Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 04, December 2024.

INTERPRETING ARABIC TRANSFORMER MODELS: A STUDY ON XAI INTERPRETABILITY FOR QUR'ANIC SEMANTIC-SEARCH MODELS

Ahmad M. Mustafa, Saja Nakhleh, Rama Irsheidat and Raneem Alruosan

(Received: 10-Jan.-2024, Revised: 24-Mar.-2024, 16-May-2024 and 8-Jul.-2024, Accepted: 13-Jul.-2024)

ABSTRACT

Transformers have shown their effectiveness in various machine-learning tasks. However, their "black box" nature often obscures their decision-making processes, particularly in Arabic, posing a barrier to their broader adoption and trust. This study delves into the interpretability of three Arabic transformer models that have been fine-tuned for semantic-search tasks. Through a focused case study, we employ these models for retrieving information from the Holy Qur'an, leveraging Explainable AI (XAI) techniques—namely, LIME and SHAP—to shed light on the decision-making processes of these models. The paper underscores the unique challenges posed by the Qur'anic text and demonstrates how XAI can significantly boost the transparency and interpretability of semantic-search systems for such complex text. Our findings reveal that applying XAI techniques to Arabic transformer models for Qur'anic content not only demystifies the models' internal mechanics, but also makes the insights derived from them more accessible to a broader audience. This contribution is twofold: It enriches the field of XAI within the context of Arabic semantic search and illustrates the utility of these techniques in deepening our understanding of intricate religious documents. By providing this nuanced approach to the interpretability of Arabic transformer models in the domain of semantic search, our study underscores the potential of XAI to bridge the gap between advanced machine-learning technologies and the nuanced needs of users seeking to explore complex texts like the Holy Qur'an. Our code is available at[1].

KEYWORDS

Explainable machine learning, Semantic search, Arabic NLP, Transformers, SHAP, LIME.

1. INTRODUCTION

Interpretive and Explainable Artificial Intelligence (XAI), in the field of machine learning, deep learning and transformer models, has seen remarkable developments. However, XAI for Arabic transformer models remains a notable challenge. The idea of Explainable AI (XAI) emerged, introducing techniques that offer a reasonable trade-off between explainability and predictive power for a variety of machine-learning (ML), deep-learning and transformer techniques [1].

Transformer models, well known for their effectiveness in natural-language processing (NLP) tasks, often operate as black-box, making it difficult to understand the decision-making process they employ. This ambiguity sheds light on significant challenges, especially in the context of Arabic models, where small linguistic differences can increase the complexity of the interpretation task. The absence of robust XAI tools hinders the examination of model outputs, leading to a potential lack of trust and liability. Bridging this gap in interpretability for Arabic transformer models makes it possible for people to comprehend, trust and manage the newest generations of AI models in the Arabic-speaking world.

Arabic transformer models, used as black-box AI systems, have gained widespread usage in domains such as social networks, medicine and scientific fields. However, the necessity to explain and interpret these models arises from their operation as opaque decision making. These reasons include the Regulatory Perspective, exemplified by the European Union's General Data Protection Regulation (GDPR), which accords users the right to explanation. Another reason is the Model Developmental Perspective, which dives into issues such as limited training data, biased data, outliers, adversarial data and overfitting leading to inappropriate results in black-box AI systems. Lastly, the end-user and

[1] https://gist.github.com/a-mustafa/51fcacf30ecdf0c13ac91ad16fecfa89

A. Mustafa (Corresponding Author), S. Nakhleh, R. Irsheidat and R. Alruosan are with the Department of Computer Information Systems, Jordan University of Science and Technology, Irbid, Jordan. E-mails: ammustafa@just.edu.jo, swnakhleh21@cit.just.edu.jo, roirsheidat21@cit.just.edu.jo and rmalrosan21@cit.just.edu.jo

social perspectives address concerns about trust in black-box AI models, shedding light on the potential for unfair decisions and biases in the data used for model development. XAI is recognized as a solution to enhance trust by providing explanations, improving interpretability, addressing fairness concerns and ensuring that the models fulfill their intended purpose [2].

This study is motivated by the absence of XAI models for Arabic transformer models. As a case study, we perform a Qur'anic semantic search using different Arabic transformer models and then interpret them using different XAI models. The Holy Qur'an is the most significant source for Arabic and Islamic sciences. The Qur'an is considered a sacred text in Arabic and contains approximately 80,000 words divided into 114 chapters; each chapter consists of a varying number of verses. It also includes knowledge of a variety of other subjects, including science and the history of humanity [3]. Classical Arabic (CA), Modern Standard Arabic (MSA) and Colloquial Arabic are the three main styles or forms of the Arabic language [4]. Qur'an is the most important source of Classical Arabic. Many tools and applications have been developed to help in Qur'anic information retrieval.

There are three main methods for information retrieval within the Qur'an: semantic-based, keyword-based and Cross-Language Information Retrieval (CLIR) [5], as shown in Figure 1. A semantic-based method searches for concepts or meanings, whereas a keyword-based method looks for exact letter matches. CLIR searches for information in a language other than the one used in the query. Most Qur'an search tools use keyword search, but some use ontology-based or synonym-set methods [6]. The ontology-based or semantic-search approach looks for concepts or subjects that fit a user request. Semantic search emphasizes the meaning of words and the intent of the user query rather than relying only on keyword matching. It analyzes the context and considers the relationships between words and their meanings to retrieve similar information. Semantic search utilizes a transformer-based model such as BERT, neural models like RNN, ML models including n-gram and Word2Vec models [7].

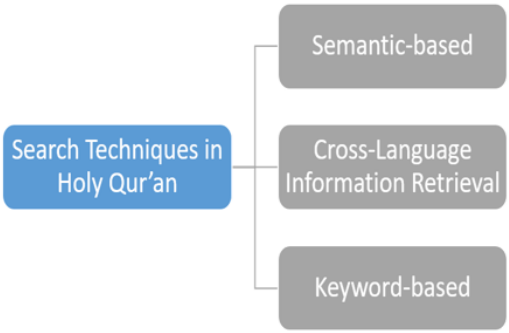


Figure 1. Classification of information-retrieval methods in the Holy Qur'an.

In this paper, we use three transformer BERT-based models for Qur'anic semantic search. Transformers are one of the most advanced techniques for many NLP problems since they were proposed by Vaswani et al. (2017) [8] for machine translation. The semantic-search models utilized in this study include: *CL-AraBERT* [7], an Arabic BERT transformer for CA. Additionally, "asafaya/bert-base-arabic" (*ArabicBERT*) developed by Safaya et al. [9] is employed. ArabicBERT is a pretrained language model based on BERT, designed for Arabic semantic-search task. Lastly, "multi-qa-MiniLM-L6-cos-v1" (*S-BERT*) model[2] [10] is used. S-BERT is a sentence-transformer model. It maps sentences and paragraphs to a 384-dimensional dense vector space and was designed for semantic search.

Although transformer-based neural networks excel at classification in various domains, they lack the capability to offer explanations for their predictions [11]. Our study shows different XAI techniques that interpret the transformers mentioned above using two SHAP [12] and LIME [13]. SHAP (SHapley Additive exPlanations) is an explainability technique that calculates the Shapley values from cooperative game theory to attribute the contributions of each feature to the model output, providing a comprehensive explanation for a given prediction. LIME (Local Interpretable Model-agnostic

[2] https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1

Explanations) is a model-agnostic technique that generates locally faithful approximations of a complex model decision boundaries by perturbing and observing input instances, facilitating interpretability for individual predictions. The main purpose of XAI is to introduce an explanation for a variety of ML, DL and transformer models that offer a reasonable trade-off between explainability and predictive power. This concept allows people to understand, trust and manage the newest generations of AI models.

1.1 Challenges of Qur'anic-text Processing

Challenges arise in the search and retrieval of relevant verses from the Holy Qur'an, due to both the search techniques and the structure of the text. The following are some of these challenges:

1. **Orthography**: The Qur'anic text employs a distinct orthography that incorporates the essential diacritics (tashkeel) and vowel marks (harakat) necessary to understand the text. Nevertheless, contemporary Arabic text frequently excludes these diacritics or stems them as part of preprocessing, posing challenges for machines to accurately recognize and handle the proper pronunciation. For example,الجَنّة (al-jannah) means heaven and الجِنّة (al-jinah) means ghosts [6]. The Qur'anic text utilizes a distinctive orthography (conventional spelling), distinct even from CA, referred to as al-rasm al-'Uthmani. This is the method of writing the Qur'anic text compiled during the reign of Caliph Uthman b. Affan, for example: أنزلناه (anzlnah \We revealed it), is written as أنزلنَاه (anzlnah\We revealed it) [14].
2. **Textual Variants**: The Qur'anic text exists in various versions that can vary in spelling, pronunciation and significance. Consequently, developing reliable and uniform computational models for processing the Qur'anic text presents a significant challenge. For example, محمد (Muhammad), أحمد (Ahmad) and المُزّمّل (Mozzammil) all refer to Prophet Muhammad [15], as following: وَمُبَشِّرًا بِرَسُولٍ يَأْتِي مِن بَعْدِي اسْمُهُ أَحْمَدُ (wamubashshiran birasul yati min baedi asmuh 'ahmad \And bringing good tidings of a Messenger who will come after me, whose name will be Ahmad), مُّحَمَّدٌ رَّسُولُ اللَّهِ (Muhammad rasul alllah \Muhammad is the Messenger of Allah), يَا أَيُّهَا الْمُزَّمِّلُ (ya 'ayuha almuzzmmil \O you who wraps himself in clothing!)

   On the other hand, there exists a disparity between user inquiry, inscribed in MSA and retrieved Qur'anic verses, written in CA [16]. For example, searching for أنزلناه (anzlnah\We revealed it), in MSA should retrieve the word أنزلنَاه (anzlnah\We revealed it) in CA. Since the vocabulary and spelling in CA differ from MSA, this makes models' selection challenging. To solve this issue, we select a multilingual S-BERT model, in addition to the CL-AraBERT model trained on CA and MSA texts, as mentioned before.
3. **Semantic Interpretation**: The Qur'anic scripture comprises numerous allegories, metaphors and parables that require deep semantic analysis and comprehension for accurate interpretation. For instance, the term الحيوان (Al-Hayawan\animal) in Arabic typically translates to 'the animal', but in this specific verse, it denotes 'the life'. و ما هذه الحياة الدنيا الا لهو و لعب و ان الدار الاخرة لهي الحيوان لو كانوا يعلمون (Wa ma hadhihi al-hayatu ad-dunya illa lahwan wa la'ab. Wa innad-dara al-akhirata lahiya al-hayawan law kanu ya'lamoon \And the worldly life is nothing but amusement and diversion. But the home of the Hereafter - that is the eternal life, if only they knew) [17]. To achieve a precise interpretation and translation of the Qur'anic text, it is imperative to understand its historical context. The text was revealed in the 7th century and the language and vocabulary used in it are indicative of the historical and cultural background of that period.
4. **Expressiveness**, which refers to rhetoric in linguistics, involves expressing meanings using fewer words. For example, the concise phrase فأسقيناكموه (Fa asqaynākumūhu), which translates to "and We have given it to you to drink" in Arabic morphology is remarkably intricate, yet follows a systematic approach [14].

Contribution: To the best of our knowledge, this study is the first to address the interpretation of Arabic semantic-search transformer models, utilizing *post-hoc* interpretation models. We propose a methodology that interprets the results obtained from three intricate transformer models; namely,

S-BERT, ArabicBERT and CL-AraBERT, which have recently been introduced for Qur'anic semantic search. Our results will help understand the inner workings of these Arabic semantic-search transformer models, facilitated by the utilization of *post-hoc* interpretation models, including LIME and two versions of SHAP, thus enhancing comprehension and insight.

2. RELATED WORK

This section provides an overview of previous studies that have explored semantic similarity in Arabic and Qur'anic text. It also delves into techniques for interpreting transformer models, such as BERT, utilizing *post-hoc* interpretation models such as SHAP and LIME. *Post-hoc* approaches refer to methods applied after a model has been trained to explain its predictions and provide insight into the decision-making process of complex models. These methods approximate the rationale of the underlying machine-learning models, proving especially valuable for the interpretation of 'black-box' models, wherein the internal mechanisms are not inherently transparent [18]-[19]. It is also worth mentioning that *ante-hoc* approaches, though not as widely recognized or discussed as *post-hoc* methods, *ante-hoc* approaches refer to techniques that are integrated during the model-development phase to ensure interpretability from the outset. These approaches are designed to build inherently explainable models, allowing for real-time interpretation of model decisions while processing data [20]. However, this approach is beyond the scope of our paper. Moreover, we evaluate and compare the *post-hoc* interpretation techniques with those used in our paper, highlighting their simplicity and informativeness.

Both topics discussed in this section are crucial to the research, as they provide the foundational knowledge and tools necessary for interpreting Arabic semantic-search transformer models, which are the main focus of this paper. Due to the lack of studies that combine Arabic semantic similarity with interpretation techniques, we have organized the related works into different sub-sections.

2.1 Semantic Similarity in Arabic and Qur'anic Texts

Several studies have employed different techniques to extract semantic similarity or relatedness from Arabic and Qur'anic texts. Alsaleh et al. (2021) [21] conducted experiments using the QurSim dataset and a fine-tuned AraBERT model, which is an Arabic-language model trained on a wide range of Arabic texts. The dataset includes pairs of verses classified into three classes: '2' for strong similarity, '1' for weak similarity and '0' for no similarity. They also filtered the dataset to eliminate repetition and create random pairs of verses. AraBERTv0.2 outperformed AraBERTv2 with an accuracy score of 92%. However, AraBERT struggled with classical-Arabic lexical synonyms and religious context, potentially due to corpus limitations. Our study utilizes AraBERT to classify pairs of Qur'anic verses as semantically related or not.

Mohamed and Shokry (2022) [6] discussed modern semantic-search techniques for the Holy Qur'an. They manually created a dataset and annotations based on Tajweed Mushaf and created an embedding matrix trained with classical Qur'anic and Arabic texts. This generated word-based feature vectors for the verses. During queries, cosine similarity was used to find the most semantically similar result. However, this approach only retrieved verses for the first query and ignored the rest of the topics, although they are also relevant to the query.

Saeed et al. (2020) [22] explored using word embeddings to identify semantically similar verses from the Holy Qur'an. Using Word2Vec and Sent2Vec models, they highlighted the importance of semantic text similarity in NLP and various fields, including religious-text analysis. They trained custom word embeddings from multiple English translations of the Holy Qur'an and compared them to pre-trained embeddings from the Spacy library. The custom-trained models showed promising performance, with Model #5 achieving the highest accuracy. The study emphasized the framework's potential to be applied to any text, contributing to a deeper understanding of sacred and literary works. Notably, their research focused on English translations of the Holy Qur'an, potentially missing nuances in the original Arabic.

Malhas and Elsayed (2022) [7] proposed the first Qur'anic Reading Comprehension Dataset (QRCD), consisting of 1,337 question-passage-answer triplets for 1,093 question-passage pairs. They introduced CLassical-AraBERT (CL-AraBERT), pre-trained on a 1.0B-word classical Arabic dataset to

complement modern standard Arabic (MSA) resources, enhancing its utility for reading comprehension tasks. Leveraging cross-lingual transfer learning from MSA to classical Arabic, they fine-tuned CL-AraBERT using MSA-based machine-reading comprehension datasets followed by QRCD. For evaluation, they used the F1-score and Partial Average Precision (pAP), integrating partial matching for multi-answer and single-answer MSA questions, thus constituting the first MRC system on the Holy Qur'an.

2.2 Interpretation Techniques

Although there are many studies related to Qur'anic semantic search, there is no previous work that interprets Arabic semantic-search models using XAI techniques. Several *post-hoc* XAI interpretation techniques are discussed here to interpret and explain different transformer models.

The first technique is LIME [13], which generates local explanations for each instance in a dataset. LIME introduces disturbances to an instance and uses the newly generated dataset to predict the class of each instance using a trained classifier. A simpler model is then used to explain the classifier's prediction. While LIME is likely to be locally faithful, it does not perfectly represent complex models.

SHAP [23], another *post-hoc* XAI technique, interprets the complex behavior of machine-learning, deep-learning and transformer models. SHAP values, based on game theory, allocate importance scores to each feature within a model to provide consistent explanations. Positive SHAP values indicate a positive contribution to the prediction, whereas negative values indicate a negative impact. The Semantic Textual Similarity (STS) explainer [24] is a SHAP-based technique designed to explain sentence-level scores by highlighting erroneous words in both source and target sentences. This method helps understand the contribution of each word using SHAP for tasks like machine translation and semantic search involving different text languages. TransSHAP, proposed by Kokalj et al. (2016) [11], adapts SHAP to provide sequential explanations for transformer models such as BERT-based text classifiers. Unfortunately, it is notable that TransSHAP is currently not compatible with semantic-search transformer models. Despite not being compatible with semantic-search transformer models, TransSHAP was found effective for tasks like sentiment analysis. It was rated better than SHAP and slightly better than LIME in overall user preferences.

Layer-wise Relevance Propagation (LRP) [1] assigns relevance scores to input features to explain machine-learning model predictions. When applied to Transformer models, LRP computes relevance scores for each input token to understand its contribution to the final prediction. Although useful, LRP, like TransSHAP, faces limitations in providing explanations for tasks involving multiple-sentence analysis, such as semantic search.

El Zini et al. (2022) [25] proposed new metrics and techniques to evaluate the explainability of Arabic Sentiment Analysis (SA) models. They assessed the accuracy of 'rationales' extracted by the model and compared the agreement between XAI techniques and human judgment on a dataset. Their results showed that transformer models have better explainability than convolutional and recurrent neural-network architectures. This research lays the foundation for designing interpretable NLP models and creating a common evaluation framework.

3. METHODOLOGY

3.1 Dataset

The Holy Qur'an, revered by 1.5 billion Muslims worldwide, is structured into 30 sections and 114 chapters, encompassing 6,236 verses, totaling approximately 78,000 words. These words are organized into verses, with sets forming parts, chapters and groups (Hizb) or Hizb quarters. Each of the 114 chapters belongs to one of the 30 sections and the text is further segmented into 60 groups (Hizb), with each section comprising two groups (Hizb) [7].

We have used a verified Qur'an dataset called Tanzil Quran text[3]. The Tanzil Quran text provides a verified digital version of the Holy Qur'an in many scripting styles, including the Uthmani style. We have utilized the normalized simple-clean text style (in Tanzil 1.0.2) to enable the use of the dataset with transformer-based language models that have already been pre-trained using normalized Arabic

[3] https://tanzil.net/download/

text. Tanzil Qur'an dataset consists of three columns, as shown in Table 1: 1. Surah ID: is an id for each Surah from 1 to 114. 2. Verse ID: is an id for each verse (ayah) from 1 to 6236 without verse Basmallah, except in Chapter 1 (Surah Al-Fatiha). 3. Verse Text: the content of verse text with diacritics. For the model evaluation using the Qur'an exegesis step, we used an official Qur'an exegesis (Tafsir) called QuranEnc[4]. Qur'anEnc is a dataset that provides an interpretation for each verse of the Qur'an. As shown in Table 2, there are three columns in Qur'an exegesis (QuranEnc) dataset: 1. Verse ID: is an id for each verse (ayah) from 1 to 6236. 2. Exegesis: the content of verse's exegesis (tafsir). 3. Verse Text: the content of verse text with diacritics.

Table 1.The-holy-Qur'an dataset.

Surah ID	Verse ID	Verse Text
1	1	بِسْمِ اللَّهِ الرَّحْمَـٰنِ الرَّحِيمِ (Bismillāhi al-Raḥmāni al-Raḥīm\In the name of Allah, the Most Gracious, the Most Merciful)
1	2	الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ (Al-ḥamdu lillāhi rabbi al-'ālamīn\Praise be to Allah, the Lord of all the worlds)
1	3	الرَّحْمَـٰنِ الرَّحِيمِ (Ar-Raḥmāni ar-Raḥīm\The Entirely Merciful, the Especially Merciful)

Table 2. Qur'an exegesis (QuranEnc) dataset.

Verse ID	Exegesis	Verse Text
3122	وإن ربك, أيها الرسول, لهو العزيز الذي ينتقم من أعدائه، الرحيم بمن تاب من عباده (Wa inna rabbaka, ayyuha ar-rasulu, lahu al-'azizu alladhi yantaqimu min 'a'ada'ihi, ar-rahīmu biman tāba min 'ibādihi\And indeed, your Lord, O Messenger, He is the Exalted in Might, the One who exacts retribution upon His enemies, yet He is the Merciful to those among His servants who repent and mend their ways)	وَإِنَّ رَبَّكَ لَهُوَ الْعَزِيزُ الرَّحِيمُ (Wa innna rabbaka lahuwa al-'azīzu ar-raḥīmu.\And indeed, your Lord is the Exalted in Might, the Merciful)
4465	في بساتين وعيون جارية (Fī basātīn wa 'uyūn jāriyah \In gardens and flowing springs)	فِي جَنَّاتٍ وَعُيُونٍ (Fi jannatin wa 'uyun \In gardens and springs)
5888	واستمعت لربها منقادة، وحق لها ذلك (Wa istam'at li rabbiha munqādah, wa ḥaqqun lahā dhālik \And she listened to her Lord obediently. It was rightful for her to do so)	وَأَذِنَتْ لِرَبِّهَا وَحُقَّتْ (Wa 'adhīnat li rabbiha wa ḥuqqat \And she listened to her Lord and fulfilled [her obligation])

3.2 Data Preprocessing

The lack of diacritics in Modern Standard Arabic (MSA) is a common issue in the Arabic language. Diacritical marks are significant, because they impact the meaning and subsequently, the comprehension of Arabic texts [26]. Although the Holy Qur'an is extensively diacritical, most NLP tasks involving digital Qur'anic text resort to normalization by eliminating diacritics during the preparation stage. In this phase, we applied several preprocessing techniques to clean the text before feeding it into models using the Holy Qur'an dataset. Firstly, we added a new column named "surah name" to the dataset, which includes the name of each surah in the Holy Qur'an. Following this addition, we removed tashkeel (diacritical marks) and tatweel (character lengthening), as well as eliminating stop words and punctuation from the verses. Lastly, we normalized certain characters to standardize the dataset. Table 3 illustrates examples of the data preprocessing steps.

Table 3. Overview of data preprocessing steps with examples.

Original Verses	بِسْمِ اللَّهِ الرَّحْمَـٰنِ الرَّحِيمِ (Bismillāhi al-Raḥmāni al-Raḥīm\In the name of Allah, the Most Gracious, the Most Merciful)	فَوَاكِهُ ۖ وَهُمْ مُكْرَمُونَ (Fawākihu, wahum mukramūn \Fruits and they are honored)
Tashkeel removing	بسم الله الرحمن الرحيم	فواكه وهم مكرمون
Tatweel removing	بسم الله الرحمن الرحيم	-
Punctuation and stop words removing	-	فواكه مكرمون

[4] https://quranenc.com/ar/browse/arabic_mokhtasar/

3.3 Workflow

As discussed earlier, the aim of this study is to interpret Arabic BERT-based semantic-search models using LIME and two SHAP techniques, the most well-known XAI techniques. Observe Figure 2.

Semantic search is designed to understand the meaning of a user query, as opposed to simply matching keywords and to return results that are relevant to the user intent. This can make search results more accurate and useful to the user. Semantic-search technology is used in a variety of applications, including search engines, e-commerce websites and voice assistants. We will evaluate the semantic-search models using two different methods, BERTScore and Cosine similarity and then interpret the model results through the XAI techniques.

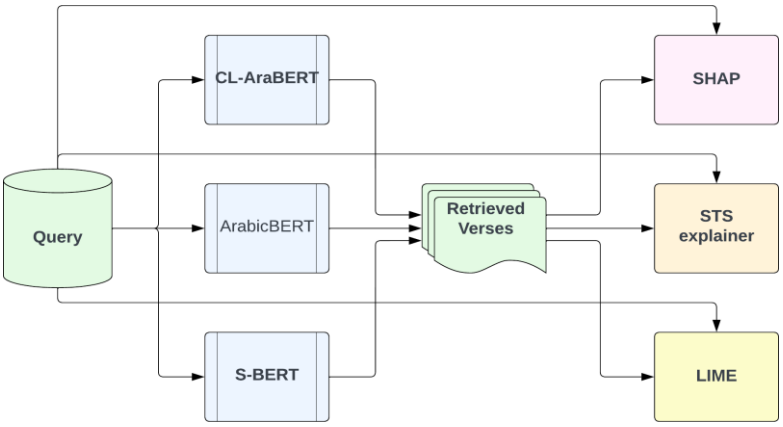


Figure 2. Proposed framework: A query is sent to semantic-search models and the retrieved verses are interpreted alongside the query using XAI techniques.

For the Qur'anic semantic search, first, we passed all 6,236 Qur'an exegesis, as queries to the three transformer BERT-based models: CL-AraBERT, ArabicBERT and S-BERT. The verses retrieved by the three models have been compared with the reference verse recorded in the Qur'anEnc Tafsir dataset. In this step, BERTScore precision, recall and F1-score were calculated for all the verses and their exegesis. BERTScore metric [27] evaluates the quality of text embeddings, particularly in the context of comparing the generated text against reference text. Specifically, it compares token-level similarity and leverages contextual embeddings from BERT or other transformer-based models.

The performance of each model has been evaluated using BERTScore precision, recall and F1-score measurements. BERTScore is an automated evaluation metric that is used to assess the quality of text-generation systems. The precision (P), expressed in Equation1, measures the mean cosine similarities between each retrieved token and its closest reference token, normalized by the number of retrieved tokens. Using contextual embeddings, tokens are represented in a reference verse $x = x_1, \ldots, x_k$ and a retrieved verse $\hat{x} = \hat{x}_1, \ldots, \hat{x}_l$. The cosine similarity $(x_i^T \hat{x}_j)$ weighs each retrieved token. Recall (R) indicates the extent of coverage completeness, as shown in Equation 2, calculated by dividing the number of relevant retrieved tokens by the number of all possible related tokens. The F1-score is the harmonic mean of precision and recall as shown in Equation 3.

$$P = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} (x_i^T \hat{x}_j) \tag{1}$$

$$R = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} (x_i^T \hat{x}_j) \tag{2}$$

$$F = \frac{2 * P * R}{P + R} \tag{3}$$

Cosine similarity [28] measures the similarity between two nonzero vectors in an inner product space. In NLP, it is commonly employed to evaluate the similarity between two pieces of text by converting each text into a vector of word counts or frequencies and finding the cosine of the angle between the vectors. The cosine similarity helps identify the degree of alignment of these vectors, indicating semantic similarity. This allows for efficient retrieval of sentences (verses) with similar meanings, making it a valuable metric for tasks such as semantic search.

In the second evaluation, samples of exact, similar and dissimilar queries were sent to these three semantic-search models and the results were evaluated based on the cosine similarity between each query result and reference verse.

Finally, sample queries, along with their respective search results, underwent three interpretation techniques: SHAP, STS explainer and LIME. Two criteria were employed in selecting the *post-hoc* interpretation techniques. Firstly, the XAI technique should support the Arabic transformer models. Secondly, it should support tasks that involve comparing two sentences, such as machine translation, question answering and semantic search. Notably, both TransSHAP and LRP were excluded from consideration, since they do not support tasks that involve comparing two sentences. The interpretation step will explain the two evaluation methods mentioned previously for semantic-search models. This process involves three steps: First, a query is sent to each semantic-search model. Then, three results are chosen (exact, similar and dissimilar) based on their scores. Finally, each resulting verse is compared with the query using the interpretation techniques. More details will be explained in the next sections.

4. EXPERIMENTS

In this section, we outline the experimental setup utilized in our study, focusing specifically on the methodology adopted for interpreting the semantic-search model outcomes and reporting the ultimate findings. To test the three semantic-search models, we have developed our testing procedure, where we use the sentences mentioned in an official Qur'an exegesis (Tafsir), Qur'anEnc. We fed all 6,236 verses of interpretation texts (Tafsir) into the semantic-search models. If the reference verse mentioned in Tafsir is retrieved among the closest five resulting verses, we consider it as the prediction; otherwise, we consider the top retrieved verse as the prediction. Subsequently, we compare these predictions with the references from the Tafsir dataset. This procedure is repeated for all semantic-search models.

4.1 Models Evaluation Using Exact, Similar and Dissimilar Sample Queries

Since BERT is a transformer-based model, its embeddings are contextual and depend on the entire input sequence, so that SHAP can be adapted to work with BERT models by approximating the Shapley values for token embeddings. Therefore, calculating Shapley values directly becomes computationally expensive. To handle this limitation, we have employed sampling sub-sets of input tokens to estimate Shapley values for the three BERT models. To do so, we have passed a set of 3 queries (samples) to three models: CL-AraBERT, ArabicBERT and S-BERT. Then, we measured the cosine similarity between the query and the retrieved verses. After that, we interpret samples of them, using SHAP and STS-Explainer interpretation techniques in sub-section 4.2. Here, as a first experiment, we have applied 3 test cases to validate each of the 3 BERT models:

1. We have passed an existing text such as علمه البيان ('Alamahu al-bayan \He taught him eloquence), as the expected results are the exact match with a similarity of 1.0. The other retrievals should be other similar sentences, but not identical. As expected, their cosine scores will be less than the exact match.
2. We have passed a text that does not exist in the Qur'an, but similar to existing words such as ابراهام (Ibrāhīm\Abraham). Here, the expected results should be verses with words similar to the query.
3. The third query uses words neither exist nor are similar to Qur'anic words such as كمبيوتر (Kumbiutir \Computer).

4.2 Model Interpretation Using SHAP

For each query, all semantic-search transformer models typically operate by retrieving a set of results, prioritizing the exact matches if they exist, followed by similar results and then possibly dissimilar ones. In this step, we interpret 3 samples from the retrieved verses for exact match, similar and dissimilar results using *post-hoc* interpretation techniques, SHAP and STS explainer.

First, we search for an existing sentence, such as: علمه البيان ('Alamahu al-bayan \He taught him eloquence), so that we could select 3 results:

1. Exact match with cosine similarity scores 1.0, such as: علمه البيان (’Alamahu al-bayan \He taught him eloquence).
2. Similar verse with high cosine similarity score, such as علم القران (’Allama al-Qur’an \He taught the Quran).
3. A verse with a very low cosine similarity score, such as إنما جزاء الذين يحاربون الله ورسوله ويسعون في الأرض فساداً أن يقتلوا (Innama jazau alladhina yuharibuna Allah wa rasulahu wa yas’awna fi al-ardi fasadan an yuqtalu \Indeed, the penalty for those who wage war against Allah and His Messenger and strive upon earth [to cause] corruption is none but that they be killed).

All these results were picked for the 3 BERT models. So in total, we have 9 results that were passed to 2 SHAP interpretation techniques. For the first SHAP model, we have tuned a Question Answering SHAP technique, to work as a semantic-search model. We assumed that the question was our query علمه البيان (’Alamahu al-bayan \He taught him eloquence) and sent it with the context, which was the Surah or verse that contained the query. Finally, we have assigned scores to the selected three search results to display their relevance.

SHAP explains the output of the semantic search by attributing the importance of each feature in the context to the model prediction (query result). The SHAP summary plots provide many visual details that thoroughly explain the machine-learning models in a simple way. First, the summary plots provide SHAP scores, f(input), which equals the summation of all the word (feature) scores in the context. They provide insights into the contribution of each feature in the context to the model prediction for a particular instance of data, in our case, the query result. The words in the context, with a positive SHAP score suggest a positive influence on the prediction (query result), while the negative score suggests a negative influence on the prediction. The magnitude of SHAP score provides a measure of the feature importance relative to other features in the input query. Features with higher SHAP values are considered more important in influencing the model prediction. Second, SHAP colors in the summary plots are important to investigate the SHAP scores. The red color means a positive effect, the blue color means a negative effect and the shade of the color indicates the amount of effect. Therefore, dark red means a high positive effect, while light red means a low positive effect.

In the "STS explanation" step, we interpret the results from another perspective, where another technique was utilized from SHAP called STS Explainer. We passed the same three search results from the "Search" step to the STS Explainer which was implemented especially for semantic-search tasks. The similarity score metric of the STS Explainer is F1 by default and we have fixed it for all the upcoming experiments. f(x) shows the similarity between the query and the result, while E(f(X)) shows the Expected SHAP score, which is calculated as the mean of all predictions. Just like the original SHAP technique, the STS Explainer provides visual interpretation using colors to indicate the positive or negative impact of the values, red for positive and blue for negative. STS Explainer also provides the SHAP score for each word in both the query and the result sentences, which indicates each word contribution to the model prediction that is figured out in the summary plot.

4.3 Interpretation Using LIME

In this study, we utilized a surrogate model, specifically LIME, to interpret the outputs generated by BERT models when analyzing the Holy Qur’an. We examined two variations of the Qur’anic text: the original text, which includes tatweel (elongation marks), tashkeel (diacritical marks), punctuation and stop words and a second version where only the tashkeel was removed (the Tanzil Qur’an dataset).

The Holy Qur’an dataset underwent a detailed normalization process, as described in sub-section 3.2, which includes steps such as stop-word removal and character normalization. The text (verses) resulting from this process is referred to as "Normalized Verses" in Table 4. In contrast, the Tanzil Qur’an version underwent a simpler process, with only the removal of diacritical marks (tashkeel). The text resulting from this process is labeled as "Normalized Verses (Tashkeel Removed)" in Table 4. The primary motivation for these different approaches was to investigate the impact of text normalization on model interpretation and similarity assessment. This methodological choice allowed us to directly compare how varying levels of text normalization influence the performance and interpretability of BERT models.

To facilitate this comparison, we encoded the verses from these two variations using the Sentence

Transformer of our transformer models (CL-AraBERT, ArabicBERT and S-BERT) to get the embedding. The Sentence Transformer is a deep-learning model that encodes text into high-dimensional vector representations (embeddings) to capture their semantic meaning, facilitating efficient comparison and analysis of text data[5]. These embeddings were then used to calculate cosine similarity with the verse علمه البيان (’Alamahu al-bayan \He taught him eloquence), serving as a benchmark for assessing verse similarity.

We adopted a binary classification approach to present these similarities, designating verses as "*Similar*" (label ’1’) or "*Not Similar*" (label ’0’) based on predefined thresholds of 0.6, 0.8 and 1.0[6]. This classification facilitated a structured analysis of the impact of text normalization at varying levels of strictness in similarity assessment. The verses were then split into training and testing datasets. The outcomes of this approach are detailed in Table 4, which presents the classification results (labels) of selected normalized and normalized (tashkeel removed) verses at different similarity thresholds.

Table 4. Classification outcomes of selected verses at varied similarity thresholds.

Verses	Similarity threshold	Similarity value (label)
Normalized Verses		
بسم الله الرحمن الرحيم - علمه البيان Bismillāhi ar-Raḥmāni ar-Raḥīm \In the name of Allah, the Most Gracious, the Most Merciful	>=0.8	0
الحمد لله ربّ العالمين - علمه البيان Al-ḥamdu lillāhi rabbi al-’ālamīn\Praise be to Allah, the Lord of all the worlds	>=0.8	0
فألقى عصاه فإذا هي ثعبان مبين - علمه البيان Fa’alqā ’aṣāhu fa’idhā hiya thu’ban mubīn \So he threw down his staff and behold! it was a manifest serpent	>=0.8	1
الذين جعلوا القران عضين - علمه البيان Alladhīna ja’alu al-qur’āna ’idīn\Those who have made the Quran burdensome	>=0.8	1
بسم الله الرحمن الرحيم - علمه البيان Bismillāhi ar-Raḥmāni ar-Raḥīm \In the name of Allah, the Most Gracious, the Most Merciful	>=0.6	0
الحمد لله ربّ العالمين - علمه البيان Al-ḥamdu lillāhi rabbi al-’ālamīn\Praise be to Allah, the Lord of all the worlds	>=0.6	0
مالك يوم الدين - علمه البيان Māliki yawmi ad-dīn \Master of the Day of Judgment	>=0.6	1
ذلك الكتاب لا ريب فيه هدى للمتقين - علمه البيان Dhālika al-kitābu lā rayba fīhi hudan lil-muttaqīn \This is the Book about which there is no doubt, a guidance for those conscious of Allah	>=0.6	1
علمه البيان - علمه البيان ’Allamahu al-bayan \He taught him eloquence	==1	1
Normalized Verses (Tashkeel Removed)		
بسم الله الرحمن الرحيم - علمه البيان Bismillāhi ar-Raḥmāni ar-Raḥīm \In the name of Allah, the Most Gracious, the Most Merciful	>=0.8	0
الحمد لله رب العالمين - علمه البيان Al-ḥamdu lillāhi rabbi al-’ālamīn\Praise be to Allah, the Lord of all the worlds	>=0.8	0
وإن عليك اللعنة إلى يوم الدين - علمه البيان Wa inna ’alayka al-la’natu ilá yawmi ad-dīn \And upon you is the curse until the Day of Recompense	>=0.8	1
وأن عذابي هو العذاب الأليم - علمه البيان Wa anna ’adhābī huwa al-’adhābu al-aleem \And indeed, My punishment is the painful punishment	>=0.8	1
بسم الله الرحمن الرحيم - علمه البيان Bismillāhi ar-Raḥmāni ar-Raḥīm \In the name of Allah, the Most Gracious, the Most Merciful	>=0.6	0
الذين يؤمنون بالغيب ويقيمون الصلاة ومما رزقناهم ينفقون - علمه البيان Alladhīna yu’minūna bil-ghaybi wa yuqīmūna aṣ-ṣalāh wa mimma razaqnāhum yunfiqūn \Those who believe in the unseen, establish prayer and spend out of what We have provided for them	>=0.6	0
الحمد لله رب العالمين - علمه البيان Al-ḥamdu lillāhi rabbi al-’ālamīn\Praise be to Allah, the Lord of all the worlds	>=0.6	1
مالك يوم الدين - علمه البيان Māliki yawmi ad-dīn \Master of the Day of Judgment	>=0.6	1
علمه البيان - علمه البيان ’Allamahu al-bayan \He taught him eloquence	==1	1

[5] https://sbert.net/

[6] If the similarity between verses and the verse (query) (’Alamahu al-bayan \He taught him eloquence) is greater than or equal to the similarity threshold, we set the similarity class to “Similar" else “Not Similar". For example, suppose that we set the similarity threshold to 0.8 and the similarity score is 0.7, the similarity value will be “Not Similar" since 0.7 is less than the threshold.

In Arabic, two words that appear different on the surface can become similar when represented using word embeddings due to the language's rich morphological structure. Arabic is characterized by its introflexive, fusional and inflectional nature, which means that words can have different surface forms, but share underlying roots and patterns that convey similar meanings or functions [29]-[30]. For instance, in verses such as فَأَلْقَىٰ عَصَاهُ فَإِذَا هِيَ ثُعْبَانٌ مُّبِينٌ (So he threw down his staff and behold, it was a manifest serpent) and عَلَّمَهُ الْبَيَانَ (He taught him eloquence), the words البيان (al-bayan) and مبين (mubīn) look different but share the same root ب– ي– ن (b-y-n), which conveys the meaning of clarity, explanation or distinction. When these words are embedded in a vector space, the embeddings capture these morphological and semantic similarities, making their vectors close to each other. This closeness can be quantified using cosine similarity.

Our analysis involved a Logistic Regression (LR) model to predict the similarity class (label) of each verse in the test dataset. This model was chosen for its simplicity and effectiveness in binary-classification problems. Specifically, we trained the LR model on the training dataset and utilized TF-IDF (Term Frequency-Inverse Document Frequency) [31] to convert the textual data into numerical vectors. Finally, the predicted probabilities from the LR model were then passed to the LIME XAI framework for interpretability and explanation of the model's decisions.

5. RESULTS AND DISCUSSION

We delve into the outcomes of the conducted experiments. To understand the intent and context of a user query, we pass two evaluation techniques to evaluate both semantic-search models retrieving the meaning and the matching keywords. First, the model evaluation using Qur'an exegesis using BERTScore and second, the model evaluation using exact, similar and dissimilar sample queries using cosine similarity.

5.1 Model Evaluation Using Qur'an Exegesis

As mentioned in the experimental-setup section, we compared the three semantic-search model predictions with the references from the Tafsir dataset. This procedure is repeated for all semantic-search models. Table 5 shows that CL-AraBERT outperformed the other two models with 0.92, 0.93 and 0.92 for Precision, Recall and F1 BERTScore measurements, respectively.

Table 5. Test scores for 3 different semantic-search models.

	CL-AraBERT	**ArabicBERT**	**S-BERT**
P-$_{\text{BERTScore}}$ Eq. 1	0.92	0.79	0.67
R-$_{\text{BERTScore}}$ Eq. 2	0.93	0.81	0.71
F1-$_{\text{BERTScore}}$ Eq. 3	0.92	0.80	0.69

5.2 Model Evaluation Using Exact, Similar and Dissimilar Sample Queries

We have passed a set of three sample queries -exact match, similar and dissimilar- to three models: CL-AraBERT, ArabicBERT and S-BERT. Then, the cosine similarity is measured between the query and the retrieved verses. The results for these experiments are shown in Table 6.

The exact-match results show identical similarity for the three models, with a cosine score of 1.0 or approximately 1 (0.99) for all of them. Even for the other retrieved results, as will be shown in experiment 2, they were close to each other. For the similar word, the expected retrieval from a linguistic perspective , is: إبراهيم (Ibrāhīm \Abraham). However, the retrieved verses contain similar sub-words (or tokens), not the same word, that got similar embeddings, such as token: "را" (ra). Their cosine scores were not very high, which indicates that they are far from the query embedding. For the third query, since we have searched for a non-existent word, the retrieved sentences contained similar tokens. Hence, the retrieved result for S-BERT, for example, contained sub-token: "بيـ" (byo) . The proposed results through this experiment have obviously shown that the BERT transformers got the same cosine similarity scores, for the exact-match results. Also, variations have been shown for the other queries due to their different embedding, even if all of their implementations were BERT-based.

Table 6. The top retrieved verse for each query using the three models along with cosine similarity.

BERT model	علمه البيان: exact match ('Alamahu al-bayan \He taught him eloquence)	إبراهام: similar (Ibrāhīm\Abraham)	كمبيوتر: not exist (Kumbiutir\Computer)
S-BERT	علمه البيان	فالمدابرات أمرا (Fāl-dabirāt amrā\So the consequences are decreed.)	كراما كاتبين (Karaman kātibīn\Honored recorders)
	0.99	0.83	0.93
ArabicBERT	علمه البيان	إلا المصلين (Illā al-muṣallīn\Except the ones who pray)	عسق ('Ayn, Seen, Qaf\letters, none but Allah (Alone) knows their meanings)
	1	0.87	0.84
CL-AraBERT	علمه البيان	إله الناس (Ilāh al-Nās\God of the people)	فيها كتب قيمة (Fīhā kutubun qīmah\In it (are) valuable books)
	1	0.76	0.7

5.3 Model Interpretation Using SHAP

In this technique, we explain the functioning of the model output. For the CL-AraBERT model, after retrieving the exact match and hovering over it among the three options, the resulting SHAP score was 0.03. This score represents the summation of all scores attributed to the contextual features (words). Referring to Figure 3, the blue words in the context negatively impacted the score of the علمه البيان ('Alamahu al-bayan \He taught him eloquence) result, while the red tokens positively influenced it.Specifically, red words force the SHAP value towards the positive side (arrow from left to right), while blue words push the SHAP values from right to left, towards the negative side. The negative and positive values shown correspond to the line, representing the magnitude of the influence exerted by each word. To compare the similar result with other results of lower and upper similarities, observe Table 7.

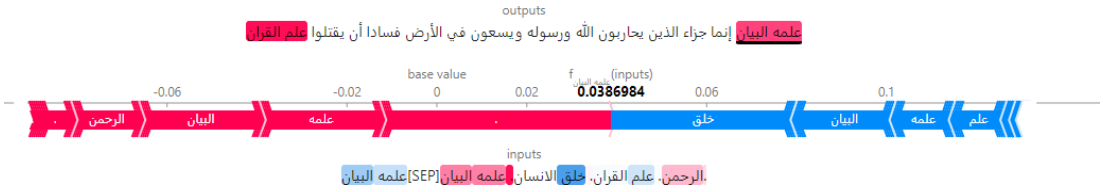


Figure 3. SHAP values for exact-match retrieval for CL-AraBERT model.

After hovering over the similar result علم القران ('Allama al-Qur'an\He taught the Quran), as shown in Figure 4, the obtained SHAP score was 0.04. This score was determined by summing the SHAP scores for both the red and blue words, which, respectively, influenced the positive and negative SHAP scores. To compare the similar result with other results of lower and upper similarities, observe Table 7.

Figure 4. SHAP values of the query similar retrieval for CL-AraBERT model.

In cases where dissimilar results were observed, it was noted that the SHAP score is 0. This suggests that the SHAP values of the words in the context were very low, less than 0.002, indicating their negligible impact on the prediction of the SHAP score for the dissimilar results. Figure 5 illustrates the

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 04, December 2024.

SHAP score of the dissimilar retrievals.

إنما جزاء الذين يحاربون الله ورسوله ويسعون في الأرض فساداً أن يقتلوا

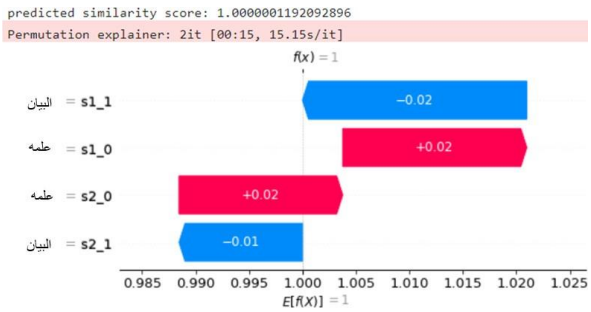
(Innama jazau alladhina yuharibuna Allah wa rasulahu wa yas'awna fi al-ardi fasadan an yuqtalu \ Indeed, the penalty for those who wage war against Allah and His Messenger and strive upon earth [to cause] corruption is none but that they be killed)

علمه البيان

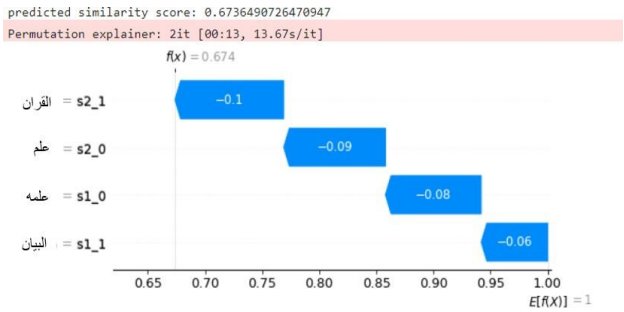
('Alamahu al-bayan \He taught him eloquence)

Figure 5. SHAP values of the query dissimilar retrieval for CL-AraBERT model.

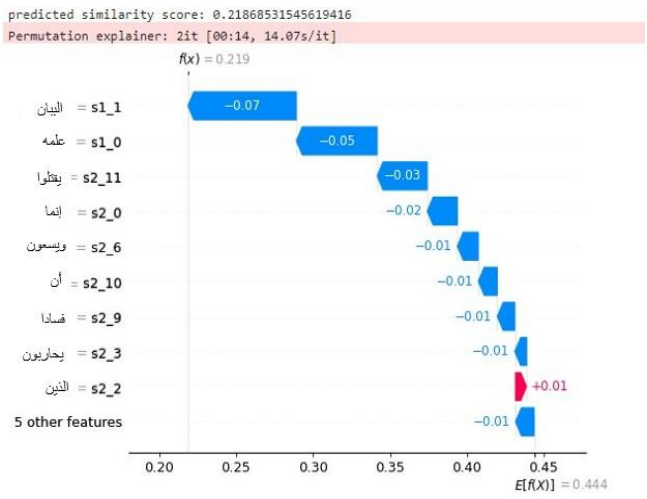
These outcomes show obviously that the SHAP explanation has fitted the semantic-search outcomes in addition to the visual interpretation for each vector effect. All details for the 3 models are mentioned in the Supporting Materials[7]. Despite these results, tuning SHAP QA technique revealed some shortcomings. For example, the word علمه ('Alamahu\He taught him) gave both positive and negative indications simultaneously in calculating the SHAP score for similar results, as shown in Figure 5. This can be attributed to the intricate nature of language and the contextual nuances present. As SHAP utilizes BERT model to score and interpret results, it is expected to reveal words with varying impacts simultaneously. BERT score tokens are based on the surrounding context and the overall tone of the text, which can result in certain words having mixed effects. Additionally, the SHAP score of the exact-match obtained values lower than those of similar results, which contrasts with the outcome of semantic-search evaluation either using the Qur'anic exegesis or using the cosine similarity.



(a) Exact match



(b) Similar retrieval



(c) Non-convergence

Figure 6. STS Explainer scores for CL-AraBERT model.

[7] https://github.com/SajaNakhleh/Quranic-semantic-search

Consequently, in the "STS explanation" step, we passed the same three search results from the "Search" step to the STS Explainer. For the exact-match results, all STS Explainer scores matched the predicted similarity, even if sometimes with a very minor difference (less than 0.01). Figure 6a shows the same scores for similar words with the same effect on the predicted score, which is the value of f(x). The expected scores E(f(x)) were close to the predicted similarity scores f(x), as shown in Fig. 6b. However, the presence of the phrase علم القران ('Allama al-Qur'an\He taught the Quran) in a sentence resulted in a lower score of 0.67 and negatively affected the model prediction. For the nonconvergence results, expected score E(f(x)) was far from the predicted score f(x) and SHAP scores, even if the nonconvergence tokens got high negative affect on the predicted scores. Observe Figure 6c.

Table 7 summarizes the results of 3 scores: cosine similarity score from "Search" step, SHAP scores from "SHAP Interpretation" step and STS Explainer similarity scores from "STS Explainer Interpretation" step. The observed results strongly support the notion that the outcomes of Arabic BERT model are consistent with SHAP results. While the positive SHAP scores (i.e., +0.04 and +0.03) indicate a direct relationship; while the neutral SHAP score indicates dissimilar results. STS Explainer has presented results that align with the cosine-similarity results of the Arabic BERT model. The relationship is direct and positive, as observed from the table (i.e., 0.67 for similar results and 0.21 for dissimilar results).

Table 7. Scores of SHAP interpretation and STS explanation experiments for CL-AraBERT.

ID	Result	Cosine Sim.	QA SHAP	STS Sim.
A	علمه البيان 'Alamahu al-bayan \He taught him eloquence	1	+0.03	1
B	علم القرآن 'Allama al-Qur'an\He taught the Qur'an	0.71	+0.04	0.67
C	إنما جزاء الذين يحاربون الله ورسوله و يسعون في الأرض فساداً أن يقتلوا	0.12	Neutral	0.21

Innama jazau alladhina yuharibuna Allah wa rasulahu wa yas'awna fi al-ardi fasadan an yuqtalu. The penalty for those who wage war against Allah and His Messenger and strive upon earth to cause corruption is none but that they be killed

5.4 Interpretation Using LIME

In this sub-section, we utilize the LIME framework to interpret the predictions made by our logistic regression model in the S-BERT model. Specifically, LIME helps explain why the model classified certain verses as similar (label '1') or not similar (label '0'). The insights gained from this interpretative step are summarized in Table 8. This table presents a comparative analysis of the S-BERT model's performance across various similarity thresholds, taking into account the effects of normalization. The model's assessment is conducted on a single verse from the Holy Quran, showcasing its normalized form, the applied similarity threshold, the predicted probability, which indicates the likelihood assigned by the model to a particular class for the given instance and the resulting similarity classification.

Preliminary findings highlight a nuanced impact of text normalization on similarity assessment. Particularly, the verses evaluated at a similarity threshold of >= 0.8 in Table 8 did not exhibit discernible differences between normalized and normalized (tashkeel removed) versions with a similarity result of 0. This can be attributed to the high similarity threshold and the significant impact of each word (feature) in predicting the "*Not Similar*" class. For instance, in the verse يدريك يزكي, which has been normalized, both words contributed to the "*Not Similar*" classification. Similarly, in the verse وما يدريك لعله يزكى, which has been tashkeel removed, the words يدريك, لعله and يزكى had the highest impact on predicting the "*Not Similar*" class.

In contrast, a slight divergence emerged among verses assessed at a similarity threshold of >= 0.6. For instance, in the verse يدريك يزكي, which has been normalized, the word يزكى contributed to the "*Not Similar*" class, achieving a prediction probability of 0.89, while the word يدريك contributed to the "*Similar*" class, achieving a prediction probability of 0.11. Similarly, in the verse وما يدريك لعله يزكى,

which has been tashkeel-removed, the words وما, يدريك and يزكى had the highest impact on prediction as "*Not Similar*", achieving a prediction probability of 0.79, whereas the word لعله contributed to "*Similar*" class, achieving a prediction probability of 0.21.

Interestingly, the analysis also indicated that text normalization significantly impacts the assessment of text similarity in Arabic. The process of normalization, which includes more than one technique such as tashkeel removing, tatweel removing and punctuation and stop-word removal, has been shown to improve determining similarity of text by ensuring that only the most distinctive lexical features are retained.

These observations underscore the complex nature of text normalization's influence on semantic analysis when employing BERT models. The findings suggest that while normalization can facilitate the identification of superficial textual similarities, it might obscure deeper semantic relationships present in the unaltered text. This insight opens up new avenues for research, particularly in the development of more nuanced-text preprocessing techniques that balance the need for normalization with the preservation of semantic richness. Future studies could explore alternative approaches to text preprocessing and their effects on model interpretability and performance, further enriching our understanding of the intricate dynamics between text normalization and NLP models.

Table 8. Comparative analysis of S-BERT model performance across various similarity thresholds considering normalization effects. Values in **bold** represent the prediction probability for class *'Not Similar* (0)', while values in parentheses indicate the prediction probability for class '*Similar* (1).'

Models	Verses	Normalized Forms	Similarity threshold	Prediction Probability	Similarity Result
S-BERT	وَمَا يُدْرِيكَ لَعَلَّهُ يَزَّكَّى Wa mā yudrikal-la'allahu yazzakkā \And what can make you know? Perhaps he [will] purify himself	يدريك يزكي	>=0.8	**1.00** (0.00)	0
	وَمَا يُدْرِيكَ لَعَلَّهُ يَزَّكَّى Wa mā yudrikal-la'allahu yazzakkā \And what can make you know? Perhaps he [will] purify himself	يدريك يزكي	>=0.6	**0.89** (0.11)	0
	وَمَا يُدْرِيكَ لَعَلَّهُ يَزَّكَّى Wa mā yudrikal-la'allahu yazzakkā \And what can make you know? Perhaps he [will] purify himself	وما يدريك لعله يزكى	>=0.8	**1.00** (0.00)	0
	وَمَا يُدْرِيكَ لَعَلَّهُ يَزَّكَّى Wa mā yudrikal-la'allahu yazzakkā \And what can make you know? Perhaps he [will] purify himself	وما يدريك لعله يزكى	>=0.6	**0.79** (0.21)	0

In our experiments, we have compared the performance of our proposed explainability technique on a semantic-search task with the previously proposed explainability techniques by El Zini et al. [25] on a similar task. Our method has used only AraBERT-based models, while the method by El Zini et al. used eight different models for English texts, two of which were BERT-based. We have found that our method achieved very close outcomes. El Zini et al. used an additional XAI technique called: anchor, while we have used STS Explainer as an additional XAI technique.

The results are different from sentiment analysis, which is a classification task. Our task measures the scores for the query *versus* the retrieved results. Furthermore, the accuracy of the model is not measured in the sentiment-analysis tasks proposed by El Zini et al. Our results demonstrate that SHAP and LIME align with the BERT transformers for the Arabic language. However, it is important to note that our experiment is limited to a specific domain (i.e., Qur'anic text) and further research is needed to generalize the findings.

6. CONCLUSION AND FUTURE WORK

We study the opacity of Arabic transformer models using SHAP and LIME interpretation techniques, applying these to benchmark Qur'anic semantic search within the Qur'anEnc dataset. Our findings reveal that SHAP interpretations align closely with BERT model predictions, highlighting their effectiveness in predicting correct results. Specifically, our experiments demonstrated that SHAP and STS Explainer scores correlate with cosine similarity in exact-match retrievals, with CL-AraBERT showing significant positive effects for exact matches. Interestingly, nonconvergence retrievals exhibited divergent scores, suggesting areas for further investigation. Additionally, our analysis using

LIME of text normalization impact on BERT models' performance revealed that unnormalized texts yield more logical similarity scores in certain instances. These insights not only shed light on the interpretability of Arabic transformer models, but also underscore the nuanced influence of text normalization on semantic-search tasks. In the future, our objective is to extend our examination to include AraT5 and AraGPT models, thereby enhancing our understanding and interpretation of Arabic transformers. This endeavor will undoubtedly contribute to the robustness and reliability of future Arabic-language processing tools.

REFERENCES

[1] A. Ali, T. Schnake, O. Eberle, G. Montavon, K.-R. Müller and L. Wolf, "XAI for Transformers: Better Explanations through Conservative Propagation," Proc. of the 39th Int. Conf. on Machine Learning, ser. Proc. of Machine Learning Research, vol. 162, pp. 435–451, PMLR, [Online], Available: https://proceedings.mlr.press/v162/ali22a.html, 17–23 Jul 2022.

[2] W. Saeed and C. Omlin, "Explainable AI (XAI): A Systematic Meta-survey of Current Challenges and Future Opportunities," Knowledge-based Systems, vol. 263, p. 110273, 2023.

[3] H. U. Khan, S. M. Saqlain, M. Shoaib and M. Sher, "Ontology Based Semantic Search in Holy Quran," International Journal of Future Computer and Communication, vol. 2, no. 6, p. 570, 2013.

[4] I. Al-Huri et al., "Arabic Language: Historic and Sociolinguistic Characteristics," English Literature and Language Review, vol. 1, no. 4, pp. 28–36, 2015.

[5] M. Mustafa, H. AbdAlla and H. Suleman, "Current Approaches in Arabic IR: A Survey," Proc. of the Int. Conf. on Asian Digital Libraries, Digital Libraries: Universal and Ubiquitous Access to Information, Part of the Book Series: Lecture Notes in Comp. Science, vol. 5362, pp. 406–407, 2008.

[6] E. H. Mohamed and E. M. Shokry, "QSST: A Quranic Semantic Search Tool Based on Word Embedding," J. of King Saud Uni.-Computer and Inform. Sciences, vol. 34, no. 3, pp. 934–945, 2022.

[7] R. Malhas and T. Elsayed, "Arabic Machine Reading Comprehension on the Holy Qur'an Using Clarabert," Information Processing and Management, vol. 59, no. 6, p. 103068, DOI:10.1016/j.ipm.2022.103068, 2022.

[8] A. Vaswani et al., "Attention Is All You Need," arXiv: 1706.03762, DOI: 10.48550/arXiv.1706.03762, 2023.

[9] A. Safaya, M. Abdullatif and D. Yuret, "KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media," Proc. of the 14th Workshop on Semantic Evaluation, Barcelona (online): Int. Committee for Computational Linguistics, pp. 2054–2059, Dec. 2020.

[10] N. Reimers and I. Gurevych, "Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation," Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, arXiv: 2004.09813, 2020.

[11] E. Kokalj, B. Škrlj, N. Lavrač, S. Pollak and M. Robnik-Šikonja, "BERT Meets Shapley: Extending SHAP Explanations to Transformer-based Classifiers," Proc. of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, Association for Computational Linguistics, pp. 16–21, [Online]. Available: https://aclanthology.org/2021.hackashop-1.3, Apr. 2021.

[12] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Proc. of the 31st Int. Conf. on Neural Information Processing Systems (NIPS'17), pp. 4768–4777, Red Hook, USA, 2017.

[13] M. T. Ribeiro, S. Singh and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD '16), pp. 1135–1144, New York, NY, USA, DOI: 10.1145/2939672.2939778, 2016.

[14] M. H. Bashir et al., "Arabic Natural Language Processing for Qur'anic Research: A Systematic Review," Artificial Intelligence Review, vol. 56, no. 7, pp. 6801–6854, 2022.

[15] A. Farghaly and K. Shaalan, "Arabic Natural Language Processing: Challenges and Solutions," ACM Trans. on Asian Language Inform. Processing, vol. 8, no. 4, DOI: 10.1145/1644879.1644881, 2009.

[16] K. Dukes and N. Habash, "Morphological Annotation of Quranic Arabic," Proc. of the 7th Int. Conf. on Language Resources and Evaluation (LREC'10), Valletta, Malta, ELRA, May 2010.

[17] S. Altammami and E. Atwell, "Challenging the Transformer-based Models with a Classical Arabic Dataset: Quran and Hadith," Proceedings of the 13th Language Resources and Evaluation Conf., pp. 1462–1471, European Language Resources Association, Marseille, France, Jun. 2022.

[18] D. Vale, A. El-Sharif and M. Ali, "Explainable Artificial Intelligence (XAI) Post-hoc Explainability Methods: Risks and Limitations in Non-discrimination Law," AI and Ethics, vol. 2, no. 4, pp. 815–826, 2022.

[19] E. M. Kenny, E. D. Delaney, D. Greene and M. T. Keane, "Post-hoc Explanation Options for XAI in Deep Learning: The Insight Centre for Data Analytics Perspective," Proc. of Int. Conf. on Pattern Recognition, ICPR Int. Workshops and Challenges, Part of the Book Series: Lecture Notes in

Computer Science, vol. 12663, pp. 20–34, 2021.
[20] A. Sarkar, D. Vijaykeerthy, A. Sarkar and V. N. Balasubramanian, "A Framework for Learning Ante-hoc Explainable Models *via* Concepts," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 10 286–10 295, June 2022.
[21] A. Alsaleh, E. Atwell and A. Altahhan, "Quranic Verses Semantic Relatedness Using AraBERT," Proc. of the 6th Arabic Natural Language Processing Workshop, pp. 185–190, Kyiv, Ukraine, [Online]. Available: https://aclanthology.org/2021.wanlp-1.19, Apr. 2021.
[22] S. Saeed, S. Haider and Q. Rajput, "On Finding Similar Verses from the Holy Quran Using Word Embeddings," Proc. of the 2020 IEEE Int. Conf. on Emerging Trends in Smart Technologies (ICETST), pp. 1–6, Karachi, Pakistan, 2020.
[23] S. M. Lundberg et al., "Explainable Machine-learning Predictions for the Prevention of Hypoxaemia During Surgery," Nature Biomedical Engineering, vol. 2, no. 10, p. 749, 2018.
[24] C. Leiter, P. Lertvittayakumjorn, M. Fomicheva, W. Zhao, Y. Gao and S. Eger, "Towards Explainable Evaluation Metrics for Machine Translation," Journal of Machine Learning Research, vol. 25, pp. 1-49, 2023.
[25] J. El Zini, M. Mansour, B. Mousi and M. Awad, "On the Evaluation of the Plausibility and Faithfulness of Sentiment Analysis Explanations," Proc. of the IFIP Int. Conf. on Artificial Intelligence Applications and Innovations, Part of the Book Series: IFIP Advances in Information and Communication Technology, vol. 647, pp. 338–349, 2022.
[26] N. Habash, Introduction to Arabic Natural Language Processing, 1st Edn., ser. Synthesis Lectures on Human Language Technologies, Morgan and Claypool Publishers, vol. 3, pp. 1-124, 2010.
[27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi, "BERTscore: Evaluating Text Generation with BERT," Towards Data Science, arXiv: 1904.09675, 2020.
[28] Y. Qiao, C. Xiong, Z. Liu and Z. Liu, "Understanding the Behaviors of BERT in Ranking," arXiv preprint, arXiv: 1904.07531, 2019.
[29] S. Yagi, A. Elnagar and S. Fareh, "A Benchmark for Evaluating Arabic Word Embedding Models," Natural Language Engineering, vol. 29, no. 4, p. 978–1003, 2023.
[30] B. Dahy, M. Farouk and K. Fathy, "Arabic Sentences Semantic Similarity Based on Word Embedding," Proc. of the 2022 20th Int. Conf. on Language Engineering (ESOLEC), vol. 20, pp. 35–40, Valencia, Spain, 2022.
[31] G. Salton, "Introduction to Modern Information Retrieval," ISBN 0-07-054484-0, McGraw-Hill, 1983.

ملخص البحث:

تغوص هذه الدّراسة في تفسير ثلاثة نماذج تحويل باللّغة العربية تمّ تطويرها وتكييفها للمهامّ المتعلّقة بالبحث في علم دلالات الألفاظ. فمن خلال دراسة حالةٍ مركّزة، نقوم بتوظيف تلك النّماذج في استرجاع معلوماتٍ من القرآن الكريم، باستخدام تقنيتي (LIME) و (SHAP)، لإلقاء الضّوء على عمليات اتّخاذ القرار في النّماذج المدروسة. وتؤكّد الدّراسة التّحدّيات الفريدة التي تفرضها النّصوص العربية، وبخاصّةٍ القرآنية، كما تُبين أنّ النّماذج المدروسة تعمل على زيادة الشّفافية وقابلية التّفسير لأنظمة البحث في علم دلالات الألفاظ، وبخاصّةٍ بالنّسبة للنّصّ القرآني. وقد أثبتت النّتائج أنّ استخدام التقنيات المذكورة من شأنه أن يعمل على توضيح الآليات الدّاخلية للنّماذج قيد البحث، بالإضافة إلى جعل الاستنتاجات المستخلصة منها متاحةً لشريحةٍ أوسع من الجمهور. ويمكن القول إنّ مساهمة هذه الدّراسة في مجال البحث هي ذات وجهين. فمن جهة، تُغني المجال فيما يتعلق بالبحث في علم دلالات الألفاظ في النّصوص باللّغة العربية، وبخاصّة النّصّ القرآني، ومن جهة أخرى فهي تبيّن الفائدة من التّقنيات المستخدمة في تعزيز فهمنا للنّصوص والوثائق الدّينية.

والجدير بالذّكر أنّ هذه الدراسة من شأنها أن تُسهم في جَسْر الهُوّة بين تقنيات تعلُّم الآلة المتقدمة واحتياجات المستخدمين الّذين يسْعَون إلى استكشاف نصوصٍ "معقّدة"، كما هو النّصّ القرآني.