Interpreting Arabic Transformer Models: A Study on XAI Interpretability for Quranic Semantic Search Models

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

January 9, 2024

*(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 " " - Query1 .The Sentence " " 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 strong 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		1	+0.03	0.99
B		0.87	+0.04	0.84
C		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:
   (" - ")

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

   Query1- result 1.B:
   (" - ")

   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.

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

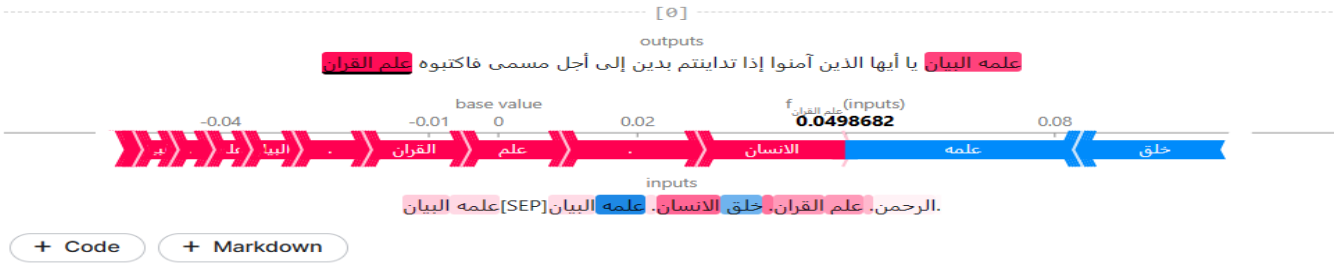

Figure 2: SHAP values of the query similar retrieval for the **S-BERT** model

Query1- result 1.C:
"
"

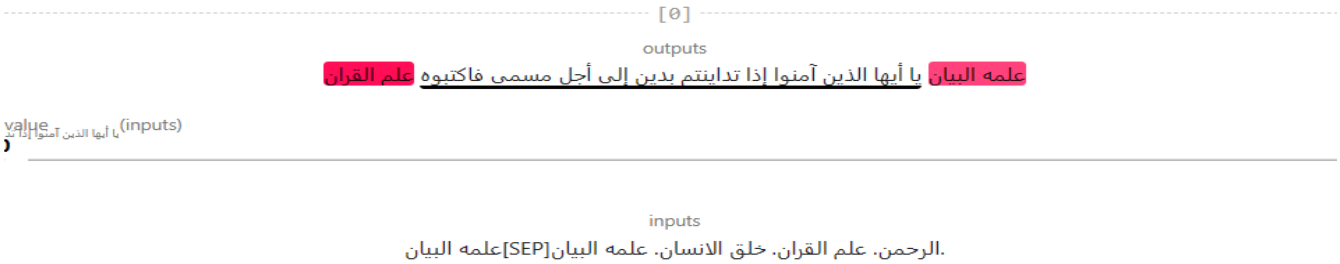

Figure 3: SHAP values of the query non-convergence retrieval for the **S-BERT** model

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:
(" - ")

For the similar retrievals, scores were close to the cosine similarity, as shown on Fig. 17, the little variation in the sentence: 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:
(" - ")

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

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

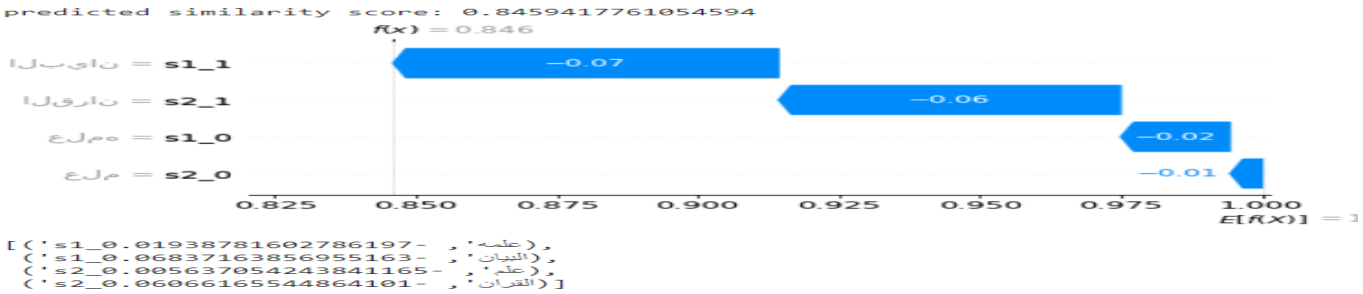

Figure 5: STS explainer scores for similar retrieval for **S-BERT** model (Query1 – result 2.B)

each token in the STS explainer score.

Query1 – result 2.C:
("

‹ ")

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 " " – Query2. The Sentence" " 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)

Table 2: Explaining the results of ArabicBERT model

ID	Result	Cosine Similarity	SHAP for QA score	STS similarity	explainer
A		0.99	+0.03	1	
B		0.87	+0.04	0.83	
C		0.16	Neutral	0.43	

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:
(" - ")

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

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

Query2 – result 1.B:
(" - ")

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:
("
- ")

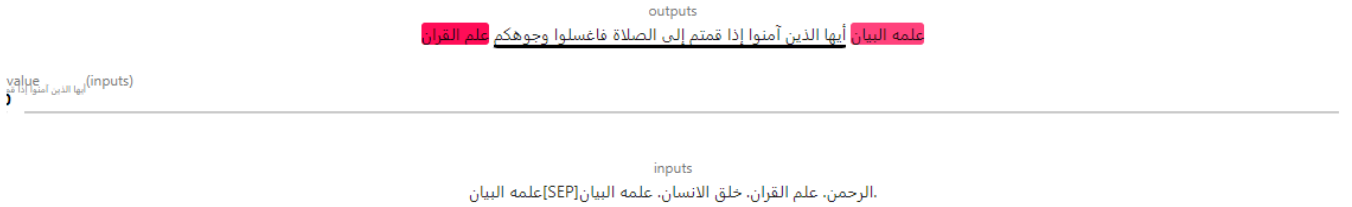

Figure 9: SHAP values of the query non-convergence retrieval for the **ArabicBERT** model

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:
   (" - ")

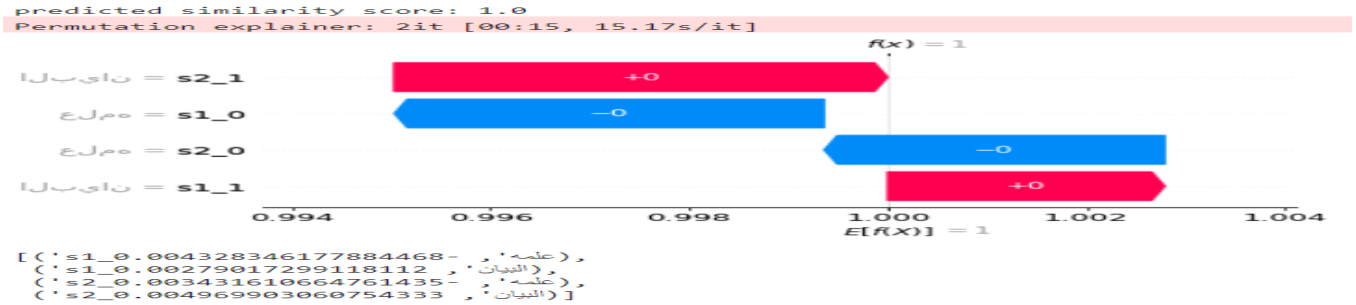

Figure 10: STS explainer scores of exact match retrieval for **ArabicBERT** model (Query2 – result 2.A)

For the similar retrievals, scores were close to the cosine similarity, as shown on Fig. 23, the little variation in the sentence: 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:
(" - ")

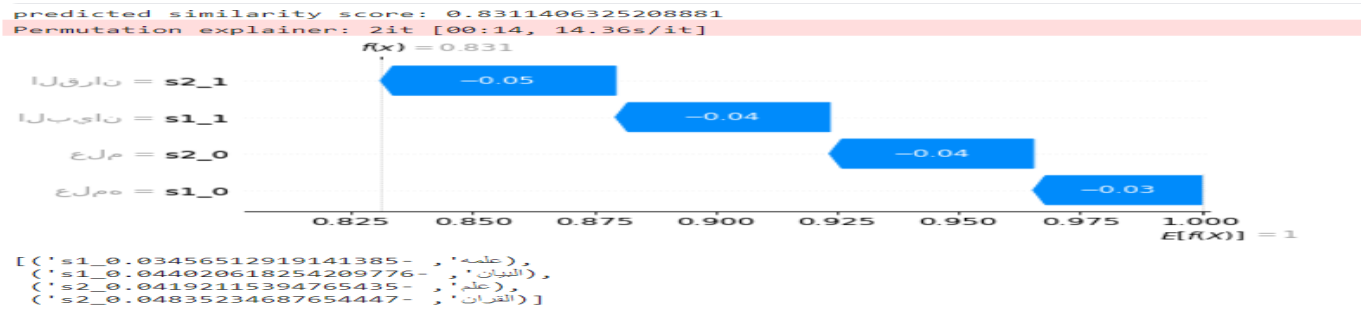

Figure 11: STS explainer scores for similar retrieval for **ArabicBERT** model (Query2 – result 2.B)

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:
("

- ")

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 " " – Query3. The Sentence " " 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)

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

ID	Result	Cosine Similarity	SHAP for QA score	STS explainer similarity
A		1	+0.03	1
B		0.71	+0.04	0.67
C		0.12	Neutral	0.21

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.

   Query3 – result 1.A:
   (" - ")

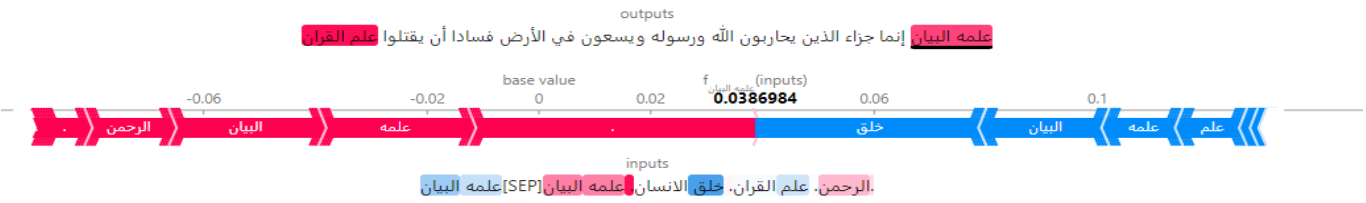

Figure 13: SHAP values for exact match retrieval for the **CL-AraBERT** model Query3- result A)

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

Query3 – result 1.B:
(" - ")

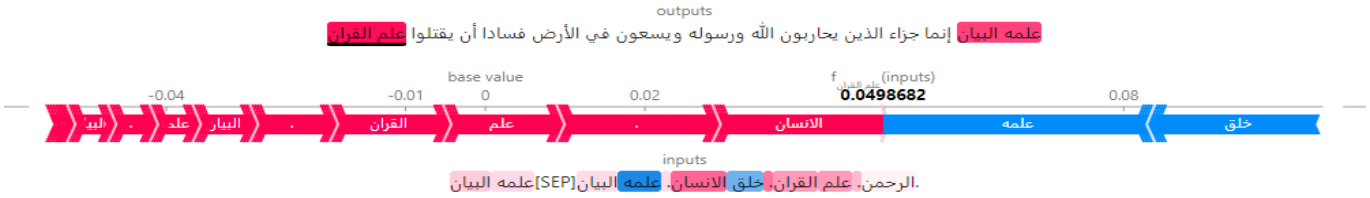

Figure 14: SHAP values of the query similar retrieval for the **CL-AraBERT** model Query3- result B)

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:
("
- ")

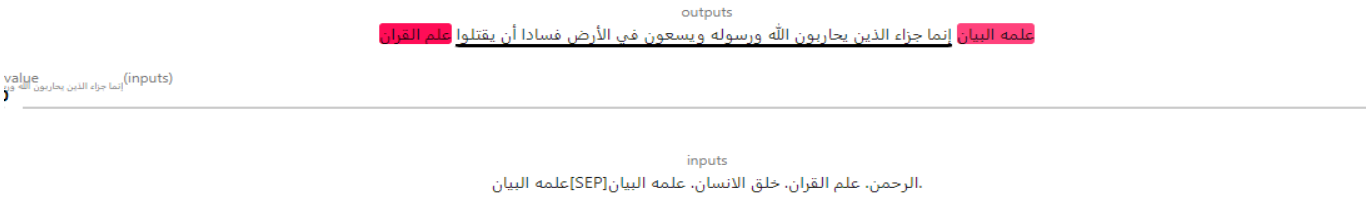

Figure 15: SHAP values of the query non-convergence retrieval for the **CL-AraBERT** model Query3- result C)

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:
(" - ")

For the similar retrievals, scores were close to the cosine similarity, as shown on Fig. 29, the little variation in the sentence: 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:
(" - ")

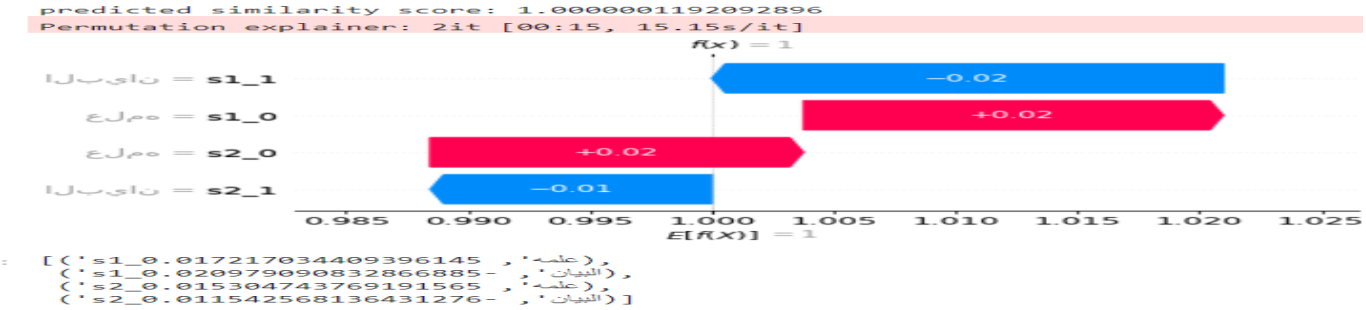

Figure 16: STS explainer scores of exact match retrieval for **CL-AraBERT** model (Query3 – result 2.A)

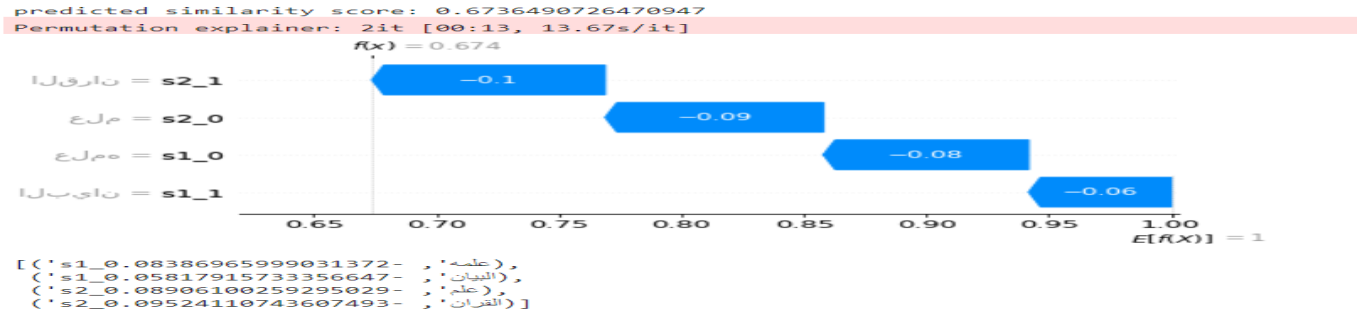

Figure 17: STS explainer scores for similar retrieval for**CL-AraBERT** model (Query3 – result 2.B)

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 high negative affect on the predicted scores. Fig. 18 shows the contribution of each token in STS explainer score.

Query3 – result 2.C:
("
- ")

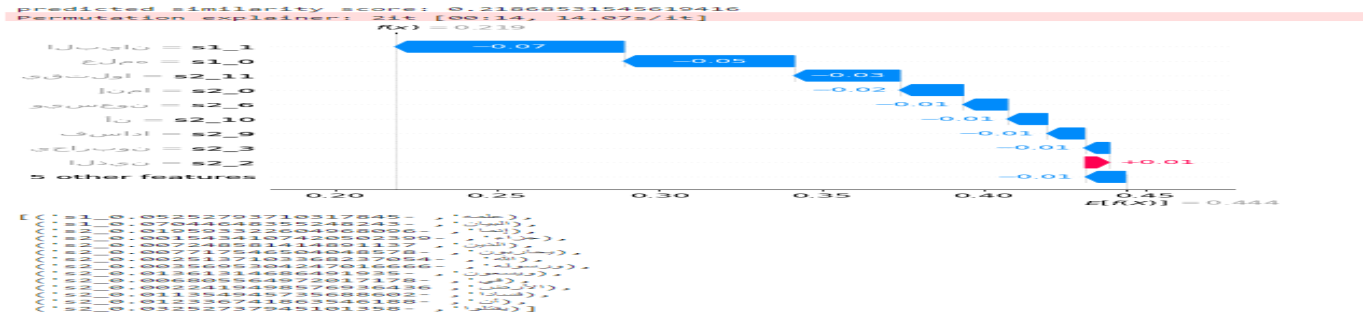

Figure 18: STS explainer scores for non-convergence retrievals **CL-AraBERT** model (Query3 – result 2.C)

B Experiment 3 results: Interpretation using Lime

In Fig. 19, we have applied the Lime explainer on ArabicBERT BERT model at threshold > = 80 and with normalized text. We have noticed that the Lime assigned the prediction probabilities to the verses correctly and make sense to humans.

In Fig. 20, we have applied the Lime explainer on ArabicBERT model at threshold > = 60 and with normalized text. We have noticed that the Lime assigned the prediction probabilities to the verses correctly and make sense to humans.

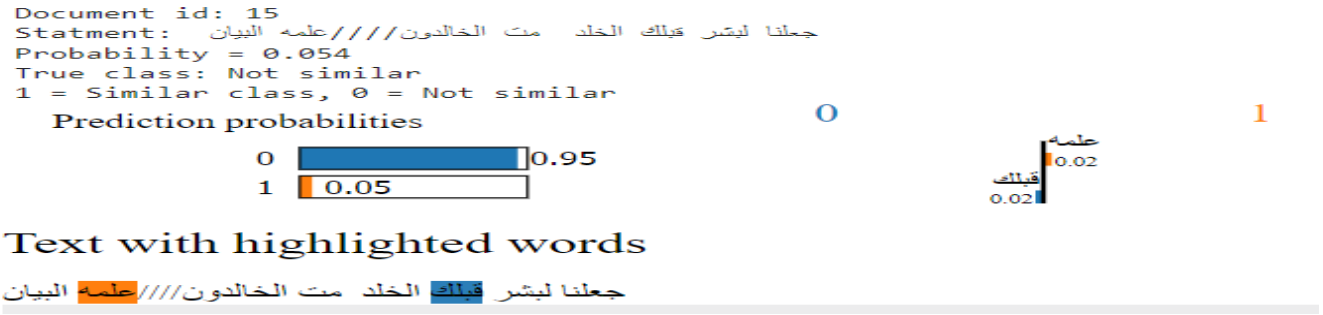

Figure 19: Lime explainer result for **ArabicBERT** model the text was with normalization and Similarity threshold >=0.8

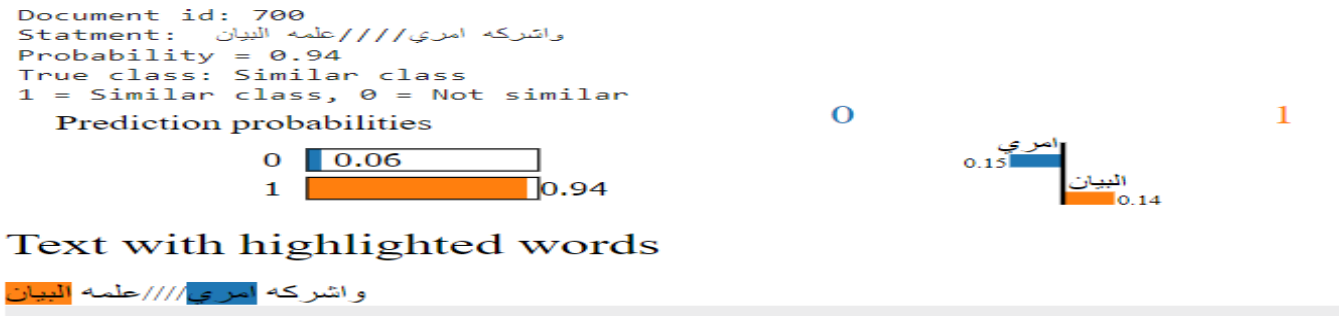

Figure 20: Lime explainer result for **ArabicBERT** model the text was with normalization and Similarity threshold >=0.6

In Fig. 21, we have applied the Lime explainer on ArabicBERT model at threshold > = 80 and with non- normalized text. We have noticed that the Lime assigned the prediction probabilities to the verses correctly and make sense to humans.

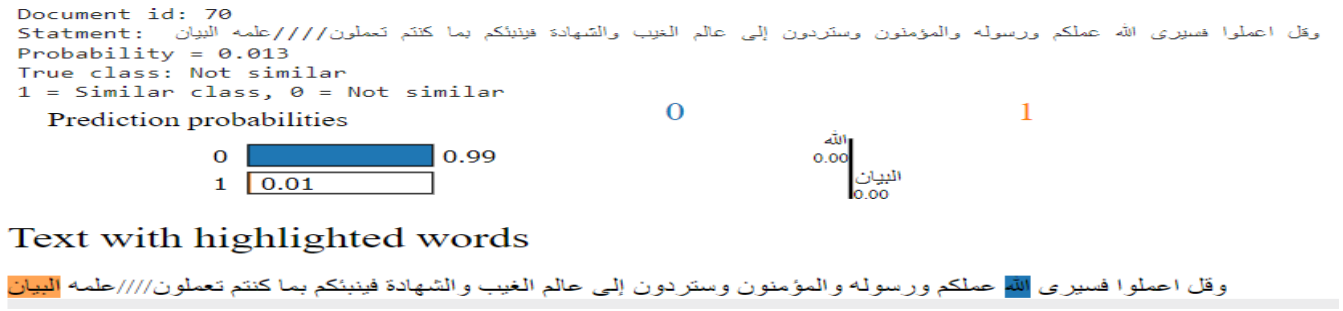

Figure 21: Lime explainer result for **ArabicBERT** model the text was without normalization and Similarity threshold >=0.8

In Fig. 22, we have applied the Lime explainer on ArabicBERT model at threshold > = 60 and with non-normalized text. We have noticed that the Lime assigned the prediction probabilities to the verses correctly and make sense to humans.

In Fig. 23, we have applied the Lime explainer on CL-AraBERT BERT model at threshold > = 60 and with normalized text. We have noticed that the Lime assigned the prediction probabilities to the verses correctly and make sense to humans.

In Fig. 24, we have applied the Lime explainer on CL-AraBERT BERT model at threshold > = 60 and with non -normalized text. We have noticed that the Lime assigned the prediction probabilities to the verses correctly and make sense to humans.

Note: we could not show the Lime explainer result for the CL-AraBERT model on Similarity threshold >=0.8

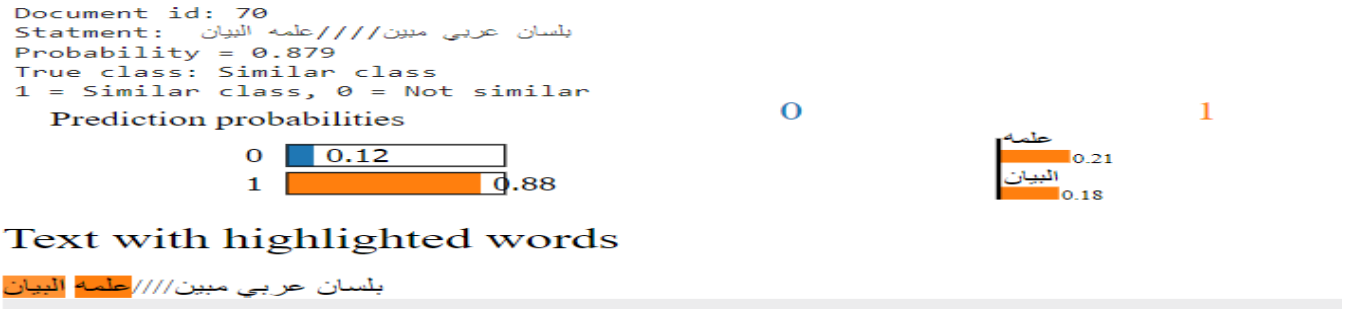

Figure 22: Lime explainer result for **ArabicBERT** model the text was without normalization and Similarity threshold >=0.6

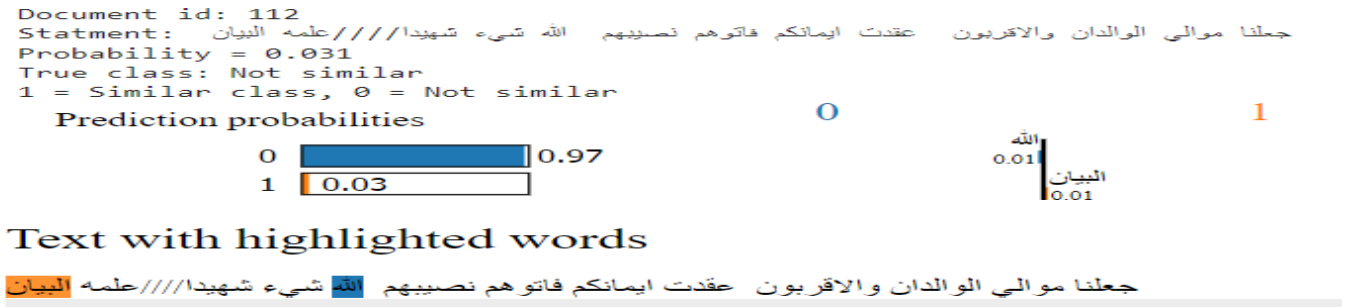

Figure 23: Lime explainer result for **CL-AraBERT** model the text was with normalization and Similarity threshold >=0.6

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.

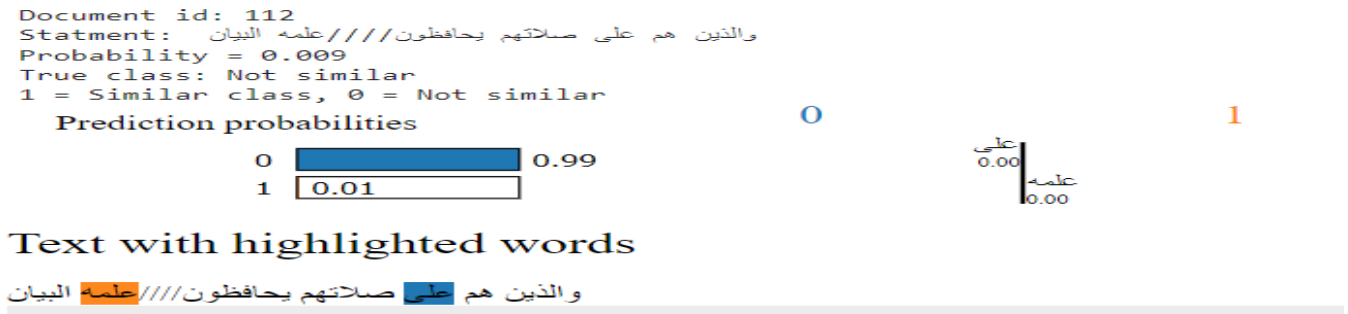

Figure 24: Lime explainer result for **CL-AraBERT** model the text was without normalization and Similarity threshold >=0.6

Table 4: Summarization for experiment ArabicBERT model

Verses	Similarity value (label)	Similarity threshold	With/Without Normalization
-	0	>=0.8	With
-	0	>=0.8	With
-	1	>=0.8	With
-	1	>=0.8	With
-	0	>=0.6	With
-	0	>=0.6	With
-	1	>=0.6	With
-	1	>=0.6	With
-	1	==1	With
-	0	>=0.8	With
-	0	>=0.8	With
-	1	>=0.8	Without
-	1	>=0.8	Without
-	0	>=0.6	Without
-	0	>=0.6	Without
-	1	>=0.6	Without
-	1	>=0.6	Without
-	1	==1	Without

Table 5: Summarization for experiment CL-AraBERT model

Verses	Similarity value (label)	Similarity threshold	With/Without Normalization
-	0	>=0.8	With
-	0	>=0.8	With
-	1	>=0.8	With
-	0	>=0.6	With
-	0	>=0.6	With
-	1	>=0.6	With
-	1	>=0.6	With
-	1	==1	With
-	0	>=0.8	Without
-	0	>=0.8	Without
-	1	>=0.8	Without
-	0	>=0.6	Without
-	0	>=0.6	Without
-	1	>=0.6	Without
-	1	>=0.6	Without
-	1	==1	Without