag of words architecture. We have noticed that changing the vector size did not have a big effect on the classifier performance. Table 40 shows the classifier performance results using different settings of the model: Distributed Bag-of-words (PV-DBOW) and Distributed Memory (PV-DM).

Train set: 80% / Test set: 20%.			
Doc2vec model	Vector size	Testing Accuracy	Testing F1score
PV-DBOW	50	0.68	0.56
	100	0.68	0.55
PV-DM	50	0.63	0.56
	100	0.63	0.57

Table 40: Classification Performance Results

8.4.2 Categories Similarity Results

To inspect relationships between the verses numerically, we calculated the cosine distances between their inferred vectors using the trained Doc2vec model. We used this information to calculate the average similarity scores and the mean difference for each category. We wanted to know how much more similar the same-category documents are to each other than to documents from other categories. Therefore, we created sets of verses pairs for all categories; which we denoted by C_1, . . ., C_{15}, where each category is a set of verses. Hence, we derived 15 average similarities per each category; one for same-category documents and 14 for cross-category documents. Finally, we calculated the mean similarity differences between the cross-category average similarities and the same-category average similarity.

A higher mean difference implies that the model is able to identify that document in one category are more distinct from those in other categories. We show the results using the two architectures of Doc2vec model, and vector size of 50. We didn't include the results using vectors size of 100 as no difference were noticed. The result of the evaluation can be summarized as in Table 41.

Arabic Category	English Category	(Mean Difference, Same-category Avg. Similarity)
أركان الإسلام	Pillars of Islam	(11%, 15%)
الإيمان	Faith	(21%, 26%)
الإنسان والعلاقات الأخلاقية	Human and Ethical Relationships	(2%, 0.74%)
القصص والتاريخ	Stories and History	(5.4%, 5.6%)
القرآن الكريم	The Holy Quran	(2.8%, 3.7%)
العمل	The Work	(2.9%, 6.6%)
الإنسان والعلاقات الاجتماعية	Human and Social Relationships	(3.8%, 7.8%)

الجهاد	Al-Jihad	(0.92%, 0.16%)
العلوم والفنون	Sciences and Arts	(7.2%, 5.3%)
الديانات	Religions	(13%, 19%)
الدعوة إلى الله	Call to Allah	(3.2%, 6.5%)
التجارة والزراعة والصناعة والصيد	Trade, Agriculture, Industry, and Hunting	(8.6%, 8.7%)
العلاقات القضائية	Judaical Relationships	(4.5%, 8.2%)
تنظيم العلاقات المالية		(2.6%, 4.9%)
العلاقات السياسية والعامة	General and Political Relationships	(1.5%, 2.7%)

Table 41: Evaluation Results using PV-DBOW, Vector-size=50

8.5 Discussion

The three top classes that achieved higher average same-category similarity and mean-difference are Faith, Pillars of Islam, and Religions. The three classes scored higher values for both metrics with different runs. The two metrics are not relatively high for some classes. It is observed that to some classes' documents being similar to those of another class. Besides, some verses in the Qur'an discuss more than one concept/topic. The uniqueness and complexity of the Qur'an language also could be a significant reason. Table 42 shows an example of instances where a verse belongs to different classes/ topics.

Verse	**Al-Nahl / سورة النحل (16, 94)**
Arabic Verse	**"وَلاَ تَتَّخِذُواْ أَيْمَانَكُمْ دَخَلاً بَيْنَكُمْ فَتَزِلَّ قَدَمٌ بَعْدَ ثُبُوتِهَا وَتَذُوقُواْ السُّوءَ بِمَا صَدَدتُّمْ عَن سَبِيلِ اللهِ وَلَكُمْ عَذَابٌ عَظِيمٌ"**
English Translation	**And make not your oaths a means of deceit between you, lest a foot should slip after its stability, and you should taste evil because you hinder (men) from Allah's way and grievous chastisement be your (lot).**
Topic	**Judicial Relationships and Jihad**

Table 42: An example of an instance where a verse belongs to different classes/ topics

More examples are shown on Table 43. The high score here could be attributed to the uniqueness and distinctiveness of words used to describe that subject. The results confirmed that the class “Faith” has achieved the highest average similarity and mean difference.

Verse	**Al- Ma'ida / سورة المائدة** **(5, 96)**
Arabic Verse	**أُحِلَّ لَكُمْ صَيْدُ الْبَحْرِ وَطَعَامُهُ مَتَاعًا لَّكُمْ وَلِلسَّيَّارَةِ ۖ وَحُرِّمَ عَلَيْكُمْ صَيْدُ الْبَرِّ مَا دُمْتُمْ حُرُمًا ۗ وَاتَّقُوا اللَّهَ الَّذِي إِلَيْهِ تُحْشَرُونَ**
English Translation	**Lawful to you is game from the sea and its food as provision for you and the travellers, but forbidden to you is game from the land as long as you are in the state of ihram. And fear Allah to whom you will be gathered.**
Topic	**- Judicial Relationships** **- Trade, Agriculture, Industry and Hunting**
Verse	**An-Nisa/ سورة النساء** **(4, 31)**
Arabic Verse	**إِن تَجْتَنِبُوا كَبَائِرَ مَا تُنْهَوْنَ عَنْهُ نُكَفِّرْ عَنكُمْ سَيِّئَاتِكُمْ وَنُدْخِلْكُم مُّدْخَلًا كَرِيمًا**
English Translation	**If you avoid the major sins which you are forbidden, We will remove from you your lesser sins and admit you to a noble entrance [into Paradise].**
Topic	**- Judicial Relationships** **- Work**

Table 43: More examples of instances where a verse belongs to different classes

8.6 Summary and Conclusion

This chapter used NLP combined with ML to classify the verses of the Qur'an into fifteen predefined classes. The semantics of the verses were captured using Doc2vec embeddings that were used to group similar documents. The model achieved a classification accuracy of 70% and an F1 score of 60%. The results confirmed that the classifier scored higher accuracy results with the distributed bags of words architecture of the Doc2vec model. Next, we evaluated the model by examining the semantic similarity of the Qur'anic verses. We calculated the mean difference and average similarity values for each category to indicate how well the model describes that category. Derived classes showed high average similarity using distributed bags of words architecture for some classes.

This chapter concluded experiments conducted using Doc2vec, a recent breakthrough in document embeddings techniques. Doc2vec was used to map the Qur'anic verses to vectorized form. The document vector approach proved to be more useful in order to retrieve the semantically close verses for a given verse. In addition, the results open many potentials for improvements on the level of pre-processing the text and the training data. The derived datasets can act as a base for future research relevant to the Qur'anic semantic analysis. Moreover, the verses' embeddings provided a rich representation for the semantic similarity (chapter 7) and the topical classification (chapter 8). Detecting semantic similarity in the Qur'an is an emerging research area and the derived results are significant towards achieving the ultimate goal.

Chapter 9
Deep Learning of Semantic Similarity in the Qur'an

9.1 Introduction

This chapter exploits trendy concepts in NLP and deep learning such as transformers and neural language models to learn deep semantic similarities between Qur'an verses. The chapter presents a novel approach for modelling the complex features of the Arabic Qur'an verses and detecting the embedded semantic relations.

This chapter presents a novel Siamese transformer-based architecture for Qur'an verse similarity. It is the most novel contribution of this dissertation. The architecture leverages both the pre-trained contextualized representations for the Arabic language and the Siamese architecture to derive semantically meaningful verse embeddings and achieve remarkable results in pairwise semantic similarity detection in the Qur'an. The F1 score 95% on the Qur'anic semantic similarity test was impressively high

BERT has achieved state-of-the-art results on a broad range of NLP tasks (Devlin et al., 2019). However, it is not suitable for various pair regression tasks due to massive computations. To address this problem, SBERT adopts a Siamese network structure that allows for generating fixed-sized vectors for sentences pair with reduced computational overhead. As a result, SBERT has proven to outperform other state-of-the-arts sentence embeddings methods on sentence-pair regression tasks like semantic textual similarity while maintaining the accuracy of BERT (Reimers and Gurevych, 2019).

The research question that will be answered here is:

Is deep learning a viable approach for modelling the complicated features of the Arabic Qur'anic text, and learning subtle semantic relations?

Hence, to answer this question, we adopt a Siamese sentence-transformer networks structure to yield useful sentence embeddings that can be compared using a similarity measure. We use a pre-trained BERT model that supports the Arabic Language as a feature extractor and fine-tunes it on semantic similarity datasets drawn from the Qur'an.

We create training and test datasets drawn from the Qur'an and exiting knowledge resources such as Qursim for training and testing the derived model. Finally, we evaluate the performance of our model for the semantic textual similarity task.

The chapter is organized as follows: Section 2 provides a review on semantic similarity approaches, and a background on Transformers and Pretrained language models and their application to the Arabic. Section 3 contains information about datasets used. Section 4 presents the novel Siamese transformer-based architecture for Qur'an verse similarity. Section 5 describe the experimental setup used to train our models. Section 6 presents our results. The architecture is evaluated in Section 7. The chapter finishes with conclusions and future directions.

9.2 Related Work

Many approaches were proposed to solve semantic similarity problem, ranging from conventional approaches that were mainly based on representing text as a vector of word features, to deep learning models that learn complicated relationships among texts (Chandrasekaran and Mago, 2021; Hadj Taieb et al., 2020; Zhao et al., 2014; Wu et al., 2017; Feng et al., 2017; Wang et al., 2017; Tan et al., 2018).

Word2vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014) were early pretrained text representation models that intended to represent words by capturing their distributed syntactic and semantic aspects. These models, on the other hand, did not include the context in which a word appears in its embedding. This issue was solved by employing models like ELMO to generate contextualised representations (Peters et al., 2018). Recently, there has been an emphasis on using transfer learning to fine-tune large pretrained language models for downstream NLP/NLU tasks using a limited number of instances,

which has resulted in significant performance improvements for these tasks (Radford et al., 2018; Devlin et al., 2018).

9.2.1 Pre-training General Language Representations

9.2.1.1 Unsupervised Feature-based Approaches

For decades, non-neural (Brown et al., 1992; Ando and Zhang, 2005; Blitzer et al., 2006) and neural (Mikolov et al., 2013; Pennington et al., 2014) methods have been used to learn broadly applicable representations of words. Modern NLP systems rely on pre-trained word embeddings, which provide considerable advantages over embeddings learned from scratch (Turian et al., 2010). Left-to-right language modelling objectives (Mnih and Hinton, 2009), as well as objectives to distinguish correct from incorrect words in the left and right contexts (Mikolov et al., 2013), were utilised to pre-train word embedding vectors.

These approaches have been extended to coarser granularities, such as sentence embeddings (Kiros et al., 2015; Logeswaran and Lee, 2018) or paragraph embeddings (Le and Mikolov, 2014). Previous work has used objective-based ranking of candidate next sentences (Jernite et al., 2017; Logeswaran and Lee, 2018), left-to-right generation of next sentence words given a representation of the previous sentence (Kiros et al., 2015), or auto-encoder derived denoising (Hill et al., 2016) objectives to train sentence representations.

Traditional word embedding research is generalised in a different dimension by ELMo and its ancestor (Peters et al., 2017, 2018a). They use a left-to-right and a right-to-left language model to extract context-sensitive characteristics. Each token's contextual representation is the sum of its left-to-right and right-to-left representations. ELMo improves the state of the art for numerous major NLP benchmarks (Peters et al., 2018a), including question answering (Rajpurkar et al., 2016), sentiment analysis (Socher et al., 2013), and named entity recognition (Tjong Kim Sang and De Meulder, 2003). Moreover, Melamud et al. (2016) proposed utilising LSTMs to learn contextual representations by performing a

task that required them to predict a single word from both left and right context. Their paradigm, like ELMo's, is feature-based rather than deeply bidirectional.

9.2.1.2 Unsupervised Fine-tuning Approaches

The early attempts in this technique, like the feature-based algorithms, used solely pre-trained word embedding parameters from unlabelled text (Collobert and Weston, 2008). More recently, sentence or document encoders that produce contextual token representations have been pre-trained from un-labelled text and fine-tuned for a supervised downstream task (Dai and Le, 2015; Howard and Ruder, 2018; Radford et al., 2018). For pre-training such models, left-to-right language modelling and auto-encoder objectives have been utilised (Howard and Ruder, 2018; Radford et al., 2018; Dai and Le, 2015). The benefit of these approaches is that only a few parameters must be learned from scratch.

Radford et al. presented a generative transformer-based-based language model trained on a diverse corpus of unlabelled text, followed by discriminative fine-tuning on diverse language tasks. They evaluated their approach on four types of language understanding tasks – natural language inference, question answering, semantic similarity, and text classification. The model obtained state-of-the-art results on a suite of language tasks such as commonness reasoning (Stories Cloze Test) (Mostafazadeh et al., 2017), question answering (RACE) (Lai et al., 2017), textual entailment (MultiNLI) (Williams et al., 2017) and the recently introduced GLUE multi-task benchmark (Wang et al., 2018). GPT was followed by a serious of impressive developments such as GPT-2 (Ziegler et al., 2019), GPT-3 (Brown et al. 2020), CLIP (Radford et al., 2021) and, recently, Whisper (Radford et al., 2022).

9.2.2 Bidirectional pre-training for language representations (BERT)

Pre-training language models have recently demonstrated a crucial impact in improving several NLP tasks. These pre-trained language representations can be applied to NLP tasks in two ways: feature-based or fine-tuning (Devlin et al., 2018).

Researchers who utilise the feature-based approach (Peters et al., 2018) employ the output of a pre-trained model as additional features in their models,

depending on the task they are attempting to solve. The fine-tuning approach (Radford et al., 2018) on the other hand, allows the model to be trained on a different task by learning task-specific parameters. Because they use left-to-right unidirectional architectures, the two techniques outlined before have limits when it comes to learning generic language representations. Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) has, on the other hand, surpassed earlier cutting-edge unidirectional models significantly.

The multi-head self-attention mechanism used by the BERT model allows it to achieve state-of-the-art accuracy on a wide range of tasks, including natural language inference, question answering, and sentence classification. The BERT model's architecture is based on the self-attention layer, which is a transformer layer. In contrast to standard unidirectional models, the representations of words are transferred from previous levels regardless of their placements for each layer. The model learns bidirectional encoder representations for each input word using the masked language model, which randomly masks some of the words from the input to contextually predict the masked word (Devlin et al., 2018).

9.2.3 The Siamese Architecture

Siamese neural networks are double networks with tied weights, in which the weights of the two sub-networks must be the same, an objective function. The purpose of training is to learn how to represent text pairs in a highly structured space that reflects complex semantic relationships (Chopra et al., 2005; Bromley et al., 1993).

Siamese neural networks are widely used in tasks that require determining similarity or a link between two similar objects (Ge, 2018). Some examples include paraphrasing, in which two sentences are entered and the output is a score indicating how similar they are; and signature verification (Bromley et al. 1993), in which two signatures are compared to see if they are from the same person.

Siamese recurrent neural networks have been recently used in STS tasks (Mueller and Thyagarajan, 2016; Pang et al., 2016; Severyn and Moschitti, 2015; Wang et al., 2017; Neculoiu et al., 2016; Ranasinghe et al., 2019). The Siamese architecture is a common framework in which the encoder, which can be either

Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN), is applied individually on the two input texts, resulting in intermediate contextual representations for both texts. Mueller and Thyagarajan (2016) presented a Siamese adaptation of the Long Short-Term Memory (LSTM) network for labelled data comprised of pairs of variable-length sequences. Their model outperformed other neural network models like Tree-LSTM (Tai et al., 2015).

9.2.3.1 SBERT

Despite the huge successes of Transformer-based language models, BERT's design makes it unsuitable for unsupervised tasks such as clustering and semantic similarity search. BERT uses a cross-encoder; it represents a single sentence or a pair of sentences in one token sequence. It is disadvantageous that no independent sentence embeddings are computed which makes it difficult to derive sentence embedding from BERT. Researchers from the UKP Lab released S-BERT (Sentence-BERT), which modifies the pre-trained BERT network to use Siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be computed using cosine similarity (Reimers & Gurevych, 2019). S-BERT uses pre-trained BERT and RoBERTa networks and then fine-tunes them to produce fixed-sized sentence embeddings (see Figure 33). This inspired us to use the Siamese architecture to detect semantic similarity in the Qur’an. This chapter, therefore, proposes a Siamese transformer networks architecture for pairwise semantic similarity detection in the Qur’an. We start with pre-trained BERT models, and a Siamese set-up is used to fine-tune the models. We use AraBERT transformer that make up the base of out model, to which a pooling layer has been appended.

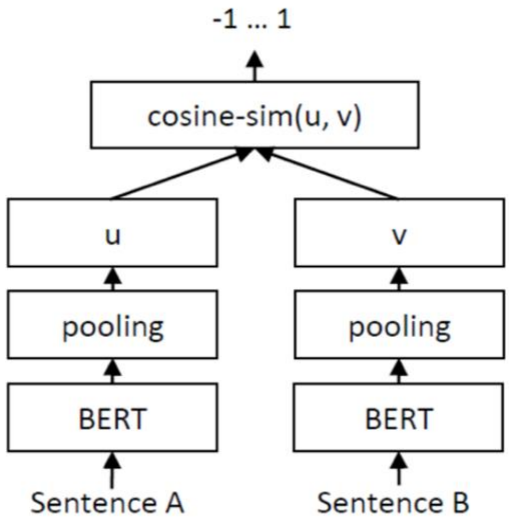


Figure 33: SBERT Siamese network architecture, with regression objective function, for fine-tuning on STS dataset (Reimers & Gurevych, 2019)

9.2.4 Semantic Similarity for the Arabic language and the Qur'an

Arabic is considered to be low-resourced language, has many dialects, and rich in morphology. Therefore, identifying semantic similarity in Arabic text is not a trivial task. The Qur'an, as a significant religious text, uses classical Arabic to its most potential. Therefore, to quantify semantic similarity between its passages, computational models need to incorporate deep semantic analysis and external domain knowledge.

Researchers have proposed numerous models to determine textual semantic similarity. However, few related works were proposed for Arabic (Alian and Awajan, 2018). Some efforts studied the semantic similarity in Arabic (Mohamed et al., 2015; Mahmoud & Zrigui, 2017; Schwab, 2017; Al-Bataineh et al., 2019), and in the Qur'an (El-Deeb et al., 2018; Alshammeri et al., 2021; Alsaleh et al., 2021). However, there is a lack of deep learning studies on the topic of STS in the Qur'an. Bashir et al. (2021) presented a thorough examination of Qur'anic Arabic NLP approaches, tools, and applications. They also outlined open research challenges and promising future research possibilities. Their survey can act as a useful reference for researchers and practitioners in the field. NLP in the Qur'an is a growing field of study. However, when compared to Arabic NLP, Qur'anic NLP research is still immature and has potential for investigation and

experimentation (Bashir et al. 2021). Our approach combines the power of transformers with the Siamese architecture for achieving impressive results in semantic similarity detection.

9.3 Dataset Description

We use a popular benchmark resource that provides pairs of similar verses from the Qur'an. QurSim (Sharaf and Atwell, 2012b) is considered a valuable resource of related pairs of the Qur'an in which semantically related pairs of verses are linked together. It is regarded as a gold standard resource in analysing relatedness in short texts. Qursim contains 7679 pairs that are related with a degree of relevance[50] 0, 1, or 2. We create a dataset[51] that contains pairs of verses from the Qur'an with binary labels for semantic similarity/ relatedness: (i) We import the related pairs from Qursim; we tried different combination of pairs given their degree of relevance when training our model. (ii)We map the verses' pairs to their text using the Tanzil[52] project. (iii) The dataset undergoes some cleaning to eliminate a total of 372 duplicate records. Examples of such pairs are shown in Table 44. (iiii) We create non-related pairs, randomly generated to be not in Qursim. And (v) We split the data into training and test sets.

No.	Verse1	No.	Verse2
30: 4	فِي بِضْعِ سِنِينَ ۗ لِلَّهِ الْأَمْرُ مِن قَبْلُ وَمِن بَعْدُ ۚ وَيَوْمَئِذٍ يَفْرَحُ الْمُؤْمِنُونَ	30: 4	فِي بِضْعِ سِنِينَ ۗ لِلَّهِ الْأَمْرُ مِن قَبْلُ وَمِن بَعْدُ ۚ وَيَوْمَئِذٍ يَفْرَحُ الْمُؤْمِنُونَ
1: 5	إِيَّاكَ نَعْبُدُ وَإِيَّاكَ نَسْتَعِينُ	73: 9	رَّبُّ الْمَشْرِقِ وَالْمَغْرِبِ لَا إِلَٰهَ إِلَّا هُوَ فَاتَّخِذْهُ وَكِيلًا
73: 9	رَّبُّ الْمَشْرِقِ وَالْمَغْرِبِ لَا إِلَٰهَ إِلَّا هُوَ فَاتَّخِذْهُ وَكِيلًا	1: 5	إِيَّاكَ نَعْبُدُ وَإِيَّاكَ نَسْتَعِينُ

Table 44: Examples of duplicated records from Qursim; we kept one of the duplicated pairs as in < 1:5,73:9>, <37:9, 1:5>, and we removed pairs like <30:4, 4: 30> where the verse is related to itself.

[50] In Qursim, pairs are assigned two levels of degree of relatedness: level 2 represents strong relation, and level 1 represents weaker relation.

[51] All produced dataset will be accessible at Githup repository: Mhalshammeri

[52] https://tanzil.net/docs/

An example for the dataset is provided in Table 45. The pairs have a strong relationship as identified by Qursim. Finally, we construct the task of predicting the semantic similarity between two verses in a binary classification task.

Location (Chapter: Verse)	Verse1	Verse2	Relevance
3:142, 29: 2	أَمْ حَسِبْتُمْ أَن تَدْخُلُوا الْجَنَّةَ وَلَمَّا يَعْلَمِ اللَّهُ الَّذِينَ جَاهَدُوا مِنكُمْ وَيَعْلَمَ الصَّابِرِينَ Or do you think that you will enter Paradise while Allah has not yet made evident those of you who fight in His cause and made evident those who are steadfast?	أَحَسِبَ النَّاسُ أَن يُتْرَكُوا أَن يَقُولُوا آمَنَّا وَهُمْ لَا يُفْتَنُونَ Do the people think that they will be left to say, "We believe" and they will not be tried?	2
19: 63, 23: 11	تِلْكَ الْجَنَّةُ الَّتِي نُورِثُ مِنْ عِبَادِنَا مَن كَانَ تَقِيًّا That is Paradise, which We give as inheritance to those of Our servants who were fearing of Allah.	الَّذِينَ يَرِثُونَ الْفِرْدَوْسَ هُمْ فِيهَا خَالِدُونَ Who will inherit Al-Firdaus. They will abide therein eternally.	2

Table 45: Examples of related pairs from Qursim

9.4 A Siamese Transformer-based Architecture for Detecting Semantic Similarity in the Qur'an

This section describes the proposed Siamese transformer-based networks architecture. It is composed of two Siamese transformer networks, each process one of the verses in a given pair. Our model incorporates AraBERT Arabic pre-trained contextualized representation to derive semantically meaningful sentence embeddings and achieve state-of-the-art binary semantic similarity classification results. Furthermore, our model benefits from the Siamese network architecture, like in SBERT to fine-tune the pre-trained model with less computational burden characterizing sentence-pair regression tasks.

We experiment with a transformer-based model that was pre-trained for the Arabic Language. The model is AraBERT, where Antoun et al. (2020) pre-trained the BERT transformer model (Devlin et al., 2019) for the Arabic Language. So,

we apply the sentence pairs classification task on Arabic Qur'an verses by fine-tuning the non-segmented AraBERT model. Our architecture is depicted in Figure 34.

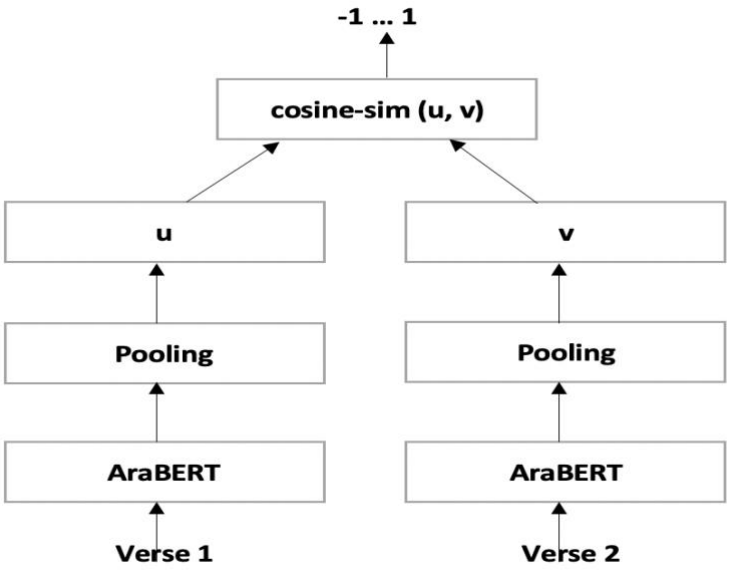


Figure 34: A Siamese transformer networks architecture, with regression objective function, for fine-tuning on the Qur'anic semantic similarity dataset

Each transformer network is composed of a transformer layer and a pooling layer, works as the following:

1. Layer 1: The input text (Verse1) is passed through a pre-trained Transformer model, AraBERT. The Transformer outputs are contextualized word embeddings for all input tokens. The same process runs in the other network using input text Verse2.
2. Layer 2: The embeddings go through a pooling layer to get a single fixed-length embedding for all the text. Here, mean pooling averages the embeddings generated by the model to generate the verse embeddings.

The structure of dataset suggests which loss function to use. The dataset is a list of pairs, each is a pair of verses and a label indicating how similar they are. Therefore, we finetune the network on our training data using the mean-squared error loss[53] as the regression objective function, where the cosine similarity between two sentence embeddings, Verse1 and Verse2, is calculated. We

[53] Most regression techniques are evaluated using the Mean Squared Error as the standard measure. The error is going down as our algorithm acquires more and more experience.

optimize to get verses embeddings as close as possible to similar verses, and far away from each other for dissimilar verses.

9.5 Experiment

We fine-tune sentence-level AraBERT model on an STS dataset. An objective regression function with mean-squared error loss is used. We use the sentence-transformers Python library from the UKP Lab[54]. We use version two of a non-segmented AraBERT[55] model "bert-base-arabertv02". While the segmented version of the model uses an external tokenizer (FARASA Segmenter), we use the model's internal tokenizer for the segmentation process. We train our model using our semantic similarity dataset consisting of verses pairs labeled as 1 or 0 for the binary semantic similarity between the two verses.

We run our experiment using a Google Colab notebook to benefit from the free-tier GPU instance to speed up the training. First, we load the training dataset. Then we split the training set for validation data during training while keeping the test set to evaluate the final model. Finally, we experiment with the transformer, bert-base-arabertv02. We fine-tune the model with an objective regression function for eight epochs. We use a batch size of 16, Adam optimizer with learning rate 2e−5 as in the SBERT paper, 10% of the training data is used for warm-up, and evaluation is set to happen every 1000 steps. Our default pooling strategy is MEAN as in the SBERT paper. The last step is to evaluate the model on the STS test set.

The Finetuning process is summarized as the following:

1. Preparing the dataset for training the transformers model: The dataset is a list of pairs; each is a pair of verses and a label indicating how similar they are.
2. Converting the dataset into a format the Sentence Transformers model can understand.

---

[54] https://github.com/UKPLab/sentence-transformers

[55] There are two versions of the model, AraBERTv0.1 and AraBERTv1, with the difference being that AraBERTv1 uses pre-segmented text where prefixes and suffixes were splitted using the Farasa Segmenter. AraBERTv2 is a recent non-segmented version of AraBERT.

3. Finetuning the model on the training dataset using the loss function: The loss function 'CosineSimilarityLoss' optimizes such that the sentences with the closest labels are near in the vector space, and the sentences with the farthest labels are as far as possible.
4. Evaluating the model's performance using the test set. We compute both the Pearson correlation and the Spearman's rank correlation between the cosine-similarity of the verse embeddings and the labels in the annotated dataset. The similarity is computed by cosine- similarity.

9.6 Results

We evaluate our model on the semantic similarly task. The evaluation results in Table 46 below are for cosine similarity, Manhattan distance, Euclidean distance, and dot-product similarity, as measured by the Pearson correlation and Spearman correlation metrics. According to Reimer et al. (2016), unlike Pearson correlation, the Spearman correlation metric is more suited for evaluating STS tasks. We, therefore, consider the Spearman correlation as representative of the ability to accurately determine whether two verses are similar. Our model achieves a score of 84.96% with AraBERT using the pairs with relevance degree 1 and 2 from Qursim corpus as the similar pairs. Table 47 shows the Spearman correlation for cosine similarity using the different combinations of relevance degree for similar pairs for our model.

Cosine-Similarity:	Pearson: 0.9163	Spearman: 0.8496
Manhattan-Distance:	Pearson: 0.8999	Spearman: 0.8477
Euclidean-Distance:	Pearson: 0.8955	Spearman: 0.8471
Dot-Product-Similarity:	Pearson: 0.9097	Spearman: 0.8495

Table 46: Evaluation of holdout test data on a Qur'anic Semantic Similarity dataset

Model	# Epochs	Relevance degree for similar pairs	Spearman correlation Cosine-Similarity
Siamese AraBERT	**4**	**2**	**79.01%**
		1 + 2	**82.32%**
		0 + 1 + 2	**80.71%**
	8	**2**	**78.32%**
		1 + 2	**84.96%**
		0 + 1 + 2	**83.35%**

Table 47: Spearman correlation scores using different settings of our model (Epochs) and different versions of data

The results confirm that a Siamese AraBERT networks architecture has the capability to accurately evaluate whether two verses are similar which is represented by the Spearman correlation. Indeed, AraBERT networks shows high performance on detecting semantic similarity in the Qur'an.

9.7 Evaluation

We evaluate our model for predicting similarity using the Qur'anic semantic similarity dataset. The evaluation metric we use for this task is accuracy and F1-Score. We compute the verses embeddings for each pair using the Siamese AraBERT networks architecture to be used to calculate the cosine distance and determine how semantically similar they are. Using a threshold of 0.60[56] for the cosine similarity, we consider the pairs with similarity equal or above the threshold to be related (1), otherwise non-related (0). We then compare the actual results (the annotation in the dataset) with the predicted ones (using our model). Our model scored 95% accuracy, and F1-score of 95%. Table 48 shows the confusion matrix and classification report.

Classification Report	Precision	Recall	F1 Score	Support

[56] Using high threshold can generate incorrect results as relatedness is based on meanings, not only lexical matching. Also, to maintain consistency as the same threshold is used in previous experiment using Dco2vec.

0	0.92	0.98	0.95	6236
1	0.98	0.92	0.95	6436
Accuracy			0.95	12672

Confusion Matrix	Actually positive (1)	Actually negative (0)
Predicted positive (1)	TP: 6114	FP: 122
Predicted negative (0)	FN: 625	TN: 5910

Table 48: Classification report and confusion matrix

We further report performance of our model (the AraBERT Siamese architecture) on the same dataset, compared against systems of earlier work based on accuracy and F1-score as shown in Table 49.

Metric	**Siamese AraBERT**	**AraBERT**[57]	**Doc2vec**[58]
Accuracy	**95%**	**92.1%**	**76%**
F1-Score	**95%**	**85%**	**54%**

Table 49: Performance of the Siamese transformer-based networks architecture

9.7.1 Qualitative Evaluation

We also qualitatively look at examples where our model made correct and wrong predictions. We picked pairs that we have already known are related with some degree of relevance 1 and 2 from Qursim. We picked non-related pairs as well; we labelled non-related pairs with the relevance of -1. We compared the actual target values with those predicted by our model to identify what type of prediction is made. The results are reported in Table 50. Studying the results provides insights on where our model succeeded and failed in prediction.

[57] We finetuned a transformer model AraBERT using our dataset. Moreover, same score was reported in a similar work by Alsaleh, et al. (2021).

[58] Doc2vec has achieved 76% accuracy and 54% F1 score as reported in previous experiment, reported in chapter 7.

No.	(Chapter: Verse)	Verse 1	Verse2	Rel	Sim Score%	Prediction
1	(8:39, 2: 193)	وقاتلوهم حتى لا تكون فتنة ويكون الدين كله لله فإن انتهوا فإن الله بما يعملون بصير	وقاتلوهم حتى لا تكون فتنة ويكون الدين لله فإن انتهوا فلا عدوان إلا على الظالمين	2	99.51	TP
		And fight them until there is no fitnah and [until] the religion, all of it, is for Allah. And if they cease - then indeed, Allah is Seeing of what they do.	Fight them until there is no [more] fitnah and [until] worship is [acknowledged to be] for Allah. But if they cease, then there is to be no aggression except against the oppressors.			
2	(6:142, 35: 6)	ومن الأنعام حمولة وفرشا كلوا مما رزقكم الله ولا تتبعوا خطوات الشيطان إنه لكم عدو مبين	إن الشيطان لكم عدو فاتخذوه عدوا إنما يدعو حزبه ليكونوا من أصحاب السعير	1	56.44	FN
		And of the grazing livestock are carriers [of burdens] and those [too] small. Eat of what Allah has provided for you and do not follow the footsteps of Satan. Indeed, he is to you a clear enemy.	Indeed, Satan is an enemy to you; so take him as an enemy. He only invites his party to be among the companions of the Blaze.			
3	(55:41,42: 39)	يعرف المجرمون بسيماهم فيؤخذ بالنواصي والأقدام	والذين إذا أصابهم البغي هم ينتصرون	-1	21.35	TN
		The criminals will be known by their marks, and they will be seized by the forelocks and the feet.	And those who, when tyranny strikes them, they defend themselves.			
4	(33:24, 47:31)	ليجزي الله الصادقين بصدقهم ويعذب المنافقين إن شاء أو يتوب عليهم إن الله كان غفورا رحيما	ولنبلونكم حتى نعلم المجاهدين منكم والصابرين ونبلو أخباركم	-1	60.00	FP

		That Allah may reward the truthful for their truth and punish the hypocrites if He wills or accept their repentance. Indeed, Allah is ever Forgiving and Merciful.	And We will surely test you until We make evident those who strive among you [for the cause of Allah] and the patient, and We will test your affairs.			
5	(8:48, 4: 120)	وإذ زين لهم الشيطان أعمالهم وقال لا غالب لكم اليوم من الناس وإني جار لكم فلما تراءت الفئتان نكص على عقبيه وقال إني بريء منكم إني أرى ما لا ترون إني أخاف الله والله شديد العقاب	يعدهم ويمنيهم وما يعدهم الشيطان إلا غرورا	2	28.06	FN
		And [remember] when Satan made their deeds pleasing to them and said, "No one can overcome you today from among the people, and indeed, I am your protector." But when the two armies sighted each other, he turned on his heels and said, "Indeed, I am disassociated from you. Indeed, I see what you do not see; indeed I fear Allah. And Allah is severe in penalty."	Satan promises them and arouses desire in them. But Satan does not promise them except delusion.			
6	(19:53, 20,39)	أن اقذفيه في التابوت فاقذفيه في اليم فليلقه اليم بالساحل يأخذه عدو لي وعدو له وألقيت عليك محبة مني ولتصنع على عيني	ووهبنا له من رحمتنا أخاه هارون نبيا	-1	61.25	FP

		[Saying], 'Cast him into the chest and cast it into the river, and the river will throw it onto the bank; there will take him an enemy to Me and an enemy to him.' And I bestowed upon you love from Me that you would be brought up under My eye.	And We gave him out of Our mercy his brother Aaron as a prophet.			
7	(57:2, 49:18)	إن الله يعلم غيب السماوات والأرض والله بصير بما تعملون	له ملك السماوات والأرض يحيي ويميت وهو على كل شيء قدير	-1	75.69	FP
		Indeed, Allah knows the unseen [aspects] of the heavens and the earth. And Allah is Seeing of what you do.	His is the dominion of the heavens and earth. He gives life and causes death, and He is over all things competent.			
8	(49:17, 41:8)	إن الذين آمنوا وعملوا الصالحات لهم أجر غير ممنون	يمنون عليك أن أسلموا قل لا تمنوا علي إسلامكم بل الله يمن عليكم أن هداكم للإيمان إن كنتم صادقين	1	21.63	FN
		Indeed, those who believe and do righteous deeds - for them is a reward uninterrupted.	They consider it a favor to you that they have accepted Islam. Say, "Do not consider your Islam a favor to me. Rather, Allah has conferred favor upon you that He has guided you to the faith, if you should be truthful."			
9	(41:21, 36:65)	وقالوا لجلودهم لم شهدتم علينا قالوا أنطقنا الله الذي أنطق كل شيء وهو خلقكم أول مرة وإليه ترجعون	اليوم نختم على أفواههم وتكلمنا أيديهم وتشهد أرجلهم بما كانوا يكسبون	1	49.52	FN
		And they will say to their skins, "Why have you testified against us?" They will say, "We were made to speak by Allah, who has made everything speak; and He created you the first time, and to Him you are returned.	That Day, we will seal over their mouths, and their hands will speak to Us, and their feet will testify about what they used to earn.			

10	(70:11, 76:27)	يبصرونهم يود المجرم لو يفتدي من عذاب يومئذ ببنيه	إن هؤلاء يحبون العاجلة ويذرون وراءهم يوما ثقيلا	-1	70.04	FP
		They will be shown each other. The criminal will wish that he could be ransomed from the punishment of that Day by his children.	Indeed, these [disbelievers] love the immediate and leave behind them a grave Day.			

Table 50: Performance of the Siamese AraBERT architecture on the Qur'anic semantic similarity dataset

Rel Column represents the degree of relevance between verses as labelled in the annotated dataset drawn from Qursim. The Prediction can be TP (true positive), FP (false positive), TN (true negative), or FN (false negative). We define positive and negative to be similar and non-similar respectively. True means the actual value matches the predicted value, and negative means the predicted and actual values do not match. The actual value represents the actual similarity from the semantic similarity dataset (it could be 2 or 1 if similar, and -1 if non-similar), while the predicted value represents the similarity score computed based on the proposed model. For example, TP means the predication was positive (similar with score >= 60) and the actual values was also positive (relevance is 2 or 1). FP means the predication was positive (similar) and the actual value was negative (non-similar).

To analyse the similarity results, we refer to Al-Tabari's and Ibn-Kathir's commentaries, as populated by King Saud Project[59]. The project is an electronic simulation of the Holy Qur'an - available in seventeen languages – with a margin for translating the meanings of the Noble Qur'an into more than twenty languages. It also supports seven interpretations from renowned commentaries. By referring to Table 50, in pair 1 and 3, the model has detected the relation.

In pairs 4 and 6, the pairs were annotated as non-related from our dataset. It is because the pairs were generated randomly; so is not in Qursim. However, our

[59] https://quran.ksu.edu.sa

model, indeed, detected the deep semantic similarity here. In pair 6, the model was able to detect the similarity between the two verses. Both verses refer to Moses[60]. Verse [19:53], refers to Moses' brother, according to Ibn Kathir commentary, "And We gave him (Moses) out of Our mercy his brother Aaron as a prophet". Verse [20:39] is interpreted in Ibn Kathir commentary as: "[Saying], 'Cast him into the chest and cast it into the river, and the river will throw it onto the bank; there will take him an enemy to Me and an enemy to him.' And I bestowed upon you love from Me that you would be brought up under My eye."

In pair 4, the similarity between the two verses [33:24, 47:31] seems ambiguous as the meanings are not close. However, our model has detected a similarity of 60%, which reveals a deep relation that commentary books have not pointed out. By referring to Al-Tabari commentary, the discussion has emerged from the previous verse [33:23] revealed about people who did not witness battle of Badr: among those, some of them have fulfilled their vow, and others await the appointed time; they have not changed in the least. The first verse [33:24] is explained as: "in order that Allah may reward the truthful for their truthfulness, and either punish the hypocrites or, if He so wills, accept their repentance." The second verse [47:31] is explained as: "We shall certainly test you until We know those of you who truly strive and remain steadfast, and will ascertain about you." Therefore, the two verses point out that God almighty tests believers by al-jihad against the enemies of Allah, to distinguish those who have insight into his religion (truthful) from those who have doubt and confusion (hypocrites).

9.8 Discussion

In pairs 7 and 10, the model assigns high similarity scores for the associated pairs while the relation is not recognised in Qursim. The model revealed a relation between verses in each pair. We referred to Al-Tabari's and Ibn-Kathir's commentary to look up verses' interpretation and examine our model's capability to detect any embedded relation. We outline our findings as the following:

---

[60] https://quran4all.net/ar/sura

1. In pair 7, both verses [57,2] and [49, 18], glorify the God and mention his attributes; he is the kingdom of the heavens and the earth and he knows the unknown of the heavens and the earth.
2. In pair 10, both verses discuss disbelievers' status. Verse [70, 11] describes their state on the day of Judgment, the sinner would like to ransom himself from the torment of that Day by offering his sons. Verse [76, 27] tells that they love the world, they love to stay in it and admire its adornment and they forget to work for the Hereafter.

On the other hand, there is some instances where our model failed to detect the similarity despite the high likelihood due to the commonality in terms used within the verses, like in examples 2, 5, 8, and 9. The similarity score in examples 2 and 9 is close to the threshold, so we may need to re-consider its value.

9.9 Chapter Summary and Conclusions

This chapter presented a novel verse-embeddings using Siamese AraBERT-networks for fast and efficient semantic similarity detection in the Qur'an. We proposed a Siamese transformer-based architecture where two networks have tied weights. To achieve high performance, our architecture starts with pre-trained AraBERT models, and a Siamese set-up is used to fine-tune the models on a semantic similarity dataset drawn from the Qur'an. Our model achieved a score of 83.56% Spearman correlation representing its ability to assess whether two verses are similar. The F1-score 95% on the Qur'anic semantic similarity test was impressively high.

This chapter presented the most novel contribution by leveraging the paradigm of pre-trained language models and the application of transformers to the Arabic language and the Qur'an. The model is a success and detected a semantic similarity between Qur'an verses with high accuracy. The relation was recognized by existing resources. In addition, the model revealed other embedded relation that was not covered before. When compared to other distributional models, like doc2vec, the transformer-based model has excelled in terms of accuracy and performance.

Deep learning techniques, indeed, can be useful in tasks where context remembering is key, such as extracting knowledge from the Qur'an. This thesis goes in line with these insights and benefits from the deep learning revolutions to help gain the teachings and knowledge from the sacred text. Distributional models, like word2vec, doc2vec, BERT, etc., have the potential to generate meaningful embedding representations for the Qur’an verses. Such representation can be a base for many ML algorithms such as classification, topic modelling, and semantic similarity.

Chapter 10
Conclusions and Future Work

10.1 Overall Conclusion

The extraordinary accuracy and elegance of the Qur'an distinguished it from other books in conveying the utmost quantity of knowledge in the fewest number of words. Moreover, the Holy Book employs innovative means to pass on subtle meanings to future generations. There are several significant themes throughout the Qur'an, and references to them appear throughout the Qur'anic text. It is challenging to distinguish between them. Each time one of these concepts is mentioned, the Qur'an emphasises a distinct aspect of it in the exact phrasing of sections. The semantic similarity is evident in the Holy Qur'an through the emergence of meanings and topics all over the text. Therefore, the study of semantic similarity can help direct verbal similarity and reveal its connotations according to its manifestation in the contexts of Qur'anic verses. Potentially, it presents evidence and compelling argument to respond to the doubts of atheists and to revoke their allegations regarding the issue of the occurrence of verbal similarities in the Holy Qur'an and the repetition of Qur'anic concepts in different verses and various combinations.

In this thesis, we examined the semantic similarity task using powerful methods to encode the verses of the Qur'an and get their embeddings, and then apply a semantic similarity metric to score the relationship between the Qur'anic verses. The framework utilizes deep learning models based on distributional semantics to achieve state-of-the-art semantic textual similarity task results.

This dissertation provides an empirical framework of computational tools to develop the Qur'anic knowledge, which should be easily reused and customized for relevant work. The framework employs the concept of distributed

representation, the basis of sophisticated deep learning models. It uses machine learning tools for NLP, which leverage the power of distributional semantics using deep learning models to generate semantically meaningful verse embeddings and potentially allow the discovery of knowledge-related connections between the Qur'anic verses. In addition, it obtained state-of-the-art results on the semantic similarity search and unsupervised clustering.

We used recent breakthroughs in document embeddings to provide a robust representation of the Qur'an verses fed to ML algorithms for classification, regression, and semantic similarity. We proposed a Siamese transformer-based architecture for pairwise semantic similarity detection in the Qur'an. Arabic pre-trained contextual representations are exploited to derive semantically meaningful verse embeddings. Then the twin transformers networks are tuned on a semantic similarity dataset drawn from the Qur'an. It is shown that the model improves the Qur'anic semantic similarity measures and performance over previous studies is indicated. These experiments resulted in a collection of resources, models, and corpora that can be a basis for potential applications and research.

10.2 Why Considering Learning Representations?

Building computational models that capture the semantics and contextual information of linguistic units in a text is essential for its understanding. The first step is projecting text features in a chosen feature space. A small and concise corpus like the Qur'an requires a multi-level automatic feature representation learning which is possible with deep learning. Traditional machine learning-based NLP systems, on the other hand, rely primarily on hand-crafted features.

Traditional or shallow learning approaches encompass all of the conventional or classical methods associated with conventional machine learning. Any approaches before neural networks in which the prediction is based on hand-engineered features are included in this group. Furthermore, this category comprises neural networks with a few hidden layers (usually 0 – 2), which are

referred to as "shallow" and serve as a bridge between this group of approaches and their deep learning-based successors (Andrea et al., 2022; Bengio et al., 2013).

Shallow learning approaches arrived after rule-based approaches, which in accuracy and stability, outperformed them. Shallow learning approaches are still prevalent in many practical applications or strong baselines. However, these traditional approaches necessitate a feature engineering step, which can be costly depending on the domain's complexity. While the computational side of this cost can be substantial, the domain knowledge requirements required for the effective use of relevant feature extraction algorithms may be more challenging to meet in practice (Andrea et al., 2022; Bengio et al., 2013).

New machine learning algorithms take advantage of recent advances in deep learning approaches to automate the extraction of expressive features. The rapid advancement of these methods has resulted in many options for converting natural language into machine-readable data. Because of their capacity to model complicated features without hand engineering, these methods have gained appeal, reducing a portion of the domain knowledge needed. Instead, efforts have been directed toward developing neural network architectures that extract appropriate representations for textual units. Recent advances have proved particularly fruitful in this regard, resulting in semantically meaningful and contextual representations. Automatic feature extraction is beneficial for modelling textual data because it may utilize a document's underlying linguistic structure (Andrea et al., 2022; Bengio et al., 2013).

10.3 Literature Review

The literature review draws from a large body of related work on several concepts that make the theoretical base of this research. It emerges in chapters 2 and 3. Chapter 2 provided a background introduction to the Qur'an as scripture, followed by a review of Qur'anic exegesis, their evolution, and trends. It then surveyed computational and corpus linguistics resources for the Qur'an. reviewed previous work in three areas: Arabic Computational Linguistics, Arabic NLP, and NLP for the Qur'an.

Chapter 3 identified the main terms employed in this research: NLP, machine learning, and deep learning. First, it introduced the concept of distributed representation, the basis of deep learning models. It then provided a review of recent deep learning models and methods that have been employed for NLP tasks. Finally, it explored the different word embedding techniques in NLP and their success in solving challenging NLP tasks, for Example, Sentence Embedding, Deep Neural Networks, and Pre-trained models.

10.4 A Review of the Research Aims

The dissertation sought to employ computational models and benefits from the power of deep learning, based on the distributional semantics, to uncover the overlooked places by digging and scrutinizing the secrets of the semantics inherent in the sacred text. We fulfilled the goal by utilizing advanced NLP techniques within the context of deep learning, such as embeddings and transformers, to generate a dense representation of the Qur'an verses and use it to achieve excellent results in classification, regression, and semantic similarity. We also developed models and language resources that would be a baseline for future research. This dissertation indeed provides corpora and software resources to a low-resource language like the classical Arabic of the Qur’an. Furthermore, we provided a collection of semantically related verses that indicate novel computational discoveries not found in any Tafsir interpretation, thus expanding the Islamic knowledge base. The dataset can serve as a valuable resource for future analysis and exploration of underlying meanings and teachings for Islamic scholars. Exegetes and scholars who use a thematic approach to interpret and understand the Qur'anic text would profit from the delivered knowledge because they examine the sacred book according to distinct themes rather than individual verses.

The ultimate goal was achieved by presenting a semantic representation of the Qur’an verses that can be a base for potential ML experiments and emerging developments. This thesis is unique in that it uses computational linguistics techniques to analyse the classical Arabic of the Qur'an, the world's most recognizable and miraculous sacred scripture. Also, this innovative research has attracted the interest of other scholars, as indicated by regular requests for

collaboration and inquiries about the published chapters from the NLP research community. Most of the findings were shared in significant conferences like IMAN and ACLing and published in flourishing Journals like IJASAT. The paper presented at the IMAN2021 conference was awarded the “Distinguished Paper." It presents the most successful novel contribution which is the Siamese Transformer model of Qur’an verse similarity presented in chapter 9. This work was also presented at ResCompLeedsCon2022. David Hogg (keynote speaker) commented that the F1-score 95% on the Qur’anic semantic similarity test was impressively high.

10.5 A Review of the Research Questions

This thesis tackled two main research questions.

10.5.1 RQ1: How to represent the semantics of the Qur’an words to capture the intangible semantic relations?

To address this question, we examined two distributional approaches to model and explain word distributions in the Qur’an. In chapter 4, we developed a statistical language model for discovering topics in the Qur'an, using a topic modelling algorithm: Latent Dirichlet Allocation (LDA). Using unsupervised learning, the model built a topic per document model and words per topic model, modelled as Dirichlet distributions. The derived clusters are not coherent and keywords, sometimes, are not interpretable. However, the derived clusters may provide scholars with directions to look at the chapters that contributed to a specific topic and explore patterns that may support the teachings.

On the other hand, chapter 5 investigated neural networks to produce semantic representation of the Qur’an’s words. It used word-level neural networks' capability to learn the semantic representation of Qur’an words. A word embeddings algorithm (Word2vec) generated distributed representations, dense real-valued vectors that encode the semantics of Qur’an words. Then cosine similarity was used to determine how similar two vectors are. These embeddings have shown their effectiveness in capturing context similarities and analogies, as well as being fast and efficient in clustering and semantic similarity due to their reduced dimensionality. One limitation of word embeddings is that two words with multiple meanings are represented in one single vector, which affects similarity

detection. Also, evaluating derived embeddings is challenging at the level of data available for training the model. and scarcity of tools devoted to classical Arabic such as stemmer.

The research question is partially addresses as LDA and word2vec captured the semantics in the Qur'an and achieved relative results, however, evaluating these methods needs improvements on the level of data and evaluation resources.

10.5.2 RQ2: Is deep learning a viable approach for modelling the complicated features of the Arabic Qur'anic text, and learning subtle semantic relations?

This thesis investigated the potentiality of deep learning approaches based on distributional semantics to model the semantics in the Qur'an. Different models with different capabilities were used and trained on the classical Arabic of the Qur'an, generating fruitful features for unsupervised clustering, classification, and semantic similarity. To address this question, the thesis investigated document embeddings and transformers to derive semantically meaningful and contextual representations of the Qur'an verse, that are useful for the semantic-based similarity task and eventually capture meanings and concepts in the Qur'an. The question is addressed which is represented by the novel findings and impressive semantic textual similarity task results. However, many potential improvements are presented, as future work, to ensure better performance and higher accuracy.

10.5.2.1 Document Embeddings (Doc2vec)

Chapter 6 exploited a recent trend in machine intelligence, which is the distributed representation of text, to learn an informative representation of the passages of the Qur'an. It used a recent breakthrough in feature embedding, Paragraph Vectors, enabling machine learning models to have an informative numerical representation of the Qur'anic passages. Unsupervised learning trained such embeddings and used as features for NLP tasks such as classification, topic modelling, and semantic textual similarity.

Chapter 7 measured similarity among those vectors. The model's performance was judged through cosine similarity between assigned semantic similarity

scores and annotated textual similarity datasets. The model scored 76% accuracy and 51% F1 score.

Chapter 8 classified the verses of the Qur'an into 15 predefined classes. The semantics of the verses were captured using Doc2vec embeddings that were used to group similar documents. The model achieved a classification accuracy of 70% and an F1 score of 60%. The results confirmed that the classifier scored higher accuracy results with the distributed bags of words architecture of the Doc2vec model. The model then was evaluated by examining the semantic similarity of the Qur'anic verses. The mean difference and average similarity values for each category were calculated to indicate how well the model describes that category. Derived classes showed high average similarity using distributed bags of words architecture for some classes.

10.5.2.2 Transformers

Chapter 9 leverages both the pre-trained contextualized representations for the Arabic language and the Siamese architecture to derive semantically meaningful verse embeddings and achieve remarkable results in pairwise semantic similarity detection in the Qur'an. The chapter proposed a novel Siamese transformer-based architecture for pairwise semantic similarity detection in the Qur'an. The architecture starts with pre-trained AraBERT models, and a Siamese set-up is used to fine-tune the models on a semantic similarity dataset drawn from the Qur'an. The model achieved a score of 83.56% for the Spearman correlation representing its ability to assess whether two verses are similar. The model obtained a high-performance 95% F1 score on the Qur'anic semantic similarity dataset. Thus, the model improved the Qur'anic semantic similarity measures and performance over previous studies.

10.6 Future Work

This research paved the road for potential studies, situated existing work within it, and evaluated which approaches appear to be the most promising to accomplish the ultimate goal. The primary tasks were detecting the semantic similarity in the Qur'an and generating datasets of related verses, and topical classification for the Qur'an verses. Moreover, machine learning experiments

were conducted on top of the dense representation of the Qur'an verses that were derived by deep learning approaches. This study established the groundwork for a variety of future extensions, applications, and experiments, as outlined in the following subsections.

10.6.1 User Evaluation

Additional research on user evaluation is presenting this work to Islamic scholars etc to get evidence that they may take up and use derived resources. The derived corpora and semantic similarity datasets[61] can be verified by Islamic scholars and then published as a significant resource for evaluating semantic similarity in the Qur'an.

Moreover, chapter 9 suggests another area of improvement by using different data, as generating non-related pairs was occurred randomly. Another potential is using Islamic scholars or human annotators with domain knowledge to annotate datasets for training and testing the model.

10.6.2 A Corpus for evaluation of semantic categorization in the Qur'an

10.6.2.1 Introduction

To determine the semantic accuracy of text representations, one can use a variety of empirical tests. Semantic categorization is one example that has been used to probe semantic accuracy of the corpus derived vectors (Bullinaria and Levy, 2007). The purpose of this test is to see how well semantic categories are represented in vector space. It evaluates how often particular word vectors are closer to their own semantic category centre than to one of the others (Patel et al., 1997). Hence, we can evaluate the efficiency of embedding models by measuring their capabilities to represent semantic categories. To perform the test, we suggest a corpus for evaluation of the semantic categorization in the Qur'an.

---

[61] All produced dataset will be accessible at Github repository: Mhalshammeri

10.6.2.2 Knowledge Resources for Semantic Categorization in the Qur’an

A range of Qur’anic corpus-based resources and ontologies have been developed to serve as a substantial resource for religious knowledge; several of them were developed by the Artificial Intelligence research group at the University of Leeds. The different resources can be considered a good base for potential research and development, to aid semantic search, querying, clustering and classification of the Qur’an. This section reviews existing knowledge resources that can be a base for our corpus.

1. Qur’anic Arabic Corpus (QAC)

The Qur’anic Arabic corpus is an annotated linguistic resource that displays the Arabic grammar, syntax, and morphology for each word in the Qur’an. Three levels of analysis are available in the corpus: morphological annotation, a syntactic treebank, and a semantic ontology. It presents an ontology of Qur’anic concepts that defines key concepts in the Qur’an. Named entities in verses are linked to concepts in the ontology using named entity tagging. The concept is the root entity in the Qur’an ontology, and divided into subcategories (Dukes, 2013).

2. Qurany

The Qurany corpus is augmented with an ontology or index of key concepts, taken from a recognized expert source ‘Mushaf Al Tajweed’. The corpus allows users to search the Qur’an corpus for abstract concepts via an ontology browser. It contains a comprehensive hierarchical index or ontology of nearly 1200 concepts in the Qur’an. Scholars can utilise the Qurany ontology browser to pinpoint a specific concept and locate verses that allude to it more precisely (Abbas, 2009).

3. The Indexed Dictionary of the meanings in the Qur’an

Another beneficial resource is "the indexed dictionary of the meanings in the Qur’an" or "المعجم المفهرس لمعاني القرآن العظيم" (Alzain, 1995). The dictionary collects the Qur’anic verses linked by one topic and classifies them into subtopics through which the subject unity and logical coherence appear. The reference is indexed alphabetically by keywords that represent the different meanings and topics in the Qur'an. Each topic is represented by a keyword, under which a list of verses

addresses that topic. And the relevant word is identified and underlined. When formulating the topic or meaning, what distinguishes this resource is that the author used keywords close to the researcher's mind. On the other hand, the reference is available in a scanned PDF format, Appendix A, which could be challenging when retrieving its contents automatically. Also, the reference is written in Arabic and published in volumes; to the extent of our knowledge, no translation is available.

10.6.2.3 Challenges with current knowledge resources

Computational and corpus linguistics resources on the Qur'an are available in different formats. They present the knowledge at different levels and in different formats. Usually, a researcher who seeks experimentation and evaluation of his own models can benefit from such contributions. However, sometimes it is challenging to mine the information and convert them in a format suitable for our needs; to train and test ML and DL models. We, therefore, create a corpus, that is drawn from existing corpora, to suit our needs and act as a benchmark evaluation resource for similar tasks.

10.6.2.4 Corpus and Data

The corpus will be created from the original Qur'anic text, and presents a semantic classification of categories in the Qur'an. Each category is represented by a key word, and linked to a list of verses belonging to that category. The corpus will be a great assessment resource for computational linguistics investigating semantic similarity and topic modelling in the Qur'an and classical Arabic texts.

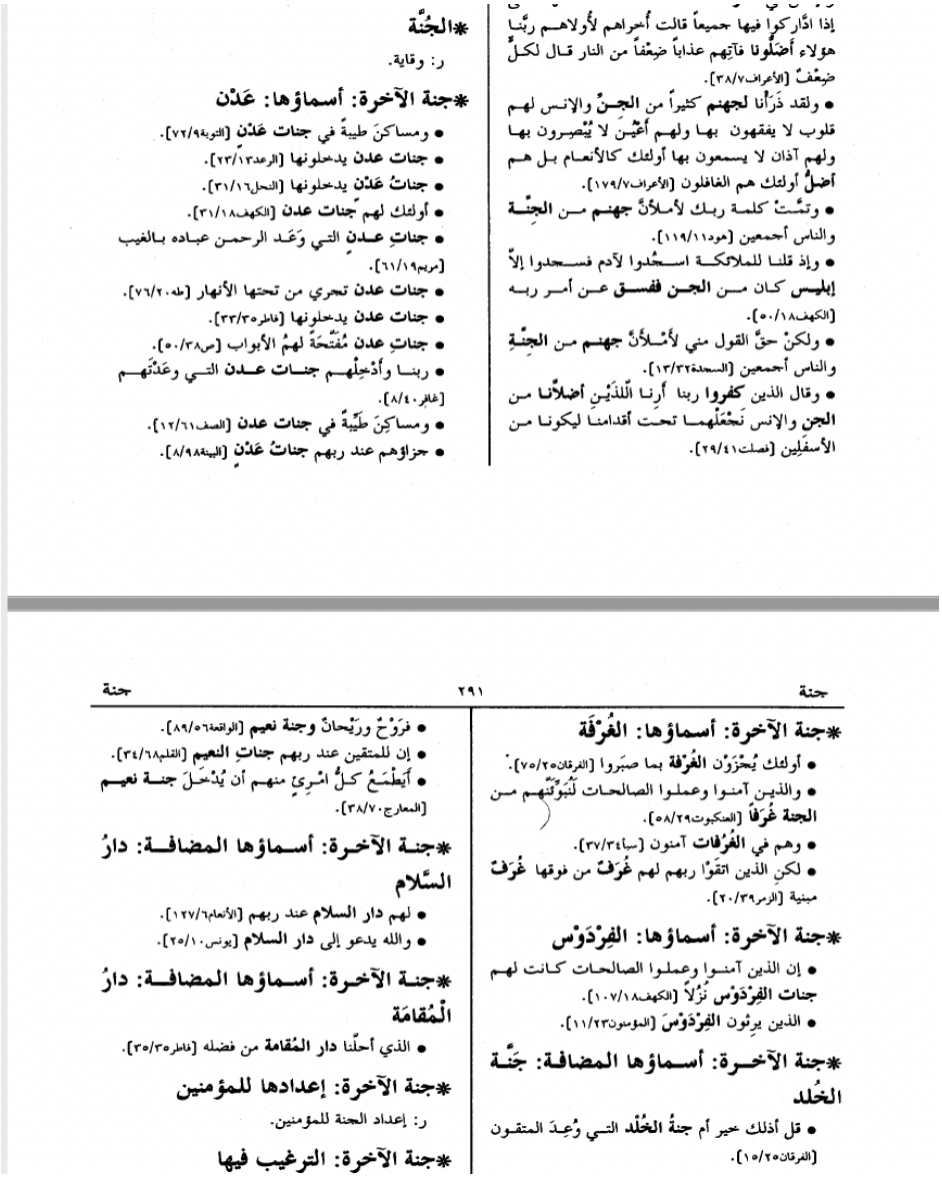
The data in the corpus is derived from two different linguistics resources on the Qur'an. The first resource is the indexed dictionary of the meanings in the Qur'an (Alzain, 1995). The book is written in the Arabic language. The reference classifies the verses into semantic categories or topics, and each category is described by a keyword. The second resource is the Qur'anic Arabic corpus (QAC) by Dukes (2013). The QAC provided a Qur'anic ontology that employs knowledge representation to define major concepts in the Qur'an and predicate logic to show relationships between them. Named entities in verses are related to concepts in the ontology, such as the names of historical persons and places referenced in the Qur'an. The ontology is expanded to include concepts as well as a list of all instances of those concepts in the Qur'an.

To build a category, for example, paradise or ‘الجنة’, we look up the subject in both resources to include all the verses discussing the topic. The topic may be addressed using different terms or keywords. Here, the paradise is referenced in the Qur’an using many names, to mention some in Table 51:

الفردوس	عدن	جنة الخلد	جنة النعيم
Firdous	Eden	Garden of Eternity	Garden of Pleasure

Table 51: Different terms describing the topic 'Paradise'

The term is mentioned in the first resource as shown in Figure 35 . The paradise is a category/ topic represented by the keyword ‘جنة الآخرة’, located within the ‘الجيم’ chapter. It is divided into subtopics that explained its different names and characteristics. Other chapters also include the same topic under alternative names to notify the reader of the link between the different terms as they refer to the same topic.



إذا ادّاركوا فيها جميعاً قالت أُخراهم لأُولاهم ربَّنا هؤلاء أَضَلُّونا فآتِهم عذاباً ضِعْفاً من النار قال لكلٍّ ضِعْفٌ [الأعراف٧/٣٨].

• ولقد ذَرَأْنا **لجهنم** كثيراً من **الجنِّ** والإنس لهم قلوب لا يفقهون بها ولهم أعيُنٌ لا يُبْصِرون بها ولهم آذان لا يسمعون بها أولئك كالأنعام بل هم **أضلُّ** أولئك هم الغافلون [الأعراف٧/١٧٩].

• وتمَّتْ كلمة ربك لأملأنَّ **جهنم** من **الجنَّة** والناس أجمعين [هود١١/١١٩].

• وإذ قلنا للملائكة اسجُدوا لآدم فسجدوا إلاّ **إبليس** كان من **الجن ففسق** عن أمر ربه [الكهف١٨/٥٠].

• ولكنْ حقَّ القول مني لأَملأنَّ **جهنم** من **الجِنَّةِ** والناس أجمعين [السجدة٣٢/١٣].

• وقال الذين **كفروا** ربنا أَرِنا اللَّذَيْنِ **أضلاّنا** من **الجن** والإنس نَجْعَلْهما تحت أقدامنا ليكونا من الأسفَلِين [فصلت٤١/٢٩].

***الجُنَّة**

ر: وقاية.

***جنة الآخرة: أسماؤها: عَدْن**

• ومساكنَ طيبةً في **جنات عَدْنٍ** [التوبة٩/٧٢].

• **جنات عدن** يدخلونها [الرعد١٣/٢٣].

• **جناتُ عَدْنٍ** يدخلونها [النحل١٦/٣١].

• أولئك لهم **جنات عدن** [الكهف١٨/٣١].

• **جناتِ عدنٍ** التي وَعَد الرحمن عباده بالغيب [مريم١٩/٦١].

• **جنات عدن** تجري من تحتها الأنهار [طه٢٠/٧٦].

• **جنات عدن** يدخلونها [فاطر٣٥/٣٣].

• **جناتِ عدن** مُفَتَّحَةً لهمُ الأبواب [ص٣٨/٥٠].

• ربنا وأَدْخِلْهم **جنات عدن** التي وعَدْتَهم [غافر٤٠/٨].

• ومساكِنَ طَيِّبةً **في جنات عدن** [الصف٦١/١٢].

• جزاؤهم عند ربهم **جناتُ عَدْنٍ** [البينة٩٨/٨].

***جنة الآخرة: أسماؤها: الغُرْفَة**

• أولئك يُجْزَوْن **الغُرْفة** بما صبَروا [الفرقان٢٥/٧٥].

• والذين آمنوا وعملوا الصالحات لَنُبَوِّئَنَّهم من **الجنة غُرَفاً** [العنكبوت٢٩/٥٨].

• وهم في **الغُرُفات** آمنون [سبأ٣٤/٣٧].

• لكنِ الذين اتقَوْا ربهم لهم **غُرَفٌ** من فوقها **غُرَفٌ** مبنية [الزمر٣٩/٢٠].

***جنة الآخرة: أسماؤها: الفِرْدَوْس**

• إن الذين آمنوا وعملوا الصالحات كانت لهم **جنات الفِرْدَوْس** نُزُلاً [الكهف١٨/١٠٧].

• الذين يرِثون **الفِرْدَوْسَ** [المؤمنون٢٣/١١].

***جنة الآخرة: أسماؤها المضافة: جَنَّة الخُلد**

• قل أذلك خير أم **جنةُ الخُلْد** التي وُعِدَ المتقون [الفرقان٢٥/١٥].

• فرَوْحٌ ورَيْحانٌ **وجنة نعيم** [الواقعة٥٦/٨٩].

• إن للمتقين عند ربهم **جناتِ النعيم** [القلم٦٨/٣٤].

• أَيَطْمَعُ كلُّ امْرِئٍ منهم أن يُدْخَلَ **جنة نعيم** [المعارج٧٠/٣٨].

***جنة الآخرة: أسماؤها المضافة: دارُ السَّلام**

• لهم **دار السلام** عند ربهم [الأنعام٦/١٢٧].

• والله يدعو إلى **دار السلام** [يونس١٠/٢٥].

***جنة الآخرة: أسماؤها المضافة: دارُ الْمُقامَة**

• الذي أحلّنا **دار المُقامة** من فضله [فاطر٣٥/٣٥].

***جنة الآخرة: إعدادها للمؤمنين**

ر: إعداد الجنة للمؤمنين.

***جنة الآخرة: الترغيب فيها**

Figure 35: The ‘Paradise’ category in the " Indexed dictionary of meanings in the Qur'an" (Alzain, 1995)

The QAC addresses “Paradise” as the main concept and each of its names, as mentioned in the Qur’an, is directly related to it. As shown in Figure 36 ,‘’Paradise’ is related to ‘Firdous’,’ Salsabil’, ‘Lote Tree’, and ‘Garden of Eden’. The link between ‘Firdous’ and ‘Salsabil’ happens through the main concept which is ‘Paradise’; they are not related directly. Moreover, for each concept in the map, there is a list of verses related to this concept, and the keyword is highlighted.

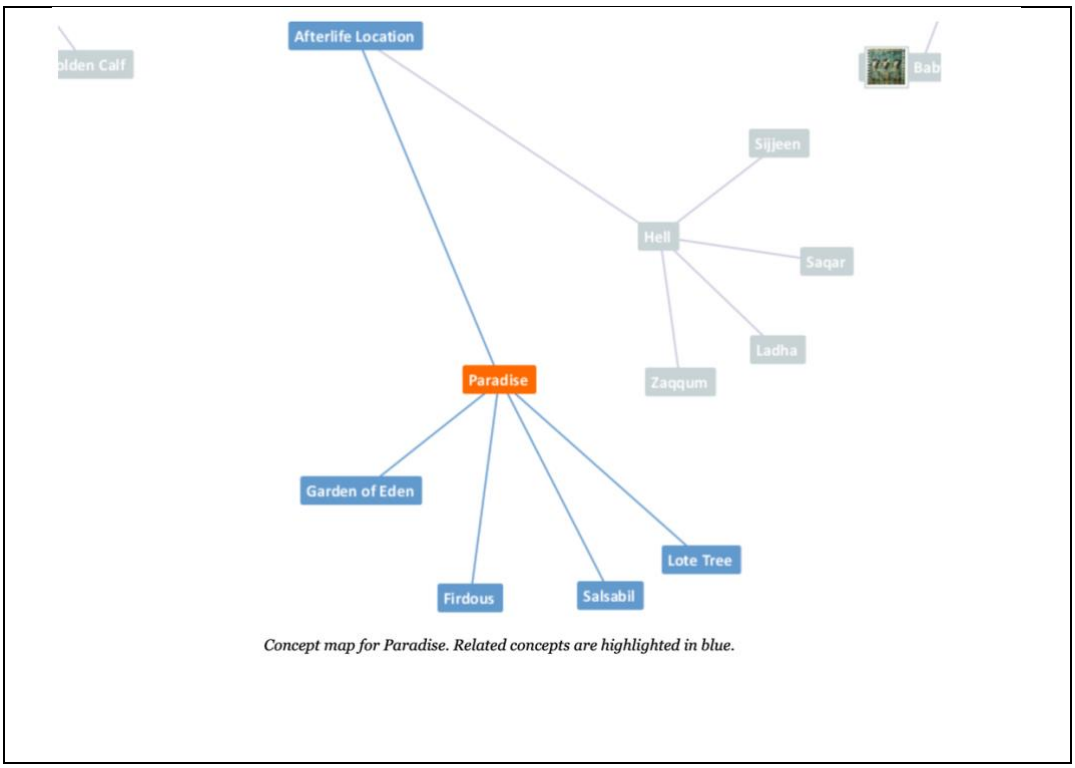


Figure 36: Concept map for concept 'Paradise" as in QAC (Dukes, 2013)

The two resources refer to the same concept or category using different structure and using different names. For example, the two concepts ‘Lot Tree’ or ‘ سدرة المنتهى’and ‘Salsabil’ or ‘ سلسبيل‘ are not linked to ‘Paradise’ in the first source, but mentioned as a separate category in different chapters without a reference to ‘Paradise’ concept as in the second resource (QAC). Also, the QAC provides an English translation to its contents, which is not the case in the first resource; in Arabic. Therefore, we would compile the proposed semantic categorization using both resources to cover nearly all semantic relations in the Qur’an.

10.6.3 Potential ML Experiment: Topic Modelling based on Transformers

The research potentials are using transformer-based sentence embedding across a wide range of NLP tasks and leveraging the paradigm of pre-trained language models and the application of transformers to the Arabic language and the Qur’an. The only limitation here is the availability of labelled datasets for fine-tuning. Therefore, one prospective extension is to create datasets drawn from the Qur’an and relevant knowledge resources to support training and fine-tuning transformers.

This work can be an excellent starting point for many potential types of research on modelling the semantic similarity/ relatedness in the Qur'an and extracting the embedded knowledge. For example, the derived model from chapter 9 can be used for topic identification and unsupervised tasks such as clustering based on the semantic similarity of sentence/verse embeddings, providing a deep learning alternative to traditional models such as Latent Dirichlet Allocation. A possible experiment would be to use tBERT, a topic-informed BERT-based architecture for pairwise semantic similarity detection, using AraBERT as the base.

10.7 Challenges and limitations

A variety of Natural language processing (NLP) tasks have benefited from the recent remarkable advancements of deep neural networks models. However, such models require large-scale annotated datasets for training. In the meantime, we are witnessing a recent trend of creating such annotated datasets for various NLP problems. Most of these annotations only exist within high-resource languages such as English, while, other languages, such as Arabic, do not have the sufficient labelled data for training modern deep neural networks for a variety of NLP tasks.

10.7.1 A Small and Specialized Corpus

In the context of this research, the small size of the training data is a significant challenge when training deep learning models. The application of deep learning techniques requires a large corpus of texts to be representative and sufficient to train word embeddings. Our target specialized text, the Qur'an, is considered a small and concise corpus. Another significant challenge is that such a text has specialized terms and terminologies. Existing word embeddings models are learnt from general corpora, and using such general models does not scale well; nor does it work for every potential need. Additionally, existing Qur'anic ontologies are considered inadequate and have different scopes, formats, and entity names for the same concept (Alrehaili & Atwell, 2014).

10.7.2 Lack of Stemmer for the Qur'an

Another significant issue is the lack of stemmer algorithms for the classical Arabic of the Qur'an, or the deficiency of existing ones. It is a grand challenge to develop a stemmer considering the unique linguistics features and complexity of the Qur'anic text. A stemming algorithm usually normalizes the words by replacing different shapes of the word with its normal form. For example, the variations of the word 'رب' or God, usually refers to the same entity; the God or 'الله', such as in the two examples here.

اتَّبِعُوا مَا أُنزِلَ إِلَيْكُم مِّن رَّبِّكُمْ وَلَا تَتَّبِعُوا مِن دُونِهِ أَوْلِيَاءَ ۗ قَلِيلًا مَّا تَذَكَّرُونَ Follow, [O mankind], what has been revealed to you from your Lord and do not follow other than Him any allies. Little do you remember.[7: 3]
وَإِنَّ رَبَّكَ هُوَ يَحْشُرُهُمْ ۚ إِنَّهُ حَكِيمٌ عَلِيمٌ And indeed, your Lord will gather them; indeed, He is Wise and Knowing. [15: 25]

However, the same word refers to a different entity in verse 42 of chapter 12 (Yusuf), which is the king or master:

{وَقَالَ لِلَّذِي ظَنَّ أَنَّهُ نَاجٍ مِّنْهُمَا اذْكُرْنِي عِندَ رَبِّكَ فَأَنسَاهُ الشَّيْطَانُ ذِكْرَ رَبِّهِ فَلَبِثَ فِي السِّجْنِ بِضْعَ سِنِينَ} And he said to the one whom he knew would go free, "Mention me before your master." But Satan made him forget the mention [to] his master, and Joseph remained in prison several years. [12: 42]

The different variations of the same word, despite the same stem, have different meanings. Applying the existing Arabic stemmers would produce the same roots or stem to the different variation of the word, ignoring the different contexts within which the word was used. Faults are bound to occur with NLP, however, applying

such algorithms to the Qur’an is inappropriate and faults are not permitted. Hence, although Arabic language stemming algorithms exist, their accuracy still needs to be improved. The algorithms should consider contexts of the words to allow for learning meaningful representations of the words and sentences in the corpus.

10.8 Implications for Future Research

Continuing the themes covered in this thesis and expanding resources pertinent to the Qur’anic semantic similarity in response to its application in recent research, are two sources of motivation for future work.

10.8.1 Building Language Resources and Tools

It may be interesting to incorporate other sources of Islamic knowledge like Hadith or ontologies from the Qur'an to build novel datasets that are useful for the Qur’anic semantic relatedness task. The aim is to enhance the learning process by extending the training data with extra features from additional knowledge resources like Hadith and ontologies, linking each verse with Hadith and ontology concepts.

Bashir et al. (2021) pointed out that because Qur'anic Arabic, or Classical Arabic, differs from Modern Standard Arabic, and the linguistic context of rare words can be found by consulting complementary resources, by referring to non-Qur’anic corpus—in particular, the Hadith corpus and other classical Arabic Islamic books—for Qur’anic research. Future Qur'anic NLP research should consider including these additional data sources to improve models and solutions. In addition, for the Qur'an, intelligent search algorithms are needed that can answer questions from Muslims and non-Muslims alike. The majority of today's semantic and concept-based search systems are built from a topic standpoint rather than a user standpoint (Bashir et al., 2021).

Another recommendation for future work is developing a stemming algorithm specific to the classical Arabic of the Qur'an. A Qur'anic stemmer is likely to produce superior results because Arabic is highly morphologically ambiguous and encompasses complex dependencies. The algorithm may also be effective for both Classical and Modern Arabic.

Finally, the derived classifications and datasets can be used as an evaluation dataset for many interesting challenges in the field of computational semantics like semantic similarity/ relatedness and question-answering from the Qur'an. Another potential idea is organizing a SemEval shared task contest on Qur'anic semantic relatedness. The goal is to encourage participation in solving the Qur'anic semantic textual similarity task and eventually establish best practices for approaching the problem.

10.8.2 Improving Relation Extraction using Syntactic Patterns and Syntactic Roles

In Chapter 6, we presented a vector representation of the Qur'anic verses in the Classical Arabic. The vectors were exploited for locating similar/related verses throughout the text. We then used ML on top of the derived vectorized verses to explore themes/concepts in the text. Eventually we put the data into groups of similar/related passages to capture the concepts/themes presented in each group. Be exploring the text within each cluster, we observed verses sharing a common topic, and identified associated keywords with significant presence. However, this work opens the horizon for further analysis in the light of the following points:

1. Identifying recurrent patterns in the verses of the Qur'an would help to explain how key concepts/ sequence of concepts are presented in the Qur'an, and how they are related and shared over its passages. For example, in Table 52, studying recurrent patterns related to "Day of Resurrection" would contribute a multitude of relevant subtopics like its time, signs of the day, punishment of those who do not believe in the Hereafter, and more. Therefore, studying all such instances would benefit understanding the teachings.

Chapter: Verse	Arabic Text	English Translation
47: 18	{فَهَلْ يَنظُرُونَ إِلَّا السَّاعَةَ أَن تَأْتِيَهُم بَغْتَةً ۖ فَقَدْ جَاءَ أَشْرَاطُهَا ۚ فَأَنَّىٰ لَهُمْ إِذَا جَاءَتْهُمْ ذِكْرَاهُمْ}	Then do they await except that the Hour should come upon them unexpectedly? But already there have come [some of] its indications. Then what good to them, when it has come, will be their remembrance?

43: 66	{هَلْ يَنظُرُونَ إِلَّا السَّاعَةَ أَن تَأْتِيَهُم بَغْتَةً وَهُمْ لَا يَشْعُرُونَ}	Are they waiting except for the Hour to come upon them suddenly while they perceive not?
4: 77	{أَلَمْ تَرَ إِلَى الَّذِينَ قِيلَ لَهُمْ كُفُّوا أَيْدِيَكُمْ وَأَقِيمُوا الصَّلَاةَ وَآتُوا الزَّكَاةَ فَلَمَّا كُتِبَ عَلَيْهِمُ الْقِتَالُ إِذَا فَرِيقٌ مِّنْهُمْ يَخْشَوْنَ النَّاسَ كَخَشْيَةِ اللَّهِ أَوْ أَشَدَّ خَشْيَةً ۚ وَقَالُوا رَبَّنَا لِمَ كَتَبْتَ عَلَيْنَا الْقِتَالَ لَوْلَا أَخَّرْتَنَا إِلَىٰ أَجَلٍ قَرِيبٍ ۗ قُلْ مَتَاعُ الدُّنْيَا قَلِيلٌ وَالْآخِرَةُ خَيْرٌ لِّمَنِ اتَّقَىٰ وَلَا تُظْلَمُونَ فَتِيلًا}	Have you not seen those who were told, "Restrain your hands [from fighting] and establish prayer and give zakah"? But then when fighting was ordained for them, at once a party of them feared men as they fear Allah or with [even] greater fear. They said, "Our Lord, why have You decreed upon us fighting? If only You had postponed [it for] us for a short time." Say, The enjoyment of this world is little, and the Hereafter is better for he who fears Allah. And injustice will not be done to you, [even] as much as a thread [inside a date seed]."

Table 52: Instances of recurrent patterns representing one concept in the Qur'an "Day of Resurrection”

2. One important feature that can be considered is the syntactic relations between words in a single verse as presented in the Quranic Arabic Corpus (QAC). There is a syntactic role to each word in a sentence. Pairs of syntactic units are related through directed binary dependencies. In the Quranic Arabic Corpus, these relations are represented as directed edges like the graph in Figure 37, and morphological annotation is described in Figure 38. The QAC defines different high-level categories of the syntactic relations in the Qur’an, which are used to relate morphological segments, words, phrases and clauses. They are nominal dependencies, verbal dependencies, Phrases and Clauses, Adverbial Dependencies, and Particle Dependencies.

Extracting relations between the Qur’anic verses may be improved using syntactic patterns. Analysing the syntactic structure of a sentence/ verse may help convey the overall meaning delivered by the sentence. For example, the existence of accusative particles and preventive particles can be evidence on the connection between the verses, and sometimes, an indication of the type of the relationship.

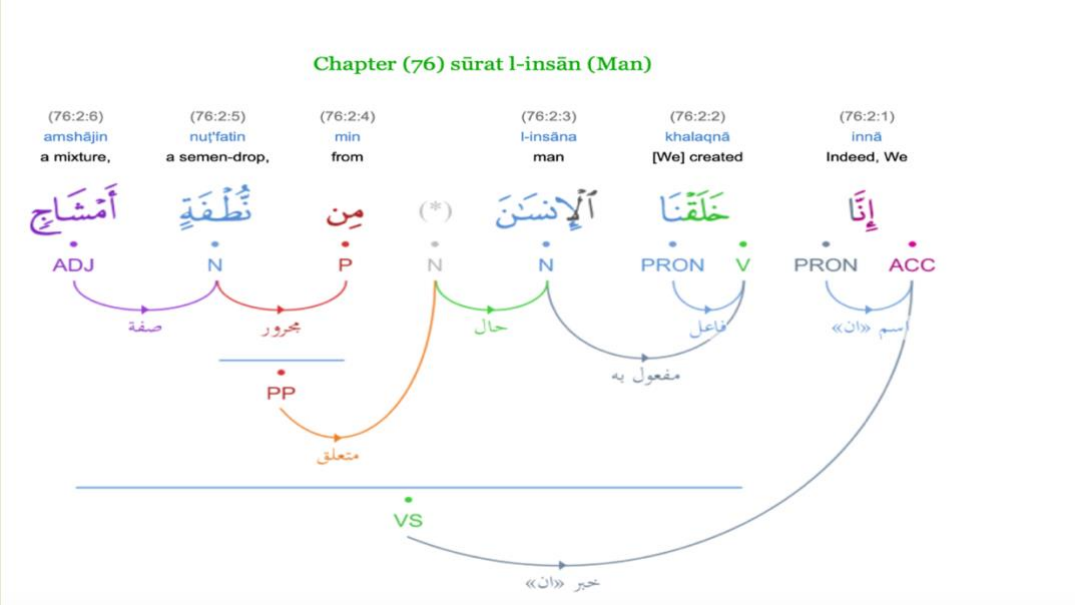


Figure 37: A dependency graph showing syntax roles and morphology for each word in the verse 76:2, as part of the Syntactic Treebank presented by QAC (Dukes, 2003)

Translation	Arabic word	Syntax and morphology
(76:2:1) innā Indeed, We	إِنَّا PRON ACC	ACC – accusative particle PRON – 1st person plural object pronoun حرف نصب و«نا» ضمير متصل في محل نصب اسم «ان»
(76:2:2) khalaqnā [We] created	خَلَقْنَا PRON V	V – 1st person plural perfect verb PRON – subject pronoun فعل ماض و«نا» ضمير متصل في محل رفع فاعل
(76:2:3) l-insāna man	ٱلْإِنسَٰنَ N	N – accusative masculine noun اسم منصوب
(76:2:4) min from	مِن P	P – preposition حرف جر
(76:2:5) nuṭ'fatin a semen-drop,	نُّطْفَةٍ N	N – genitive feminine singular indefinite noun اسم مجرور
(76:2:6) amshājin a mixture,	أَمْشَاجٍ ADJ	ADJ – genitive feminine plural indefinite adjective صفة مجرورة

Figure 38: The Morphological annotation for each word in the verse 76:2, as part of the Syntactic Treebank presented by QAC (Dukes, 2003)

10.8.3 Potential Enhancements on Detecting Pairwise Similarity in the Qur'an

In machine learning, each problem is unique in terms of techniques employed and data attributes. Modelling the semantics in the Qur'an is challenging, and detecting pairwise similarity between its passages is a unique task. In chapter 7, we have created our dataset, and the task itself is unlikely tested before given our data. Therefore, the first step was to look for algorithms that best model our prediction problem. Also, we needed a basis for the comparison of results.

When trying different algorithms, a baseline result can inform us whether a change is bringing value. Once we have established a baseline, we can add or change the data attributes, the algorithms we are testing, or the settings of the algorithms. Eventually, we know whether we have improved our approach to the problem.

In our case, using state-of-the-art results is not possible, and we need to calculate a baseline result. It is a straightforward prediction that could be a random result or, in some cases, the most prevalent prediction. There are many ways to calculate a baseline result for a prediction problem. For example, for a classification problem, one can select the class with the most observations and use that class as the result of all predictions. Here, we have a total of 9315 pairs in the test set. There are 3079 related pairs and 6236 non-related pairs. So, the baseline was computed as (6236/ 9315), which is 0.669 (67%).

Our baseline seems poor; however, it could imply that the task is very complex or that our algorithm has much room for improvement. For example, we can test different settings of the algorithm, try various features of the data, or even try other algorithms.

10.8.4 A Unified Topical Classification from the Qur'an

Potential work is to build a unified and unique topical classification for the Qur'an verses using existing knowledge resources. The initial stage was performed in chapter 7, and we plan to extend the experiment by incorporating subtopic chains

from the Qurany corpus and testing our derived model using them. In addition, we may consider other approaches for computing the semantic similarity, investigate their performance, and how they compare to our approach. Moreover, the derived classification was developed using scholarly sources; therefore, they are likely to be reliable and verified. They must, however, be subjected to additional manual authentication by Qur'anic scholars before being used in larger applications.

References

"Qurʾān: Tradition of Scholarship and Interpretation." Encyclopedia of Religion. Retrieved June 03, 2022 from Encyclopedia.com: https://www.encyclopedia.com/environment/encyclopedias-almanacs-transcripts-and-maps/quran-tradition-scholarship-and-interpretation.

Al-Bataineh, H., Farhan, W., Mustafa, A., Seelawi, H. and Al-Natsheh, H.T., 2019, November. Deep contextualized pairwise semantic similarity for arabic language questions. In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 1586-1591). IEEE.

Al-Kabi, M.N., Ata, B.M.A., Wahsheh, H.A. and Alsmadi, I.M., 2013, December. A topical classification of Quranic Arabic text. In Proceedings of the 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences (pp. 252-257).

Al-Kabi, M.N., Kanaan, G., Al-Shalabi, R., Nahar, K. and Bani-Ismail, B., 2005. Statistical classifier of the holy Quran verses (Fatiha and Yaseen chapters). Journal of Applied Sciences, 5(3), pp.580-583.

Al-Mahmoud, H., and M. Al-Razgan. 2015. "Arabic text mining a systematic review of the published literature 2002-2014." In 2015 International Conference on Cloud Computing (ICCC). IEEE. pp. 1-7.

Al-Sallab, A., Baly, R., Hajj, H., Shaban, K.B., El-Hajj, W. and Badaro, G., 2017. Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 16(4), pp.1-20.

Abbas, N.H. 2009. Quran'search for a concept'tool and website (Doctoral dissertation, University of Leeds (School of Computing)).

Abbas, N., & Atwell, E. 2013, July. Annotating the Arabic Quran with semantic web content tags. In Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics. Leeds.

Abdul-Raof, H. (2013). Schools of Qur'anic exegesis: genesis and development. routledge.

Abu A. Al-Khair, and M. Kabbani. "Koran teacher intonation." The Tunisian Company for Distribution, 2003.

Adeleke, A.O., Samsudin, N.A., Mustapha, A. and Nawi, N.M., 2017. Comparative analysis of text classification algorithms for automated labelling of Quranic verses. International Journal on Advanced Science, Engineering and Information Technology, 7(4), p.1419.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M. and Soroa, A., 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Citeseer, 19.

Agirre, E., Cer, D., Diab, M. and Gonzalez-Agirre, A., 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012) (pp. 385-393).

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A. and Guo, W., 2013, June. * SEM 2013 shared task: Semantic textual similarity. In Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity (pp. 32-43).

Akila, M.D. and Jayakumar, C., 2014. Semantic Similarity-A Review of Approaches and Metrics. International Journal of Applied Engineering Research, 9(24), pp.27581-27600.

Akour, M., Alsmadi, I.M. and Alazzam, I., 2014. MQVC: Measuring quranic verses similarity and sura classification using N-gram. WSEAS Transactions on Computers, vol. 13, pp. 485-491, 2014.

Ali, A.S., 2017. A Brief Introduction to Qur'anic Exegesis. International Institute of Islamic Thought (IIIT).

Alhawarat, M., 2015. Extracting topics from the holy Quran using generative models. International Journal of Advanced Computer Science and Applications, 6(12), pp.288-294.

Alian, M. and Awajan, A., 2018, November. Arabic semantic similarity approaches-review. In 2018 International Arab Conference on Information Technology (ACIT) (pp. 1-6). IEEE.

Alian, M. and Awajan, A., 2020. Semantic similarity for english and arabic texts: a review. Journal of Information & Knowledge Management, 19(04), p.2050033.

Aliguliyev, R.M., 2009. A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Systems with Applications, 36(4), pp.7764-7772.

Alqahtani, M. and Atwell, E., 2016, June. Arabic quranic search tool based on ontology. In International Conference on Applications of Natural Language to Information Systems (pp. 478-485). Springer, Cham.

Alrehaili, S.M., and E. Atwell. 2014. "Computational ontologies for semantic tagging of the Quran: A survey of past approaches." In LREC 2014 Proceedings. European Language Resources Association.

Alsaleh, A.N., Atwell, E. and Altahhan, A., 2021, April. Quranic Verses Semantic Relatedness Using AraBERT. In Proceedings of the Sixth Arabic Natural Language Processing Workshop (pp. 185-190). Leeds.

Alsaleh, A., Althabiti, S., Alshammari, I., Alnefaie, S., Alowaidi, S., Alsaqer, A., Atwell, E., Altahhan, A. and Alsalka, M.A., 2022, June. LK2022 at Qur'an QA 2022: Simple Transformers Model for Finding Answers to Questions from Qur'an. In Proceedings of the OSACT 2022 Workshop (pp. 120-125). ELRA European Language Resources Association.

Alshammeri, M., Atwell, E. and Alsalka, M.A., 2020, September. Quranic Topic Modelling Using Paragraph Vectors. In Proceedings of SAI Intelligent Systems Conference (pp. 218-230). Springer, Cham.

Alshammeri, M., Atwell, E. and Ammar Alsalka, M., 2021. Detecting Semantic-based Similarity Between Verses of The Quran with Doc2vec. Procedia Computer Science, 189, pp.351-358.

Altınel, B. and Ganiz, M.C., 2018. Semantic text classification: A survey of past and recent advances. Information Processing & Management, 54(6), pp.1129-1153.

Altszyler, E., Sigman, M., Ribeiro, S. & Slezak, D. F., 2016. Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. arXiv preprint arXiv:1610.01520.

Anderson, A.J., Kiela, D., Clark, S. and Poesio, M., 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. Transactions of the Association for Computational Linguistics, 5, pp.17-30.

Ando, R.K., Zhang, T. and Bartlett, P., 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research, 6(11).

Andrews, M., Vigliocco, G. and Vinson, D., 2009. Integrating experiential and distributional data to learn semantic representations. Psychological review, 116(3), p.463.

Antoun, W., Baly, F. and Hajj, H., 2020. Arabert: Transformer-based model for arabic language understanding. arXiv preprint arXiv:2003.00104.

Asda, T.M.H., Gunawan, T.S., Kartiwi, M. and Mansor, H., 2016. Development of Quran reciter identification system using MFCC and neural network. Indonesian Journal of Electrical Engineering and Computer Science, 1(1), pp.168-175.

Atwell, Eric. 2018. "Using the Web to Model Modern and Qurʾanic Arabic." Arabic Corpus Linguistics (Edinburgh University Press) 100.

Atwell, E., Brierley, C., Dukes, K., Sawalha, M. and 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 (pp. 1-8). Leeds.

Atwell, E.S., Dickins, J. and Brierley, C., 2013. Natural Language Processing Working Together with Arabic and Islamic Studies. Engineering and Physical Sciences Research Council (EPSRC). EP/K015206/1. Online. Accessed: 29.06.2014. http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/K015206/1.

Atwell, E., Habash, N., Louw, B., Abu Shawar, B., McEnery, T., Zaghouani, W. and El-Haj, M., 2010. Understanding the Quran: A new grand challenge for computer science and artificial intelligence. ACM-BCS Visions of Computer Science 2010.

Alzain, M. (1995). The indexed Dictionary of the meanings in the Qur'an. Available at: https://waqfeya.net/book.php?bid=1393. (Accessed 06 June 2022).

Babić, K., Martinčić-Ipšić, S. and Meštrović, A., 2020. Survey of neural text representation models. Information, 11(11), p.511.

Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Baroni, M. and Lenci, A., 2010. Distributional memory: A general framework for corpus-based semantics. Computational Linguistics, 36(4), pp.673-721.

Baroni, M. and Zamparelli, R., 2010, October. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In Proceedings of the 2010 conference on empirical methods in natural language processing (pp. 1183-1193).

Baroni, M., Bernardi, R. and Zamparelli, R., 2014. Frege in space: A program for compositional distributional semantics. Linguistic Issues in language technology, 9, pp.241-346.

Bashir, M.H., M Azmi, A., Nawaz, H., Zaghouani, W., Diab, M., Al-Fuqaha, A. and Qadir, J., 2021. Arabic Natural Language Processing for Qur'anic Research: A Systematic Review.

Battig, W.F. and Montague, W.E., 1969. Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. Journal of experimental Psychology, 80(3p2), p.1.

Bechara, H., Costa, H., Taslimipoor, S., Gupta, R., Orăsan, C., Pastor, G.C. and Mitkov, R., 2015, June. Miniexperts: An svm approach for measuring semantic textual similarity. In Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015) (pp. 96-101).

Bender, Emily. 2009. Linguistically naive != language independent: Why NLP needs linguistic typology. In Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics, pages 26–32. Athens, Greece.

Bengio, Y., Courville, A. and Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), pp.1798-1828.

Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C., 2003. A neural probabilistic language model. Journal of machine learning research, Feb, Volume 3, pp. 1137-1155.

Bentivogli, L., Bernardi, R., Marelli, M., Menini, S., Baroni, M. and Zamparelli, R., 2016. SICK through the SemEval glasses. Lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. Language Resources and Evaluation, 50(1), pp.95-124.

Blei, D. M., Ng, A. Y. & Jordan, M. I., 2003. Latent Dirichlet allocation. Journal of machine Learning research, 3(Jan), pp. 993--1022.

Blei, D.M. and Jordan, M.I., 2003, July. Modeling annotated data. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 127-134).

Blei, D.M., 2012. Probabilistic topic models. Communications of the ACM, 55(4), pp.77-84.

Blitzer, J., McDonald, R. and Pereira, F., 2006, July. Domain adaptation with structural correspondence learning. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 120-128).

Blutner, R., Hendriks, P. and Hoop, H.D., 2003. A new hypothesis on compositionality.

Bordes, A., Chopra, S. and Weston, J., 2014. Question answering with subgraph embeddings. arXiv preprint arXiv:1406.3676.

Boyd-Graber, J., Blei, D. and Zhu, X., 2007, June. A topic model for word sense disambiguation. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL) (pp. 1024-1033).

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. and Shah, R., 1993. Signature verification using a" siamese" time delay neural network. Advances in neural information processing systems, 6.

Brown, P.F., Della Pietra, V.J., Desouza, P.V., Lai, J.C. and Mercer, R.L., 1992. Class-based n-gram models of natural language. Computational linguistics, 18(4), pp.467-480.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. Advances in neural information processing systems, 33, pp.1877-1901.

Brownlee, J., 2017. Deep learning for natural language processing. Machine Learning Mystery, Vermont, Australia, 322.

Brownlee, J., 2019. Deep learning for computer vision: image classification, object detection, and face recognition in Python. Machine Learning Mastery.

Bullinaria, J.A. and Levy, J.P., 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior research methods, 39(3), pp.510-526.

Bullinaria, J.A. and Levy, J.P., 2012. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. Behavior research methods, 44(3), pp.890-907.

Camacho-Collados, J. and Pilehvar, M.T., 2018. From word to sense embeddings: A survey on vector representations of meaning. Journal of Artificial Intelligence Research, 63, pp.743-788.

Cambria, E. and White, B., 2014. Jumping NLP curves: A review of natural language processing research. IEEE Computational intelligence magazine, 9(2), pp.48-57.

Cer, D., Yang, Y., Kong, S.Y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C. and Sung, Y.H., 2018. Universal sentence encoder. arXiv preprint arXiv:1803.11175.

Chandrasekaran, D. and Mago, V., 2021. Evolution of semantic similarity—a survey. ACM Computing Surveys (CSUR), 54(2), pp.1-37.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. and Blei, D., 2009. Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems, 22.

Chen, Y., 2015. Convolutional neural network for sentence classification (Master's thesis, University of Waterloo).

Chiu, B., Crichton, G., Korhonen, A. and Pyysalo, S., 2016, August. How to train good word embeddings for biomedical NLP. In Proceedings of the 15th workshop on biomedical natural language processing (pp. 166-174).

Chomsky, N. 1957. Aspect of Syntax Theory. Cambridge, MA: MIT Press.

Chopra, S., Hadsell, R. and LeCun, Y., 2005, June. Learning a similarity metric discriminatively, with application to face verification. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) (Vol. 1, pp. 539-546). IEEE.

Clark, S., Coecke, B. and Sadrzadeh, M., 2008. A compositional distributional model of meaning. In Proceedings of the Second Quantum Interaction Symposium (QI-2008) (pp. 133-140).

Collins, A. M. & Quillian, M. R., 1969. Retrieval time from semantic memory. Journal of verbal learning and verbal behavior, 8(2), pp. 240--247.

Collobert, R. and Weston, J., 2008, July. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning (pp. 160-167).

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P., 2011. Natural language processing (almost) from scratch. Journal of machine learning research, 12(ARTICLE), pp.2493-2537.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L. and Bordes, A., 2017. Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364.

Cook, M., 2000. The Koran: A very short introduction. OUP Oxford.

Dagan, I., Roth, D., Sammons, M. and Zanzotto, F.M., 2013. Recognizing textual entailment: Models and applications. Synthesis Lectures on Human Language Technologies, 6(4), pp.1-220.

Dai, A.M. and Le, Q.V., 2015. Semi-supervised sequence learning. Advances in neural information processing systems, 28.

Dai, A.M., Olah, C. and Le, Q.V., 2015. Document embedding with paragraph vectors. arXiv preprint arXiv:1507.07998.

Deng, L. and Liu, Y. eds., 2018. Deep learning in natural language processing. Springer.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. arXiv preprint arXiv:1810.04805.

Dieng, A.B., Ruiz, F.J. and Blei, D.M., 2020. Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics, 8, pp.439-453.

Dinu, G. and Lapata, M., 2010, August. Topic models for meaning similarity in context. In Coling 2010: Posters (pp. 250-258).

Dukes, K. (2009- 2011). The Quranic Arabic Corpus. 2009-2011; Available at: https://corpus.quran.com/. (Accessed 2022 06 June).

Dukes, K., E. Atwell, and N. Habash. 2013. "Supervised Collaboration for Syntactic Annotation of Quranic Arabic." Language resources and evaluation (Springer) 47 (1): pp.33-62.

DURAKOVIC, E. (2005). Stylistic Potentials of the Elative in the Qur'ān. Islamic studies, 44(3), 313-325.

El-Deeb, R., Al-Zoghby, A.M. and Elmougy, S., 2018. Multi-corpus-based model for measuring the semantic relatedness in short texts (SRST). Arabian Journal for Science and Engineering, 43(12), pp.7933-7943.

Edunov, S., Baevski, A. and Auli, M., 2019. Pre-trained language model representations for language generation. arXiv preprint arXiv:1903.09722.

Fabre, C. and Lenci, A., 2015. Distributional semantics today.

Farghaly, A., and K. Shaalan. 2009. "Arabic natural language processing: Challenges and solutions." ACM Transactions on Asian Language Information Processing (TALIP) (ACM New York, NY, USA) 8 (4):

pp.1-22.

Farghaly, A., 2010. The Arabic language, Arabic linguistics and Arabic computational linguistics. Arabic computational linguistics, pp.43-81.

Fedus, W., Goodfellow, I. and Dai, A.M., 2018. Maskgan: better text generation via filling in the_. arXiv preprint arXiv:1801.07736.

Feldman, R. and Sanger, J., 2007. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press.

Feng, W., Wu, Y., Wu, W., Li, Z. and Zhou, M., 2017, August. Beihang-msra at semeval-2017 task 3: A ranking system with neural matching features for community question answering. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 280-286).

Ferrone, L. and Zanzotto, F.M., 2020. Symbolic, distributed, and distributional representations for natural language processing in the era of deep learning: A survey. Frontiers in Robotics and AI, p.153.

Fernando, S. and Stevenson, M., 2008, March. A semantic similarity approach to paraphrase detection. In Proceedings of the 11th annual research colloquium of the UK special interest group for computational linguistics (pp. 45-52).

Filice, S., Da San Martino, G. and Moschitti, A., 2017, August. Kelp at semeval-2017 task 3: Learning pairwise patterns in community question answering. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 326-333).

Frege, G., 1884. Die Grundlagen der Arithmetik (‘Foundations of Arithmetic’). Breslau, Germany: Wilhelm Koebner.

Gan, Z., Pu, Y., Henao, R., Li, C., He, X. and Carin, L., 2016. Learning generic sentence representations using convolutional neural networks. arXiv preprint arXiv:1611.07897.

Ge, J., 2018. Measuring Short Text Semantic Similarity with Deep Learning Models.

Gelder, V., 1990. “Compositionality: A Connectionist Variation on a Classical Theme”. Cognitive Science, (14), pp.355-3.

Goldberg, Y., 2016. A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research, 57, pp.345-420.

Goldberg, Y. 2017. Neural network methods in natural language processing. Morgan & Claypool Publishers.

Gómez-Adorno, H., Pinto, D. and Vilarino, D., 2013, June. A question answering system for reading comprehension tests. In Mexican Conference on Pattern Recognition (pp. 354-363). Springer, Berlin, Heidelberg.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. Deep learning. MIT press.

Grefenstette, E. and Sadrzadeh, M., 2011. Experimental support for a categorical compositional distributional model of meaning. arXiv preprint arXiv:1106.4058.

Grefenstette, E., Dinu, G., Zhang, Y.Z., Sadrzadeh, M. and Baroni, M., 2013. Multi-step regression learning for compositional distributional semantics. arXiv preprint arXiv:1301.6939.

Griffiths, T.L. and Steyvers, M., 2004. Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl_1), pp.5228-5235.

Griffiths, T. L., Steyvers, M. & Tenenbaum, J. B., 2007. Topics in semantic representation. Psychological review, 114(2), p. 211.

Guellil, I., Saâdane, H., Azouaou, F., Gueni, B. and Nouvel, D., 2021. Arabic natural language processing: An overview. Journal of King Saud University-Computer and Information Sciences, 33(5), pp.497-507.

Habash, N.Y., 2010. Introduction to Arabic natural language processing. Synthesis lectures on human language technologies, 3(1), pp.1-187.

Hadj Taieb, M.A., Zesch, T. and Ben Aouicha, M., 2020. A survey of semantic relatedness evaluation datasets and procedures. Artificial Intelligence Review, 53(6), pp.4407-4448.

Haleem, M.A. and Haleem, M.A., 2010. Understanding the Qur'an: themes and style. Bloomsbury Publishing.

Hamed, S.K. and Ab Aziz, M.J., 2016. A Question Answering System on Holy Quran Translation Based on Question Expansion Technique and Neural Network Classification. J. Comput. Sci., 12(3), pp.169-177.

Hamed, S.K. and Ab Aziz, M.J., 2018. Classification of holy Quran translation using neural network technique. Journal of Engineering and Applied Sciences, 13(12), pp.4468-4475.

Harris Zellig S. 1954. Distributional structure. Word, X/2-3, 1954. 146-62 [reprinted in Harris Zellig S. 1970. Papers in Structural andTransformational Linguistics. Dordrecht: Reidel. 775-794].

Hill, F., Cho, K. and Korhonen, A., 2016. Learning distributed representations of sentences from unlabelled data. arXiv preprint arXiv:1602.03483.

Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. 1986. “Distributed representations,” in Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations, eds D. E. Rumelhart and J. L. McClelland (Cambridge, MA: MIT Press), 77–109.

Hinton, G.E. and Shallice, T., 1991. Lesioning an attractor network: investigations of acquired dyslexia. Psychological review, 98(1), p.74.

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.

Hofmann, T., 1999, August. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57).

Howard, J. and Ruder, S., 2018. Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

Hu, B., Lu, Z., Li, H. and Chen, Q., 2014. Convolutional neural network architectures for matching natural language sentences. Advances in neural information processing systems, 27.

Huang, P.S., He, X., Gao, J., Deng, L., Acero, A. and Heck, L., 2013, October. Learning deep structured semantic models for web search using clickthrough data. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (pp. 2333-2338).

Husain, K.M., 1983. "Islam and Modernity" by Fazlur Rahman (Book Review). Arab Studies Quarterly, 5(3), p.306.

Ichida, A.Y., Meneguzzi, F. and Ruiz, D.D., 2018, July. Measuring semantic similarity between sentences using a siamese neural network. In 2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.

Iyyer, M., Manjunatha, V., Boyd-Graber, J. & Daumé III, H., 2015. Deep unordered composition rivals syntactic methods for text classification. s.l., s.n., pp. 1681-1691.

Jernite, Y., Bowman, S.R. and Sontag, D., 2017. Discourse-based objectives for fast unsupervised sentence representation learning. arXiv preprint arXiv:1705.00557.

Jiao, X., Wang, F. and Feng, D., 2018, August. Convolutional neural network for universal sentence embeddings. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 2470-2481).

Jones, K.S., 2007. Last words: computational linguistics: what about the linguistics? Computational linguistics, 33(3), pp.437-441.

Jones MN, Kintsch W, Mewhort DJK. 2006. High-dimensional semantic space accounts of priming. J. Mem. Lang. 55:534–52.

Kay, Martin. 1973. The mind system. In R. Randall, ed., Natural Language Processing, pages 155–188. New York, NY: Algorithmic Press.

Kalchbrenner, N. and Blunsom, P., 2013. Recurrent convolutional neural networks for discourse compositionality. arXiv preprint arXiv:1306.3584.

Kim, Y., 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Kim, S., Fiorini, N., Wilbur, W.J. and Lu, Z., 2017. Bridging the gap: Incorporating a semantic similarity measure for effectively mapping PubMed queries to documents. Journal of biomedical informatics, 75, pp.122-127.

Kim, Y., Jernite, Y., Sontag, D. and Rush, A.M., 2016, March. Character-aware neural language models. In Thirtieth AAAI conference on artificial intelligence.

Kalchbrenner, N., Grefenstette, E. and Blunsom, P., 2014. A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188.

Kintsch, W., 2001. Predication. Cognitive science, 25(2), pp.173-202.

Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A. and Fidler, S., 2015. Skip-thought vectors. Advances in neural information processing systems, 28.

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

Lacoste-Julien, S., Sha, F. and Jordan, M., 2008. DiscLDA: Discriminative learning for dimensionality reduction and classification. Advances in neural information processing systems, 21.

Lai, G., Xie, Q., Liu, H., Yang, Y. and Hovy, E., 2017. Race: Large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683.

Lai, S., Liu, K., He, S. and Zhao, J., 2016. How to generate a good word embedding. IEEE Intelligent Systems, 31(6), pp.5-14.

Landauer, T.K. and Dumais, S.T., 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2), p.211.

Lastra-Díaz, J.J., Goikoetxea, J., Taieb, M.A.H., García-Serrano, A., Aouicha, M.B. and Agirre, E., 2019. A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art. Engineering Applications of Artificial Intelligence, 85, pp.645-665.

Le, Q. and Mikolov, T., 2014, June. Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. nature, 521(7553), pp.436-444.

Lenci, A. 2008. Distributional approaches in linguistic and cognitive research. Italian Journal of Linguistics20(1). 1–31.

Lin, Z., Feng, M., Santos, C.N.D., Yu, M., Xiang, B., Zhou, B. and Bengio, Y., 2017. A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130.

Liou, C.Y., Cheng, W.C., Liou, J.W. and Liou, D.R., 2014. Autoencoder for words. Neurocomputing, 139, pp.84-96.

Lofi C. 2015. Measuring semantic similarity and relatedness with distributional and knowledge-based approaches. Inf. Media Technol. 10:493–501.

Logeswaran, L. and Lee, H., 2018. An efficient framework for learning sentence representations. arXiv preprint arXiv:1803.02893.

Lopez-Gazpio, I., Maritxalar, M., Gonzalez-Agirre, A., Rigau, G., Uria, L. and Agirre, E., 2017. Interpretable semantic textual similarity: Finding and explaining differences between sentences. Knowledge-Based Systems, 119, pp.186-199.

Lu, Y., Mei, Q. and Zhai, C., 2011. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. Information Retrieval, 14, pp.178-203.

Majumder, G., Pakray, P., Gelbukh, A. and Pinto, D., 2016. Semantic textual similarity methods, tools, and applications: A survey. Computación y Sistemas, 20(4), pp.647-665.

Mandera P, Keuleers E, Brysbaert M. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: a review and empirical validation. J. Mem. Lang. 92:57–78.

Mahmoud, A., Zrigui, A. and Zrigui, M., 2017, April. A text semantic similarity approach for Arabic paraphrase detection. In International conference on computational linguistics and intelligent text processing (pp. 338-349). Springer, Cham.

Mahmoud, A. and Zrigui, M., 2017, November. Semantic similarity analysis for paraphrase identification in Arabic texts. In Proceedings of the 31st Pacific Asia conference on language, information and computation (pp. 274-281).

Manning, Christopher and Hinrich Schuetze. 1999. Foundations of Statistical Natural Language Processing. Massachusetts: MIT Press.

Magueresse, A., Carles, V. and Heetderks, E., 2020. Low-resource languages: A review of past work and future challenges. arXiv preprint arXiv:2006.07264.

McRae, K., De Sa, V.R. and Seidenberg, M.S., 1997. On the nature and scope of featural representations of word meaning. Journal of Experimental Psychology: General, 126(2), p.99.

Mei, Q., Ling, X., Wondra, M., Su, H. and Zhai, C., 2007, May. Topic sentiment mixture: modeling facets and opinions in weblogs. In Proceedings of the 16th international conference on World Wide Web (pp. 171-180).

Melamud, O., Goldberger, J. and Dagan, I., 2016, August. context2vec: Learning generic context embedding with bidirectional lstm. In Proceedings of the 20th SIGNLL conference on computational natural language learning (pp. 51-61).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013a. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26, 3111-3119.

Mikolov, T., Chen, K., Corrado, G. & Dean, J., 2013b. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Yih, W.T. and Zweig, G., 2013c, June. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: Human language technologies (pp. 746-751).

Mitchell, J. and Lapata, M., 2008, June. Vector-based models of semantic composition. In proceedings of ACL (pp. 236-244).

Mitchell, J. &. L. M., 2010. Composition in distributional models of semantics. Cognitive science, 34(8), pp. 1388-1429.

Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, et al. 2008. Predicting human brain activity associated with the meanings of nouns. Science 320:1191–95.

Mnih, A. and Hinton, G.E., 2008. A scalable hierarchical distributed language model. Advances in neural information processing systems, 21.

Mohamed, R., Ragab, M., Abdelnasser, H., El-Makky, N.M. and Torki, M., 2015, June. Al-Bayan: A knowledge-based system for Arabic answer selection. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (pp. 226-230).

Montague, R., 1974. English as a Formal Language. Reprinted in Formal Philosophy.

Mostafazadeh, N., Roth, M., Louis, A., Chambers, N. and Allen, J., 2017, April. Lsdsem 2017 shared task: The story cloze test. In Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics(pp. 46-51).

Mueller, J. and Thyagarajan, A., 2016, March. Siamese recurrent architectures for learning sentence similarity. In Proceedings of the AAAI conference on artificial intelligence (Vol. 30, No. 1).

Muneeb, T.H., Sahu, S. and Anand, A., 2015, July. Evaluating distributed word representations for capturing semantics of biomedical concepts. In Proceedings of BioNLP 15 (pp. 158-163).

Nadkarni, P.M., Ohno-Machado, L. and Chapman, W.W., 2011. Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), pp.544-551.

Neculoiu, P., Versteegh, M. and Rotaru, M., 2016, August. Learning text similarity with siamese recurrent networks. In Proceedings of the 1st Workshop on Representation Learning for NLP (pp. 148-157).

Nordström, B., Ranta, A., (Eds.), GoTAL 2008: 6th International Conference on Natural Language Processing, Gothenburg, Sweden, August 25-27, Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence (LNCS/LNAI): Advances in Natural Language Proceedings, Vol. 5221, PP. 440-451, Springer-Verlag, Berlin, Germany, 2008.

Norman, D. A., 1973. Memory, knowledge, and the answering of questions. In R. L. Solso (Ed.), Contemporary issues in cognitive psychology: The Loyola Symposium (pp. 135–165). Washington, DC: Winston.

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

Oepen, Stephan et al. 2000. Introduction. Natural Language Engineering 6:1–14.

Olivas, E.S., Guerrero, J.D.M., Martinez-Sober, M., Magdalena-Benedito, J.R. and Serrano, L. eds., 2009. Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques: Algorithms, methods, and techniques. IGI Global.

Ouda, K., 2015. QuranAnalysis: A semantic search and intelligence system for the Quran. UK: Leeds University.

Pagliardini, M., Gupta, P. and Jaggi, M., 2017. Unsupervised learning of sentence embeddings using compositional n-gram features. arXiv preprint arXiv:1703.02507.

Pang, L., Lan, Y., Guo, J., Xu, J., Wan, S. and Cheng, X., 2016, March. Text matching as image recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1).

Panju, M.H., 2014. Statistical extraction and visualization of topics in the Qur'an corpus. Student. Math. Uwaterloo. Ca.

Panjwani, F., 2012. Fazlur Rahman and the search for authentic Islamic education: A critical appreciation. Curriculum Inquiry, 42(1), pp.33-55.

Patel, M., Bullinaria, J.A. and Levy, J.P., 1998. Extracting semantic representations from large text corpora. In 4th Neural Computation and Psychology Workshop, London, 9–11 April 1997 (pp. 199-212). Springer, London.

Pennington, J., R. Socher, and C.D. Manning. 2014. "Glove: Global vectors for word representation." In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. (EMNLP). pp.

1532-1543.

Pereira, Fernando and D. Warren. 1980. Definite clause grammars for language analysis. Artificial Intelligence 13:231–278.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Phang, J., Févry, T. and Bowman, S.R., 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. arXiv preprint arXiv:1811.01088.

Plate, T.A., 1995. Holographic reduced representations. IEEE Transactions on Neural networks, 6(3), pp.623-641.

Putra, S.J., Mantoro, T. and Gunawan, M.N., 2017, November. Text mining for Indonesian translation of the Quran: A systematic review. In 2017 International Conference on Computing, Engineering, and Design (ICCED) (pp. 1-5). IEEE.

Qahl, S.H.M., 2014. An automatic similarity detection engine between sacred texts using text mining and similarity measures. Rochester Institute of Technology.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding with unsupervised learning. OpenAI.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I., 2022. Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356.

Rahman, Fazlur. Major Themes of the Qur'an. Minneapolis: Bibliotheca Islamica, 1980; repr. 1989, 1994.

Rahman, F., 1985. My belief in action. The courage of conviction, pp.153-159.

Rahman, F., 2009. Major Themes of the Qur'an. University of Chicago Press.

Radovanović, M., & Ivanović, M. (2008). Text mining: Approaches and applications. Novi Sad J. Math, 38(3), 227-234.

Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P., 2016. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Ranasinghe, T., Orăsan, C. and Mitkov, R., 2019, September. Semantic textual similarity with siamese neural networks. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019) (pp. 1004-1011).

Reimers, N., Beyer, P. and Gurevych, I., 2016, December. Task-oriented intrinsic evaluation of semantic textual similarity. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers(pp. 87-96).

Reimers, N. and Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Reiter, Ehud. 2007. The Shrinking Horizon of Computational Linguistics. Computational Linguistics 33(2):283–287.

Rimell, L., Maillard, J., Polajnar, T. and Clark, S., 2016. RELPRON: A relative clause evaluation data set for compositional distributional semantics. Computational Linguistics, 42(4), pp.661-701.

Riordan B, Jones MN. 2011. Redundancy in perceptual and linguistic experience: comparing feature-based and distributional models of semantic representation. Top. Cogn. Sci. 3:303–45.

Ritter, A. and Etzioni, O., 2010, July. A latent dirichlet allocation method for selectional preferences. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 424-434).

Rogers, A., Kovaleva, O. and Rumshisky, A., 2020. A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics, 8, pp.842-866.

Rolliawati, D., Rozas, I. and Khalid, K., 2020, May. Text Mining Approach for Topic Modeling of Corpus Al Qur'an in Indonesian Translation. In Proceedings of the 2nd International Conference on Quran and Hadith Studies Information Technology and Media in Conjunction

with the 1st International Conference on Islam, Science and Technology, ICONQUHAS & ICONIST, Bandung, October 2-4, 2018, Indonesia.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, pp.53-65.

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6), p.386.

Saeed, A., 2004. Fazlur Rahman: A Framework for Interpreting the Ethico-Legal Content of the Qur'an Dalam Suha Taji-Farouki. Modern Muslim Intellectuals and the Qur'an, pp.37-66.

Saeed, A., 2008. The Qur'an: an introduction. Routledge.

Sahlgren, M., 2008. The distributional hypothesis. Italian Journal of Disability Studies, 20, pp.33-53.

Salloum, W. and Habash, N., 2014. ADAM: Analyzer for dialectal Arabic morphology. Journal of King Saud University-Computer and Information Sciences, 26(4), pp.372-378.

Salton, G., Wong, A. and Yang, C.S., 1975. A vector space model for automatic indexing. Communications of the ACM, 18(11), pp.613-620.

Salton, G., 1989. Automatic text processing: The transformation, analysis, and retrieval of. Reading: Addison-Wesley, 169.

Sang, E.F. and De Meulder, F., 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050.

Schwab, D., 2017, April. Semantic similarity of Arabic sentences with word embeddings. In Third Arabic natural language processing workshop (pp. 18-24).

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural networks, 61, pp.85-117.

Schuster, M. and Paliwal, K.K., 1997. Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 45(11), pp. 2673-2681.

Schwab, D., 2017, April. Semantic similarity of arabic sentences with word embeddings. In Third arabic natural language processing workshop (pp. 18-24).

Shaalan, K., and Raza, H, 2009. NERA: Named Entity Recognition for Arabic, the Journal of the American Society for Information Science and Technology (JASIST), John Wiley & Sons, Inc., NJ, USA, 60, 7,1–12.

Shao, Y., 2017, August. 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) (pp. 130-133).

Sharaf, A.B.M., and E. Atwell. 2012a. "QurAna: Corpus of the Quran annotated with Pronominal Anaphora." In LREC. Citeseer. pp. 130-137.

Sharaf, A.B.M., and E. Atwell. 2012a. "QurAna: Corpus of the Quran annotated with Pronominal Anaphora." In LREC. Citeseer. pp. 130-137.

Sharaf, A.B.M., and E. Atwell. 2012b. "QurSim: A corpus for evaluation of relatedness in short texts." In LREC. pp. 2295-2302.

Severyn, A. and Moschitti, A., 2015, August. Learning to rank short text pairs with convolutional deep neural networks. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval (pp. 373-382).

Shieber, Stuart. 1986. Introduction to Unification-based Approaches to Grammar. CSLI.

Shoufan, A. and Alameri, S., 2015, July. Natural language processing for dialectical Arabic: A Survey. In Proceedings of the second workshop on Arabic natural language processing (pp. 36-48).

Siddiqui, M.A., Faraz, S.M. and Sattar, S.A., 2013, December. Discovering the thematic structure of the Quran using probabilistic topic model. In 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences (pp. 234-239). IEEE.

Smith, E.E. and Medin, D.L., 1981. Categories and concepts (Vol. 9). Cambridge, MA: Harvard University Press.

Socher, R., Huang, E., Pennin, J., Manning, C.D. and Ng, A., 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. Advances in neural information processing systems, 24.

Socher, R., Huval, B., Manning, C.D. and Ng, A.Y., 2012, July. Semantic compositionality through recursive matrix-vector spaces. In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (pp. 1201-1211).

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y. and Potts, C., 2013, October. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).

Srinivasa-Desikan, B., 2018. Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing Ltd.

Ta'a, A., Abdullah, M.S., Ali, A.B.M. and Ahmad, M., 2014, October. Themes-based classification for Al-Quran knowledge ontology. In 2014 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 89-94). IEEE.

Tai, K. S., Socher, R. & Manning, C. D., 2015. Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075.

Tan, M., Dos Santos, C., Xiang, B. and Zhou, B., 2016, August. Improved representation learning for question answer matching. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 464-473).

Tan, C., Wei, F., Wang, W., Lv, W. and Zhou, M., 2018, January. Multiway Attention Networks for Modeling Sentence Pairs. In IJCAI (pp. 4411-4417).

Titov, I. and McDonald, R., 2008, June. A joint model of text and aspect ratings for sentiment summarization. In proceedings of ACL-08: HLT (pp. 308-316).

Torrey, L. and Shavlik, J., 2010. Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques (pp. 242-264). IGI global.

Turian, J., Ratinov, L. and Bengio, Y., 2010, July. Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 384-394).

Turney, P.D. and Pantel, P., 2010. From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 37, pp.141-188.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

Verspoor, K. and MacKinlay, A., 2012. Simple similarity-based question answering strategies for biomedical text.

Vinyals, O., Kaiser, L. u., Koo, T., Petrov, S., Sutskever, I., and Hinton, G. (2015a). “Grammar as a foreign language,” in Advances in Neural Information Processing Systems 28, eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Montreal, QC: Curran Associates, Inc.), 2755–2763.

Vinyals, O., Toshev, A., Bengio, S. and Erhan, D., 2015b. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S.R., 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

Wang, S., Zhang, J. & Zong, C., 2016. Learning sentence representation with guidance of human attention. arXiv preprint arXiv:1609.09189.

Wang, T., Yuan, X. and Trischler, A., 2017. A joint model for question answering and question generation. arXiv preprint arXiv:1706.01450.

Wang, Z., Hamza, W. and Florian, R., 2017. Bilateral multi-perspective matching for natural language sentences. arXiv preprint arXiv:1702.03814.

Wei, X. and Croft, W.B., 2006, August. LDA-based document models for ad-hoc retrieval. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 178-185).

Weiss, D., Alberti, C., Collins, M. and Petrov, S., 2015. Structured training for neural network transition-based parsing. arXiv preprint arXiv:1506.06158.

Weiss, K., Khoshgoftaar, T.M. and Wang, D., 2016. A survey of transfer learning. Journal of Big data, 3(1), pp.1-40.

Werbos, P., 1974. Beyond regression:" new tools for prediction and analysis in the behavioral sciences. Ph. D. dissertation, Harvard University.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang, K.J., 1989. Phoneme recognition using time-delay neural networks. IEEE transactions on acoustics, speech, and signal processing, 37(3), pp.328-339.

Wielandt, R., 2002. Exegesis of the Qur'an: Early modern and contemporary. Encyclopaedia of the Qur'an: EQ, pp.124-142.

Wieting, J., Bansal, M., Gimpel, K. and Livescu, K., 2015. Towards universal paraphrastic sentence embeddings. arXiv preprint arXiv:1511.08198.

Wieting, J. and Gimpel, K., 2017. Revisiting recurrent networks for paraphrastic sentence embeddings. arXiv preprint arXiv:1705.00364.

Williams, A., Nangia, N. and Bowman, S.R., 2017. A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426.

Wu, G., Sheng, Y., Lan, M. and Wu, Y., 2017, August. ECNU at semEval-2017 task 3: using traditional and deep learning methods to address community question answering task. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 365-369).

Xu, D. and Tian, Y., 2015. A comprehensive survey of clustering algorithms. Annals of Data Sciences, 2, pp. 165-193.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y., 2015, June. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.

Yi, X. and Allan, J., 2009. A comparative study of utilizing topic models for information retrieval. In Advances in Information Retrieval: 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings 31 (pp. 29-41). Springer Berlin Heidelberg.

Yin, W., Schütze, H., Xiang, B. and Zhou, B., 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. Transactions of the Association for Computational Linguistics, 4, pp.259-272.

Young, T., Hazarika, D., Poria, S. and Cambria, E., 2018. Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 13(3), pp.55-75.

Yu, M. & Dredze, M., 2015. Learning composition models for phrase embeddings.

Transactions of the Association for Computational Linguistics, Volume 3, pp. 227-242.

Zaenen, Annie. 2006. Mark-up Barking Up the Wrong Tree. Computational Linguistics 32(4):557–580.

Zanzotto, F.M., Korkontzelos, I., Fallucchi, F. and Manandhar, S., 2010. Estimating linear models for compositional distributional semantics. In International conference on computational linguistics (COLING).

Zeiler, M.D. and Fergus, R., 2014, September. Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.

Zhao, J., Zhu, T. and Lan, M., 2014, August. ECNU: One Stone Two Birds: Ensemble of Heterogenous Measures for Semantic Relatedness and Textual Entailment. In SemEval@ COLING (pp. 271-277).

Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P. and Irving, G., 2019. Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.

Zou, W.Y., Socher, R., Cer, D. and Manning, C.D., 2013, October. Bilingual word embeddings for phrase-based machine translation. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1393-1398).

Resources imported from external websites:

Sarkar, Dipanjan. (2018). A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning. Available at: https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a. (Accessed: 06 June 2022).

Goergen, Adrian. (2022). What is transfer Learning and Why Does it Matter? Available at: https://levity.ai/blog/what-is-transfer-learning. (Accessed: 06 June 2022).

7wData. (2021). Natural Language Processing: Taking your Business to the Next Level. Available at: https://7wdata.be/article-general/natural-language-processing-taking-your-business-to-the-next-level/ . (Accessed: 06 June 2022).

Xantaro, Deutschland & GmbH. (2018). Artificial Intelligence, Machine Learning and Deep Learning? Available at: https://www.xantaro.net/en/tech-blogs/machine-and-deep-learning-cybersecurity/ . (Accessed 06 June 2022).

List of Abbreviations

DS – Distributional Semantics

DSMs – Distributional Semantics Models

CA – Classical Arabic

MSA – Modern Standard Arabic

DA – Dialect Arabic

NLP – Natural Language Processing

DL – Deep Learning

ML – Machine Learning

LDA – Latent Dirichlet Allocation

LM – Language Modelling

STS – Semantic Textual Similarity

RNN – Recurrent Neural Networks

LSTM – Long-Short Term Memory

BiLSTM – Bidirectional LSTM

CNN – Convolutional Neural Networks

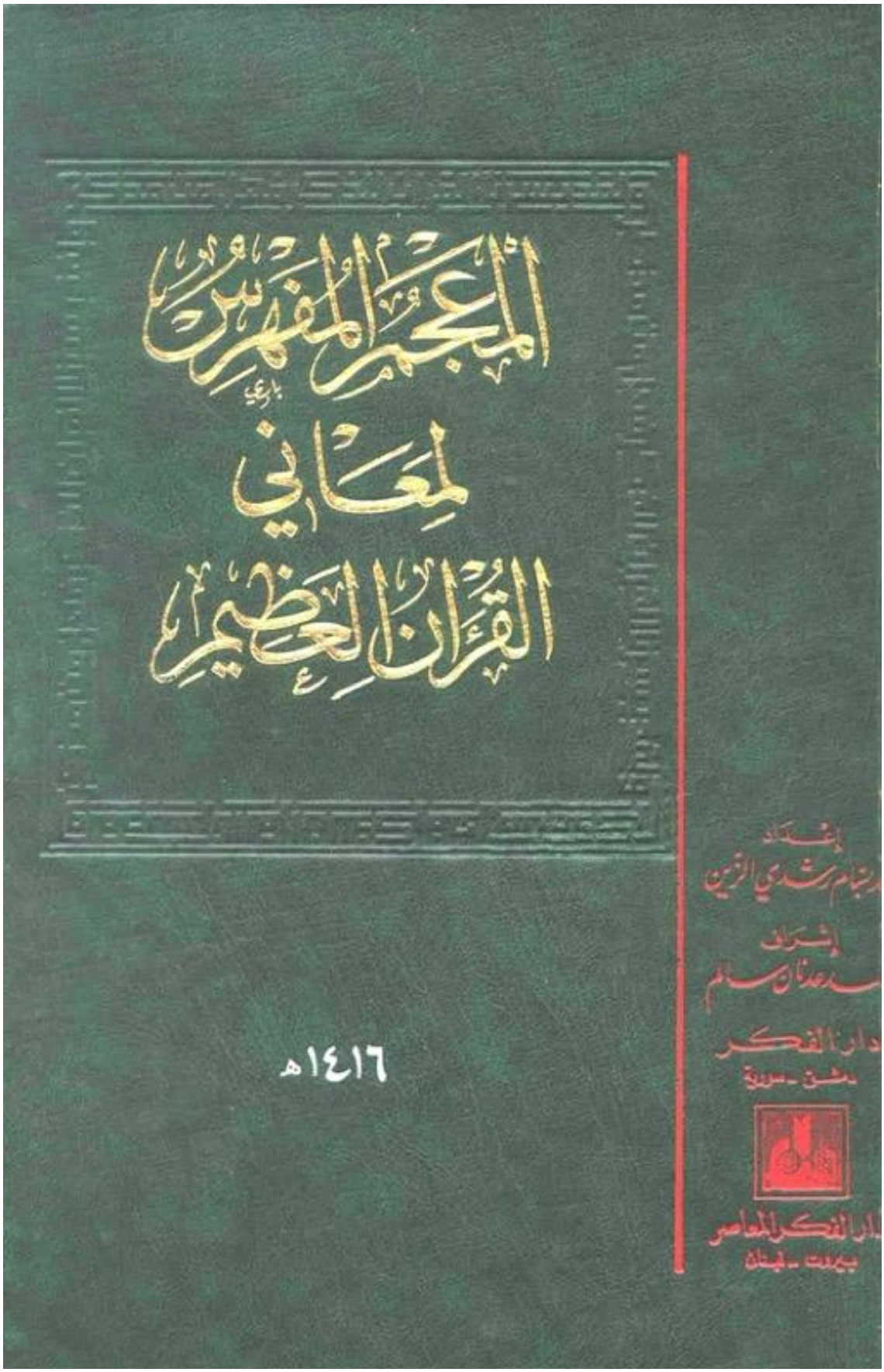
BERT - Bidirectional Encoder Representations from Transformers

AraBERT – Arabic pretrained language model based on BERT architecture

QAC – Qur’anic Arabic Corpus

Qursim – Corpus of Qur’anic Similar/Related Verses

Appendix A


المعجم المفهرس
لمعاني
القرآن العظيم
إعداد
د. بسام رشدي الزين
إشراف
د. عدنان سالم
دار الفكر
دمشق - سورية
دار الفكر المعاصر
بيروت - لبنان
١٤١٦هـ

المعجم المفهرس لمعاني القرآن العظيم

باري

﴿ولقد يسرنا القرآن للذكر
فهل من مدّكر؟!﴾
[القمر ٥٤/١٧]

المجلد الأول
أ ـ ص

إعداد
محمد بسام رشدي الزين

إشراف
محمد عدنان سالم

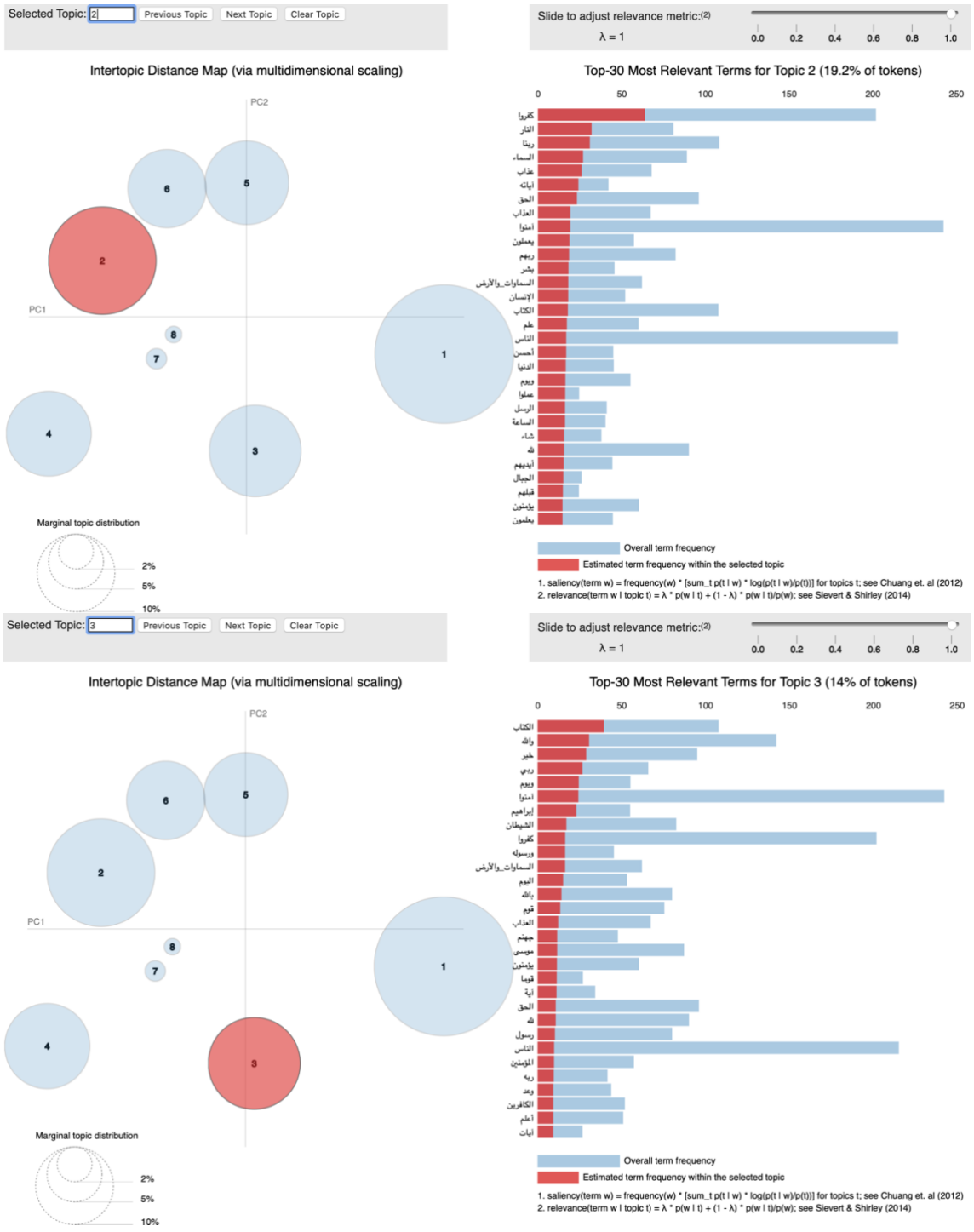
دار الفكر المعاصر
بيروت ـ لبنان

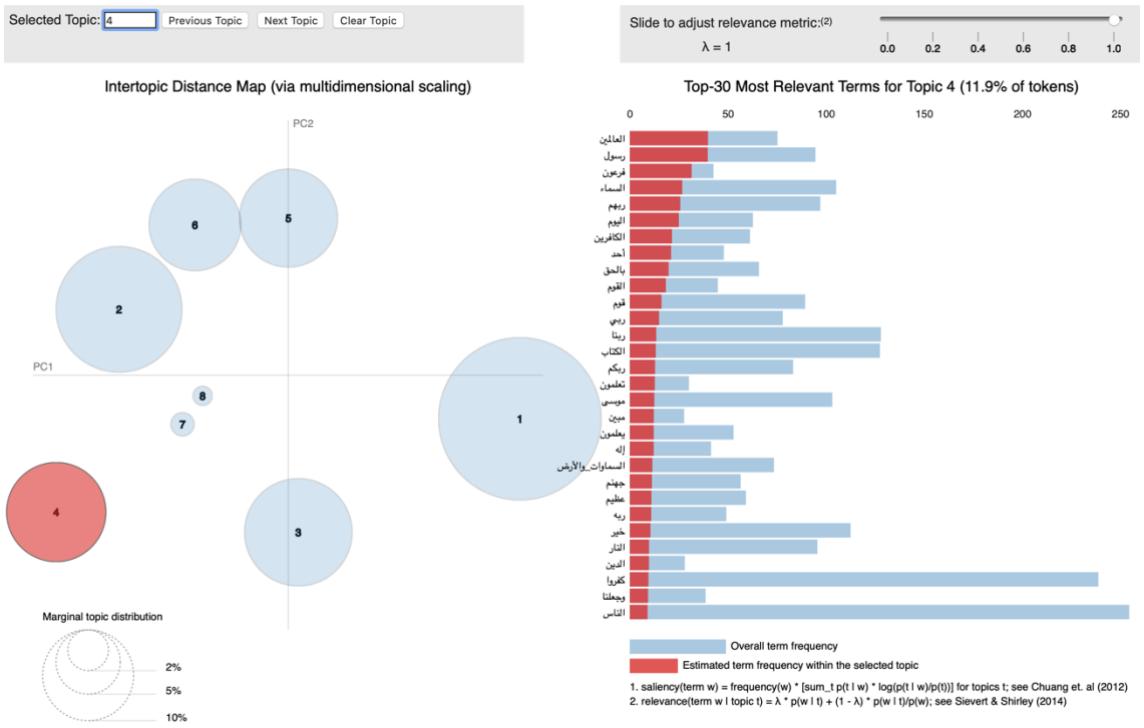
دار الفكر
دمشق ـ سورية

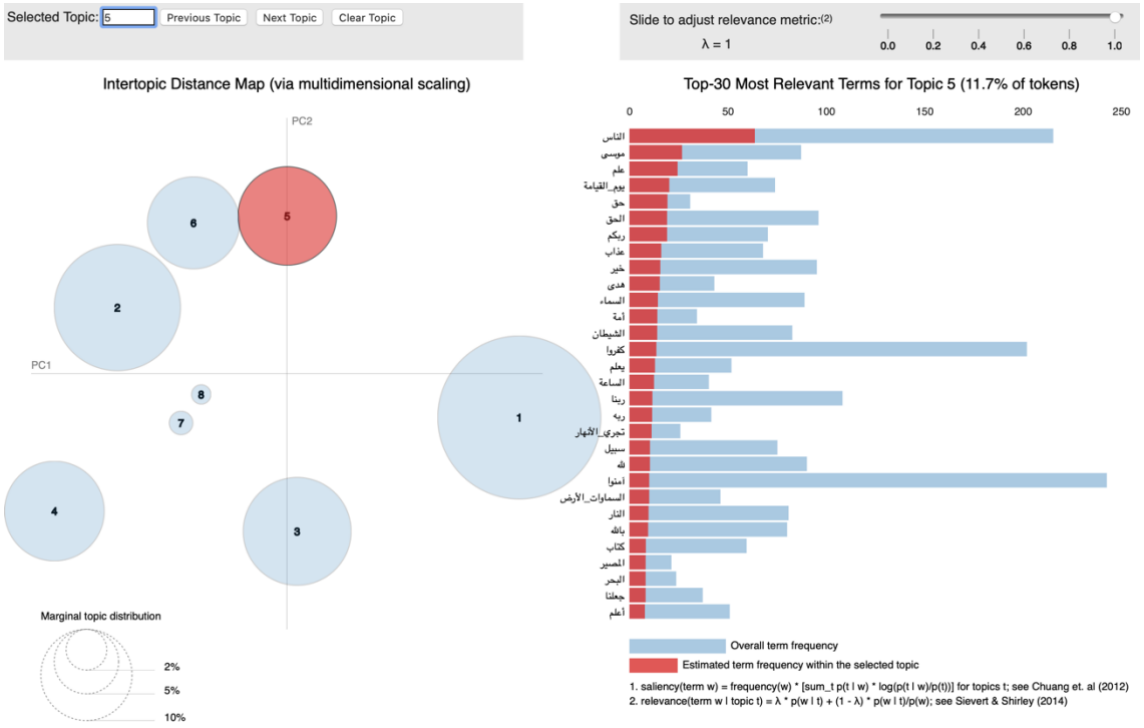
A scanned copy of The indexed dictionary of the meanings in the Qur’an.

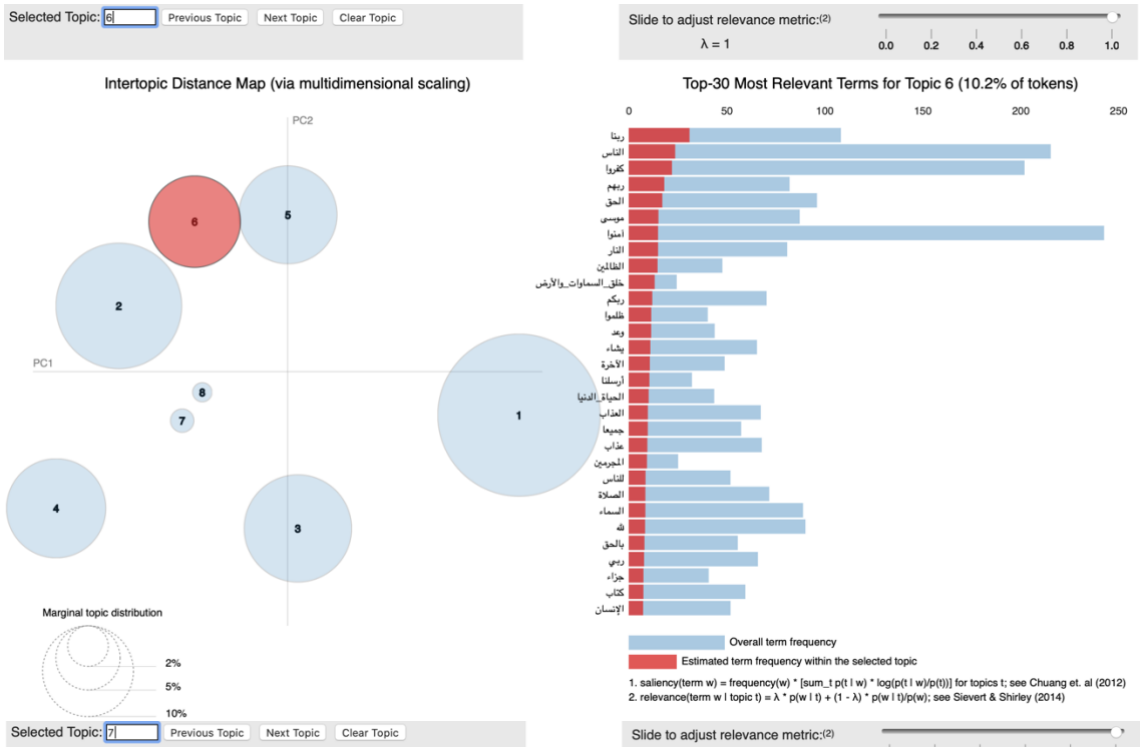
Appendix B

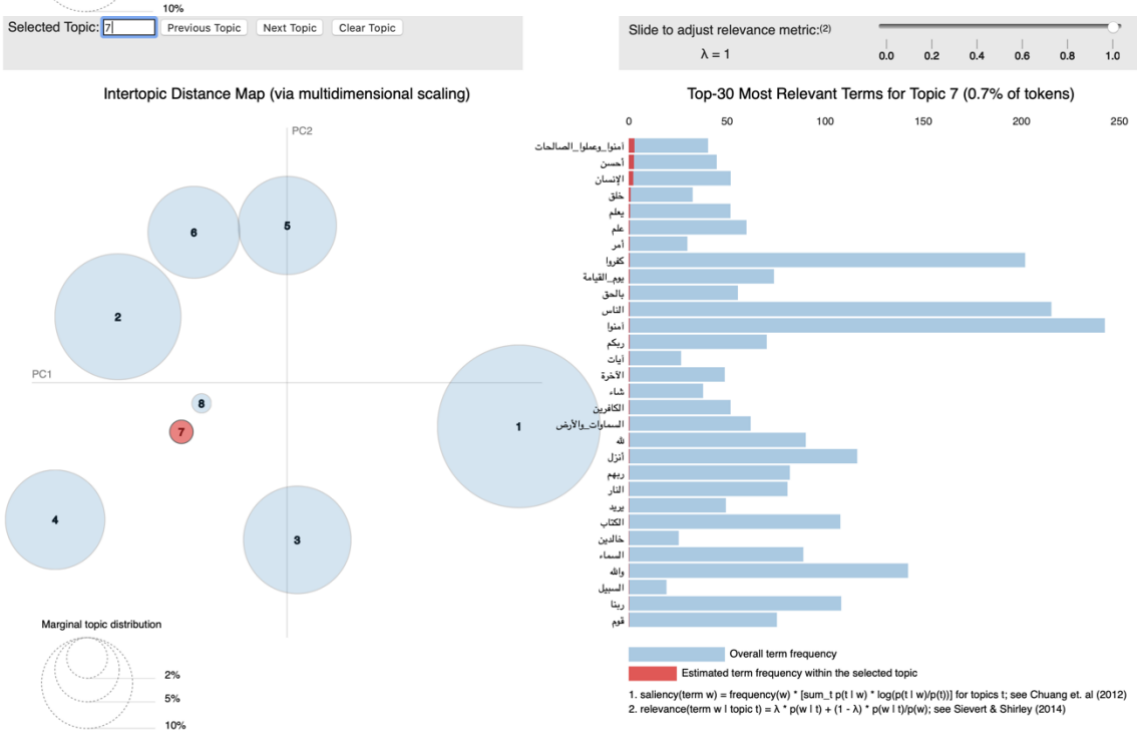
The output of pyLDAvis illustrating the topics and associated keywords, from Chapter 4.

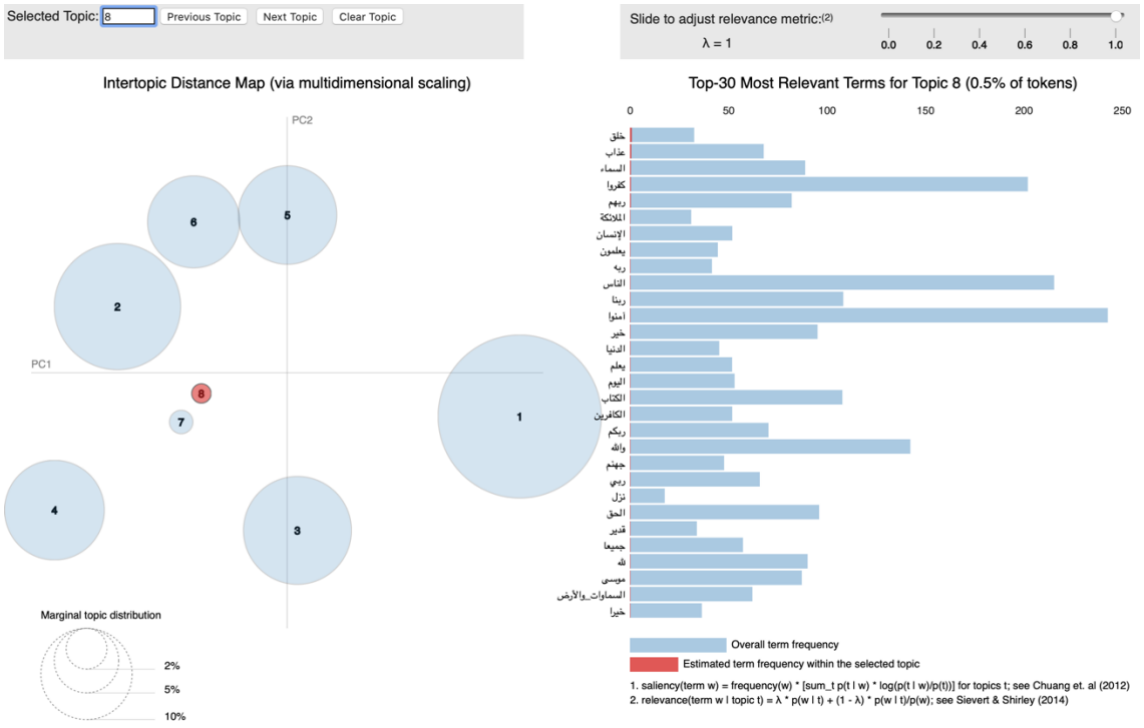



Selected Topic: 4
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 4 (11.9% of tokens)
0 50 100 150 200 250
العالمين
رسول
فرعون
السماء
ربهم
اليوم
الكافرين
أحد
بالحق
القوم
قوم
ربي
ربنا
الكتاب
ربكم
تعلمون
موسى
مبين
يعلمون
إله
السماوات_والأرض
جهنم
عظيم
ربه
خير
النار
الدين
كفروا
وجعلنا
الناس
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Selected Topic: 5
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 5 (11.7% of tokens)
0 50 100 150 200 250
الناس
موسى
علم
يوم_القيامة
حق
الحق
ربكم
عذاب
خير
هدى
السماء
أمة
الشيطان
كفروا
يعلم
الساعة
ربنا
ربه
تجري_الأنهار
سبيل
لله
آمنوا
السماوات_الأرض
النار
بالله
كتاب
المصير
البحر
جعلنا
أعلم
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Selected Topic: 6
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 6 (10.2% of tokens)
0 50 100 150 200 250
ربنا
الناس
كفروا
ربهم
الحق
موسى
امنوا
النار
الظالمين
خلق_السماوات_والأرض
ربكم
ظلموا
وعد
يشاء
الآخرة
أرسلنا
الحياة_الدنيا
العذاب
جميعا
عذاب
المجرمين
للناس
الصلاة
السماء
لله
بالحق
ربي
جزاء
كتاب
الإنسان
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Selected Topic: 7
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 7 (0.7% of tokens)
0 50 100 150 200 250
امنوا_وعملوا_الصالحات
أحسن
الإنسان
خلق
يعلم
علم
أمر
كفروا
يوم_القيامة
بالحق
الناس
امنوا
ربكم
آيات
الآخرة
شاء
الكافرين
السماوات_والأرض
لله
أنزل
ربهم
النار
يريد
الكتاب
خالدين
السماء
والله
السبيل
ربنا
قوم
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Selected Topic: 8
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 8 (0.5% of tokens)
0
50
100
150
200
250
خلق
عذاب
السماء
كفروا
ربهم
الملائكة
الإنسان
يعلمون
ربه
الناس
ربنا
امنوا
خير
الدنيا
يعلم
اليوم
الكتاب
الكافرين
ربكم
والله
جهنم
ربي
نزل
الحق
قدير
جميعا
لله
موسى
السماوات_والأرض
خيرا
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Appendix C

Results of the distance comparison test to evaluate embeddings' capability to capture the semantics of the Qur'an words from chapter5.

Target Word	Related Word	Cosine Distance	Cosine distance between target word and each control word									
			Control Word1	Control Word2	Control Word3	Control Word4	Control Word5	Control Word6	Control Word7	Control Word8	Control word9	Control Word10
الكتاب	التوراة	0.481	الشيطان	النار	الصلاة	خالدون	جنات	ترتيلا	الضالون	فاعبدوه	السماء	الفقراء
			1.125	0.975	0.940	1.145	0.923	0.825	0.850	0.926	0.967	0.877
الشيطان	الغاوين	0.268	المكذبين	حافظون	الفقراء	الناصحين	تنصرون	البنون	الأرحام	يخشى	الصادقون	ترتيلا
			0.873	0.769	0.895	0.706	0.785	0.972	0.951	0.858	0.828	0.933
خلق	وصوركم	0.338	النار	الناس	الكتاب	المؤمنون	خشية	يعلمون	عذاب	آمنوا	ريب	الصلاة
			0.941	0.842	1.051	0.952	0.752	0.756	0.957	0.930	0.541	0.561
القرآن	ترتيلا	0.322	ريب	الخلد	الأرض	علم	موسى	تنبت	الصالحات	قولا	الحياة	الآخرة
			0.788	1.179	1.111	0.806	0.781	1.009	0.977	0.966	1.318	0.432
النار	خالدون	0.228	نعيم	الآخرة	الصلاة	السماوات	رحمة	ميثاق	البينات	اتبعوا	أسلم	الخلد
			0.782	0.728	1.040	0.899	0.844	1.125	0.997	0.709	1.008	0.319
الصالحات	مغفرة	0.353	النار	الكفر	الأرض	الكتاب	الضالين	اليهود	الزير	تتبعون	الرسل	السماء
			0.837	1.053	0.956	0.834	0.804	0.954	0.697	1.256	1.225	0.916
الصابرين	ولنبلوكم	0.238	سيروا	الخيرات	كفروا	النار	يختلفون	الجحيم	أقيموا	أولياء	الأمثال	الزير
			0.818	0.790	1.015	0.956	1.077	0.960	1.027	0.768	0.812	0.634

الجحيم	ادعوا	النساء	ينتصرون	الباطل	المعتدين	يتوب	الصيام	الدنيا	الموت	0.353	المأوى	الجنة
0.712	0.807	1.171	0.755	1.059	0.813	0.935	1.069	0.869	0.881			
الحياة	الأنعام	السحاب	ينفقون	ينزغنك	الملك	إسرائيل	عيسى	الحياة	شركاءكم	0.340	نعيم	الفردوس
0.790	0.702	0.660	0.681	0.654	1.107	0.881	0.996	0.817	0.588			
الملك	المنافقين	الذكر	الشيطان	القيامة	يعذبكم	المؤمنون	المؤمنات	أنفقتم	معروف	0.252	رواسي	الجبال
0.936	1.147	0.842	0.935	0.970	1.056	1.211	0.966	0.802	0.832			