Journal of Interdisciplinary Qur'anic Studies

Journal of Interdisciplinary Qur'anic Studies Vol.4, Issue 1, June 2025, 53-73

Semantic Abductive Network Construction for the Holy Qur'an: A Hybrid Ontology-Based Approach

Nafiseh Jafari [1]

Master's degree in Artificial Intelligence, Faculty of Electrical and Computer Engineering, Malek Ashtar Industrial University, Tehran, Iran.

Maryam Hourali

Assistant Professor, Faculty of Electrical and Computer Engineering, Malek Ashtar Industrial University, Tehran, Iran.

Article History: Received 27 March 2025; Accepted 21 May 2025

ABSTRACT:

Original Paper

The Engineering and constructing semantic networks constitute one of the foundational technologies in the fields of cognitive processing, natural language processing, the semantic web, and the development of artificial intelligence-based systems. Consequently, expertise in the design, construction, engineering, maintenance, evolution, and optimization of ontologies has played a crucial role in advancing intelligent technologies in recent years, and this trend, particularly in the context of dependable and responsible artificial intelligence, is expected to continue in the coming years. The Holy Qur'an, as the sacred book of Muslims worldwide and the primary source of Islamic religion, civilization, and culture, has consistently served as a principal resource in the humanities and Islamic studies, as well as in socio-religious service applications, within Muslim communities.

In this paper, a semantic network is automatically constructed using a hybrid approach that integrates multiple technological solutions, including ontologies, word embeddings, co-occurrence analysis, and Arabic root extraction. After the construction of the semantic network and through the application of clustering techniques, several semantic frames were automatically extracted and designated as "abduction frames." To evaluate the proposed approach, a questionnaire-based assessment was conducted, in

1. Corresponding Author. Email Address: nafisehjafari75@gmail.com

http://dx.doi.org/10.37264/JIQS.V4I1.8

which 1,295 individuals participated voluntarily. The results yielded a precision of 69.47% and a recall of 85.35%. Additionally, a mixed quantitative–qualitative evaluation conducted by a panel of experts rated the validity and innovation of the proposed method's outputs as "good."

KEYWORDS: The Qur'an, Ontology, Word embedding, Co-occurrence, Semantic frame, Abduction, Association, Collocation.

1. Introduction

Constructing a semantic network requires the integration of multiple components. The approach employed in this paper focuses on the automatic construction of a semantic network. Modern approaches to computerized Qur'an mining were established in the second half of the 20th century by Muslim scholars, in parallel with advancements in computational processing equipment. In the early years of the 21st century, as natural language processing and semantic computing methods flourished, these approaches gained greater coherence and diversity. From 2010 onward, with the emergence of novel artificial intelligence techniques such as deep neural networks, these methods were increasingly employed to further expand this research domain. Following the widespread introduction and success of large language models (LLMs) between 2017 and 2021, and continuing to the present, researchers have adopted methods based on this computational infrastructure—such as machine comprehension of texts, automatic machine tagging, and prompt engineering techniques—within the field of computerized Qur'an mining.

The knowledge of Qur'anic exegesis encompasses a substantial corpus of profound insights and wisdom preserved in interpretive texts. Semantic—and even cognitive—processing of this valuable corpus can lead to significant advancements in understanding the Holy Qur'an through computerized Qur'an mining. To achieve this objective, robust semantic processing infrastructures must be applied to existing exegetical texts, whether through human, semi-automatic, or fully automatic methods. Among the most important of these semantic infrastructures are semantic networks and ontologies. In the present paper, we propose a multifaceted solution for the automatic generation of an abductive semantic network for the Holy Qur'an, employing a hybrid engineering approach. This solution integrates advances from information engineering, information retrieval, natural language processing, word embedding systems, and semantic-cognitive computing.

2. Theoretical Framework

2.1. Ontology

Ontologies serve as formal and explicit representations of conceptual structures and are essential tools in semantic networks. An ontology comprises several components, including concepts (classes) (C), relations (properties) (R), instances (I), axioms (A), data types (T), and values (V) (Gruber 1993). Ontology relations are divided into two main categories: taxonomic relations and non-taxonomic relations. Non-taxonomic relations encompass several subcategories, including Part-of, Antonymy, Synonymy, Possession, Causality, Hypernymy, and Hyponymy.

In the Persian language, a number of ontologies have been developed. Some of these are domain-specific, as outlined below:

- IrGo: A general ontology of Iranian traditional medicine based on *Makhzan al-Adwiyah*. This ontology includes 3,521 classes, 15 properties, and 20,903 axioms (Naghizadeh et al. 2021).
- PMD: An ontology of diseases in Persian medicine, designed to classify diseases in traditional Persian medicine. This ontology contains 529 classes and 41 properties (Persian Medicine Diseases Ontology 2019).
- QuranJooy: Developed by the Iran Telecommunication Research Center, this ontology includes more than 69,000 concepts and 8,000 instances (Mirarab & Khorram 2022).
- Borhan: This ontology focuses on Islamic Fiqh (jurisprudence) principles and includes more than 6,000 classes and 2,000 relations (Mirarab & Khorram 2022).
- FarsNet: In addition to these domain-specific ontologies, FarsNet has been developed as a general-purpose ontology, containing over 100,000 entries (Shamsfard et al. 2009).

The table below (Table 1) lists several prominent studies on computational ontology development for the Holy Qur'an conducted between 2023 and 2025. These studies emphasize ontology engineering, ontology modeling, text mining, and natural language processing. They originate from universities and research institutes in Germany, Indonesia, Egypt, Turkey, Iraq, Pakistan, Iran, China, Malaysia, Japan, and the UK, demonstrating the global scope of research efforts in this field. Specific references corresponding to each study title are provided in the references section.

Table 1. Several prominent studies on computational ontology development for the Qur'an (2023–2025)

No.	Research Title	Ontology Engineering Approach	Ontology Modeling Approach	NLP Techniques	Text Mining	Key Results	Evaluation Criteria
1	Investigating Retrieval-Augmented Generation in Qur'anic Studies (Khalila et al. 2025)	RAG	Creating Semantic Ontologies of Surahs	LLM, BERT, Embeddings	Text and Meaning Extraction	Analysis of 13 LLM Models	Relevance, Faithfulness, Correlation
2	Computational Analysis of Qur'an Text using ML and LLMs (Shahid et al. 2025)	Extraction of Qur'anic Patterns	Knowledge Graph of Verses	text-embedding-3-large, GPT-4	Analysis and Clustering	Implicit Semantic Patterns	Coherence, Clustering
3	Exploring Thematic Structures in Surah al-Baqarah (Salah et al. 2025)	Semantic Clustering	Key Concept Extraction	TF-IDF, SVD, UMAP, K-means	Theme Analysis	10 Distinct Clusters	Davies–Bouldin, Silhouette
4	QuranMorph: Morphologically Annotated Corpus (Akra et al. 2025)	Morphological Tagging	Lemmatization and POS Tagging	SAMA/Qabas, Lemmatization	Morphological Parsing	77,429 Words, 40 Tags	IAA, Cohen's Kappa
5	LLM-Based Approach for Qur'anic QA (Saadaoui et al. 2024)	Question Concept Extraction	Semantic Response Representation	GPT-4, BART, LLAMA, CoT	Question Analysis	F1 = 95% with BART	BLEU, ROUGE, F1 Score
6	Developing a Qur'anic QA System (Urdu) (Tariq et al. 2025)	Extraction of Qur'anic Entities	Urdu Semantic Modeling	RoBERTa, TF-IDF, SBERT	Semantic Search	MAP = 0.85, F1 = 0.88	Precision, Recall@K
7	Qur'anic Conversations: Semantic Search (Shohoud et al. 2023)	Search Concept Extraction	Arabic Semantic Ontology	Arabic NLP, Semantic Search	Phrase Analysis	Improved Precision	Precision, Recall, MRR
8	Linking Qur'an and Hadith Topics through Ontology (Alshammari et al. 2024)	Extraction of Cross-Topics	Connected Qur'an – Hadith Ontology	BERT Embeddings, Cellfie	Relation Extraction	Linking Two Islamic Sources	Topic Alignment
9	Enhancing Qur'anic Ethics through NLP-Based Search (Aftab et al. 2025)	Extraction of Ethical Concepts	Representation of Qur'anic Virtues	Sentence Transformers	Semantic Analysis	High Precision and Recall	F1-Score, User Evaluation

2.2. Abduction (Tadāʿī)

In the *Amid Dictionary* (2010), three meanings are provided for the term *tadāʿī* (abduction): A. The principle or state in which thoughts, ideas, emotions, and experiences become interconnected, such that they emerge sequentially in the mind; a chain of thoughts or notions; B. Calling upon one another and gathering together; C. Recalling or remembering. Moreover abduction (*tadāʿī*) is defined as calling upon one another, and in psychology, it refers to the relationship between a phenomenon and its associated thoughts (Pournamdarian & Tehrani Sabet 2010). Lobo describes abduction as a process for explaining the observable influences of phenomena in the universe (Lobo & Uzcátegui 1997). Ernest defines abductive relations as the chain of antecedents and consequents of an expression (Ernest 2023).

As is evident from the meaning of the term "abduction" (*tadāʿī*), the

intended sense in this paper is the recalling of a word as a result of seeing or hearing another word. In English, two equivalents exist for *tadāʿī*: in the context of logic, the term "abduction" is used, whereas in the context of psychology, the term "association" is employed.

3. Scope of Prior Approaches

Numerous methods have been employed to construct or develop ontologies, semantic networks, and associative structures. Mohammadi and Badie, in their article, proposed a method for extracting concept chains and assigning scores to them. First, the text is segmented and semantically parsed. Then, concept chains are extracted. Finally, key concepts are identified using the assigned scores and a predefined threshold (Mohammadi & Badie 2017). Ahmadi et al. (2017) utilized a lexical co-occurrence method to extract concept hierarchies in the field of scientometrics. Mousavi et al. (2017) presented a method for constructing a Persian ontology by establishing links between Persian words and PWN (Princeton WordNet). This WordNet contains 16,000 words and 22,000 synsets. The accuracy reported in this study was 91.18%. Mousavi and Faili (2021) improved upon their previous work by adding Persian compound verbs to the existing ontology. The method employed in this study also relied on supervised learning.

Humans typically present their observations based on their background knowledge. This process resembles abduction more closely than deductive reasoning, as it requires assumptions that are not explicitly present in the observation itself. Humans possess the ability to comprehend complex situations, a capability that does not necessarily arise directly from observation (Langley & Meadows 2019). Al-Salhi and Abdulla (2022) introduced a method based on domain-specific language mapping for the automatic construction of a Qur'anic ontology. Ghayoomi (2019) proposed a method for automatically determining word meanings based on word-embedding vectors. For each target word, two vectors were constructed: one representing the word itself and another representing the contextual environment in which the word appears. Soliman et al. (2017) presented a trained word-embedding model using a dataset that included Twitter data, web data, and Wikipedia content. The technique employed in this study was Word2Vec.

4. The Proposed Solution and Its Method Engineering

In this section, after examination of the scope and limitations of prior

approaches, the proposed technical solution is introduced. It is noteworthy that the rationale underlying the engineering decisions made during the development of the proposed solution, which constitutes a significant portion of this paper's technical contribution from a method-engineering perspective, is explained at each step of the processing pipeline in this section.

Figure 1 illustrates the overall workflow of the proposed approach. First, the text of Tafsir Nūr is converted into a structured dataset, and its constituent words are extracted. Then, TF-IDF is applied, and a threshold is used to select the final set of relevant words. In the next stage, and in parallel for these selected words, a co-occurrence matrix, word-embedding vectors, Persian FarsNet ontology relations, and connections to Arabic root extraction are obtained. Finally, based on the relations derived in the previous stage, abduction pairs and abduction frames are generated.

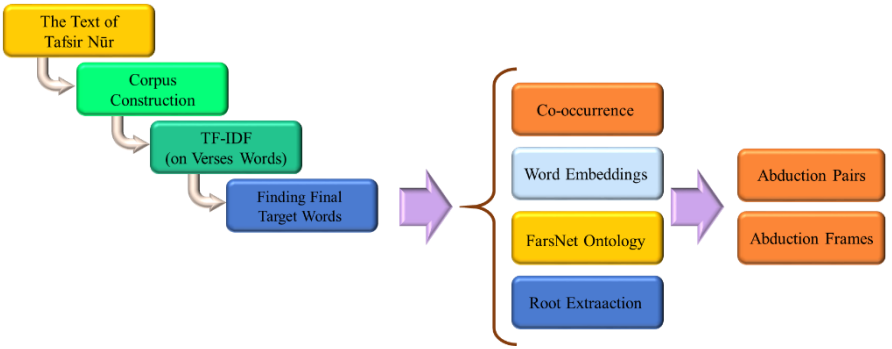

Figure 1. General Method of the Proposed Approach

This process (co-occurrence matrix, word embedding vector, Persian FarsNet ontological relations, connection to Arabic text root-finding, and finally clustering) was conducted in three cases:

1. Identifying final words using TF-IDF thresholding
2. Identifying final words using a thesaurus
3. Identifying final words using normalization, word embedding, and FarsNet

4.1. Corpus Construction

To prepare the corpus, the Tafsir Nūr (Qara'ati 2004) was selected. The reason for using this Tafsir is its simple and fluent text, which makes the concepts of the Qur'an accessible to the general public. The source of the

corpus is the Islamic Encyclopedia website (https://wiki.ahlolbait.com).

The Qur'an Comprehensive Database website (https://quran.inoor.ir) provides thematic categorization of verses. Using this site, verses related to the Hereafter were selected, comprising approximately 1600 records.

4.2. Preprocessing

After constructing the corpus from the Tafsir texts, preprocessing was performed. Preprocessing consisted of four stages: normalization, removal of extra punctuation marks, stop word removal, and lemmatization. The Hazm library was used for the preprocessing stages. After preprocessing, the number of words decreased from approximately 19000 to 17000. Subsequent processing stages were conducted separately for the three word categories. After removing low-frequency words (and applying TF-IDF thresholding), the number of words reduced to approximately 4000.

4.3. Co-Occurrence

The co-occurrence matrix, a square matrix representing the probability of words appearing together within a window size of 5, was computed.

4.4. Word Embedding

The two main approaches in word embedding systems are CBOW and Skip-gram (Hinton 1986; Mikolov et al. 2013). Commonly used models include word2vec, GloVe, and FastText (Chawla 2018). This study utilized two pre-trained word embedding models: AminMozhgani and FarsiYar. The Mozhgani model (n.d.) is based on the word2vec word embedding trained on the Persian Wikipedia 2016 corpus (wikipedia_fa_all_nopic_2016-12.zim). Approximately 2000 words were extracted from this model. The FarsiYar (n.d.) corpus collection provides several services for Persian language processing (https://text-mining.ir/corpus). Among the multiple word embedding models available in this corpus, the GloVe model was selected. Approximately 2900 words were extracted from this model.

After extracting the word embedding vectors (separately for each of the two models), cosine similarity measurement was performed. The meaningfulness of high-scoring word pairs indicated the validity of the proposed models (e.g., the word pair "Hazrat" and "Hassan" from the Mozhgani model, and the pair "adab" and "rusum" from the GloVe model had high scores).

4.5. FarsNet

FarsNet was selected as the best and most comprehensive Persian WordNet. This ontology offers several advantages over other Persian ontologies:

- FarsNet covers more than one million nodes (words).
- It is general-purpose.
- It was manually produced (non-automatically), thus having very high reliability. Additionally, it is continuously updated with the help of crowd-sourcing.
- It provides a suitable application programming interface (API), available both online and offline.
- It has been practically evaluated in various applications and is recognized as a key technical infrastructure for Persian NLP.

Using the FarsNet web-service, synsets exactly matching the words were retrieved and stored.

4.6. FarsNet Usage in the Proposed Approach

For each word, the corresponding synset ID, sense ID, synset text, noun category, verb past stem, verb present stem, verb type, semantic category, and part of speech were stored. Through matrix processing and utilization of the FarsNet ontology, relations among words in the Tafsir texts of the verses were derived (Figure 2).

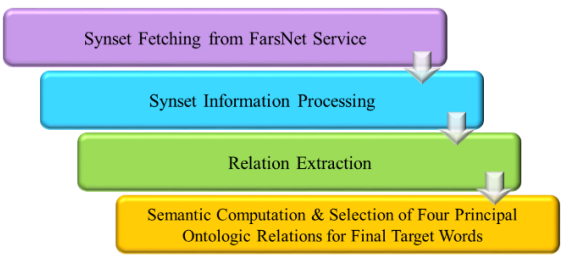

Figure 2. Method of Extracting Relations Using the FarsNet Ontology

4.7. Word Roots in the Qur'anic Text

The text and wording of the Qur'an are highly sacred and meticulously structured. Many Qur'anic scholars believe that the Qur'anic text possesses linguistic inimitability. Accordingly, the Arabic Qur'anic text was also

employed in this study.

This research proposes a method to establish connections between two texts in different languages that are related (not necessarily translations of each other). Here, the Arabic text of the Qur'an and the Persian text of Tafsir Nūr constitute these two texts. Since they are not direct translations, one-to-one mapping cannot be applied. Instead, this study proposes a TF-IDF-based method applied separately to both texts to computationally link them.

4.8. Forqan Corpus

Using the Forqan corpus (Estiri et al. 2013), the roots of Arabic words in each verse were first extracted, followed by computation of TF-IDF scores for these roots across the verses (Figure 3).

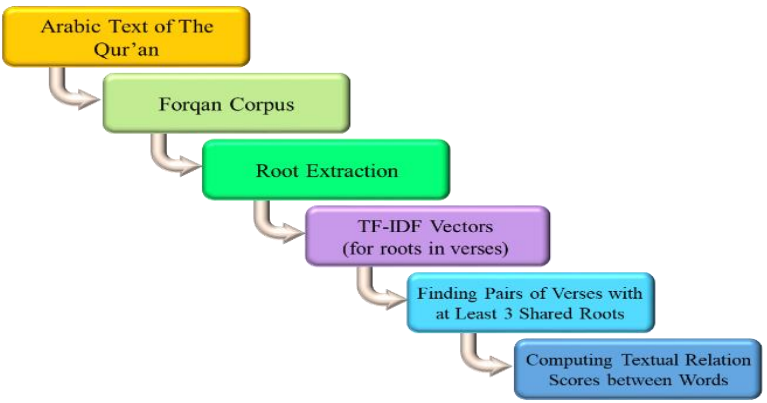

Figure 3. Method of Calculating Relations between Words Based on Verse Word Roots

4.9. TRR Relation Calculation

This relation is derived based on the previously computed TF-IDF scores of roots and words.

1. Identifying pairs of verses that share more than 3 common roots. If a pair of verses shares fewer than 3 common roots, they are considered unrelated.
2. Calculating the RR relation (Root Relation): For all common roots in each verse pair.

$$RR_{x,y} = \sum_{i=1}^{n} T_{r_i,x} * T_{r_i,y}$$

n : Number of Common Roots Between Verses x and y

$T_{r_i,x}$: TF-IDF Value for Root r (the i-th Common Root) in Verse x

$T_{r_i,y}$: TF-IDF Value for Root r (the i-th Common Root) in Verse y

x,y : Index of Verse Pairs with More Than Three Common Roots (If x,y Share Fewer Than Three Common Roots, It Is Zero)

The matrix below is constructed using the root relations derived from the verses.

$$RR_{Mat} = \begin{bmatrix} RR_{1,1} & \dots & RR_{1,m} \\ \vdots & \ddots & \vdots \\ RR_{m,1} & \dots & RR_{m,m} \end{bmatrix}$$

The matrix is a square matrix with rows and columns equal in number to the verses.

3. Calculation of TRR for all words:

$$TRR_{w_\alpha w_\beta} = \sum_{i,j=1}^{m} T_{w_{\alpha_i}} * T_{w_{\beta_j}} * RR_{Mat_{i,j}}$$

w_α , w_β: Persian word α and Persian word β

$T_{w_{\alpha_i}}$: TF-IDF vector value of the word in the interpretation text of i-th verse

$T_{w_{\alpha_j}}$: TF-IDF vector value of the word in the interpretation text of j-th verse.

$RR_{Mat_{i,j}}$: Entry corresponding to verses \(i\) and \(j\) in the matrix.

Some of the results from these calculations with high TRR values are presented in Table 2.

Table 2. Some of the results with high TRR values.

First Word	Second Word
معاد	متاع
هستی	مکان
موسی	صالح
معاد	الدنیا
قیامت	اتقوا
ینظرون	شاهد

4.10. Matrix Weighting, Summation, and K-Means Clustering Method

Using the matrices obtained from the previous sections, a square matrix

was constructed where the rows and columns represented the final words. The K-Means clustering method was applied to the resulting matrix. The clusters obtained are referred to as "abduction frames." Some of these clusters are presented in the table 3. Note that only a portion of the cluster members are shown for display purposes.

Table 3. Some Clusters obtained from K-Means method.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
تفکیک	بیهوده	بخشش	ارائه
جداگانه	جایز	خرج	استفاده
جرم	روا	طلب	اصلاح
دسته	زشت	شهادت	اعمال
جزء	سزاوار	شعار	انجام
ذیل	غفلت	هدیه	ایجاد
قاعده	قاتل	وعده	برداشت
مطابق	مجرم	پاداش	تقویت
معین	مقصر		حرکت
وجه	گناهکار		فراهم

Figure 4 shows the Davies-Bouldin index values for different numbers of clusters.

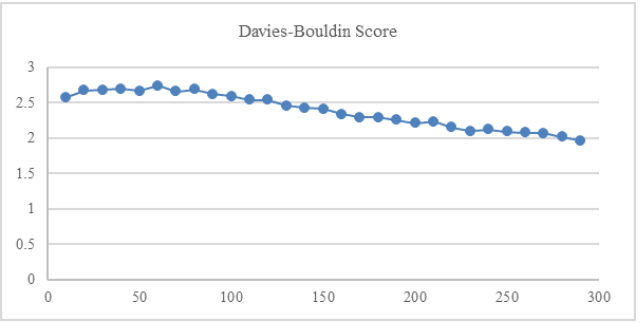

Figure 4. The Davies-Bouldin Score for different numbers of clusters

As shown in Figure 4, although the criterion continues to decrease after 200 clusters, suggesting apparent improvement, an excessively large number of clusters reduces the interpretability of the clusters and may lead to overfitting of these semantic clusters. It is noteworthy that other clustering methods (such as agglomerative and divisive clustering and DBSCAN) were also employed, but they yielded significantly lower output quality compared to K-Means.

4.11. Thesaurus Method

One effective approach for selecting appropriate words in constructing a semantic network is the use of a thesaurus. In this study, the *Thesaurus of Qur'anic Concepts* (2007) was utilized, which is available digitally through the Noor website (https://noorlib.ir/book/info/5028).

This thesaurus contains approximately 3000 single-word Qur'anic terms. Of these, around 1200 words were found in the corpus used in this research (i.e. mentioned in the resurrection verses). All previous steps were performed separately for these words. First, words not present in the thesaurus were removed. Then, the remaining steps were applied to the surviving words.

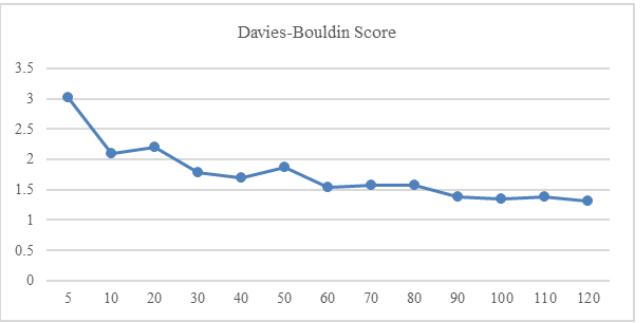

Figure 5. Davies-Bouldin Index Chart for thesaurus Word Clustering

The above figure displays the Davies-Bouldin index values for various numbers of clusters. Based on the chart, 90 clusters were selected for use. In Table 4, some resulting clusters by this method are presented.

Table 4. Some of the clusters obtained using the K-Means method in the thesaurus section with 90 clusters

Cluster 21	Cluster 11	Cluster 7	Cluster 4
آهن	هنر	غرور	وفات
طلا	اعتقاد	عذاب	زادگاه
قلم	اخلاق	محبت	نسب
برق	علم	نفرت	وارث
زره	مطالعه	رنج	بستگان
مس	دانش	فقر	باردار
سلاح	شناخت	احساس	تولد
سپر	آموزش	آزار	ارث
سنگ	پژوهش	اضطراب	اجداد
زر	علوم	لذت	نوه

4.12. Normalization

The total number of words used when IDF values determined the threshold was approximately 4,000 words. Additionally, among these words were meaningless terms, errors (typos), and words that had been incorrectly stemmed. In the thesaurus section, the total selected words numbered around 1,200. Solutions were proposed to address each of these issues.

Problem 1: Some words that had been incorrectly stemmed. These words were divided into two categories: Persian and Arabic.

Solution: The Persian text processing library AIPA (https://aipaa.ir) and the Arabic text processing library Qalsadi (https://pypi.org/project/qalsadi) were utilized. Qalsadi is a Python library used for Arabic language text processing. This library also features Arabic stemming capabilities (Zerrouki 2020).

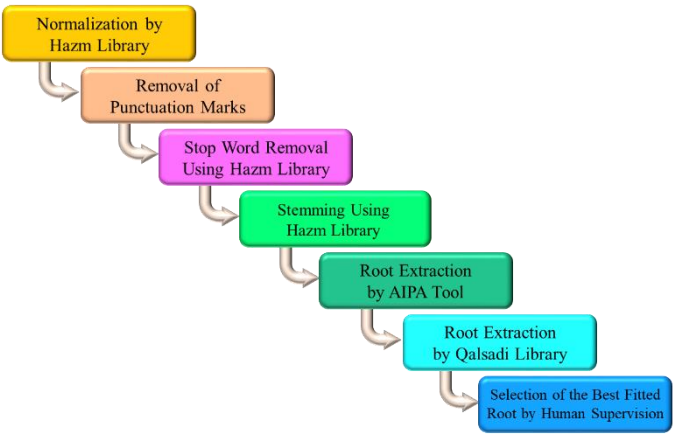

Figure 6. Root Extraction Process Using AIPA and Qalsadi Tools

After initial normalization, words were listed and then root-extracted using AIPA. In the next stage, these words (which had undergone Persian stemming) were processed by Qalsadi for Arabic Root Extraction. Given that both tools (Arabic and Persian) have errors, human judgment determined which stemming\Root-Extraction was more accurate. This file was then used for root extraction of the words. From approximately 17000 words stemmed using the Hazm library, around 15000 words remained after all.

Problem 2: Words that were meaningless or erroneous (typos).

Solution: To exclude meaningless or erroneous words, a condition was established. For a word to be included in the final word list, it must belong

to at least one of the following categories:

1. The word exists in the AminMozhgani or FarsiYar word embedding models.
2. Co-occurrences of the word are present in FarsNet.

Words not falling into these categories were removed from the corpus, leaving approximately 7000 words. For these remaining words, TF-IDF vectors, co-occurrence matrices, and TRR (Arabic stemming) were computed. The matrices were then combined according to the previous priorities. Following the formation of the final matrix, the words were clustered using the K-Means method.

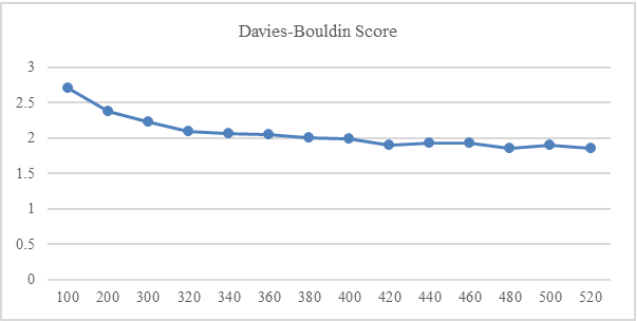

Figure 7. Davies-Bouldin Index Chart for Word Clustering in the Normalization Section

Figure 7 illustrates the Davies-Bouldin index for various numbers of clusters. Based on the chart, 420 clusters were selected. Note that in both the TF-IDF threshold method for final word selection and the thesaurus-assisted selection, a large number of words clustered together in a single group. In the TF-IDF threshold approach, approximately 1300 words fell into one cluster, while in the thesaurus method, around 300 words did so. This phenomenon generally stems from two factors:

1. Low informativeness of the word (absence from FarsNet or word embedding models).
2. Low connectivity of the word (very low scores, near zero, with other words).

Due to this issue, a significant portion of words (about 30%) lacked interpretability and were not included in abduction frames. However, this problem was less pronounced in the normalization section, where the largest clusters contained around 400 members each, comprising less than 14% of the total words.

Table 5. Some of the clusters obtained using the K-Means method in the normalization section with 420 clusters.

Cluster 33	Cluster 28	Cluster 10	Cluster 6
نصب	خوک	کلاه	شعور
هدایت	اسب	چنگال	شهود
پرتاب	سگ	چرخ	معنویت
ذخیره	پرنده	سوزن	تجسم
پیاده	شیر	شلاق	ذات
تعویض	مار	شمشیر	انسانیت
تولید	حیوان	میخ	بینش
تخریب	گاو	سپر	تخیل
متصل	عقاب	آهنین	عواطف
تخلیه	اژدها	زره	وجدان

5. Evaluation

5.1. Questionnaire

For the external evaluation criterion, 10% of the clusters from each section were randomly selected, with the condition that they contain at least 10 and at most 40 members. For the TF-IDF threshold condition, 20 clusters; for the thesaurus condition, 12 clusters; and for the normalization condition, 42 clusters were randomly chosen. After cluster selection, 10 members were randomly picked from each cluster. These 10 members are identified as positive. An additional 10 members were randomly selected from other clusters. These are identified as negative. For each user, a form is displayed that asks about the degree of belonging of a word to a group of words. Users must rate the degree of a word's belonging to a group on a scale from very high, high, medium, low, and very low. Two scenarios exist for the queried word:

1. A word is randomly selected from the cluster-related words (true positive) and becomes the query word. The ideal response here is "very high".
2. A word is selected from unrelated words (true negative) and becomes the query word. The ideal response here is "very low".

Two questions are asked per cluster: one for true positive and one for true negative. Each user receives a total of 6 questions. Additionally, age and gender are collected from each participant. To gather data from various individuals, advertisements were posted three times (once per experiment) in a Telegram channel (https://telegram.me/OfficialPersianTwitter) with approximately 500,000 members. Multiple advertisements were placed at

meaningful time intervals to balance the data collection.

5.2. Evaluation Method

If a respondent selects "very high" or "high," it is considered a positive evaluation. If "low" or "very low" is selected, it is considered negative. "Medium" responses are excluded due to respondent uncertainty. Four values are defined for evaluation using metrics such as accuracy:

- True Positive (TP): The model agrees with human judgment, both indicating high word belonging.
- True Negative (TN): The model agrees with human judgment, both indicating low word belonging.
- False Positive (FP): The model indicates high belonging contrary to human low assessment.
- False Negative (FN): The model indicates low belonging contrary to human high assessment.

Five metrics—accuracy, precision, recall, specificity, and F-measure—were computed for evaluation.

5.3. Experiments Conducted

Multiple experiments were conducted at time intervals for data collection. Across three experiments, 1,295 volunteers participated in the dynamic online questionnaire, each answering 6 questions, yielding 7,770 responses stored. These questions evaluated a total of 71 clusters.

It can be observed that there is no significant difference between the results of the first and second experiments and the combined results of the first, second, and third experiments, with outcomes showing convergence. The overall results from the three methods—TF-IDF threshold, thesaurus, and normalization—are presented in the table 6, 7, 8.

Table 6. Number of Participating Persons in each Experiment

Method / Experiment Order	Normalization	Thesaurus	TF-IDF Threshold
Number of Participants in the First Experiment	180	53	83
Number of Participants in the Second Experiment	236	65	116
Number of Participants in the Third Experiment	348	81	133

Table 7. Evaluation Results for the Combined First, Second, and Third Experiments (values in percentages)

Criteria / Experiment	F	Specificity	Recall	Precision	Accuracy
TF-IDF Threshold	62.09	81.34	73.37	53.81	67.89
Thesaurus	62.5	83.61	75	53.57	69.23
Normalization	62.44	87.49	79.22	51.52	70.20

Table 8. Overall Results of Accuracy, Precision, Recall, Specificity, and F-Measure.

F	Specificity	Recall	Precision	Accuracy
62.35	85.35	76.93	52.42	69.47

5.4. Qur'anic Experts Evaluation

A separate questionnaire was designed for Qur'anic experts. 12 clusters were randomly selected, and from each cluster, 7 words were chosen to ensure cluster coverage and diversity. Two questions were asked per cluster:

1. To what extent are these concepts semantically related or associated? This question had 5 options: very high, high, medium, low, very low. Very high was scored as 5, and very low as 1.
2. If the computational Qur'an mining detects associations between these concepts, does it identify innovative connections (from a research perspective) among some concepts? This question had 4 options:
   - Yes, the connection is highly innovative research-wise.
   - Yes, the connection is innovative research-wise.
   - No, it is relatively established in prior research.
   - No, the connection is entirely obvious and well-known.

   The first option scored 4, and the last scored 1.

Seven experts participated. The average score for the first question was 3.74, and for the second, 2.6.

6. Conclusion

This study presented and evaluated a technical approach for creating association semantic networks for the Holy Qur'an. Results indicate that natural language processing methods can automatically and semi-automatically generate abduction semantic networks for the Qur'an using interpretive data, base ontologies like FarsNet, and word co-occurrences in verses.

Evaluation by Qur'anic experts shows that the outputs exhibit strong semantic connection accuracy and innovation discovery. Thus, generating abduction semantic networks for the Qur'an holds practical significance for Qur'anic and Islamic studies communities.

A key outcome of this research is the recommendation that producing abduction semantic networks using various technical methods be considered a vital and practical topic in future computational Qur'an mining research, particularly in semantics, interpretation, and cross-lingual studies application domains.

Acknowledgements

This paper is extracted from the thesis prepared by Nafiseh Jafari under the supervision of Dr. Maryam Hourali, in fulfilment of the requirements for the degree of Master of Computer Engineering in Artificial Intelligence. The authors extend their gratitude to the Faculty of Electrical and Computer Engineering at Malek Ashtar Industrial University and to the Research Deputy for their support of this research.

Declarations

Funding: No funding was received for conducting this study.
Conflict of Interest: The authors declare no competing interests.

References

Aftab, Y., Awan, M. A., Khaleeq, D., & Ismail, T. (2025). Enhancing Qur'anic Ethics and Morality: An NLP-Based Semantic Search Model for Urdu Translation. *International Journal of Innovations in Science & Technology*, 7(1), 651–663.

Ahmadi, H., Osareh, F., Hosseini Beheshti, M. S. and Heidari, G. (2017). Designing Semiautomatic System in Ontology Structure by To Co-occurrence word Analysis and C-value Method (Case Study: The field of Scientometrics of Iran). *Iranian Journal of Information Processing and Management*, *33*(1), 185-216. https://doi.org/10.35050/JIPM010.2017.008

AminMozhgani (n.d.). *Persian Word2Vec: A Persian Word2Vec model trained by Wikipedia articles*. GitHub repository. https://github.com/AminMozhgani/Persian_Word2Vec

Akra, D., Hammouda, T., & Jarrar, M. (2025). QuranMorph: Morphologically Annotated Qur'anic Corpus. *arXiv preprint arXiv:2506.18148*. https://doi.org/10.48550/arXiv.2506.18148

Al-Salhi, R. Y., & Abdulla, A. M. (2022). Building Qur'anic stories ontology using MappingMaster domain-specific language. *International Journal of Electrical and Computer Engineering*, 12(1), 684-693.

https://dx.doi.org/10.11591/ijece.v12i1.pp684-693

Alshammari, I. K., Atwell, E., & Alsalka, M. A. (2024). Linking Qur'an and Hadith Topics in an Ontology using Word Embeddings and Cellfie Plugin. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)* (449-455).

Amid, H. (2010). *Amid Dictionary*. Tehran: Amirkabir.

Chawla, J. S. (2018). *What is GloVe?* https://medium.com/analytics-vidhya/word-vectorization-using-glove-76919685ee0b

Ernest, P. (2023). Abduction and Creativity in Mathematics. In: Magnani, L. (eds) *Handbook of Abductive Cognition*. Springer, Cham. https://doi.org/10.1007/978-3-031-10135-9_38

Estiri, A., Kahani, M. and Ghaemi, H. (2013). Creating and publishing a semantic web infrastructure for the Holy Quran. In *the 5th Conference on Information Technology and Knowledge*, Shiraz University.

FarsiYar corpus (2019). In (Text-Mining.ir) GitHub repositories. https://github.com/Text-Mining/Useful-Corpora-for-Text-Mining-in-Persian-Language

Ghayoomi, M. (2019). Identifying Persian Wordsâ Senses Automatically by Utilizing the Word Embedding Method. *Iranian Journal of Information Processing and Management*, 35(1), 25-50. https://doi.org/10.35050/JIPM010.2019.001

Gruber, T. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2), 199-220. https://doi.org/10.1006/knac.1993.1008

Hinton, G. E. (1986). Learning distributed representations of concepts. In: *Proceedings of Eighth Annual Conference of the Cognitive Science Society*, 1-12.

Hinton, G. E. (1986). Learning Distributed Representations of Concepts. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 8. https://escholarship.org/uc/item/79w838g1

Khalila, Z., et al. (2025). Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models. *International Journal of Advanced Computer Science and Applications (IJACSA)* 16(2). http://dx.doi.org/10.14569/IJACSA.2025.01602134

Langley, P., & Meadows, B. (2019). Heuristic Construction of Explanations through Associative Abduction. *Advances in Cognitive Systems*, 8, 93-112.

Lobo, J., & Uzcátegui, C. (1997). Abductive consequence relations. *Artificial Intelligence*, 89(1-2), 149-171. https://doi.org/10.1016/S0004-3702(96)00032-X

Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). Efficient Estimation of

Word Representations in Vector Space. *International Conference on Learning Representations.*

Mirarab, A., & Khorram abadi arani, M. (2022). Analytical review of Iranian Ontographs for semantic Persian information retrieval. *Sciences and Techniques of Information Management*, *8*(3), 201-232. https://doi.org/10.22091/stim.2022.8274.1798

Mohammadi, S., & Badie, K. (2017). Key concept extraction using frame networks and concept chains. *Iranian Journal of Electrical and Computer Engineering*, 15(1), 64-72.

Mousavi, Z., & Faili, H. (2017). Persian wordnet construction using supervised learning. *International Journal of Information & Communication Technology Research*, 9(2), 35-44. https://doi.org/10.48550/arXiv.1704.03223

Mousavi, Z., & Faili, H. (2021). Developing the Persian Wordnet of Verbs Using Supervised Learning. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(4), 1-18. https://doi.org/10.1145/3450969

Naghizadeh, A., Salamat, M., Hamzeian, D., Akbari, S., Rezaeizadeh, H., Vaghasloo, M. A., Karbalaei, R., Mirzaie, M., Karimi, M., & Jafari, M. (2021). IrGO: Iranian traditional medicine General Ontology and knowledge base. *Journal of biomedical semantics*, 12(1), 9. https://doi.org/10.1186/s13326-021-00237-1

Persian Medicine Diseases Ontology (PMD). (2019). [Ontology]. NCBO BioPortal. https://bioportal.bioontology.org/ontologies/PMD?p=summary

Pournamdarian, T. & Tehrani Sabet, N. (2010). Associations and Rhetorical Figures. *Literary Arts*, 1(1), 1-13.

Qara'ati, M. (2004). *Tafsīr Nūr*. Tehran: Cultural Center of Lessons from the Qur'an.

Saadaoui, Z., Tlig, G., & Jarray, F. (2024). LLMs Based Approach for Quranic Question Answering. *International Conference on Web Information Systems and Technologies*. https://doi.org/10.5220/0013012900003825

Salah, A., Mahdi, M., & badawy, A. (2025). Exploring Thematic Structures in Surah Al-Baqarah using TF-IDF, Dimensionality Reduction, and K-means Clustering. *Journal of Computing and Communication*, 4(2), 1-12. https://doi.org/10.21608/jocc.2025.446634

Shahid, U., Hussain, M.Z., & Sayers, W. (2025). Computational Analysis of Quran Text Using Machine Learning and Large Language Models. In *8th International Conference on Data Science and Machine Learning Applications (CDMA)*, 18-24.

Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Noor, P., Famian, A.R., Bagherbeigi, S., Fekri, E., & Monshizadeh, M. (2009). *Semi Automatic Development of FarsNet: The Persian WordNet*.

http://farsnet.nlp.sbu.ac.ir/Site3/Modules/Public/Default.jsp

Shohoud, Y., Shoman, M., & Abdelazim, S. (2023). Qur'anic Conversations: Developing a Semantic Search tool for the Quran using Arabic NLP Techniques. *ArXiv, abs/2311.05120*. https://doi.org/10.48550/arXiv.2311.05120

Soliman, A. B., Eissa, K., & El-Beltagy, S. R. (2017). AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP. In *Procedia Computer Science*, 117, 256-265. https://dx.doi.org/10.1016/j.procs.2017.10.117

Tariq, M., Awan, M. A., & Khaleeq, D. (2025). Developing a Qur'anic QA System: Bridging Linguistic Gaps in Urdu Translation Using NLP and Transformer Model. *International Journal of Innovations in Science & Technology*, 7(1), 493-505. https://doi.org/10.33411/ijist/202571492505

Thesaurus of Qur'anic Concepts (2007). Qom: Bustān Ketāb. https://noorlib.ir/book/info/5028

Zerrouki, T. (2020). *Qalsadi, Arabic morphological analyzer Library for python*. https://pypi.python.org/pypi/qalsadi/