Journal of Interdisciplinary Qur'anic Studies

Journal of Interdisciplinary Qur'anic Studies Vol.2, Issue 2, December 2023, 69-85

Study of the Organization of the Qur'anic Surahs Using the Similarity-Based Approach in Deep Learning

Ehsan Khadangi [1]

Assistant Professor, Department of Computer Science, Shahed University, Tehran, Iran

Mohsen Shabani [2]

Master student in Artificial Intelligence, Faculty of Electrical and Computer Engineering, North Tehran Branch, Islamic Azad University, Tehran, Iran

Article History: Received 10 July 2023; Accepted 2 October 2023

ABSTRACT:

Original Paper

According to numerous studies, the Qur'anic surahs exhibit internal structure and organization, with each surah serving a distinct purpose. Although each surah focuses on a specific theme and the Qur'an identifies 114 broad themes, the arrangement of the surahs and the remarkable similarity between adjacent surahs (neighbors) underscores the chain-link and deliberate positioning of the surahs within the Qur'an. To investigate this phenomenon, a multifaceted and compound model was developed, comprising two main parts: embedding and autoencoding. The first part was carried out by preparing the words and roots of the Qur'anic text using the BERT model for meaning-topic representation. In the second part, the data was clustered in a soft labeling mode by the autoencoder. Analysis of the distribution of surahs within clusters revealed that neighboring surahs exhibited an average similarity of 80, while surahs with greater distance showed an average similarity of 20. The findings support the placement of similar surahs in close proximity, substantiating the organized sequence of Qur'anic surahs. To conclude, the results provide compelling evidence for the structured arrangement of Qur'anic surahs.

KEYWORDS: The Holy Qur'an, deep learning, deep neural network, clustering, surah similarity, natural language processing.

[1] Corresponding Author. Email Address: khadangi@shahed.ac.ir
[2] Email Address: ali110.moh158.sha@gmail.com

http://dx.doi.org/10.37264/JIQS.V2I2December2023.4

1. Introduction

The Qur’an, which is the primary source of the elevating school of Islam, is known for its ultra-human and miraculous features. The miracle of this book lies in the fact that all its aspects are flawless. This includes the miraculous nature of its words, grammar, phrases, verses, and surahs. One of the remarkable theories proposed by Qur’an scholars is regarding the organization of the Qur’anic surahs. According to this theory, the arrangement of the verses within the surahs, and likewise the arrangement of the surahs within the entire text, follow a specific organization.

Demonstrating the existence of a specific organization of the verses and surahs through natural language processing methods is a scholarly approach to uncovering the miraculous nature of the Qur’an. This endeavour also serves to refute assertions suggesting that the surahs were tampered with by the writers or collectors during the revelation era. Furthermore, this analytical method encourages researchers to delve into the arrangement of verses and surahs collectively, in addition to examining individual verses, in order to derive meaningful insights regarding the formation of intra- and inter-surah structures.

Several studies have explored the structural characteristics of the Qur’anic surahs using methodologies from the field of digital humanities. For instance, Zowqi provided a summary of research efforts aimed at establishing the structured-ness of Qur’anic surahs, particularly focusing on the theory of thematic unity. In a more recent study, Tavakoli Mohammadi and Makvand (2019) investigated the coherence of the Holy Qur’an's text through an analytical-historical approach. In this context, it is essential to employ methods and techniques that minimize human intervention. Many data mining approaches rely on expert-labeled data, introducing the potential for calculation errors. Therefore, there is a need for a model that can autonomously generate rules and accommodate the utilization of unlabeled data.

This paper aims to investigate the organization of the arrangement of Qur’anic surahs by employing similarity detection methods across the initial style of surahs, entire surahs, and the words and roots within surahs. The goal is to identify similarities between neighboring surahs as opposed to those that are farther apart by measuring the similarity of all verses and surahs. One of the innovative features of this paper is the deep compound network, which involves stages such as semantic embedding, autoencoding, and topic-meaning clustering.

1.1. Terminology

The field of knowledge discovery in databases (KDD) emerged in response to the abundance of data, the need for systematic analysis, and the limitations of traditional statistical methods. This process involves several stages: data cleansing, data integration, data selection, data reformation, data mining, model assessment, and knowledge presentation.

Data mining is the process of discovering knowledge and significant models from mass data. It is considered fundamental within the knowledge discovery framework as it reveals hidden models for assessment (Han et al. 2011). Clustering is a data mining technique employed to segregate unlabeled data, with the objective of separating clusters based on their density and distance from each other. Text mining refers to the automated analysis of text to extract or discover new, reliable, novel, previously unknown, useful, and understandable knowledge from unstructured and semi-structured data. Text mining utilizes methods such as information recovery, information extraction, and natural language processing (Keivanpour, Hasanzadeh, and Moradi, 2014).

Neural networks represent a branch of computational intelligence that aims to address problems through abstraction structure. The performance of neural networks is achieved through training and sampling information. Neuronal units play a crucial role in neural networks, as they follow simple computational transformation. Despite the simplicity of individual neurons, the network structure gains applicability in both basic and complex systems through the aggregation of these neurons (Teshnelab and Jafari 2015). Deep learning involves a function that maps input to output, with deep neural networks identifying the relationship between input and output data. The "deep" aspect signifies the presence of multiple layers in these networks, where each layer consists of nodes that function as computational units, in which the input data is multiplied by a weight, similar to neurons in the human brain (Jalili 2020).

Bojanowski et al. (2014) proposed the Skip-gram model for embedding fastText, representing each word as an n-gram character bag. Each n-gram character is associated with a vector representation, and words are defined as the sum of these representations. This method enables efficient training of large models and allows for the calculation of words not present in the training data.

The GloVe model, introduced by Pennington et al. (Imenpour and Me’amariani, 2019), where global vectors were designed to overcome some limitations of word2vector by training representations focused on general texts. GloVe is a device to represent words based on the global statistics of words frequency, which explains semantic information of words by modeling the textual link of words. The principal idea is that words with similar meaning often appear in similar content.

The ELMO model was proposed by Peters et al. (2018) to model the complex features of word usage, such as syntax and semantics, and to elaborate on how these applications differ in linguistic domains. Unlike more traditional embeddings such as word2vector and GloVe, there are different representations for a word, and the words with different meanings in different sentences have different vectors in the ELMO model.

The BERT model was propsoed by Devlin et al in 2019 (Bsoul et al. 2021). The most important component of the BERT algorithm is the transformer concept, which has been presented in this paper. Transformer neural networks are a modern generation of neural networks, which simultaneously hold features of recurrent neural networks (RNN) and convolution neural networks (CNN), and cover their problems. In recurrent neural networks, information disappears in the process of sequential calculation, while sequence of data is not considered in convolution neural networks. To solve this problem, the attention mechanism is used to decrease the distance between both of the sequence positions. Similarly, this is not a sequential-recurrent structure; it does not suffer from information disappearance.

1.1.1. Similarity Criteria

Similarity criteria could be proposed by two approaches: textual distance and textual representation (Wand and Dong 2020). The textual distance approach describes semantic similarity (closeness) of two words in the text from the distance viewpoint. There are three ways for measuring the distance with regard to the length, distribution, and semantics of the topic: length distance, distribution distance, and semantic distance.

Among the methods for length distance, the cosine, Manhattan, and Euclid methods are worth mentioning. The distribution distance is used for the comparison whether documents have been made from the same distribution. JS divergence and KL divergence are currently the most interesting methods for assessing the distribution distance. Kusner et al. (2015) suggested that the similarity resulting from measuring the distance based on length or distribution might be relatively low when there was no common word in the text; therefore, the distance measurement might be done at the semantic level. The distance of the Mover’s word is the most important method for defining the semantic distance, the most important examples of which are Word Mover’s Distance and Word Mover’s Distance

Extension.

In this section, the text is converted into numerical features, which might directly be calculated. Texts might be similar from two viewpoints: word and meaning viewpoint. The words that make the text are similar from the word viewpoint if they hold the same character sequence. Words are similar from the meaning viewpoint if they hold the same topic, field, and position. The words similarity in this study is used through the string method, the corpus-based method, meaning-oriented text matching, and the diagram-structure method. Most representation methods are placed in this category, such as Word2vec, ELMO, BERT, LDA, LSA, and LDA2Vec.

2. Literature Review

The paper by Slamet et al. (2016) presented a work of text extraction which is as an initial path and initiates with clustering verses upon some of the Qur’an structure phenomena. The k-means algorithm has been used for testing the clustering in a framework of data mining. According to this study, overall, 6236 verses (data set) were divided into three clusters using unsteamed and steamed words. Several models for clustering Qur’an verses of the surah al-Baqarah have been made using three methods: k-means, bisecting k-means, and k-medoid.

In another study by Huda et al. (2019), each verse of the surah al-Baqarah has been presented as a document taken from the Qur’an translated into English. The three similarity criteria cosine, Jaccard, and correlation coefficient were used. Afterward, different compounds from the clustering technique were evaluated by measuring similarity using the average distance between the cluster and the Davies Bouldin criterion. The result showed that the best performance is gained by Hemodoidal together with the cosine similarity.

In the paper by Khadangi et al. (2018), the topic sameness theory has been proposed which asserts that the inner elements of surahs are tightly related and each surah of the Qur’an holds one major topic. In this research, topic sameness of Qur’an surahs has been evaluated. The results showed that the title of the surah has been defined based on rational logic and that this could not happen by the public of the initial Islamic era.

In the research carried out by Sa’eedzadeh et al. (2020) entitled “Meaning Similarity Detection in Qur’an Verses by the Recurrent Network Architecture Siamese”, English datasets related to sentence similarity detection were used. Then a model was made to be used for detecting the similarity among the English translation of Qur’an verses. In this work, the

Siamese recurrent architecture was used for similarity detection.

In the study of Imenpour and Me'mariani (2019), "Qur’an text mining by deep learning: CNN and RNN networks", mainly deep convolutional networks and deep recurrent networks were investigated. The final goal was to train deep networks with the collected dataset and then use this trained model to detect whether other verses are Tohidi or not. To evaluate the proposed model, the three criteria of quality, readability, and time were used. Moreover, these evaluations were performed both within the surahs and between surahs. The evaluations showed that the proposed model for the Qur’an data set improved the prediction accuracy by 4%. The effect of the layers added to the model was also analyzed. The use of normalization and embedding vector generators improved the accuracy of the model by 10%. Finally, considering the properties of two-dimensional and recurrent networks could improve the accuracy of the model by 5.2%.

Similarly, in a study entitled "Intelligent Detection of Similar Meaning Verses in Holy Qur’an" by Tahanian et al (2020), the intelligent methods of deep learning supervision were used as a solution for similarity detection in Qur’anic verses. To evaluate the performance of these methods, the standard criterion dataset containing 250 pairs of verses was provided along with the labels of their similarity degrees. The evaluation results showed that the deep learning method Paragraph-Vector performed the best in detecting the meaning similarity between verses. The degree of Peterson correlation for this method was proved to be 70% by the criterion data set.

In the research conducted by Khadangi et al. (2022), there were two objectives: to examine the internal organization of each surah according to the theories of Topic Sameness, Introduction and Explanation, as well as surahs' ordering in the whole Qur’an. In this regard, the similarity of Qur’anic roots was computed based on the three methods of tf-idf, word2vec and roots' accompaniment in verses. Then, the degree of similarity of the concepts within the surahs to each other was calculated and compared with the random mode. Finally, by comparing the similarity of surahs to each other with their order distance in the Qur’an and their revelation time distance, it was revealed that the whole Qur’an is also relatively organized in terms of the order of surahs.

Bsoul et al. (2021) aimed to utilize Arabic text clustering to cluster Qur’an themes. Hence, the study reviewed the necessary improvements to Arabic text clustering, and suggested possible research directions for improving Arabic text clustering with respect to extraction, feature selection, and clustering. In this review paper, the limitations related to Arabic text clustering were discussed and the limitations of existing works were demonstrated through experiments. They identified a new problem related to extraction depending on the clustering algorithm used, so they suggested the use of BH as feature selection and as clustering to evaluate the proposed extraction. They also noted the limitations of using AP clustering on a big dataset, and their suggestion was to use BH as feature selection. For the dataset, they generated and published real Qur’an theme dataset and Arabic text dataset, which are freely available online.

3. Proposed method

The proposed method of this research, to briefly state, comprises 3 stages: A. pre-processing, B. autoencoder and C. calculating similarities. The methodology process is shown in figure 1 and the explanation of each stage is presented in the following sections.

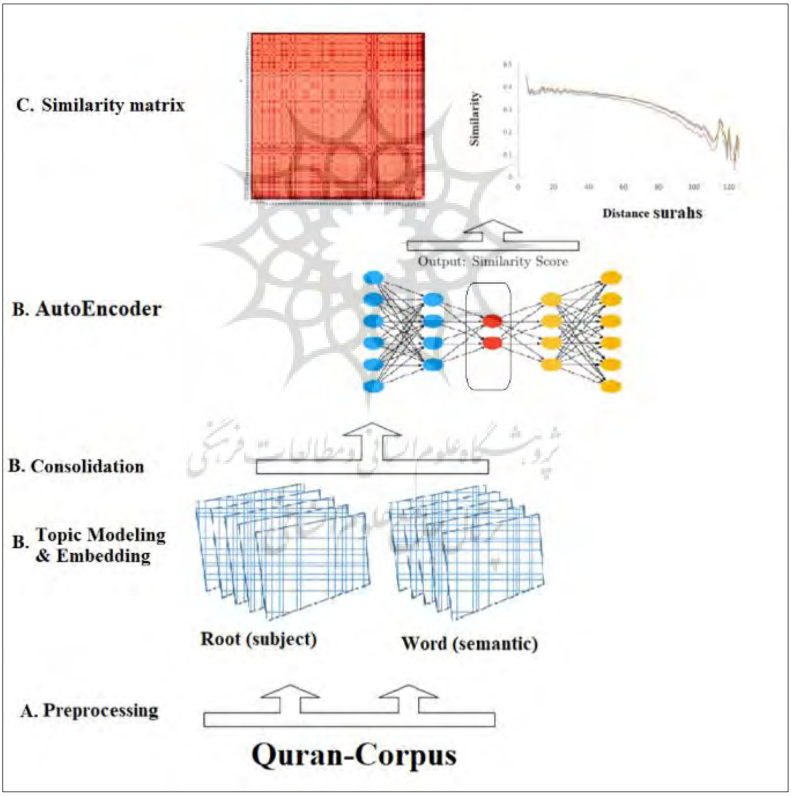


Figure 1. Proposed model process: A. Pre-processing B. Topic modeling and Embedding C. Similarity calculation

3.1. Pre-processing

The corpus used in this project belongs to the website http://corpus.quran.com, which was created under the supervision of the College of Leeds in England. This corpus contains several tables including "roots", "terms", "verses", "names of surahs", and "the role and position of words in each verse". To use this corpus in the project, it was necessary to edit the tables to calculate the desired data in the form of two tables. Table 1 is the original text of the Qur’an, the Arabization of the words was removed by the Python program. Table 2 is the text of the Qur’an, but roots are used instead of words.

Table 1. Original text of the Qur’an in the corpus

Surah number	Verse number	Verse
1	1	بسم، الله، الرحمن، الرحیم
1	2	الحمد، الله، رب، العملین
1	3	الرحیم، الرحیم
1	4	ملک، یوم، الدین
1	5	ایاک، نعبد، ایاک، نستعین
1	6	اهدنا، الصراط، المستقیم
1	7	صراط، الذین، انعمت، علیهم، غیر، المغضوب، علیهم، ولاضالین
2	1	الم
2	2	ذلک، الکتب، لا، ریب، فیه، هدی، للمتقین

Table 2. Roots of the verses in the surahs

Surah number	Verse number	Roots
1	1	سمو، اله، رحم، رحم
1	2	حمد، اله، ربب، علم
1	3	رحم، رحم
1	4	ملک، یوم، دین
1	5	عبد، عون
1	6	هدی، صرط، قوم
1	7	صرط، نعم، غیر، غضب، ضلل
2	1	الم
2	2	کتب، ریب، فیه، هدی، وقی

3.2. Processing

In the processing phase, a hybrid model was used which includes three steps: topic modeling, combination and autoencoding. The goal of these actions is to achieve a representation of surahs that has the largest inter-cluster distance and the smallest intra-cluster distance. Each step is described below.

3.2.1. Topic Modeling

In this step, the main objective is to find the main themes of each surah and find out which words are used to express each theme. Topic modeling is important because, in most embedding models, the general meaning of the sentence is determined, but the characteristic theme that distinguishes the sentence from other sentences is not identified. BERT Topic modeling is used in this project. Also, roots are used as input instead of words.

3.2.2. Embedding

First, the text (roots) was embedded using the BERT model. In this process, each surah becomes an array of length 768 with decimal numbers between 0 and 1. Since the BERT model has multiple versions, the sentence transforms version was used because it performs embedding and representation at high speed.

3.2.3. Clustering

The purpose of this step is the thematic clustering of the surahs. Considering that the embedding of the BERT model has large dimensions, it is necessary to reduce the dimensions because clustering algorithms have a large error when the dimensions are large. Among the dimension reduction methods (such as PCA, T-SNE, etc.), the UMAP method was used because it largely preserves the local structure and the global structure of the data. After the dimension reduction, the k-means algorithm was used to cluster the surahs. In the following, the c- TF-IDF algorithm was used to find the keywords of the text. This algorithm is similar to TF-IDF, but instead of performing the calculations for each surah, each cluster (including multiple surahs) is processed as one document and the calculations are performed. The formula for this method is as follows:

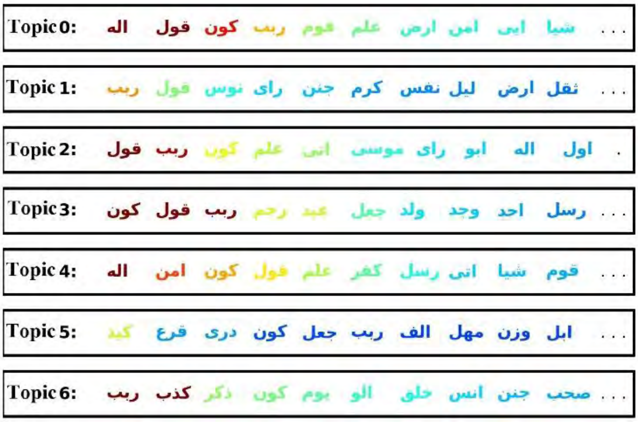
$$\mathrm{c-TF-ID}F_i = \frac{t_i}{w_i} * \log \frac{m}{\sum_j^n t_j}$$

where the frequency of each word t for each cluster i is derived and divided by the number of all words w. This can be considered as a regularization of the frequent words in the class. Then, the total number of unrelated documents is divided by the total frequency of word t in the entire class n. Since the clustering may not have been performed with high accuracy or there are clusters with many similarities, the degree of mutual similarity between the clusters is calculated and the clusters that have high similarity are merged. By performing the above actions, a new clustering is performed and the number of clusters is reduced from 30 clusters to 6 clusters. Figure 2 shows the important words of each cluster (topic). The importance (weight) of each word is indicated by a color, i.e., the richer red the color is, the greater the weight of that word in that topic (cluster). For each cluster, one important word can be selected. In this step, the keywords of each cluster (topic) were identified along with their weights. Now, to obtain an embedding based on the topic model BERT, the following actions are performed:

- The vector c-tf-idf counts the keywords of each cluster.
- The vectors from the previous step are averaged using the p-mean method and the vector of each cluster is determined as the embedding for each cluster (topic).
- The cluster number of each surah is determined and the embedding of this topic is assigned to the corresponding surah.
- Finally, the thematic embedding for each surah is obtained. Each embedding vector has an array of length 10, whose size depends on the number of keywords.

3.2.4. Consolidation

In parallel with topic modeling, the text (words) of the Qur’an is embedded by the BERT model, where each surah is converted into a 768-dimensional array of decimal numbers. BERT Topic modeling and BERT model embedding both have their importance and different functions, but when combined, a more comprehensive and versatile representation is obtained. That is, attention is paid to the "text meaning" and the "text topic" at the same time. Note that topic modeling looks at the roots of the words in the surahs and BERT embedding looks at the words in the surahs. To balance the effects of the above two approaches, we give a factor of 15 to the topic modeling dimensions, since they are much smaller than the 768 dimensions of the BERT model. Then, the topic embedding is associated with the end of the BERT model embedding.

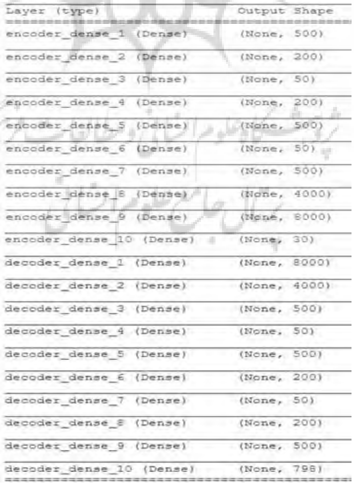


Topic 0:	اله	قول	کون	ربب	قوم	علم	ارض	امن	ابی	شیا	. . .
Topic 1:	ربب	قول	نوس	رای	جنن	کرم	نفس	لیل	ارض	ثقل	. . .
Topic 2:	قول	ربب	کون	علم	اتی	موسی	رای	ابو	اله	اول	.
Topic 3:	کون	قول	ربب	رحم	عبد	جعل	ولد	وجد	احد	رسل	. . .
Topic 4:	اله	امن	کون	قول	علم	کفر	رسل	اتی	شیا	قوم	. . .
Topic 5:	کید	قرع	دری	کون	جعل	ربب	الف	مهل	وزن	ابل	. . .
Topic 6:	ربب	کذب	ذکر	کون	یوم	الو	خلق	انس	جنن	صحب	. . .

Figure 2. Thermal table (importance level) of words in each topic (The more important roots are shown in dark red, and unimportant roots are shown in light blue)

3.2.5. Autoencoder

The output of the second and third steps is passed to the deep network of autoencoder to perform clustering. In this step, the data is mapped into high dimensions and gradually classified into smaller dimensions. Then, the dimension is reduced from about 760 to 50 to start the process of soft tagging, where the verses and surahs are placed in dense clusters far from each other. In the training phase, the above process is repeated many times to allow the deep network to perform the best clustering.



```
Layer (type)                 Output Shape
=============================================
encoder_dense_1 (Dense)      (None, 500)
_____________________________________________
encoder_dense_2 (Dense)      (None, 200)
_____________________________________________
encoder_dense_3 (Dense)      (None, 50)
_____________________________________________
encoder_dense_4 (Dense)      (None, 200)
_____________________________________________
encoder_dense_5 (Dense)      (None, 500)
_____________________________________________
encoder_dense_6 (Dense)      (None, 50)
_____________________________________________
encoder_dense_7 (Dense)      (None, 500)
_____________________________________________
encoder_dense_8 (Dense)      (None, 4000)
_____________________________________________
encoder_dense_9 (Dense)      (None, 8000)
_____________________________________________
encoder_dense_10 (Dense)     (None, 30)
_____________________________________________
decoder_dense_1 (Dense)      (None, 8000)
_____________________________________________
decoder_dense_2 (Dense)      (None, 4000)
_____________________________________________
decoder_dense_3 (Dense)      (None, 500)
_____________________________________________
decoder_dense_4 (Dense)      (None, 50)
_____________________________________________
decoder_dense_5 (Dense)      (None, 500)
_____________________________________________
decoder_dense_6 (Dense)      (None, 200)
_____________________________________________
decoder_dense_7 (Dense)      (None, 50)
_____________________________________________
decoder_dense_8 (Dense)      (None, 200)
_____________________________________________
decoder_dense_9 (Dense)      (None, 500)
_____________________________________________
decoder_dense_10 (Dense)     (None, 798)
=============================================
```

Figure 3. Layers in deep network: Autoencoder's layers whose input layer is 500-dimensional array, and clustering them down to 30 dimensions. 9 layers for dimensionality reduction and 9 layers for reconstruction, and 1 middle layer for clustering.

In this autoencoder (*Figure 3*), nine layers are used for encoding, consisting of two stacked layers. The middle layer has 50 dimensions, which is the number of clusters. In the following, 9 layers are used to reconstruct the input. The total number of parameters trained is about 69 million. The parameters learned from the input for the middle layer are stored to be used for clustering. The whole network is trained for 100 epochs so that the samples have the least inter-cluster dispersion and the largest inter-cluster distance.

4. Similarity matrix

The result of the previous step is a new embedding of the Qur’anic surahs and verses. By calculating the embedding distance of the surahs and verses using the Minkowski method, the similarity distance matrix is obtained. After summing the diagonal values of the similarity matrix (upper triangle only), the average similarity of the surahs considering the position (distance) of the surahs is obtained and presented in the form of a linear graph on the two axes of the similarity of the surahs and the distance of the surahs. This means that the similarity between each two-surah to another two-surah has been calculated and inserted into a matrix. The diameter of the similarity matrix with a light color shows full similarity of each surah with itself. The farther we go from the diameter, the color gets stronger and the similarity, less (Figure 4).

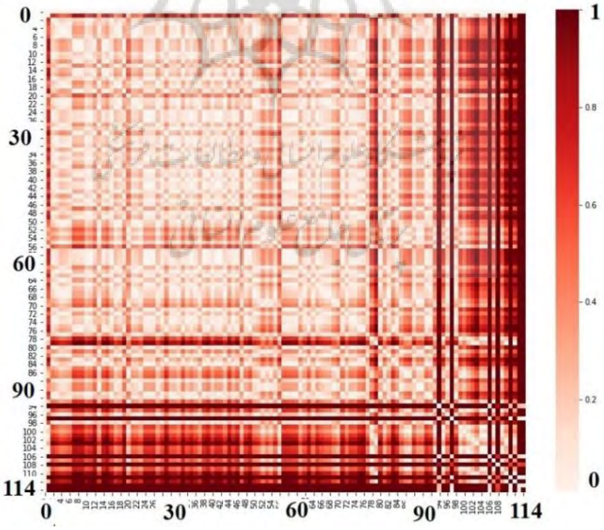


Figure 4. Similarity matrix: the upper left corner represents the semantic difference of neighboring surahs, and as we move to the right of the lower corner, the cells become bolder, and the more distant surahs are less similar.

5. Results

For evaluating methods, the diagram of the average distance of surahs is used. To make this diagram, the similarity matrix of surahs is initially calculated according to the Minkowski equation. Then the triangle-bottom diameters of the matrix are added up and their average is gained and the linear diagram is finally drawn. The hypothesis proposed in this research is the organized arrangement of Qur’an surahs.

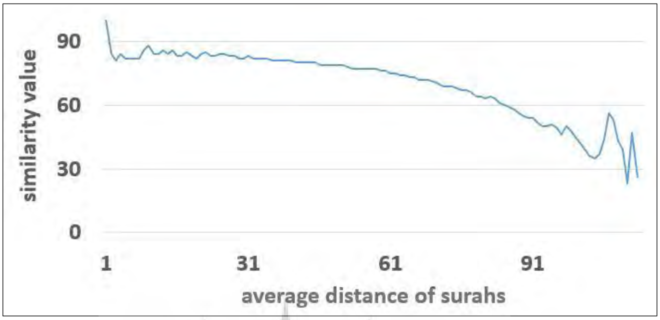


Figure 5. Chart of similarity of surahs in the proposed model

As shown in figure 5, the vertical axis shows the similarity between surahs with 0 as the least and 100 as the most. The horizontal axis shows the amount of adjacency or the distance of surahs from each other. In this diagram, the lines are with the descending trend, that is, each surah is similar to its neighbor by 80 units on the average, and as the distance increases, the similarity decreases.

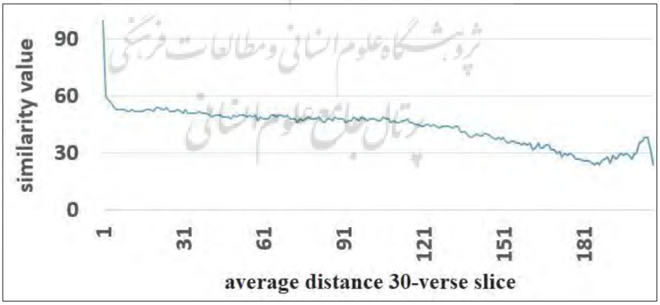


Figure 6. Similarity diagrams in clustered verses

To prove this, the similarity of surahs versus their position in the Holy Qur’an is examined. To put it more clearly, it is presumed that close surahs are more similar than the far ones. For instance, the similarity value of the surah al-Baqarah to the surah Āli ʿImrān is more than that of the surah al-

Nās. As shown in figure 5, the diagram trend starts from 100 and declines to around 10. The slope of the diagram is almost descending.

The other approach for examining the similarity and inner integrity of the Qur’an script is consideration of the verses. In this regard, verses are clustered into 20 topics by the BERTtop method, meaning that each verse is labeled with a topic. Then regardless of the surahs, all verses are categorized into 30 slices. For example, the 7 verses of the surah al-Fātiḥah plus the 23 verses of the surah al-Baqarah are put into a 30-verse slice. The descending trend of figure 6 is totally visible. The similarity of verses based on their topic has been drawn in different slices. Interesting about the 30-verse slice is that the more the distance is, the less the similarity becomes. In other words, the smaller the verse slices become, the more their difference becomes versus their position.

According to the results (Table 3), the similarity of surahs decreased by the increase in their distance by all embedding methods, and the descending trend was seen in the photos of the previous section. It is also worth mentioning that the diagrams of the other methods in machine learning and deep learning had a lighter slope. Therefore, the results state that similar surahs have been ordered beside each other accurately.

Table 3. Comparison of proposed method with other methods (The columns are the placement distance of surahs in the Qur’an, and the rows determine the method).

	Neighboring	20	50	70	100
Word2vec	100	83	80	76	72
Doc2vec	100	94	90	90	88
tf-idf	21	17	12	8	6
Bahamaee	77	75	69	64	45
Autoencoder	100	86	80	75	65
LDA	100	39	32	25	4
BERTtop	100	65	65	66	59
Proposed method	100	85	80	70	50

In a research carried out by Khadangi et al. (2022), the assortment of surahs as they are placed in the Qur’an and the order of their revelation has been studied. They have also utilized the methods tf-idf, and word2vec. Based on the correlation of the order of surahs in the whole Qur’an and the order of their revelation with their similarity, Khadangi et al. concluded that the order of surahs in the whole Qur’an also enjoy relative organization. Our method confirms the results obtained by Khadangi et al. and emphasizes that the spatial order of the Qur’anic surahs is very systematic and structured.

The obtained results can be a powerful proof to show the arrangement order of the Qur'anic surahs.

6. Conclusions

The results show that adjacent surahs hold the most similarity, and as the distance of adjacency increases, the similarity decreases considerably. This rule applies to verses as well. In other words, near surahs hold common topics and concepts, and the verses and surahs farther from each other hold more differences in terms of their meaning and subjects. Therefore, the assortment of verses and surahs of the Qur'an is organized.

We conducted this research to explore the organization of the Qur'anic surahs based on their constituent concepts in a fully automated manner. To achieve this, we employed a deep neural network for content clustering. The input data comprised the text of the Qur'an, word roots, and subjects. For the implementation, we utilized a textual autoencoder deep network. The findings revealed that neighboring surahs exhibit the highest similarity, and as the distance of adjacency increases, the similarity diminishes significantly. This pattern also holds true for individual verses. Essentially, adjacent surahs share common themes and concepts, while surahs and verses that are farther apart exhibit greater differences in terms of their meanings and subjects.

Consequently, it can be said that the verses and surahs in the Qur'an appears to have deliberately organized arrangement based on their content. This endeavour also serves to refute assertions suggesting that the surahs were tampered with by the writers or collectors during the revelation era.

6.1. Suggestions

The Holy Qur'an is like an infinite ocean that no effort in this heavenly book might meet its full dignity. However, to prove the miracles of the Holy Qur'an and to continue this research, the following are suggested:

- Using an Arabic-Qur'anic dictionary for embedding.
- Attention to the Arabic vowel signs within the embedding words.
- Simultaneous attention to the word, root, and the role of the word within the sentence for embedding.
- Using the information of the dependency tree and the analysis tree of Qur'an text.

- Using labeled data such as verses with their time, and incidence.
- Using supervised learning methods and adding more meta-data to surahs.
- Forming a deep network for detecting phrases and slices of Qur’an text regardless of the surah and verses so that the new slice is considered an independent meaningful unit.

Acknowledgements

We would like to express our sincere gratitude to Dr. Ehsaneddin Asgari for his exceptional guidance and help in providing the dataset.

This research has not been the subject of any specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. http://dx.doi.org/10.1162/tacl_a_00051

Bsoul, Q., Abdul Salam, R., Atwan, J. and Jawarneh, M. (2021). Arabic Text Clustering Methods and Suggested Solutions for Theme-Based Qur’an Clustering: Analysis of Literature. *Journal of Information Science Theory and Practice*, 9(4), 15-34. https://doi.org/10.1633/JISTaP.2021.9.4.2

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

Han, J., Kamber, M. and Pei, J. (2011). *Data mining: Concepts and Techniques.* Elsevier. https://doi.org/10.1016/C2009-0-61819-5

Huda, A. F., Deyana, M. R., Safitri, Q. U., Darmalaksana, W. and Rahmani, U. and Mahmud (2019). Analysis Partition Clustering and Similarity Measure on Al-Qur’an Verses. *2019 IEEE 5th International Conference on Wireless and Telematics (ICWT)*, Indonesia. http://dx.doi.org/10.1109/icwt47785.2019.8978215

Imenpour, R. and Me'amariani, A. A. (2019). Qur’an Text Mining by Deep Learning: CNN and RNN Networks. PhD thesis, North Tehran Branch, Islamic Azad University.

Keivanpour, M. R., Hasanzadeh, F. and Moradi, M. (2014). *Advanced Topics in Data Mining*. Kian Collegiate Publication.

Khadangi, E., Fazeli, M., Naghavi, M. (2022). The Study on Qur’anic Surahs' Structured-ness and their Order Organization Using NLP Techniques. *Journal of Interdisciplinary Qur’anic Studies (JIQS),* 1(2), 29-56.

http://dx.doi.org/10.37264/jiqs.v1i2.3

Khadangi, E., Fazeli, M. M. and Shahmohammadi, A. (2018). The Study on Qur’anic Surahs' Topic Sameness Using NLP Techniques. *2018 8th International Conference on Computer and Knowledge Engineering (ICCKE)*, 298-302. Mashhad, Iran. http://dx.doi.org/10.1109/iccke.2018.8566248

Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K. (2015). From Word Embeddings to Document distances. *Proceedings of the 32nd International Conference on Machine Learning (PMLR)*, 37, 957-966. https://proceedings.mlr.press/v37/kusnerb15.html

Pennington, J., Socher, R. and Manning, C. D. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. Doha, Qatar. http://dx.doi.org/10.3115/v1/d14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, 2227–2237. New Orleans, Louisiana. http://dx.doi.org/10.18653/v1/n18-1202

Slamet, C., Rahman, A. M., Ramdhani, A. and Darmalaksana, W. (2016). Clustering the Verses of the Holy Qur’an using K-means Algorithm. *Asian Journal of Information Technology*, 15(24), 5159-5162.

Sa’eedzadeh, M. J., Sarabadani, A. and Torabiian, N. (2020). Meaning Similarity Detection in Qur’an Verses by the Recurrent Network Architecture Siamese, *National Conference of Artificial Intelligence and Islamic Sciences*. Qom.

Tahanian, S., Borhani, M. and Minaei, B. (2020). Intelligent Detection of Similar Meaning Verses in the Holy Qur’an. *National Conference on Artificial Intelligence and Islamic Sciences*. Qom.

Tavakoli Mohammadi, N. and Makvand, M. (2019). Historical analysis of the idea of the coherence of the text of the Holy Qur’an. *Researches of Qur’an and Hadith Sciences*, 16(3), 91-125. http://dx.doi.org/10.22051/tqh.2019.23560.2270

Teshnelab, M. and Jafari, P. (2015). *Neural Networks and Advanced Neural Controllers*. K. N. Toosi University of Technology.

Wang, J. and Dong, Y. (2020). Measurement of Text Similarity: a Survey. *Information*, 11(9), 421. https://doi.org/10.3390/info11090421

Zowqi, A. (2013). A New Approach to the Study of the Textual Cohesion of the Surahs of the Holy Qur’an. *Qur’an and Hadith Studies*, 6(2), 151-177. https://doi.org/10.30497/quran.2013.1471