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[‡]

March 22, 2024

[*]A. Mustafa, S. Nakhleh, R. Irsheidat, and R. Alruosan are with the Department 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

[†](Corresponding author: Ahmad M. Mustafa)

1. APPENDIX

A Experiment 2 results: Interpretation using SHAP

A.1 S-BERT model results

1. **S-BERT** results for searching "علمه البيان" - Query1 .The Sentence "علمه البيان" has been passed to the model: S-BERT. Then the observed results A, B and C (Table 1 – results column) were passed to SHAP and STS explainer XAI models to explain the BERT results.

   The observed results strongly support the notion that the outcomes of S-BERT model are consistent with SHAP results. As the positive SHAP scores (i.e. +.04 and +.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. 0.99 for the exact match, .84 for the similar results, and the lowest value was 0.6 for dissimilar results)

Table 1: Explaining the results of: S-BERT

ID	Result	Cosine Similarity	SHAP for QA score	STS explainer similarity
A	علمه البيان	1	+0.03	0.99
B	علم القران	0.87	+0.04	0.84
C	يا أيها الذين آمنوا إذا تداينتم بدين إلى أجل مسمى فاكتبو	0.16	Neutral	0.60

SHAP for QA: The following Query1 - result, Query2 – result, and Query3 – result are explained with the SHAP for QA model. Each result has been shown with its SHAP score and its effect on the model's overall score. Red color means a positive effect, blue color means a negative effect, and the color degree means the effect degree, so dark red means a large positive effect, while light red means a low positive effect. The non-colored results have no effect on the SHAP score (neutral score) if it does not exist in the context or in the query. The results with very low cosine similarity scores mean non-convergence results which go neutral SHAP for QA score and the lowest STS explainer score also.

2. **SHAP for QA- Query 1** using **S-BERT** model
   For the **"S-BERT"** model, once the exact match was hovered, the obtained, the SHAP score will +0.03, which indicated height positive affect on the overall SHAP score, observe Fig. 1 SHAP results for the exact match.

   Query1- result 1.A:
   ("علمه البيان ـ علمه البيان")

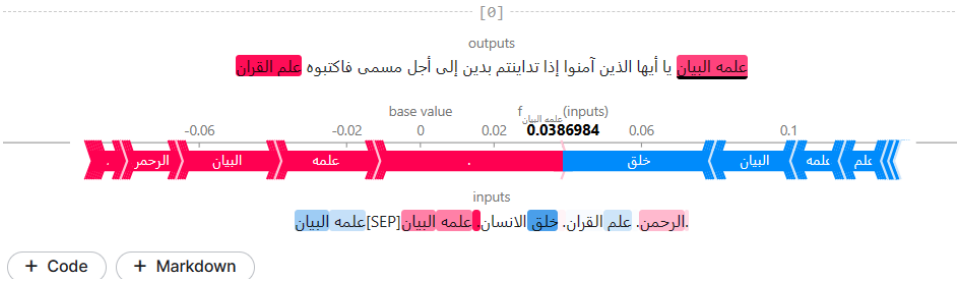


Figure 1: SHAP values for exact match retrieval for the **S-BERT** model

For the similar result "علم القران ", the SHAP score was also positive and directly affected the prediction of the SHAP score. Observe Fig. 2.

Query1- result 1.B:
("علم القرانعلمه البيان")

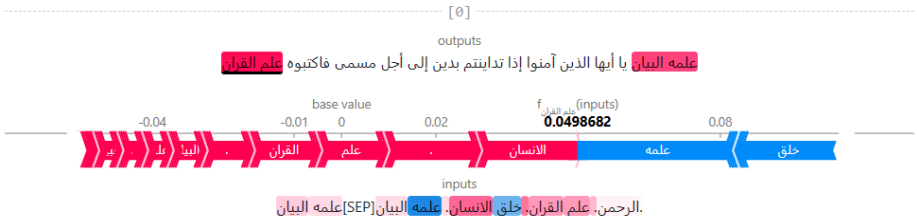


Figure 2: SHAP values of the query similar retrieval for the **S-BERT** model

And finally, for the nonconvergence results, the SHAP score was neutral, which means that it had no effect for the SHAP score prediction. Fig. 3 shows the SHAP score of the non-convergence retrievals.

Query1- result 1.C:
("يا أيها الذين آمنوا إذا تداينتم بدين إلى أجل مسمى
فاكتبوه ، علمه البيان")

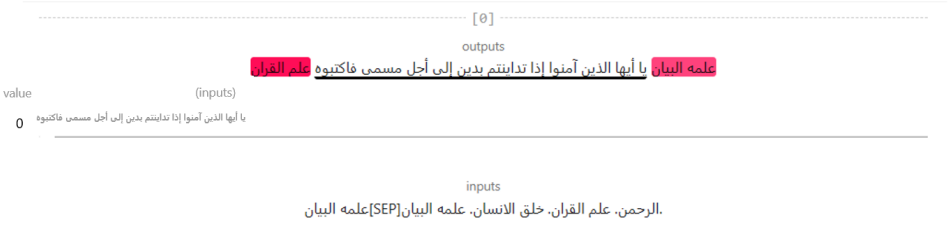


Figure 3: SHAP values of the query non-convergence retrieval for the **S-BERT** model

3. **STS explainer**:
For the exact match results, all the STS explainer scores have matched the Cosine similarity even if sometimes with very minor variety (.01 or less). Fig. 4, shows the same scores for the similar word with equal effect on the predicted score.

Query1 – result 2.A:
("علمه البيان ـ علمه البيان")

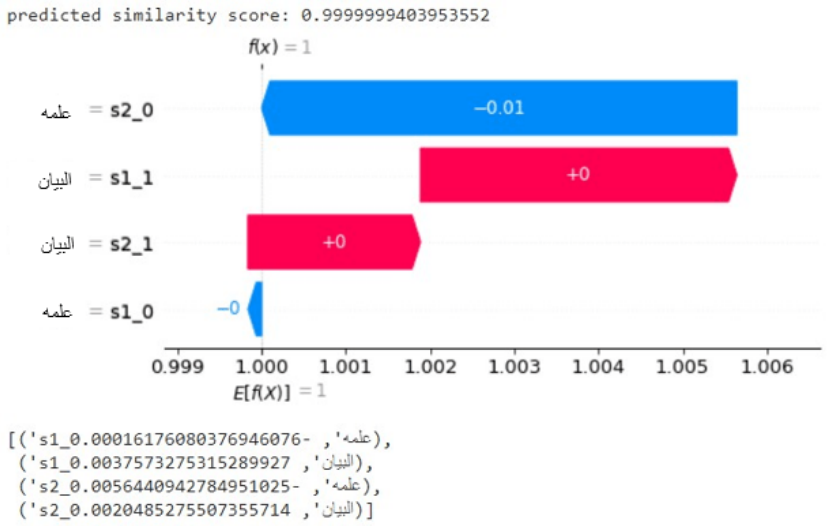


Figure 4: STS explainer scores of exact match retrieval for **S-BERT** model (Query1 – result 2.A)

For the similar retrievals, scores were close to the cosine similarity, as shown on Fig. 17, the little variation in the sentence: علم القران caused less score, .84, with negative affect on the model prediction. Observe Fig. 5 which shows the contribution of each token in the STS explainer score.

Query1 – result 2.B:
("علم القران ـ علمه البيان")

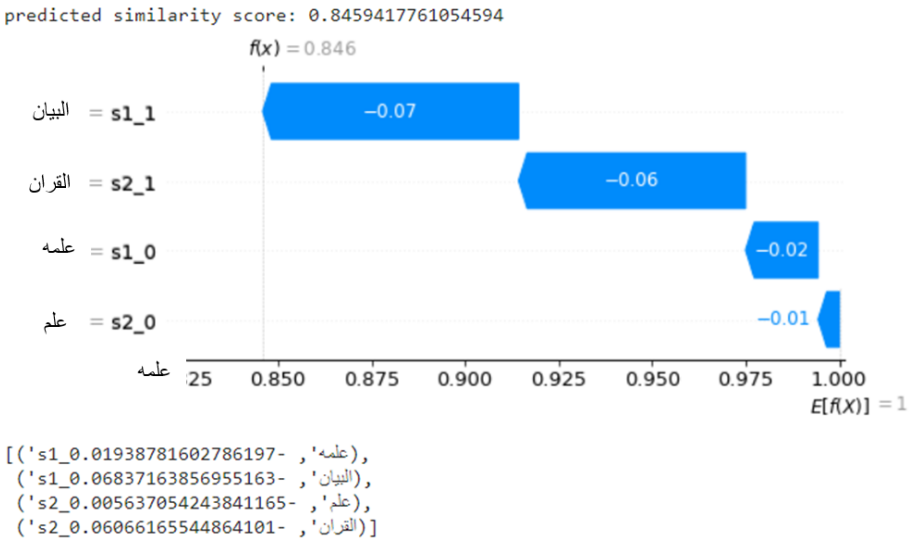
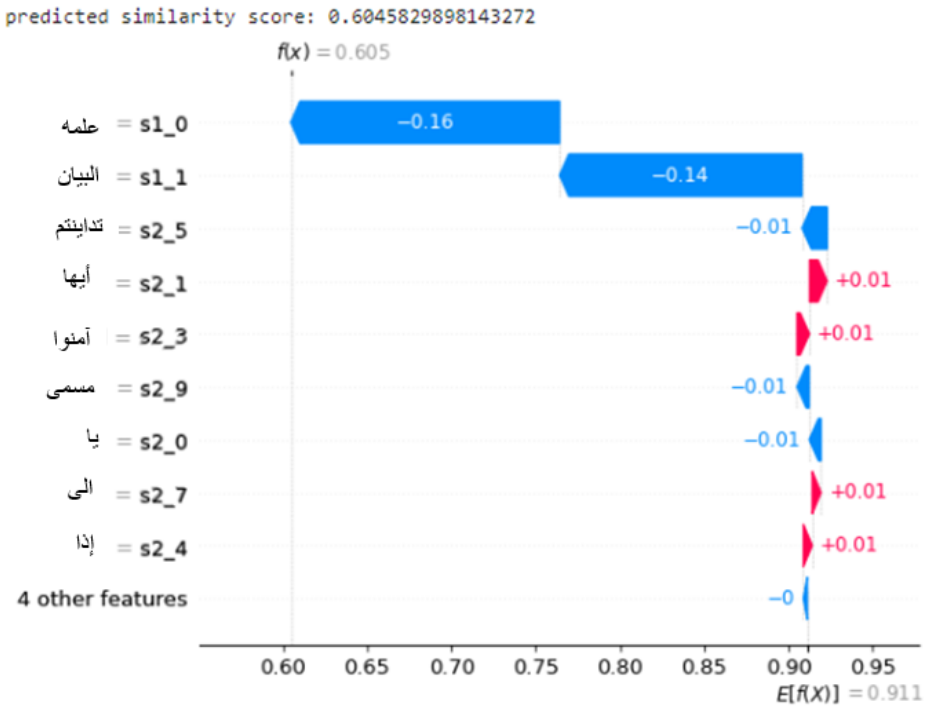


Figure 5: STS explainer scores for similar retrieval for **S-BERT** model (Query1 – result 2.B)

For the nonconvergence results, the STS explainer score was far from the cosine similarity and SHAP scores, even if the nonconvergence tokens got a high negative affect on the predicted scores. Fig. 6 shows the contribution of each token in the STS explainer score.

Query1 – result 2.C:
("يا أيها الذين آمنوا إذا تداينتم بدين إلى أجل مسمى
فاكتبوه ، علمه البيان")




```
[('s1_0.15950797063749814- ,'علمه'),
 ('s1_0.14396453164076578- ,'البيان'),
 ('s2_0.007063249305526482- ,'يا'),
 ('s2_0.010539364141229182 ,'أيها'),
 ('s2_0.003192417622455488 ,'الذين'),
 ('s2_0.0074042277132727735 ,'آمنوا'),
 ('s2_0.005220412314416823 ,'إذا'),
 ('s2_0.014943316977225491- ,'تداينتم'),
 ('s2_0.005105463353910097- ,'بدين'),
 ('s2_0.0052235301947105295 ,'إلى'),
 ('s2_0.002630361112718755 ,'أجل'),
 ('s2_0.007191340152928483- ,'مسمى'),
 ('s2_0.0030340309179007117- ,'فاكتبوه')]
```

Figure 6: STS explainer scores for non-convergence retrievals **S-BERT** model (Query1 – result 2.C)

A.2 ArabicBERT model results

1. **ArabicBERT (semantic search model)** results for searching "علمه البيان" – Query2. The Sentence"علمه البيان" was passed to the model: ArabicBERT. Then the observed results A, B and C (Table 2 – results column) were passed to SHAP and STS explainer XAI models to explain the BERT results.

   The observed results strongly support the notion that the outcomes of ArabicBERT model are consistent with SHAP results. As the positive SHAP scores (i.e. +.04 and +.03) indicate a direct strong relationship; while the neutral SHAP score indicates dissimilar results. STS Explainer has presented results that align with the cosine similarity results of the ArabicBERT model. The relationship is direct and positive, as observed from the table (i.e. .83 for the similar results and .43 for dissimilar results)

2. **SHAP for QA-Query 2 using** ArabicBERT model
   For the **"ArabicBERT"** model, once the exact match was hovered, the obtained, the SHAP score was +.03, which has indicated height positive affect on the overall SHAP score, observe Fig. 7.

   Query2 – result 1.A:
   ("علمه البيان ـ علمه البيان")

Table 2: Explaining the results of ArabicBERT model

ID	Result	Cosine Similarity	SHAP for QA score	STS explainer similarity
A	علمه البيان	0.99	+0.03	1
B	علم القران	0.87	+0.04	0.83
C	أيها الذين آمنوا إذا قمتم إلى الصلاة فاغسلوا وجوهك	0.16	Neutral	0.43

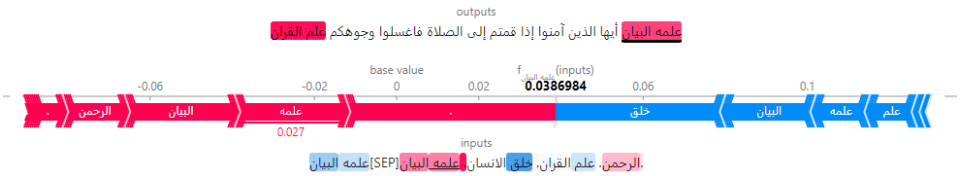


Figure 7: SHAP values for exact match retrieval for the **ArabicBERT** model

For the similar result "علم القران", the SHAP score was also positive and strongly affected SHAP score prediction. observe Fig. 8.

Query2 – result 1.B:
("علم القران ـ علمه البيان")

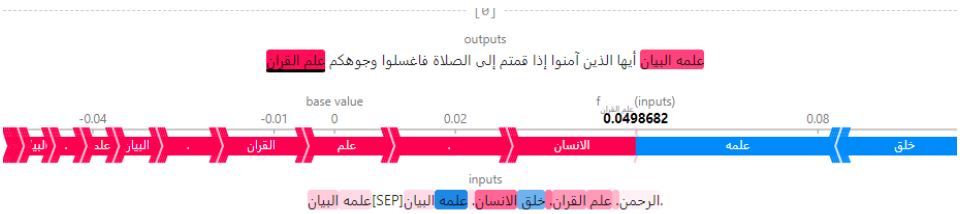


Figure 8: SHAP values of the query similar retrieval for the **ArabicBERT** model

And finally, for the nonconvergence results, the SHAP score was nuetral, which means it had no effect for the SHAP score prediction. Fig. 9 shows the SHAP score of the non-convergence retrievals.

Query2 – result 1.C:
("أيها الذين آمنوا إذا قمتم إلى الصلاة فاغسلوا
وجوهكم ـ علمه البيان")

Figure 9: SHAP values of the query non-convergence retrieval for the **ArabicBERT** model

3. **STS explainer:**
For the exact match results, all the STS explainer scores have matched the Cosine similarity even if sometimes with very minor variety (.01 or less). Fig. 10, shows the same scores for the similar word with equal effect on the predicted score.

Query2 – result 2.A:
("علمه البيان ـ علمه البيان")

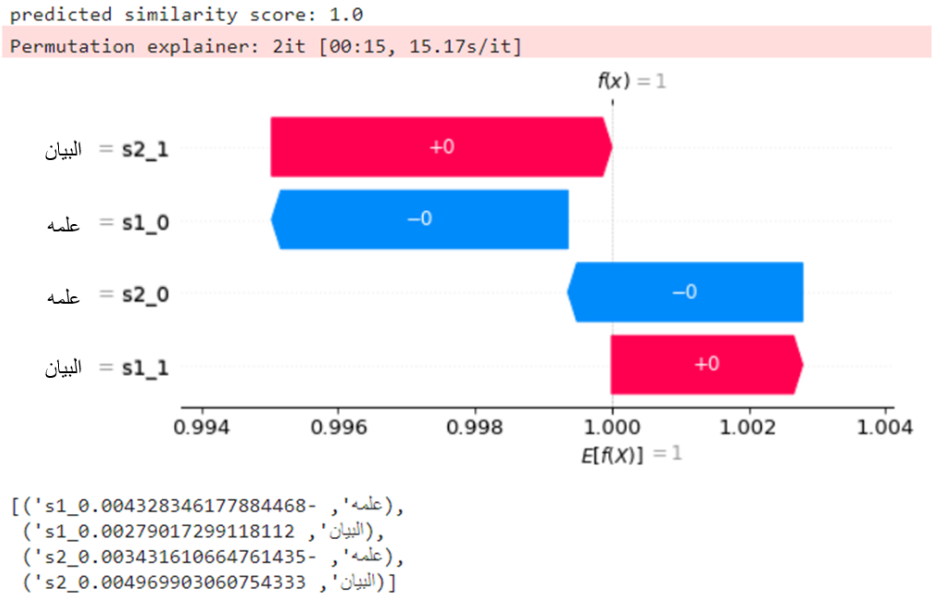


Figure 10: STS explainer scores of exact match retrieval for **ArabicBERT** model (Query2 – result 2.A)

For the similar retrievals, scores were close to the cosine similarity, as shown on Fig. 23, the little variation in the sentence: علم القران caused less score, .83, with negative affect on the model prediction. Observe Fig. 11 that shows the contribution of each token in STS explainer score.

Query2 – result 2.B:
("علم القران ـ علمه البيان")

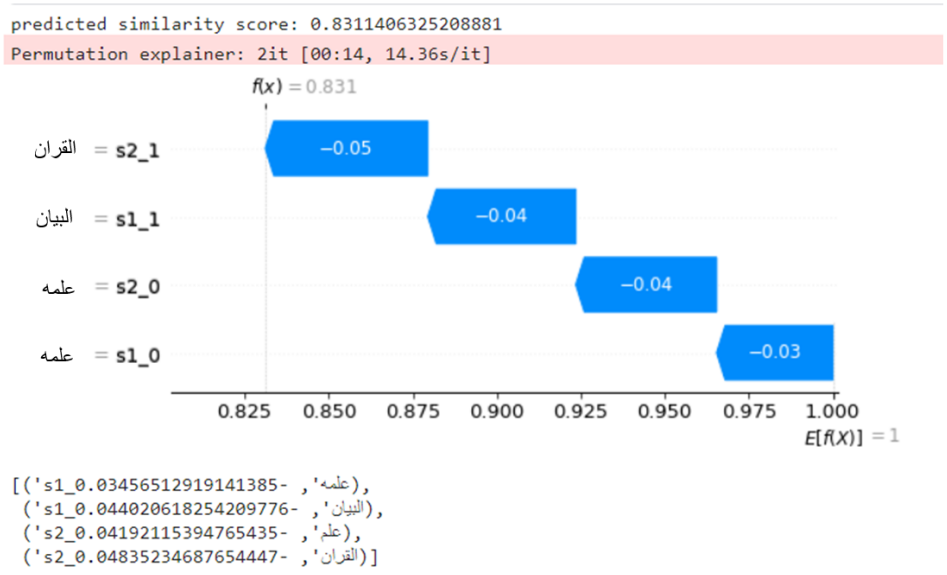


Figure 11: STS explainer scores for similar retrieval for **ArabicBERT** model (Query2 – result 2.B)

For the non-convergence results, the STS explainer score was far from the cosine similarity and SHAP scores, even if the non-convergence tokens got a high negative affect on the predicted scores. Fig. 12 shows the contribution of each token to STS explainer score.

Query2 – result 2.C:
("أيها الذين آمنوا إذا قمتم إلى الصلاة فاغسلوا
وجوهكم ـ علمه البيان")

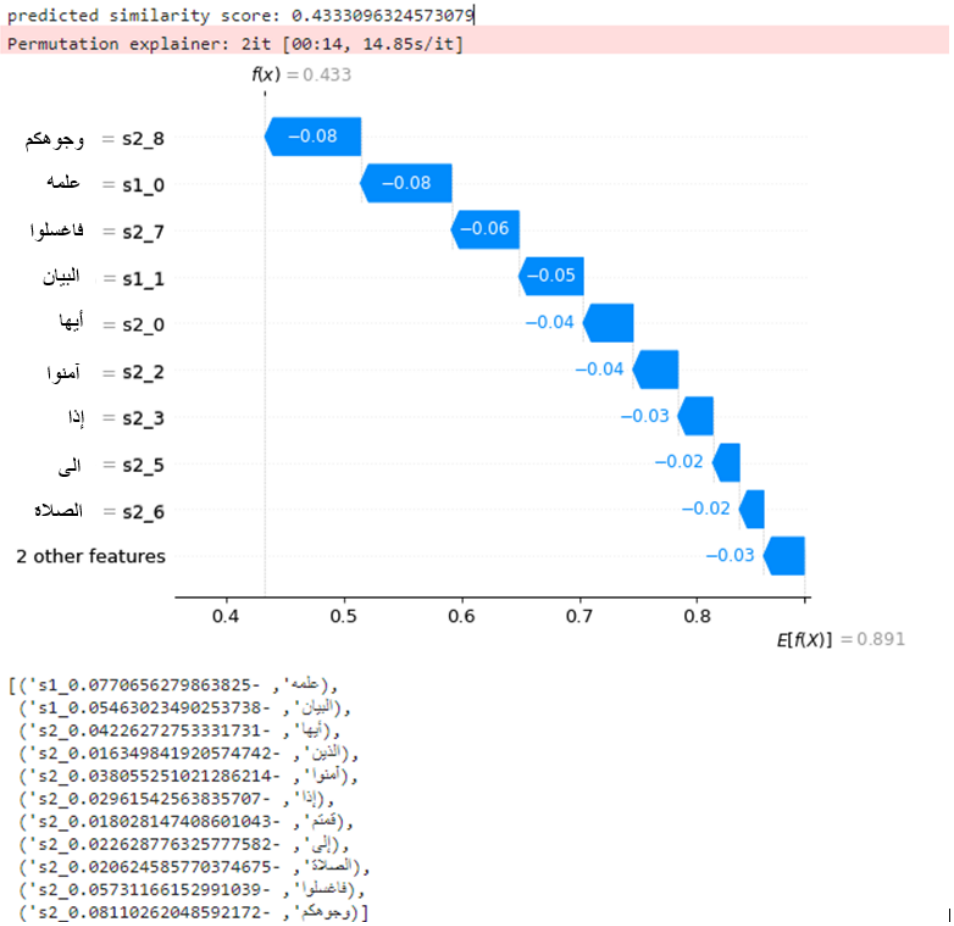


Figure 12: STS explainer scores for non-convergence retrievals **ArabicBERT** model (Query2 – result 2.C)

A.3 CL-AraBERT model results

1. **CL-AraBERT (Arabic bert model for classical texts)** results for searching "علمه البيان" – Query3.

   The Sentence "علمه البيان" was passed to the model: CL-AraBERT. Then the observed results A, B and C (Table 3 – results column) were passed to SHAP and STS explainer XAI models to explain the BERT results.

   The observed results strongly support the notion that the outcomes of CL-AraBERT model are consistent with SHAP results. As the positive SHAP scores (i.e. +.04 and +.03) indicate a direct strong relationship; while the neutral SHAP score indicates dissimilar results. STS Explainer has presented results that align with the cosine similarity results of the CL-AraBERT model. The relationship is direct and positive, as observed from the table 3 (i.e. .67 for the similar results and .21 for dissimilar results)

2. **SHAP for QA**
   For the **"CL-AraBERT"** model, once the exact match was hovered, the obtained, the SHAP score will +.03, which indicated height positive affect on the overall SHAP score, observe Fig. 13.

Table 3: Explaining the results of 3. CL-AraBERT model

ID	Result	Cosine Similarity	SHAP for QA score	STS similarity	explainer
A	علمه البيان	1	+0.03	1	
B	علم القران	0.71	+0.04	0.67	
C	إنما جزاء الذين يحاربون الله ورسوله ويسعون في الأرض فسادا أن يقتلوا	0.12	Neutral	0.21	

Query3 – result 1.A:
("علمه البيان ـ علمه البيان")

Figure 13: SHAP values for exact match retrieval for the **CL-AraBERT** model Query3- result A)

For the similar result علم القران, the SHAP score was also positive and strongly affected SHAP score prediction. observe Fig. 14.

Query3 – result 1.B:
("علم القران ـ علمه البيان")

Figure 14: SHAP values of the query similar retrieval for the **CL-AraBERT** model Query3- result B)

And finally, for the non-convergence results, the SHAP score was nuetral which means it had no effect for the SHAP score prediction. Fig. 15. shows the SHAP score of the non-convergence retrievals.

Query3 – result 1.C:
("إنما جزاء الذين يحاربون الله ورسوله
ويسعون في الأرض فسادا أن يقتلوا ـ علمه البيان")

Figure 15: SHAP values of the query non-convergence retrieval for the **CL-AraBERT** model Query3- result C)

3. **STS explainer**:

For the exact match results, all the STS explainer scores matched the Cosine similarity even if sometimes with very minor variety (.01 or less). Fig. 16, shows the same scores for the similar word with equal effect on the predicted score.

Query3 – result 2.A:
("علمه البيان ـ علمه البيان")

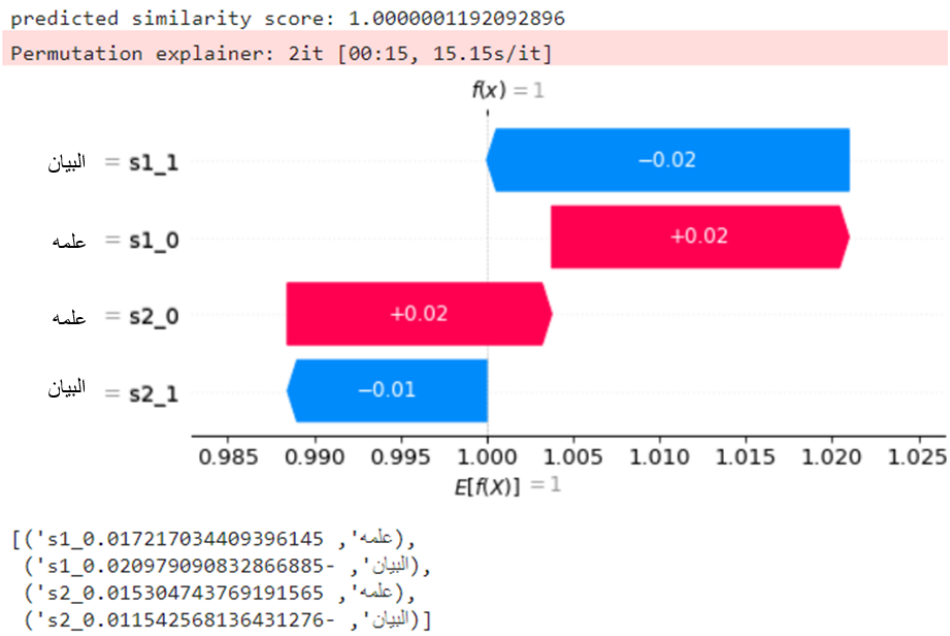
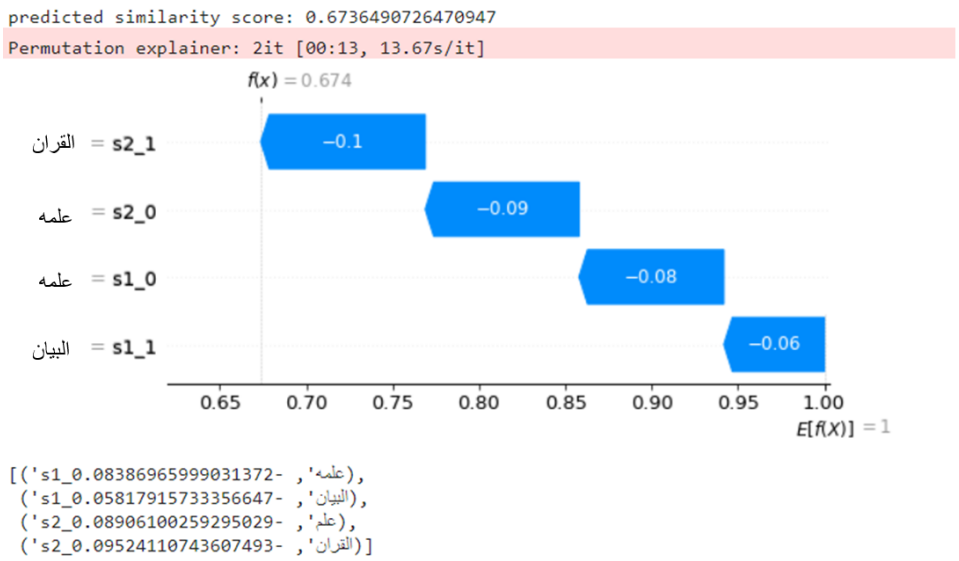


Figure 16: STS explainer scores of exact match retrieval for **CL-AraBERT** model (Query3 – result 2.A)

For the similar retrievals, scores were close to the cosine similarity, as shown on Fig. 29, the little variation in the sentence: علم القران caused less score, .67, with a negative affect on the model prediction. Observe Fig. 17 that shows the contribution of each token in STS explainer score.

Query3 – result 2.B:
("علم القران ـ علمه البيان")




```
[('s1_0.08386965999031372- ,'علمه),
 ('s1_0.05817915733356647- ,'البيان),
 ('s2_0.08906100259295029- ,'علم),
 ('s2_0.09524110743607493- ,'القران)]
```

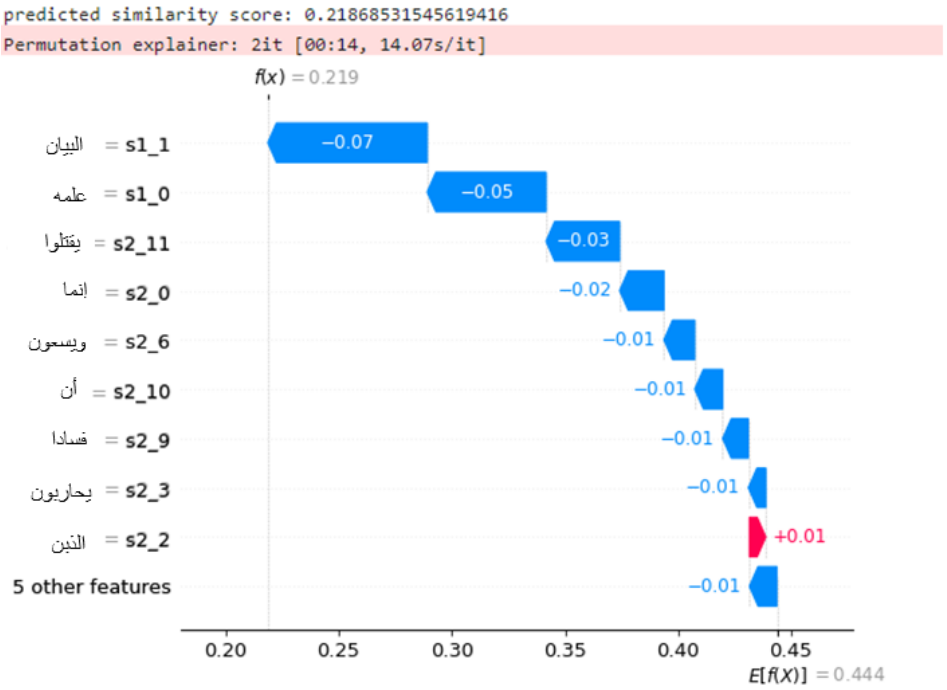
Figure 17: STS explainer scores for similar retrieval for**CL-AraBERT** model (Query3 – result 2.B)

For the non-convergence results, the STS explainer score was far from the cosine similarity and SHAP scores, even if the non-convergence tokens got a high negative effect on the predicted scores. Fig. 18 shows the contribution of each token in STS explainer score.

Query3 – result 2.C:

(" إنما جزاء الذين يحاربون الله ورسوله

" ويسعون في الأرض فسادا أن يقتلوا ـ علمه البيان)




```
[('s1_0.05252793710317845- ,'علمه),
 ('s1_0.07044648355248243- ,'البيان),
 ('s2_0.019593322604968096- ,'إنما),
 ('s2_0.0015434107420502399- ,'جزاء),
 ('s2_0.007248581414891137 ,'الذين),
 ('s2_0.007717546504048578- ,'يحاربون),
 ('s2_0.0025137103368237054- ,'الله),
 ('s2_0.0035695304247016666- ,'ورسوله),
 ('s2_0.01361314686491935- ,'ويسعون),
 ('s2_0.006805564972017178- ,'في),
 ('s2_0.0022419498576936436 ,'الأرض),
 ('s2_0.011354945735688602- ,'فسادا),
 ('s2_0.012336741863546188- ,'أن),
 ('s2_0.03252737945101358- ,'يقتلوا)]
```

Figure 18: STS explainer scores for non-convergence retrievals **CL-AraBERT** model (Query3 – result 2.C)

B Experiment 3 results: Interpretation using LIME

The subsequent set of results showcases the outcomes of the classification approach employing various models based on different embeddings. These results are detailed in Tables 4 and 5. In these tables, both normalized and unnormalized selected verses are presented alongside their classification results at different similarity thresholds.

Table 4: ArabicBERT Model Classification Outcomes of Selected Verses at Varied Similarity Thresholds.

Verses	Similarity value (label)	Similarity threshold	With/Without Normalization
بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ ـ علمه البيان	0	>=0.8	With
الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ ـ علمه البيان	0	>=0.8	With
الرَّحْمَنِ الرَّحِيمِ ـ علمه البيان	1	>=0.8	With
مَالِكِ يَوْمِ الدِّينِ ـ علمه البيان	1	>=0.8	With
صِرَاطَ الَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْرِ الْمَغْضُوبِ عَلَيْهِمْ وَلَا الضَّالِّينَ ـ علمه البيان	0	>=0.6	With
الذين يؤمنون بالغيب ويقيمون الصلاة ومما رزقناهم ينفقون ـ علمه البيان	0	>=0.6	With
بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ ـ علمه البيان	1	>=0.6	With
الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ ـ علمه البيان	1	>=0.6	With
عَلَّمَهُ الْبَيَانَ ـ علمه البيان	1	==1	With
بسم الله الرحمن الرحيم ـ علمه البيان	0	>=0.8	With
الحمد لله رب العالمين ـ علمه البيان	0	>=0.8	With
الرحمن الرحيم ـ علمه البيان	1	>=0.8	Without
مالك يوم الدين ـ علمه البيان	1	>=0.8	Without
إياك نعبد وإياك نستعين ـ علمه البيان	0	>=0.6	Without
صراط الذين أنعمت عليهم غير المغضوب عليهم ولا الضالين ـ علمه البيان	0	>=0.6	Without
بسم الله الرحمن الرحيم ـ علمه البيان	1	>=0.6	Without
الحمد لله رب العالمين ـ علمه البيان	1	>=0.6	Without
علمه البيان ـ علمه البيان	1	==1	Without

Table 5: CL-AraBERT Model Classification Outcomes of Selected Verses at Varied Similarity Thresholds.

Verses	Similarity value (label)	Similarity threshold	With/Without Normalization
بسْمِ اللَّهِ الرَّحْمَن الرَّحِيمِ ـ علمه البيان	0	>=0.8	With
الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ ـ علمه البيان	0	>=0.8	With
عَلَّمَهُ الْبَيَانَ ـ علمه البيان	1	>=0.8	With
بسْمِ اللَّهِ الرَّحْمَن الرَّحِيمِ ـ علمه البيان	0	>=0.6	With
الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ ـ علمه البيان	0	>=0.6	With
الرَّحْمَن الرَّحِيمِ ـ علمه البيان	1	>=0.6	With
مَالِكِ يَوْمِ الدِّينِ ـ علمه البيان	1	>=0.6	With
عَلَّمَهُ الْبَيَانَ ـ علمه البيان	1	==1	With
بسم الله الرحمن الرحيم ـ علمه البيان	0	>=0.8	Without
الحمد لله رب العالمين ـ علمه البيان	0	>=0.8	Without
علمه البيان ـ علمه البيان	1	>=0.8	Without
بسم الله الرحمن الرحيم ـ علمه البيان	0	>=0.6	Without
الحمد لله رب العالمين ـ علمه البيان	0	>=0.6	Without
الرحمن الرحيم ـ علمه البيان	1	>=0.6	Without
مالك يوم الدين ـ علمه البيان	1	>=0.6	Without
علمه البيان ـ علمه البيان	1	==1	Without

Note: we could not show the LIME explainer result for the CL-AraBERT model on Similarity threshold >=0.8 because the CL-AraBERT returned one Similarity value (label) on Similarity threshold >=0.8, which makes the split of the data into training and testing thing impossible to do.

The subsequent set of results showcases the outcomes of the LIME on the different models under various similarity thresholds with and without normalization applied. These results are detailed in Tables 14 and 15. In these tables, both normalized and unnormalized verses are presented alongside their prediction probability and similarity results at different similarity thresholds.

Table 6: LIME Results on the ArabicBERT model under various similarity thresholds with and without normalization applied.

Models	Verses	Normalized Forms	Similarity threshold	Prediction Probability	Result
ArabicBERT	وَمَا جَعَلْنَا لِبَشَرٍ مِّن قَبْلِكَ ٱلْخُلْدَ أَفَإِيْن مِّتَّ فَهُمُ ٱلْخَٰلِدُونَ	جعلنا لبشر قبلك الخلد مت الخالدون	>=0.8	0.95	0
	وَأَشْرِكْهُ فِي أَمْرِي	واشركه امري	>=0.6	0.94	1
	وقل اعملوا فسيرى الله عملكم ورسوله والمؤمنون	-	>=0.8	0.99	0
	وستردون إلى عالم الغيب والشهادة فينبئكم بما كنتم تعملون				
	بلسان عربي مبين	-	>=0.6	0.88	1

Table 7: LIME Results on the CL-AraBERT model under various similarity thresholds with and without normalization applied.

Models	Verses	Normalized Forms	Similarity threshold	Prediction Probability	Result
CL-AraBERT	وَلِكُلٍّ جَعَلْنَا مَوَالِيَ مِمَّا تَرَكَ الْوَالِدَانِ وَالْأَقْرَبُونَ وَالَّذِينَ عَقَدَتْ يْمَانُكُمْ فَآتُوهُمْ نَصِيبَهُمْ إِنَّ اللَّهَ كَانَ عَلَىٰ كُلِّ شَيْءٍ شَهِيدًا	جعلنا موالي الوالدين و الاقربون عقدت ايمانكم فاتوهم نصيبهم الله شيء شهيدا	>=0.6	0.97	0
	والذين هم على صلاتهم يحافظون	-	>=0.6	0.99	0