Assessment of performance of machine learning based similarities calculated for different English translations of Holy Quran

Norah Mohammad Al Ghamdi and Muhammad Badruddin Khan,
NMAlGhamdi@imamu.edu.sa mbkhan@imamu.edu.sa
Information Systems Department,
College of Computer and Information Sciences,
Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, KSA

Summary
This research article presents the work that is related to the application of different machine learning based similarity techniques on religious text for identifying similarities and differences among its various translations. The dataset includes 10 different English translations of verses (Arabic: Ayah) of two Surahs (chapters) namely, Al-Humazah and An-Nasr. The quantitative similarity values for different translations for the same verse were calculated by using the cosine similarity and semantic similarity. The corpus went through two series of experiments: before pre-processing and after pre-processing. In order to determine the performance of machine learning based similarities, human annotated similarities between translations of two Surahs (chapters) namely Al-Humazah and An-Nasr were recorded to construct the ground truth. The average difference between the human annotated similarity and the cosine similarity for Surah (chapter) Al-Humazah was found to be 1.38 per verse (ayah) per pair of translation. After pre-processing, the average difference increased to 2.24. Moreover, the average difference between human annotated similarity and semantic similarity for Surah (chapter) Al-Humazah was found to be 0.09 per verse (Ayah) per pair of translation. After pre-processing, it increased to 0.78. For the Surah (chapter) An-Nasr, before preprocessing, the average difference between human annotated similarity and cosine similarity was found to be 1.93 per verse (Ayah), per pair of translation. And. After pre-processing, the average difference further increased to 2.47. The average difference between the human annotated similarity and the semantic similarity for Surah An-Nasr before preprocessing was found to be 0.93 and after pre-processing, it was reduced to 0.87 per verse (ayah) per pair of translation. The results showed that as expected, the semantic similarity was proven to be better measurement indicator for calculation of the word meaning.

Keywords:
Text Mining, Semantic similarity, Human-annotated similarity, Cosine Similarity, Religious text, Holy Quran.

1. Introduction

The existence of religion throughout the history of the humanity is an undeniable fact and the religious text has remained the basis for understanding the meaning of life for its believers for thousands of years now and has provided answers to different difficult questions about the existence of humans. A religion can be accepted by many nations speaking different languages; therefore its religious text is translated to numerous languages to guide its adherent. The translation of a religious text can be performed by different people based on their understanding in different manner. The translation depends on the background and understanding of the person who translates and interprets the text. Therefore, the scholars of a religion want to find a simple way to easily discover the similarity and different interpretations of the religious text.

The interpreter that can be either a translator or commentator must know the accurate knowledge of the religion and its sciences, languages, vocabulary and derivatives where the meaning of interpretation is "*statement and clarification*" [1]. One methodology to interpret religious text is mentioned in [2] where author states that "*it reveals finding of the inner meanings and the hidden facts in a particular subject or science, whereby the facts are described, the reasons are explained, and everything related to it, and then the details are presented accurately to reach a correct understanding of the material interpreted, resulting in the abolition of all Incorrect facts about scientific, educational, religious or behavioral phenomena*." It should be noted that the above referenced article and the quoted statement was written in Arabic language and was translated by author of this paper [2]. The

Manuscript received April 5, 2022
Manuscript revised April 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.4.15

objective of the research is to explore whether different machine learning based similarities can be successfully used in discovering different scholarly opinions in interpretations of religious texts. The main goal of this paper is to use the machine learning technique in the domain of English religious text to identify various similarities and differences.

The scholars face difficulty when working in this field because it is 'a divine text' and does not accept any distortion. Additionally, the fact that the Noble Qur'an carries many meanings and has different connotations that become difficult for scholars and translators to translate. Hence, the religious text and related books of the Holy Quran as Tafseer or Al-Hadith in Islam became a challenge for a translator to translate. Therefore, translators are required to find and follow the best way or clear rules for translating texts to any language without any problem or mistake. The biggest challenge that is faced by translators is the equivalency at the word level and at the grammatical level to gain acceptance by the target audience. They should use the best word that can appropriately provide a target-language text.

Table 1 illustrates the 10 famous English translations of the Holy Qur'an that were selected as an input data where pre-processing is applied to analyze the text. Translations were selected from the "Al Quran English" website [1].

Table 1: English Translators and their translations Information

Translators Name	Is English speaker?	The translations and their years of publications
Mohammed Marmaduke William Pickthall	Yes	The Meaning of the Glorious Koran: An Explanatory Translation (1930)
Abdullah Yousuf Ali	No	The Holy Qur'an: Text, Translation and Commentary (April 1937)
A J Arberry	Yes	The Koran Interpreted (1955)
Mohammad Habib Shakir	No	The Holy Quran (1974)
Muhammad Taqi-up-Din alHilali and Muhammad Muhsin khan	No	Noble Quran (1977)
Ahmed Ali	No	Al-Qur'an: A Contemporary Translation (1984)
Mohammad Mahmoud Ghali (Dr. Ghali)	No	Towards Understanding the Ever-Glorious Quran (1996)
Ali Unal	No	The Quran with annotated interpretation in modern English (Nov. 1 2008)
Hamid Abdul Aziz	No	English Translation of the Holy Quran(2009)
Literal (Mohamed Ahmed & Samira)	No	The Koran: Complete dictionary and literal translation (1994)

2. Literature Review

The Qur'an is the dominant, popular and the most famous name among the names of the religious texts and the most frequently used [2]. The Holy Qur'an is considered as the greatest miracle of the Prophet Muhammad, which is called the Holy book. It includes the main instructions and guidelines for Muslims [3]. When the researchers need to perform experiments on a religious text, it should respect it and save the contents them from tampering or sabotage. "*This means that any Arab translator who wants to render the English version of the Quran into Arabic has to be careful and, at the same time, has to take into consideration the basic ethics of the Islamic religion which forbid the falsification and the misinterpretation of the Quranic verses*." [4] The researchers and scholars were trying to use and test different text mining methods in religious texts for two reasons: The first reason refers to the proof of divinity of the Qur'an and the second reasons was the competitiveness for comparing between the religion's interfaith and disseminate via the Internet by making everything simple for them to be invited to convert to Islam [5].

Since the entry of computers in the field of the storage and retrieval of information, the researchers started to find ways to analyze the texts that aim to save time and retrieve the related documents. They worked to analyze different resources like healthy, religious, novel books, etc. Further, there are different forms such as emails, text, voice, etc. One field can be used for it, which is text mining as it deals with text data and provides a lot of information and knowledge correctly way. It starts the work plan by entering words/phrases and similar documents are retrieved by using the famous technique in Text mining that is clustering [6].
Mohd Murah started using computation methods in the Quran translations in 2013. He used four different methods from computational linguistics to calculate the similarity measures by building a dataset from twenty-one translation pairs for seven English translators (Hilali, Yusuf Ali, Sahih, Shakir, Arberry, Pickthall, Maududi). [7]

However, Huda, Wahyudin, Moch, Ulfa, Safitri, and Mahmud, applied the clustering technique on Al-Baqarah surah to extract the similarity between the verses. The reason to choose Al-Baqarah is that it is the longest surah in the Qur'an and they think will find all the themes inside it. [6] They chose the clustering method because it extracts a set of similar texts in the documents that helps to classify Qur'anic verses. They used three clustering techniques: K-means, bisecting K-means, and k-medoid. Also, they used three similarity measures cosine similarity, Jaccard similarity, and correlation coefficient to find the similarity between documents, and the results have taken two values: zero (if not similar) or one (if identical) [6]. To evaluate,

they used clustering validity by calculating cluster distances and Davis Bouldin index. They found the results were difficult among the methods and techniques because every verse doesn't include the same vocabulary. Moreover, they were not sure whether the vocabulary contained other vocabularies of the same meaning or not. And, if they found the similarity by considering the length of documents or not?

Researchers weren't limited to using similarity in the religious texts but used it in the words/sentences semantics to compare among the languages. In 2019, Md. Shajalal and Masaki Aono published a paper aimed to compute the semantic textual similarity between two sentences in (both English and Bengali languages). They found the related work of their paper useless in computing the similarity which (beyond a trivial level) as they think. It's just can capture textual similarity but cannot measure the semantic similarity [8]. They estimated the similarity between sentences by using word level. Also, they have provided three semantic similarity measures exploiting word-embedding and WordNet. The methods used in their paper achieved a performance percentage compare with related methods 77.13% difference of 4.44 points difference from the previous highest performance [8].

Muhammed published a paper in 2020 that depended on the quantification of literary works for a purpose to measure the style and comparison of translations was performed using internationally recognized metrics [9]. He built the English dataset of the Holy Quran for 13 translations: Ahmed Ali, Arthur John Arberry, Abdul Majid Daryabadi, Maulana Muhammad Ali, Muhammad Sarwar, Hamid S. Aziz, Faridul Haq, Mohammad Habib Shakir, Abdullah Yusuf Ali, Muhammed Marmaduke, William Pickthall, Ali Unal, Amatul Rahman Omar, Nooruddin, and Muhamed Ahmed & Samira. He compared 13 translations based on Type Token Ration, Uber Index, Yule's K index, and HD-D index. He found the Uber Index provided a better measure of lexical diversity with respect to the length of text [9]. Moreover, the result of this study had detected which pairs for the translations have the same lexical diversity indexes depending on their similarity results for all types of models which earned a few interesting similarities [9].

In 2021, a study was published that used the semantic text similarity for facilitating the process of knowledge extraction from a religious text [10]. In particular, this study used the Holy Quran in seven translations as the dataset. These translations were built by the five different models. The first model used the cosine similarity. The second used the Continuous Bag Of Words (CBOW) architecture. The third used the same architecture that is used in the second model with applied pre-processing steps on the translations. The fourth used the Skip-Gram architecture, and the last model used the same architecture that is used in the fourth model, but with pre-processing steps. Three performance measures were selected with the window sizes (3, 5, and 10) [10]. However, they chose window size 5 across all models since they found that the size of Window 5 achieves the highest accuracy in all three performance measures. Moreover, it was found that the Skip-Gram provides the best result in comparison with Spacy [10].

In the same year, a group of researchers has published a paper written on the Enhanced Confix Stripping Stemmer by building keywords for juz 30 of the Quran. After that, they used a searching engine system that depended on using the information retrieval tools. They worked on retrieving the verse depending on the keyword of query that is in compliance with the keyword of a given dataset. The researchers found that the precision reached 71.96%, but without stemming the precision reached 82.95%. The obtained result reaches 98.64% but without stemming it reached 76.70%. The results were better when we need to find the similarity among the ayahs and retrieve which similar ayah relies on the query. [11].

Further, the research has been measuring the semantic similarity throughout the Arabic/English language on documents, sentences, and words. Such research has been carried out based on the use of different similarity techniques and based on comparing any useful measures that offer the highest similarity. The best method that was applied for measuring document similarity is the Latent Semantic Analysis Approach (LSA). While the better result was used to measure sentences, it is the hybrid approach. It joined the word embedding with a feature-based approach. Nonetheless, it was found that the best method in word similarity is the feature-based approach [12].

3. Methodology

3.1 Main Phases

The input to the work presented in this research article is different English translations of Holy Quran. The target of the work is to find extent of similarities using machine learning techniques and compare their performance with human annotated similarity.

The experiments were conducted on English translations of Holy Quran to find extent of similarity using the cosine similarity and semantic similarity. Since the text data has a free structure, pre-processing was conducted first. The experiments included four main phases, which were corpus construction, pre-processing of document, similarities measurements and validation.

Fig. 2 shows the flowchart of the experiments that were carried out and depicts different activities that were performed during each phase

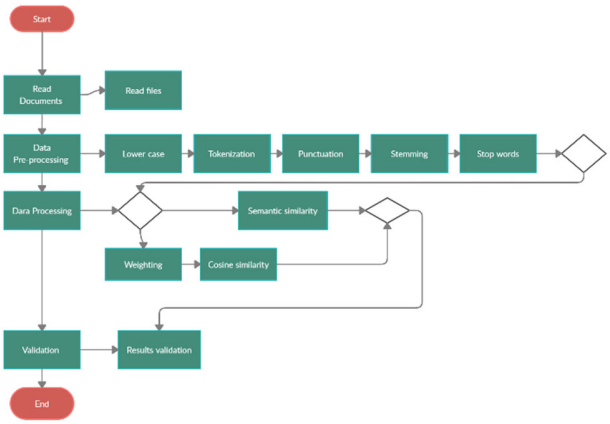

Figure 1: The flowchart of the experiments

3.2 Corpus Construction:

In this step, corpus with 10 different translations of selected part of Holy Quran was constructed. The selected part consisted of two chapters (Surah) namely: Al Humazah, and Al Nasr.

3.3 Data Pre-Processing:

In this step, the corpus went through preprocessing stage. It is significant to ensure that data processing must be performed correctly so as not to negatively affect the experiments outcome. Data pre-processing went through several stages, namely :

- Tokenize
- Transform Cases
- Punctuation
- Vowelisation
- Filter stop words
- Stemming

3.4 Similarities measurements:

Two similarities measurements namely Cosine and Semantic similarities were calculated for each pair of translation of each verse. For similarity calculation, the preprocessed data went through weighting step.

3.4.1 Cosine Similarity

It is defined as a measurement of cosine from an angle between two vector documents. Given two vectors A and B with length n [6]. It aims to find the degree of relevance by matching words.

Cosine similarity Algorithm [13]***:***

- Compute on TF-IDF Algorithm.
- 'a' refers to first document and 'b' refers to the second document.
- Then calculate the cosine similarity values are [13]: $\cos(\emptyset) = (\frac{\vec{a}\times\vec{b}}{\|\vec{a}\|\times\|\vec{b}\|}) = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n {a_i}^2}\sqrt{\sum_1^n {b_i}^2}}$ where $\vec{a}\times\vec{b} = \sum_1^n a_i b_i = a_1b_1 + a_2b_2 + \ldots + a_nb_n$ is the dot product of the two vectors.

In **Python**, the cosine similarity function is used. In the python model inside the 'sklearn' library that is used to the result of TF/IDF and find the similarity between the words in the documents [14].

3.4.2 Semantic similarity

This term is defined as the ability to detect and select the similarity degree among various terms such as words, sentences, documents, concepts, or instances [14]. Data-similarity is used where the type is numerical measures. The mathematical method that is required for calculation is based on the similarity that aims to find the degree of relevance by matching the meaning of words.

In **Python,** the similarity function is used a model inside 'spacy' library that is used to compare pairs of input sequences to find the similarities meaning between them.

```
Semantic_Similarity= Doc1.similarity (Doc2)
```

3.5 Validation

In this research, the Human-annotated similarity was used as a ground truth to evaluate the performance of machine learning based similarities. Without comparison with human annotated similarity, it is not possible to identify **quantitatively** which similarity measure is suitable to find the similar translations.

Since the human annotated similarity was based on a scale from 0 to 5, whereas the machine learning similarities ranged from 0 to 1, scaling was used to transform the machine-generated similarities so that they could be transformed in the range of 0 to 5. In the Microsoft Excel's environment, the IF formula was used to perform scaling based on the ranges.

Table 2 shows the similarity ranges and scaled classification value.

Table 2: Similarity ranges and scaled classification value

Similarity value range	Scaled Classification value
[0.8-1]	5
[0.6-0.8)	4
[0.4-0.6)	3
[0.2-0.4)	2
(0-0.2)	1
0	0

Since the similarities were calculated on the verse-level, therefore, in order to calculate the similarity on Surah-level, **averages** were calculated for Surah (chapter) Al Humazah, and Al Nasr.

In order to demonstrate how the performance of machine learning similarities is evaluated, a hypothetical example is presented in Table 3. Two phrases with similar meanings, but different words are used.

Table 3: Illustrative example of how to calculate the Difference between scaled-value for the cosine similarity, semantic similarity and Human-annotated similarity.

Phrase 1	Phrase 2	Cosine similarity from Python	Scaled value (from 0 to 5)	Semantic Similarity from Python	Scaled value (from 0 to 5)	Human Annotated similarity on scale from 0 to 5
Nice boy	Good Kid	0	0	0.83	5	5
Difference between scaled human annotated similarity and scaled cosine similarity		$5-0=5$				
Difference between scaled human annotated similarity and scaled semantic similarity		$5-5=0$				

Table 3 shows a simple example on how to calculate the difference scaled-value for the cosine similarity, semantic similarity and human-annotated similarity. Based on the obtained results, the value of the difference between the scaled human annotated similarity and the scaled cosine similarity is found 5, and the value of the difference between the scaled human annotated similarity and the scaled semantic similarity is found 0. This in fact implies that if the difference is small, the machine learning similarity is efficient. If the difference is high, the machine learning similarity is inefficient for that purpose.

4. Results and Discussion

Table 4 presents one Quranic verse and its two English translations along with cosine, semantic and three human annotated similarities to make readers understand how machine generated similarities performance was analyzed on verse level for a pair of translation.

Table 4: Analysis result on Surah Al Humazah (English Dataset) between Hamid and Ahmed for verse 3

Ayah No.	Ayah	Hamid	Ahmed	Scaled Cosine similarity	Scaled Semantic similarity	Human 1 annotated similarity	Human 2 annotated similarity	Human 3 annotated similarity
Verse 3	يَحْسَبُ أَنَّ مَالَهُ أَخْلَدَهُ	He thinks that his wealth can immortalize him.	Does he think his wealth will abide forever with him?	2	5	3	3	4
		After Pre-Processing						
		thinks wealth immortalize	think wealth abide forever	1	4	3	3	3

In Table 4, two translations of Verse 7 of Surah Al-Humazah's translation are presented to show performance of machine generated translations. Translator 1 (Hamid) translated the verse as an affirmative sentence, whereas translator 2 (Ahmed) translated the verse as an interrogative sentence. One can see that cosine similarity was unable to

recognize the similarity of translations before pre-processing as well as after pre-processing.

Tables 5 – 8 present different steps for calculating the difference between the human annotated similarity and the cosine similarity for a single verse. Table 5 shows the cosine similarity for different translations of Verse 1 of Surah Al-Humazah. Table 6 shows the transformation of Table 5 into scaled values from 0 to 5 by using ranges from Table 3. Table 7 represents the average human annotated similarity for the same verse. The average comes from similarities, which are annotated by three human annotators and operation of "Mode" was applied on the three annotations to obtain the average human annotated similarity. Table 8 illustrates the difference between Table 6 and Table 7 to show the extent of similarity between the machine annotation and human annotation.

Table 5: Cosine Similarity table for Surah Al Humazah verse 1

Verse 1	Hilali	Pickthall	Shakir	Literal	Hamid	Ghali	Unal	Ahmed	Abdullah	Arberry
Hilali	#	1.66	1.66	3.11	2.33	2.22	3.11	2.11	1.66	2.33
Pickthall		#	2.22	3.11	2.11	2.44	2.77	2.22	2.33	1.66
Shakir			#	2.55	1.55	1.55	1.88	1.77	1.44	1.88
Literal				#	2.88	2.88	2.22	2.55	2.44	3
Hamid					#	2.88	1.77	3.11	2.11	2.11
Ghali						#	2.11	2.55	3	1.88
Unal							#	1.88	2.22	2
Ahmed								#	1.55	2.22
Abdullah									#	1.11
Arberry										#

Table 6: Scaled Cosine similarity table for Surah Al Humazah verse 1

Verse 1	Hilali	Pickthall	Shakir	Literal	Hamid	Ghali	Unal	Ahmed	Abdullah	Arberry
Hilali	#	2	4	0	3	3	2	5	2	5
Pickthall		#	2	0	2	0	2	2	2	3
Shakir			#	0	2	4	2	4	2	3
Literal				#	0	0	0	0	0	0
Hamid					#	0	2	3	3	3
Ghali						#	0	3	0	2
Unal							#	2	2	2
Ahmed								#	2	5
Abdullah									#	2
Arberry										#

Table 7: Average Human annotated similarity for Surah Al Humazah verse 1

Verse 1	Hilali	Pickthall	Shakir	Literal	Hamid	Ghali	Unal	Ahmed	Abdullah	Arberry
Hilali	#	2	5	3	3	1	5	4	4	4
Pickthall		#	4	2	4	3	4	4	4	4
Shakir			#	2	4	4	5	5	3	2
Literal				#	3	3	2	3	4	3
Hamid					#	4	2	4	3	4
Ghali						#	2	3	3	3
Unal							#	3	3	4
Ahmed								#	3	2
Abdullah									#	3
Arberry										#

Table 8: Difference between Average human annotated similarity and scaled cosine similarity for Surah Al Humazah verse 1

Verse 1	Hilali	Pickthall	Shakir	Literal	Hamid	Ghali	Unal	Ahmed	Abdullah	Arberry
Hilali	#	0	1	3	0	2	3	1	2	1
Pickthall		#	2	2	2	3	2	2	2	1
Shakir			#	2	2	0	3	1	1	1
Literal				#	3	3	2	3	4	3
Hamid					#	4	0	1	0	1
Ghali						#	2	0	3	1
Unal							#	1	1	2
Ahmed								#	1	3
Abdullah									#	1
Arberry										#

Table 8 represents the difference between Average human annotated similarity and scaled cosine similarity for Surah Al Humazah verse 1. It can be seen that cosine similarities between certain pairs of translation were same as that of human annotated similarities. The entries of table with value "0" are the places where average annotated similarity and scaled cosine similarity were same. The worst difference that was observed was "4" that shows that the generated cosine similarity was far away from human annotated similarity.

Table 9 represents the average difference between different translations of the full Surah Al Humazah. The obtained

results ranged from 1.4 to 3.11 per verse per pair of translation

Table 9 : Average Difference between Average human annotated similarity and scaled cosine similarity for Full Surah Al Humazah

Verse 1	Hilali	Pickthall	Shakir	Literal	Hamid	Ghali	Unal	Ahmed	Abdullah	Arberry
Hilali	#	0.31	0.66	0	0.57	0.40	0.31	1	0.31	0.86
Pickthall		#	0.30	0	0.25	0.18	0.24	0.31	0.24	0.50
Shakir			#	0	0.31	0.60	0.30	0.66	0.30	0.57
Literal				#	0	0	0	0	0	0
Hamid					#	0.19	0.25	0.57	0.48	0.49
Ghali						#	0.18	0.40	0.18	0.34
Unal							#	0.31	0.24	0.27
Ahmed								#	0.31	0.86
Abdullah									#	0.27
Arberry										#

In order to quantify the performance of cosine similarity for Surah Al-Humazah, the entire 45 similarities were averaged and the average difference per verse per pair of translation for Surah Al-Humazah was found to be 2.24, which indicates that the cosine similarity is not an efficient measure for finding translations with similar meaning that use different words. For semantic similarity, same series of tasks was performed. In table 10, the difference between the Human-annotated similarity and Cosine similarity as well as Human-annotated similarity and Semantic similarity are presented for English translations of two chapters of Holy Quran namely Surah Al-Humazah and Surah Al-Nasr.

Table 10: Differences between the Human-annotated similarity and machine generated similarities for two chapters of Holy Quran

Before pre-processing		
Differences of human annotated similarities and machine generated similarities	**Chapter(Surah):Al Humazah**	**Chapter (Surah):Al Nasr**
Human-Cosine similarity difference	1.38	1.93
Human-Semantic similarity difference	0.09	0.93
After pre-processing		
Human-Cosine similarity difference	2.24	2.47
Human-Semantic similarity difference	0.78	0.87

From Table 10, we can conclude that the semantic similarity outperformed the cosine similarity. The difference between human annotated similarity and scaled semantic similarity was found to be very low and was around 0. The scaled cosine similarity result differed from the human-annotated similarity with high values. Following Tokenization process, the obtained results of difference of Human-annotated similarity and scaled cosine similarity of surah Al-Humazah reached to 1.37 and reached around 1.93 for surah Al-Nasr. After the pre-processing phase, the value increased because the words among translations do not have the same characters. This thing impacted the scale of cosine similarity values.

For the difference between Human-annotated similarity and scaled semantic similarity, the obtained results were the smallest compared to the results for the scaled cosine similarity. Before pre-processing, the difference between the human-annotated similarity and the semantic similarity was below 1. For Surah Al-Humazah, the difference was found to be 0.08, and for Surah Al-Nasr, the results were found to be 0.93. Nevertheless, the obtained results are better which implies that the semantic similarity is close to the human's concept that helps to find sentences/word similarity easily. Following the completion of the pre-processing stage, the obtained results were 0.77 for Surah Al-Humazah, and for Surah Al Nasr, the difference was 0.86.

5. Conclusion

In this research article, machine learning based similarities were used on the religious text to find the similarities among different English translations. For this purpose, we built the dataset in the English language. It was used to calculate two similarity measures: cosine similarity and Semantic similarity. The Python language was used to work on the dataset.

The difference between Human-annotated similarity and cosine similarity was found to be high. Whereas the difference between human-annotated similarity and semantic similarity was found to be very low. Based on achieved results, it can be inferred that the semantic similarity outperformed cosine similarity in finding the similarity between different English translations. The semantic similarity considers the different words with the same meaning as one word. The scaled semantic similarity values were almost equal to human annotated values. Thus, semantic similarity can be very helpful to automate the process of finding similarity-level of translations of same text. The cosine similarity can be helpful when same words are used in different translations. However, in the current

study, different words were used to present the same idea therefore, cosine similarity is not helpful in this domain.

References

[1]M. F. Al Nabhan ،"The need to translate the Qur'an تأليف"،*Introduction to the sciences of the Noble Qur'an* ،Aleppo ،Dar of Quran world ،1426 ، p. 286.

[2] H. Al Jazi ،"Interpretation controls،" mawdoo3 11 ،Oct 2021 . [متصل].Available: https://bit.ly/3IDP30p.

[3] Ahadi ،"alquran english،" 1 July 2009 .[متصل] .Available: https://www.alquranenglish.com.[6 1 2022 تاريخ الوصول] ./

[4] K. M. karbia ،"Reading of Religious Text "،*Journal of Humanities and Social Sciences* ،pp. 94 - 110 .2019 11 30 ،

[5] A. Jauhari ،I. O. Suzanti ،Y. D. Pramudita ،Nourma Pangestika Wulan Diantisari و Husni ،"Enhanced Confix Stripping Stemmer And Cosine Similarity For Search Engine in The Holy Qur'anTranslation "، *Information Technology International Seminar (ITIS)* 16-14 ،October 2020 .

[6] R. Agliz ،"Translation of Religious Texts: Difficulties and Challenges "،*Arab World English Journal (AWEJ) Special Issue on Translation* ، pp. 182-193 4 ،May 2015 .

[7] D. K. M. S. Al-Faqih ،"A Mathematical Phenomenon in the Quran of Earth-Shattering Proportions: A Quranic Theory Based on Gematria Determining Quran Primary Statistics (words, verses, chapters) and Revealing its Fascinating Connection with the Golden Ratio "،*Journal of Arts and Humanities* ،pp. 52-73. ،June .2017

[8] A. F. Huda ،D. R. Moch ،S. U. Q. ،W .Darmalaksana ،U . Rahmani و M. ،"Analysis Partition Clustering and Similarity Measure on Al-Quran Verses 18 "،July .2020

[9] M. Z. Murah ،"Similarity Evaluation of English translations of the Holy Quran تأليف "،*Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences* ، Al-Madina Al-Monawara .2013 ،

[10] M. Shajalal و M. Aono ،"Semantic textual similarity between sentences using bilingual word semantics "،*Progress in Artificial Intelligence* 9 ،March .2019

[11] M. B. Khan ،"Application of Computational Stylistics and Text Mining Techniques to Identify and Compare Salient Features of Different English Translations of the Holy Quran "،*International Journal of Computer Science and Network Security* ،pp. 147 - 154 20 ،Fabruary .2020

[12] S. Saeed ،S. Haider و Q. Rajput ،"On Finding Similar Verses from the Holy Quran using Word Embeddings "،*IEEE Xplore* 8 ،April .2021

[13] A. Jauhari ،I. O. Suzanti ،Y. D. Pramudita ،Husni و N. P. W. Diantisari ،"Enhanced Confix Stripping Stemmer And Cosine Similarity For Search Engine in The Holy Qur'an Translation "،*Information Technology International Seminar (ITIS)* ،pp. 207-212 ،October 2020 .

[14] M. Alian و A. Awajan ،"Arabic Semantic Similarity Approaches – Review "،*IEEE* .2018 ،

[15] S. Prabhakaran ،"Cosine Similarity – Understanding the math and how it works (with python codes)،" machine learning 22 ،Oct 2018 . [متصل].Available: https://www.machinelearningplus.com/nlp/cosine-similarity.[2021 12 7 تاريخ الوصول] ./

[16] S. Chouksey ،"Demonstrating Calculation of TF-IDF From Sklearn،" Analytics Vidhya .[متصل] .2020 04 21 ،Available: https://medium.com/analytics-vidhya/demonstrating-calculation-of-tf-idf-from-sklearn-4f9526e7e78b.

[17] A. Ali ،F. Alfayez و H. Alquhayz ،"SEMANTIC SIMILARITY MEASURES BETWEEN WORDS: A BRIEF SURVEY "،*International Center for Advanced Interdisciplinary Research (ICAIR)* ،pp. 907-914 ، 18December 2018 .

Norah Mohammad Al Ghamdi has obtained her bachelor's degree in information studies from Al-Imam Muhammad Ibn Saud Islamic University and is currently preparing for her MSc. degree in Information systems at the same university. Her research interests include, but are not limited to, Arabic NLP, information management and text mining.

Dr. Muhammad Badruddin Khan obtained his doctorate in 2011 from Tokyo Institute of Technology, Japan. He is a full-time a professor in department of Information Systems of Al-Imam Muhammad Ibn Saud Islamic University since 2012. The research interests of Dr. Khan lie mainly in the field of data and text mining. He is currently involved in number of research projects related to machine learning and Arabic language including pandemics prediction, Arabic sentiment analysis, improvement of Arabic semantic resources, Stylometry, Arabic Chatbots, trend analysis using Arabic Wikipedia, Arabic proverbs classification, cyberbullying and fake content detection, and violent/non-violent video categorization using YouTube video content and Arabic comments, and has published number of research papers in various conferences and journals. He is also co-author of a book on machine learning.