Comparative Study of Verse Similarity for Multi-lingual Representations of the Qur’an

A. Basharat\textsuperscript{1}, D. Yazdansepas\textsuperscript{1}, and K. Rasheed\textsuperscript{1}

\textsuperscript{1}Department of Comp. Sci, University of Georgia, Athens, GA, USA

Abstract—Text similarity is a subject that has received great attention in recent years. However, the application of text similarity tools to Semitic languages such as Arabic faces unique challenges. Moreover, the increasing number of texts being made available online, not only in native languages but also in translation, adds further challenge to identifying similar portions of texts across different documents. In this paper, we explore the problem of text similarity in the context of multi-lingual representations of the Qur’an. Particularly, we use Arabic and English datasets of the Qur’an for comparative study and analysis of several similarity measures applied across different representations of the verses in the Qur’an. We provide useful insights into the impact of using different similarity measures applied to different features across different representations and linguistic characteristics of similar text.

Keywords: similarity, document similarity, Qur’an, Arabic, cosine, Jaccard

1. Introduction

The Qur’an, considered as a concise data set, consists of less than 80,000 words, sequenced in 114 chapters (Surahs) and 6,236 verses (Ayahs) \cite{1}. The original data format was spoken Classical Arabic. We treat the problem of computing similarity between the verses of the Qur’an as a special case of computing document similarity, a widely studied subject in literature. Document similarity measures are often utilized for the purpose of automatic text classification, and clustering \cite{2} \cite{3}. We treat a verse somewhat similar to a document. However, some verses may be as short as a single or few words long. The longest verse in the Qur’an does not span more than a single page in the standard manuscript writing. Researchers recognize that the determination of a pair of documents being similar or different is not always clear and is often context dependent \cite{3}. The problem at hand bears resemblance to the context of short text classification and clustering \cite{4} \cite{5}.

We are not only interested in analyzing verse similarity using the original Arabic script, but also take considerable interest in undertaking this similarity study within and across different languages. Qur’an is one of the most widely translated texts, translated into numerous languages. A recent open linked dataset called ‘Semantic Quran’ has been published \cite{6}, with more than 48 known translations obtained from Tanzil\textsuperscript{1}. Research in other domains has revealed interesting findings when cross-language similarity studies were undertaken \cite{7}.

The problem of computing the similarity between the verses has not been widely studied, despite the fact, that the Qur’anic text has attracted much attention in recent years from the computational research, artificial intelligence and natural language processing researchers in particular. There have been studies conducted using various corpuses in Arabic, some of which are \cite{8}, \cite{9}. Sharaf et al.\cite{10} are among the pioneers to have conducted some similarity and relatedness studies on the the Qur’anic text. The QurSim corpus is a corpus that is marked with relatedness information judged by a human domain expert. The authors highlight the challenges the variation in the degree of relatedness between texts. While there are instances where lexical matching is evident between the terms, the authors believe that the majority of the related pairs require a deeper and more semantic analysis and domain specific world knowledge in order to relate the two texts in the pair. For a more detailed discussion on the relatedness and the extent of its broadness, we refer the reader to the original work of Sharaf and his colleagues.

Our objective here is to use the Qur’an’s text to compare and contrast the different classes of similarity that may result using different representations of the text, and also using different similarity measures. We believe that a benchmarking study towards standardizing similarity measures for the Qur’anic text would pave the way towards the vision of achieving the ongoing efforts of standardizing the Qur’anic knowledge map as described by Atwell and colleagues \cite{11}.

We intend to use this study as the basis of developing semantic networks of similarity and relatedness between not only verses of the Qur’an but also extending to other contextually relevant texts, which are considered indispensable when it comes to developing a comprehensive and coherent understanding of the Qur’anic verses.

In this paper, we evaluate a variety of document similarity measures, using multi-lingual, heterogeneous representations of the Qur’an. In order to do this, we develop a similarity computation framework for the verses in the Qur’an (Section 2). We experiment with four different dataset representations, four similarity measures and three different feature repre-
sentations and provide comparative insights into the results obtained (Section 3 and 4). We also shed light towards how we plan to extend this study further (Section 5).

2. Verse-to-Verse Similarity Computation Framework for the Qur'an

In this section, we present the design of our developed framework, which is meant to facilitate the process of similarity evaluation between the verses of the Qur'an, from the available datasets. The framework we have developed is generic enough to be readily used with other texts.

2.1 Datasets and their Characteristics

We have created a custom database for the experimentation which consists of the Qur'anic text in its original Arabic script in two forms: Firstly, with diacritics, i.e. case markings, and secondly, without diacritics, i.e. plain, simple and clean arabic script. In addition, the database also contains translations in several languages. We used the translation of the Qur'an to the English language (translation by Yusuf Ali) for this study. The last type of data representation we used was a dataset generated for the Arabic roots (stems). A summary of the datasets used is shown in Table 1.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q-DIAC</td>
<td>Arabic text with diacritics (or case markings)</td>
<td>وَلاَ أَقْرَأْنَ بِذَٰلِكَ رَبَّ اللّٰهُ</td>
</tr>
<tr>
<td>Q-NODIAC</td>
<td>Arabic text without diacritics</td>
<td>وَلاَ أَقْرَأْنَ بِذَٰلِكَ رَبَّ اللّٰهُ</td>
</tr>
<tr>
<td>Q-ROOTS</td>
<td>Arabic roots dataset</td>
<td>كون ، دعو ، ريب ، شفوي</td>
</tr>
</tbody>
</table>
| Q-ENG        | English Text | "...and never have I been in my supplication to You, my Lord, unhappy."

2.2 Preprocessing of Verses

For this purpose, we used the raw data for each dataset without any linguistic pre-processing, except tokenization. We created a workflow mechanism such that we specify an input configuration that encodes the input data representation, the feature selection method to be used, and the similarity method to be used for an experiment. The Verse-Set Generator as shown in Figure 1, uses the Query Processor to generate the dataset for experimentation from the database, and prepares the appropriate input representation to be used in later stages.

2.3 Feature Selection and Verse Representation

In order to adequately compare and contrast the different representations and the multi-lingual texts, we adopt a vector space model as means of verse representation for our experiments. For preliminary analysis we did not use any specific pre-processing techniques such as stemming. We relied on the traditional term-vector approach, which is the conventional approach adopted in most text mining applications.

We consider \( V = \{ v_1, v_2, v_3, \ldots, v_n \} \) to denote the set of Verses in the Qur'an, and \( \{ T = t_1, t_2, \ldots, t_m \} \) as the set of distinct terms occurring in \( V \). We represent each verse, \( v \), in the Qur'an data set as an \( m \)-dimensional vector \( \vec{v} = \{ a_1, a_2, a_3, \ldots, a_m \} \), where \( a_i \) is the weight of the \( i \)-th feature in the vector \( \vec{v} \).

The framework we developed primarily operates on a term-verse matrix (\( tv - matrix \)), which is generated using the Verse Vector Set generator. It is a compilation of all the verse vectors into one matrix where the columns represent the terms (the vocabulary of the dataset) and the rows represent the verses. First the vocabulary is compiled from the verses, and the frequency of terms across all verses is recorded. Then the \( tv - matrix \) is developed based on the verse order, in which the elements of the matrix are calculated according to one of the three term weighting methods namely: Boolean(B), Frequency(F) or TF-IDF(TF).

In traditional document similarity studies the \( tf - idf \) method of term weighting is most recommended. However, Qur'an is considered as a text with several unique characteristics. Arabic linguists consider it as a profound piece of text, where each letter and word is relevant. Therefore we choose to experiment with different term-weighting measures to analyze the impact on the similarity computation. The feature selector is responsible for applying the feature selection method specified in the input configuration. The verse modeler then applies this selection to each verse before passing it to the verse vector set generator, which generates the \( tv - matrix \).

We start with the most simple measure i.e. the Boolean. Using this measure the weight of the term \( t_i \), would have the value 1 if the term is present in the verse and 0 otherwise.

Equation (1) shows the term weighting using the term-frequency.

\[
 a_i = tf(v, t_i) \tag{1}
\]

Equation (2) shows the term weighting using the term-frequency, inverse document frequency approach. In this scheme, \( N_v \) is the total number of verses in the dataset, \( vf_j \) is the verse frequency, and \( tf_{ij} \) is the number of occurrences of the feature \( a_i \) in the verse \( v_f_j \).

\[
 a_i = tf_{ij} \times \log(N_v/vf_j) \tag{2}
\]

2.4 Similarity Computation

With the verses represented as vectors, we measure the degree of similarity between the two verses as the correlation between their corresponding vectors using the Similarity computation module. The measures described below reflect the degree of closeness or separation of the target verses. Choosing an appropriate measure of similarity is crucial.
We believe that understanding the effects of different similarity measures when applied to the Qur’an to be of great importance in helping to choose the best one. We adopt four of the measures described in [12] and [3] and map them to the problem of computing similarity for the Qur’anic verse vectors. The output of this stage is a Verse-Verse Similarity matrix, which provides a similarity measure on a scale of 0 to 1 for each verse pair in the Qur’an.

2.4.1 Euclidean Distance

Given the two verses $v_a$ and $v_b$, represented by their term vectors $\vec{v}_a$ and $\vec{v}_b$ respectively, the Euclidean distance of the two verses is defined in Equation (3).

$$D_E(\vec{v}_a, \vec{v}_b) = \left( \sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2 \right)^{1/2}$$ (3)

where the term set is $T = \{t_1, t_2..., t_m\}$. As mentioned earlier, we use different weighting measures. Therefore, the $w_{t,a}$ may be $tfidf(d_a, t)$ or $tf(d_a, t)$ or $tb(d_a, t)$.

2.4.2 Cosine Similarity

Cosine similarity is one of the most popular similarity measure applied to text documents. Some notable studies that report the use of this distance measure include [3] [12]. Given the two verses $v_a$ and $v_b$, represented by their term vectors $\vec{v}_a$ and $\vec{v}_b$ respectively, their cosine similarity is given by Equation (4)

$$S_C(\vec{v}_a, \vec{v}_b) = \frac{\vec{v}_a \cdot \vec{v}_b}{|\vec{v}_a| \times |\vec{v}_b|}$$ (4)

where $v_a$ and $v_b$ are m-dimensional vectors over the termset $T = \{t_1, t_2..., t_m\}$. Each dimension represents a term with its weight in the verse, which is non-negative. Therefore, the cosine similarity is non-negatively bounded between 0 and 1.

2.4.3 Jaccard Similarity

The Jaccard coefficient measures similarity as the intersection divided by the union of the entities. For the verses, the Jaccard coefficient computes the ratio between the dot product and the sum of the squared norms minus the dot product of the given verse vectors. The definition is given in Equation (5).

$$S_J(\vec{v}_a, \vec{v}_b) = \frac{\vec{v}_a \cdot \vec{v}_b}{|\vec{v}_a|^2 + |\vec{v}_b|^2 - \vec{v}_a \cdot \vec{v}_b}$$ (5)

The Jaccard coefficient is a similarity measure and ranges between 0 and 1. It is 1 when the $v_a = v_b$ and 0 when $v_a$ and $v_b$ are disjoint, where 1 means the verses are the same and 0 means the verses are completely different.

2.4.4 Pearson Correlation Coefficient

Pearson’s correlation coefficient is another measure of the extent to which two vectors are related. There are different forms of this coefficient. Given the term set $T = \{t_1, t_2..., t_m\}$, a commonly used form is given in Equation (6).

$$S_P(\vec{v}_a, \vec{v}_b) = \frac{m \sum_{t=1}^{m} w_{t,a} \times w_{t,b} - TF_a \times TF_b}{\sqrt{(m \sum_{t=1}^{m} w_{t,a}^2 - TF_a^2)(m \sum_{t=1}^{m} w_{t,b}^2 - TF_b^2)}}$$ (6)

where $TF_a = \sum_{t=1}^{m} w_{t,a}$ and $TF_b = \sum_{t=1}^{m} w_{t,b}$.

This is also a similarity measure. When $v_a = v_b$, the value will be 1.
2.5 Similarity Analysis

We devised three similarity classes for the purpose of analysis as shown in Table 2. We automate some validation measures for determining the results. We use the Verse-Verse Similarity matrix for computing the statistics on similarity reported in Section 3.

<table>
<thead>
<tr>
<th>Class</th>
<th>Description</th>
<th>Similarity Range</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Identical</strong></td>
<td>Identical Verses</td>
<td>Verse Pairs with Similarity Value 1</td>
</tr>
<tr>
<td><strong>High</strong></td>
<td>Almost Identical, Near Identical</td>
<td>Verse Pairs with Similarity Value &gt; 0.9</td>
</tr>
<tr>
<td><strong>Medium-High</strong></td>
<td>With identifiable similarity</td>
<td>Verse Pairs with Similarity Value between 0.75 and 0.9</td>
</tr>
<tr>
<td><strong>Low</strong></td>
<td>Minimal semantic similarity</td>
<td>Verse Pairs with Similarity less than 0.75</td>
</tr>
</tbody>
</table>

3. Experimentation and Results

3.1 Evaluation Measures

We devised the accuracy measure, for the sake of quantification, for the **Identical** similarity class using the evaluation measures shown in equations (7 - 9).

\[
P(\text{Precision}) = \frac{TP}{TP + FP} \tag{7}
\]

\[
R(\text{Recall}) = \frac{TP}{TP + FN} \tag{8}
\]

\[
F1(\text{F-Measure}) = \frac{2 \times P \times R}{P + R} \tag{9}
\]

For the sake of analysis, we limit our quantification to those verse pairs with similarity value 1. Therefore, we treat those verse pairs with similarity value 1 as true positives (TP). A TN is therefore a verse pair with similarity value that does not equal one. The reason for limiting this analysis is that establishing the precise similarity between two verses has not been done in a standardized manner and there are no existing benchmark datasets available. The only dataset available QurSim [10] employs a different approach, i.e. it classifies verse pairs according to their relatedness measure, using human subjective judgement, based on the verse pairs compiled from a famous source of Qur’anic Exegesis by Ibn Kathir. For our study, we rely on pure measures of distance, as described earlier in this section. Therefore, establishing a similarity benchmark is not possible against which evaluation may be carried out. The only pairs of verses we are able to consider as ground truth are the identical verse pairs.

3.2 Experiments

The experiments are applied to every combination of using each of the four dataset representations, each of the similarity measures and each of the term-weighting methods. We therefore had 3 × 4 × 4 = 48 experiments in total. An abbreviation scheme is used to denote the experiments. A sample experiment scheme indicates the dataset used, term-weighting applied and the distance measure. E.g. Q-DIAC-C-B indicates, Boolean(B) term weighting is used, and Cosine(C) similarity measure is used for the Q-DIAC dataset, which is the Qur’an representation in Arabic with diacritics (case markings). The experimental configurations are summarized in Table 3.

3.3 Results and Analysis for Identical Verses

The results for the 48 experiments are shown in Table 3. We chose the F1 measure in order to provide a general measure of performance of the similarity configurations. We highlight the performance of each dataset below and then provide an overall comparison:

3.3.1 Q-NODIAC Representation

The Q-NODIAC representation provides the most accurate results when dealing with the Identical similarity class, with a perfect recall and almost perfect precision, and therefore the F1 measure being close to 1. The best result is obtained using C-F, E-TFIDF, E-F, and P-TFIDF combinations (indicated by 1 pair of verses classified as False Positive). All other combinations also perform almost as well. The difference is indicated by the number of False Positives (FP)’s, in this case 2 pairs of verses are classified as FPs. The two FPs are the verse pairs [23:83, 27:68] and [22:62, 31:30]. Looking into the reasons for misclassification, we notice that the former pair includes a word that occurs twice in the verse 23:83 but occurs only once in 27:68. This is correctly captured using some of the combinations but missed with others such as the ones involving Boolean term weighting measure. The later verse pair, is a unique case, where the words are similar in both verses, however, the order is different. This is a unique case indeed, and perhaps, the only one of its kind in the Qur’an.

3.3.2 Q-DIAC Representation

The Q-DIAC performs considerably close to Q-NODIAC in terms of precision, however, the recall suffers due to the strikingly large number of verse pairs classified as False Negatives (FN)’s. The verse pairs classified as FPs are the same as those in Q-NODIAC, apart from one additional verse pair i.e. [30:52, 27:81]. The reason for this verse pair being mis-classified in the difference in orthography of the words, which are not captured in the original dataset. By manual inspection we analyzed the verse pairs in the

\[23:83 \text{ indicates Chapter 23, Verse 83 in the Qur’an}\]
FN category, and we discovered that although the verses are actually similar, the orthography of the words, which is preserved in the diacritics as per the original manuscript, causes them to appear different. For these experiments we preserved this in the original dataset. However, some of these orthographical cases may be removed without loss in any phonetic or linguistic characters of the word. Doing so, we expect that the Q-DIAC results would follow closely the results of Q-NODIAC dataset.

### 3.3.3 Q-ENG Representation

The Q-ENG representation, amongst all those approaches which use the raw Qur’anic representation, performs with a relatively low recall of around 85%. There is very little difference in either of the methods as indicated by the F1 measure which comes out to be the same for all. This representation reports the greatest number of False Negatives (FN) as indicated in Table 3. This is a significant indicator especially when it comes to the identical verses. A FN implies that the verse pair which was expected to be classified as identical fails to do so. This has implications towards the number of translations available. It is clear that the identical verses in different occurrences within the Qur’an are being translated differently. To verify this we took a few sample cases and verified this implication. The word ساحر and صارم are highlighted in the original text. The Arabic text clearly distinguishes ساحر and صارم.

The two words actually provide a different connotation, which is not captured by the English translation. These cases confirm this aspect of the translations that are available for the Qur’an. This can be used as a good measure for assessing the quality of translations. Another case is that of verse 37:80 [إِنَّا كَذَٰلِكُمْ نَحْفِي الْخَيْمَاتِ] and verse 37:110 [إِنَّا كَذَٰلِكُمْ نَحْفِي الْخَيْمَاتِ]. The word ساحر is well known in Arabic for a particle of extreme emphasis; this word is present in the former verse, however it is absent in the later. The English translation does not capture the difference in emphasis present in the arabic speech, as indicated by the highlighted word, in the context of these two verses.

Although, for this study we only used one translation of the Qur’an, available in English, our framework can prove useful in analyzing the quality of translations. An examination of those cases in other translations revealed that there are translations that distinguish clearly between near identical verses. This indeed is a reflection of the quality of the deliberation and effort that has gone into the translation process and thus may be measured to some extent.

<table>
<thead>
<tr>
<th>Table 3: Precision and Recall Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cosine (C)</strong></td>
</tr>
<tr>
<td><strong>B</strong></td>
</tr>
<tr>
<td>Q-NODIAC</td>
</tr>
<tr>
<td>TP</td>
</tr>
<tr>
<td>FP</td>
</tr>
<tr>
<td>FN</td>
</tr>
<tr>
<td>P</td>
</tr>
<tr>
<td>R</td>
</tr>
<tr>
<td>F</td>
</tr>
<tr>
<td>Q-DIAC</td>
</tr>
<tr>
<td>TP</td>
</tr>
<tr>
<td>FP</td>
</tr>
<tr>
<td>FN</td>
</tr>
<tr>
<td>P</td>
</tr>
<tr>
<td>R</td>
</tr>
<tr>
<td>F</td>
</tr>
<tr>
<td>Q-ENG</td>
</tr>
<tr>
<td>TP</td>
</tr>
<tr>
<td>FP</td>
</tr>
<tr>
<td>FN</td>
</tr>
<tr>
<td>P</td>
</tr>
<tr>
<td>R</td>
</tr>
<tr>
<td>F</td>
</tr>
<tr>
<td>Q-ROOTS</td>
</tr>
<tr>
<td>TP</td>
</tr>
<tr>
<td>FP</td>
</tr>
<tr>
<td>FN</td>
</tr>
<tr>
<td>P</td>
</tr>
<tr>
<td>R</td>
</tr>
<tr>
<td>F</td>
</tr>
</tbody>
</table>

The two verses are translated the same.

Analyzing the FPs, we discovered that in some cases, the English translation is the same but the Arabic terms used are different. This is significant when analyzing similar verses. Arabic is a rich morphological language. In particular the language of the Qur’an is considered to be precise and slightest variation or alteration in the arrangement or morphological manifestation of the word implies something significant, which often the translation fails to capture. E.g. in Table 4, the two verses are compared using the original text. The arabic text clearly distinguishes ساحر and صارم.

The two words actually provide a different connotation, which is not captured by the English translation. These cases confirm this aspect of the translations that are available for the Qur’an. This can be used as a good measure for assessing the quality of translations. Another case is that of verse 37:80 [إِنَّا كَذَٰلِكُمْ نَحْفِي الْخَيْمَاتِ] and verse 37:110 [إِنَّا كَذَٰلِكُمْ نَحْفِي الْخَيْمَاتِ]. The word ساحر is well known in Arabic for a particle of extreme emphasis; this word is present in the former verse, however it is absent in the later. The English translation does not capture the difference in emphasis present in the arabic speech, as indicated by the highlighted word, in the context of these two verses.

Although, for this study we only used one translation of the Qur’an, available in English, our framework can prove useful in analyzing the quality of translations. An examination of those cases in other translations revealed that there are translations that distinguish clearly between near identical verses. This indeed is a reflection of the quality of the deliberation and effort that has gone into the translation process and thus may be measured to some extent.
3.3.4 Q-ROOTS Representation

For the Q-ROOTs dataset, while recall is high, precision is quite low compared to the other datasets. In terms of the overall F1 measure, the Q-ROOTs and the Q-ENG datasets perform close to one another. However, the recall is much higher for the Q-ROOTs dataset compared to the Q-ENG dataset. In terms of the similarity measures, the pattern is not as consistent as with the other datasets. The C, J and P similarity measures perform closely. However, the Euclidean measure has a much lower F1 measure in comparison to the others. Interestingly, the recall for Euclidean comes out as 1. However, the precision is very low compared to the others. An investigation into this revealed some interesting results. An analysis of the FN cases showed that all 37 verse pairs are actually special verses which occur at the beginning of some chapters, and are special compositions of letters which are said to have no known roots or meanings. The Q-ROOTs dataset thus does not assign these terms any weight. The Euclidean measure, however, counts them as equivalent, even though the verse vectors are empty. Nevertheless, due to the raw nature of that similarity measure it computes them as equal. This explains the rather high number of FPs using the Euclidean measure, as compared to the other three. We manually inspected the verse pairs in this category to verify this. If we discount those verses, the measures would report equivalent. A future improvement to this measure could disregard those verses, or make a special case for the other similarity measures. If this is the case, all other methods would also report a perfect recall. However, for the initial experimentation, we preserved this result, because it indeed highlights the unique nature of those verses. Another interesting observation is that the Q-ROOTs returns the highest number of verse pairs with similarity value 1. This is indicative of high semantic similarity.

3.3.5 Overall Comparative Analysis

In general, we can conclude that using the Q-NODIAC representation provides the most accurate results when dealing with the high similarity class. In terms of the similarity methods used, the P-TFIDF configuration, i.e. the Pearson method, with TFIDF as the weighting measure and the C-TFIDF, the cosine method, with TFIDF as the weighting method perform close and comparable. The analysis of the Similarity class where the verse pairs are identical is not sufficient to provide a reasonable estimate of the precision of the method that would perform the best overall. We therefore also look into some of the verse pairs that fall in the next similarity range to give us a reasonable estimate.

3.4 Results and Analysis for Verses with High Similarity

While the similarity method performance results are similar for the identical cases, this is not the case for the similarity ranges other than identical. This is shown in Figure 2. An investigation into the verse pairs retrieved with similarity values falling in the range of more than 0.9 and less than 1 reveals the numbers shown in Figure 2.

The Pearson method returns the most number of verses in that range. This is an important finding, as the results are meaningful. Verses with reasonably high similarity, which are expected to fall in this range are not falling in that range when using Cosine, Jaccard or Euclidean. We manually inspected some verses, which bear high semantic similarity. The Q-ROOTs dataset returned a large number of FPs, some of which are actually verses with a difference of only a single literal. When we looked into the similarity values of such pairs, and compared these values across the different experimental configurations, we observed that for a highly similar verse pair, the Pearson method returns the highest similarity values, whereas the similarity values returned by other methods are much lower. This pattern is reflective in Figure 2 which displays the number of verse pairs in the similarity ranges from 0.75 to 1. As an example we took the verse pair [94:5, 94:6], which differs by a single literal. For both Q-DIAC and Q-NODIAC the values returned by Cosine, Jaccard and Euclidean are lower than 0.75, whereas the Pearson method, the highest values are above 0.95. Therefore it may be concluded that Pearson performs best. As for the weighting measures the TFIDF measure with pearson returns the highest value. However, the performance of the term weighting measure may not be generalized from these instances alone and would require further investigation.

4. Discussion and Analysis

Our study, as outlined and described in this paper, set out to quantify similarity between text segments, which is considered an important step in authorship attribution, corpus comparison, information retrieval, text mining and other fields. We attempted to provide an essential foundation by providing a means to choose an effective text similarity measuring scheme especially across cross-lingual texts of the same origin. We concluded that for the identical verses, the best results are obtained using the Q-NODIAC dataset, and all similarity measures perform comparably well. The Q-ROOTs dataset shows relatively low precision.

Challenges in measuring Text Similarity vs. Text Relatedness: It is worth mentioning that the notion of both text similarity and text relatedness are of strong importance when...
studied in the context of the Qur’anic scripture. The vector space model (VSM) or the bag of words approach imposes certain limitations and has several implications. The order of words is ignored which may be of value. The issue of synonymy is ignored. In addition, the issue of polysemy is not addressed because the VSM model only considers word forms. These limitations have been highlighted in previous research such as [8].

However, we feel that, the Qur’an is a text that deserves considerable attention when it comes to similarity at different levels. The results from this study will provide sound grounds to build upon and enhance the proposed similarity measures and methods with stem based methods to provide for more enhanced similarity classes. A future comparative study is planned to this end.

5. Conclusion

In this paper, we investigated the application of various similarity measures to heterogenous representations of the Qur’an. The results focused on an analysis of the identical verses, delineating upon some interesting findings. As a natural next step, we aim to obtain some highly reliable similarity measures for non-identical verses, especially in the higher similarity ranges in order to establish ground truth for verse similarity in the Qur’an. We believe our study will prove to be a useful step in this direction. We aim to involve users and experts through established crowdsourcing approaches. We hope to devise an improved, hybrid similarity measure in the future for establishing a more precise estimate of semantic similarity. We also plan to extend our work to include other translations of the Qur’an and carry out a comparative analysis, as our study revealed some interesting findings that can help assess the quality of translations, and help standardize the existing translations.

References


