

# A Comparative Evaluation of Ontology Learning Techniques in the Context of the Qur'an

Rohana Ismail<sup>1</sup>, Mokhairi Makhtar<sup>2</sup>, Hasni Hasan<sup>3</sup>, Nurnadiah Zamri<sup>4</sup>, Azilawati Azizan<sup>5</sup>

Department of Computer Science-Faculty of Informatics and Computing,  
Universiti Sultan Zainal Abidin, Campus of Besut, Terengganu, 22200, Malaysia<sup>1,2,3,4</sup>  
Universiti Teknologi MARA (UiTM), Cawangan Perak, Kampus Seri Iskandar,  
32610, Seri Iskandar, Perak Darul Ridzuan, Malaysia<sup>5</sup>

**Abstract**—Ontology Learning refers to the automatic or semi-automatic process of creating ontologies by extracting terms, concepts, and relationships from text written in natural languages. This process is essential, as manually building ontologies is time-consuming and labour-intensive. The Qur'an, a vast source of knowledge for Muslims, presents linguistic and cultural complexities, with many words carrying multiple meanings depending on context. Ontologies offer a structured way to represent this knowledge, linking concepts systematically. Although various ontologies have been developed from the Qur'an for purposes such as advanced querying and analysis, most rely on manual creation methods. Few studies have examined the use of Ontology Learning for Qur'anic ontologies. Thus, this study evaluates three Ontology Learning techniques: Named Entity Recognition (NER), statistical methods, and Quranic patterns. The NER aims to find names represented by entity, statistical techniques aimed at finding frequently occurring words, and pattern-based techniques aim to identify complex relationships and multi-word expressions. The Ontology Learning techniques were evaluated based on precision, recall, and F-measure to assess extraction accuracy. The NER technique achieved an average precision of 0.62, statistical methods of 0.45, and pattern-based techniques of 0.58, indicating the strengths and weaknesses of each approach for extracting relevant terms as concepts, instances, or relations. This indicates that improvements or enhancements to the existing techniques are necessary for more accurate results. Future work will focus on refining or adapting patterns based on the structure of the Qur'an translation using LLMs.

**Keywords**—Ontology learning; Qur'an; NER; statistical; pattern-Based; hajj

## I. INTRODUCTION

The Qur'an serving as a comprehensive knowledge source, provides guidance on various facets of life for Muslims. For instance, it provides clear principles of justice, such as in Surah Al-Baqarah, which emphasizes fair trade and the prohibition of usury. It also gives ethical guidance on personal behavior, as seen in the verses on charity and kindness to others, particularly towards parents and orphans. Given the large global Muslim population, the need to access the knowledge contained within the Qur'an has grown. Muslims around the world turn to the Qur'an for guidance in daily life, from the proper conduct of prayers to complex societal issues such as governance and finance. However, the Qur'an is written in Classical Arabic, which is syntactically and semantically complex. Classical Arabic features a rich system of morphology, where the same root word can have multiple meanings depending on its context.

This linguistic complexity makes it challenging to access the knowledge within the Qur'an in a systematic and efficient way. Creating an ontology can address this challenge. An ontology, in this context, is a structured framework that organizes the concepts and relationships within the Qur'an in a way that makes them easier to understand, search, and interpret. The ontology is especially useful for managing scattered knowledge within the Qur'an. The Qur'an contains knowledge that is spread across various chapters (Surahs) and verses (Ayahs), often with different verses addressing the same or related concepts in different contexts. Moreover, an ontology can help address the challenge of semantic interpretation by ensuring that terms are consistently understood in their full context.

By definition ontology is a formal, explicit statement that captures a shared understanding of a domain [1]. It provides a well-organized framework to help us understand the different elements within that area. The application of ontology extends across a spectrum of domains, contributing significantly to areas such as Information Retrieval, Information Extraction, Knowledge Representations and Query Answering Systems. It helps in efficiently retrieving relevant information for different domains of studies. In the Quranic study, Information Extraction enables the extraction of meaningful insights and relationships embedded within the Quranic text, contributing to a more nuanced understanding [2]. The ontology plays a crucial role in Knowledge Representations, where it serves as a structured framework for organizing and representing the complex relationships and concepts within the Quranic domain. In Knowledge Management Systems, Ontology acts as a foundational element for effective organization, storage, and retrieval of Quranic knowledge, facilitating seamless access for scholars, students, and researchers. Additionally, Ontology contributes to Intelligent Query Answering Systems by enabling more sophisticated and context-aware responses to queries related to the Quran [3]. This enhances the overall efficiency of querying systems, providing users with accurate and relevant information tailored to their specific inquiries. The integration of Ontology through Semantic Web technologies not only aids in capturing and representing disseminated knowledge within the Quran but also extends its benefits to diverse applications, including efficient retrieval, meaningful extraction, structured representation, and intelligent querying of Quranic knowledge.

However, challenges arise for creating an ontology. Since the Qur'an is written in Classical Arabic, which is syntactically and semantically complex, creating ontology manually requires

deep knowledge of language and is very time-consuming. Manual ontology methods struggle to capture these intricate patterns without extensive linguistic expertise. Furthermore, Classical Arabic is significantly different from Modern Standard Arabic, which makes it even more challenging for non-experts to accurately identify and relate concepts. Manual ontology development is prone to biases, human error, and interpretive subjectivity, particularly when dealing with a sacred text like the Qur'an. Different scholars may interpret concepts differently, leading to inconsistencies in how relationships are defined and organized within the ontology. In particular, manually creating ontologies takes a lot of time, making it hard to expand or update when new interpretations appear. Since an ontology needs to grow and stay current, manual methods are often too slow and require too many resources to maintain.

Because of these challenges, there's a growing interest in automatic or semi-automatic methods for creating ontologies, known as Ontology Learning (OL), which can help speed up the process and reduce inconsistencies. The OL is a process of either automatic or semi-automatic creation of ontologies from a corpus of natural language text [4]. This involves extracting relevant domain terms and relations between these concepts. Later, the terms are encoded using an ontology language such as OWL. Ontology learning encompasses various techniques, for example, Named Entity Recognition (NER), Machine learning, statistical-based techniques, and pattern-based techniques [5].

Ontology Learning (OL) leverages these automated techniques to extract concepts and relationships in several ways. First, by applying statistical methods, OL can identify patterns and term frequencies within the text. These methods help to spot recurring concepts and likely relationships, providing a more systematic and consistent basis for creating an ontology compared to manual methods. Second, this study can use Named Entity Recognition (NER) to automatically identify specific entities (e.g., locations, persons, events) within the Quranic text. Automated NER processes can be fine-tuned to the Quran's unique vocabulary and context, improving precision in capturing entities related to Hajj and other topics, which is often limited in manual approaches. Third, the study can extract complex relationships that are often too subtle for manual annotation by implementing pattern-based approaches. Pattern-based extraction can detect sequences or structures indicative of certain relationships, even if they aren't explicitly named, enhancing the ability to capture deeper connections between concepts. The automated approaches can maintain a high level of consistency by applying rules uniformly across the text. They reduce human error and bias, creating a more accurate and reliable ontology that can be expanded upon as new linguistic insights develop. On the other hand, it allows for faster ontology construction and allows for easy incorporation of new texts or insights. This adaptability is crucial for creating a comprehensive and continually updated representation of Quranic concepts, making it possible to refine and expand the ontology efficiently as new interpretations emerge.

Previous research on OL for Quranic knowledge, particularly in structured domains like Hajj and Umrah, has encountered several limitations, including inconsistent concept extraction methods, a lack of automation, and difficulties in handling Quranic linguistic complexity. While efforts have been

made to construct Quranic ontologies, many existing approaches rely on manual annotation or semi-automated techniques, leading to inefficiencies and inconsistencies in knowledge representation [6]. Additionally, previous studies have not fully explored the potential of advanced Natural Language Processing (NLP) techniques, such as Named Entity Recognition (NER) and statistical methods, for automating OL in religious texts [7]. Existing OL models also struggle with extracting structured knowledge from Quranic verses, particularly when capturing non-taxonomic relationships and context-specific meanings [8]. Furthermore, there is limited research evaluating different OL techniques for Quranic texts, leaving a gap in understanding which methods yield the most effective results [9]. Addressing these gaps is crucial for improving automated OL frameworks in Islamic knowledge representation.

Therefore, this paper introduces different techniques in Ontology Learning. This study specifically focuses on extracting ontological elements from a few chapters and verses that are related to Hajj and Umrah, as these domains contain structured ritual knowledge that can benefit from automated Ontology Learning. The paper also presents results from concept extractions employing the Named Entity Recognition (NER) technique, statistical techniques, and pattern-based techniques. By evaluating these techniques, the study aims to provide insights into more efficient and accurate methods for constructing ontologies in the context of Quranic knowledge. This paper is organized as follows; Literature Review, Methodology, Result, Discussion, and Conclusion

## II. LITERATURE REVIEW

Ontology Learning (OL) refers to the automated or semi-automated process of constructing ontologies by extracting terms, formation concepts, identification relations, and developing axioms within a given domain from textual sources [4]. This reduces manual effort and enhances consistency in ontology development. Subsequently, these extracted terms and relationships are transformed to build an ontology. The OL integrates techniques from diverse domains such as Information Retrieval (IR), Information Extraction (IE), Natural Language Processing (NLP) and Machine Learning (ML) [10], [11], [12], [13], [14]. It can be classified into shallow learning methods, which have linguistic techniques, statistical-based techniques, and logic-based techniques [5]. These shallow learning techniques could perform tasks such as term extraction, concept formation, taxonomy discovery, non-taxonomic relation extraction, and axiom extraction. Meanwhile, the deep learning methods can be classified into concept extraction and relation extraction, which need to have deeper analysis in understanding texts compared to shallow learning.

The linguistics are based on characteristics of languages such as Part of Speech (POS) tagging and sentence parsing and also rely on thesaurus such as WordNet [15]. Based on linguistics, patterns can be generated to perform many extraction functions. The NLU-based method uses soft pattern matching to extract contextual definitions of concepts from a domain-specific corpus of the Building Information Model and then applies deep NLU models to convert these concept names and definitions into dense vector representations [12]. The field of text pattern extraction has evolved significantly with

advancements in computational linguistics and machine learning. Recent research by Jung, Zhou, and Smith (2024) introduces the Word-Text-Topic Extraction (WTT) approach, which integrates word embedding techniques, collocation processes, and topic modeling to enhance the efficiency of text pattern extraction for theoretical research [16]. Additionally, Hua et al. (2024) proposed an automated pattern generation model for Open Information Extraction (OIE), which autonomously identifies extraction patterns in natural language text, offering improved generalization across domains [17]. These advancements underscore the growing reliance on AI-driven techniques to improve the accuracy and scalability of text pattern extraction in various applications. The widely used Hearst patterns, also known as lexico-syntactic patterns, have been utilized to extract taxonomic relations [18]. In the Quranic study, patterns from the Qur'an domain structure have been proposed to extract relations such as part of relations, definition relations, and synonym relations [19]. The pattern is crafted from the structure of the Qur'an using Hillali Khan's translation version of the Qur'an. The patterns are inspired by Lexico-syntactic patterns by Hearst. It is simple yet able to extract relations in the Qur'an related to the Solat domain.

Named Entity Recognition (NER) is an NLP method that involves in extracting and classifying relevant information within the Information Extraction field. The NER technique relies on the characteristics of linguistics syntax. The NER is significant for identifying and classifying proper names based on the type of entity or predefined categories in unstructured text within a domain such as people, organizations, locations, and other entities [20]. It also extracts relations between entities [21]. It aimed at identifying names and classifying them. The NER system such as ANNIE (A Nearly-New Information Extraction System) which is a module in GATE (General for Text Engineering architecture, marks up entities present in the text, categorizing them into predefined categories such as persons, organizations, locations, dates, and others, following the original Message Understanding Conference (MUC) entity types ) [22]. Concept extraction a main task within the OL. The task of concept extraction using NER has been accomplished in the realm of the Quran, where names often signify concepts; the NER technique has been accomplished by Dukes to extract ontological elements for the development of Quran ontologies [23]. Leveraging NER enables the automatic extraction of names from Quranic verses, categorizing them into historical places or individuals. The NER significantly contributes to constructing the Quran ontology, covering 300 concepts with 350 relations.

On the other hand, the statistical base relies on the statistics of the underlying corpus. Statistical techniques are employed to measure the most pertinent phrases according to their frequency and significance within a text corpus [24]. These techniques contain measurements such as term frequency ( $tf$ ) and term frequency-inverse document frequency ( $t$  subsumption, and so forth are examples of common techniques [5]  $tfidf$ ). Contextual, heuristic clustering, association rules, contrastive analysis, latent semantic analysis (LSA), term [24]. The statistical measurement identifies relevant domain terms by calculating their frequency in a text, with frequent terms likely being more pertinent. This measurement determines concepts by identifying single-term

occurrences in a text. Meanwhile, the Logic-based approaches are based on formal logic and reasoning. Typical methods include inductive logic programming and logical inference [5], [2]. An approach has been developed for automatically constructing axioms for concepts and relations by recognizing semantics in natural language texts and representing them in description logic [25]. The latest research addressing the automatic construction of axioms for concepts and relations in description logic [26]. The study introduces Box<sup>2</sup>EL, a novel ontology embedding method that represents both concepts and roles as boxes (i.e., axis-aligned hyperrectangles). This approach models inter-concept relationships using a bumping mechanism, aiming to enhance ontology completion performance by ensuring adherence to the semantics of the underlying description logic.

There are also numerous frameworks have been suggested to streamline the process of ontology construction and assessment. For example, the OLAF framework provides a structured approach to ontology learning, focusing on the identification and extraction of concepts from text, and has been applied successfully in various domains [27]. The framework is implemented in a search engine system for technical products. The Text2Onto [28] is a well-known flexible framework for OL. Text2Onto introduces probabilistic ontology models that consider uncertainty in the construction of ontology.

Recent advancements in ontology learning frameworks have focused on enhancing adaptability by integrating various natural language processing (NLP) techniques and learning algorithms for effective concept extraction and modeling. A notable development is the LLMs4OL approach, which leverages large language models (LLMs) to automatically extract and structure knowledge from natural language text, demonstrating significant improvements in OL tasks. A language model has been introduced to explore an approach for inserting new concepts extracted from text into an ontology by leveraging language models, embedding-based methods, and contrastive learning. The framework integrates pre-trained language models (PLMs) like BERT for edge search and large language models (LLMs) such as GPT, FLAN-T5, and Llama 2 for concept placement, making it highly adaptable for ontology learning and NLP-based concept extraction [29]. These frameworks collectively represent the latest advancements in adaptable OL systems.

Concerning texts specific to a domain, like Quranic translations, there has been limited exploration of Ontology Learning (OL). The research concentrated on the OL methodology for extracting concepts and relationships, particularly within the realm of prayer [19], employing a combination of statistical and linguistic methods. Unlike other initiatives for Quran ontology development, a substantial portion of the ontology construction is conducted through manual processes.

Several studies have explored OL in the context of the Quran. For instance, the Semantic Hadith ontology by study [14] was devised to articulate and correlate fundamental structural concepts from the hadith. Subsequently, they published the six well-known hadith collections as an RDF-Based hadith knowledge graph, which was a step towards making hadith

content accessible to both machines and humans. This project is the first step towards annotating and linking the hadith corpus. Its goal is to make semantic search capabilities easier for academics, scholars, and students who are working on developing, updating, and using a digital repository of Islamic knowledge. Moreover, automated ontology construction using mapping techniques, such as the MappingMaster domain-specific language, can facilitate efficient knowledge representation while reducing manual effort [8].

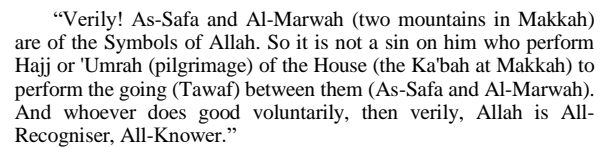
M. Alshammeri et al.[30], has employed an NPL method to identify semantic-based similarity between Quranic verses. They mapped these verses to numerical vectors encoding the semantic properties of the text. In another study, F. Beirade et al.[31], has developed a Quran semantic search engine using Quranic ontology. The semantic fields of words that present word meanings and their relationships in the holy Quran have been determined, and it is able to enrich queries for the Quranic ontology. S. Zouaoui et al. [32], presented AraFamOnto, an Arabic ontology-based inheritance calculation system. This application of ontology is crucial for storing knowledge about familial relationships, facilitating research, information processing, and accurate calculation of Islamic inheritance. Rostam et al.[33], suggested a technique for classifying some categories using text categorization. It can determine how different resources relate to one another. The study used several Islamic resource collections, such as the Quran and Hadiths, to replicate multiple relevant scenarios. The three classification algorithms (Support Vector Machine, Naïve Bayes, and K-Nearest Neighbour) with term weighting (TF-IDF) have been used to examine the three categories: Hajj, Prayer, and Zakat.

The existing ontologies for the Quranic study focus on specific domains like Quran stories, Food in the Quran, Miracles and natural science, names of God, health, time, nature, and also Quran ontology [34][35]. To the best of our knowledge, prior investigations have limited delved into the use of OL for Hajj and Umrah in the Context of the Quran. One application of Hajj ontology has been manually developed to locate verses that contain Hajj in Surah Al-Hajj [36]. Yet, the query just displays verses related to Hajj and not the other important information of the Hajj domain. Other than that, a brief of Hajj ontology has been developed for experimenting with Spatio-Temporal Database Modelling and not focusing on the Quran [37]. The modeling is used to assist huge crowds in Hajj events in such a way as to help and provide quality services. Another ontology of Hajj presents the hierarchical relationship between the categories that exist in the Hajj domain [38]. The ontology doesn't cover the Qur'an that relates to Hajj. Similar ritual to Hajj, Umrah is also a domain of study done by Sharef [39]. The ontology has been manually developed to study the semantic-based question-answering system. Pilgrims can post any question about Umrah in natural language format, then the ontology will provide specific answers to the query. Based on the study, it shows that the Hajj ontology can be extended by combining the general knowledge of Hajj, Umrah, and from the Qur'an that mentions the Hajj and Umrah. The developed ontology will facilitate more precise, contextual, and meaningful application of Qur'anic knowledge in areas ranging from education and research to AI applications. The ontology could be used by the Hajj planner application that could guide pilgrims

through the steps of Hajj based on their personal circumstances, helping them understand the rituals and their spiritual significance at each stage. In terms of AI systems, the developed ontology could assist in answering religious questions, providing context-aware advice, or even supporting legal interpretations with Verses.

### III. METHODOLOGY

This section outlines the establishment of three experiments to evaluate the performance of extracting ontological elements from Quranic texts. The three experiments are Experiment 1 (NER), Experiment 2 (Statistical-based) and Experiment 3 (Quranic pattern-based). Experiments 1 and 3 rely on Natural Language Processing based techniques, while Experiment 2 is based on statistical techniques. Primarily, the test collection involves Quran text translation from Hillali Khan [37]. The following Fig 1. shows the sample of data from Hillali Khan. To ensure accurate Part-of-Speech (POS) tagging and ontology extraction, the textual resources have been preprocessed to handle hyphenated terms appropriately. Specifically, terms such as "As-Safa" have been replaced with "AsSafa" to prevent incorrect tokenization and POS tagging.



“Verily! As-Safa and Al-Marwah (two mountains in Makkah) are of the Symbols of Allah. So it is not a sin on him who perform Hajj or 'Umrah (pilgrimage) of the House (the Ka'bah at Makkah) to perform the going (Tawaf) between them (As-Safa and Al-Marwah). And whoever does good voluntarily, then verily, Allah is All-Recogniser, All-Knower.”

Fig. 1. The input sample.

The input data is different according to experiments. In Experiment 1 and 2, the selected data input are chapters, i.e. Al-Maarij, Al-Muddathir, Al-Jinn, and Al-Muzammil, with a total of 148 verses and 2704 words. Meanwhile, in Experiment 3, the experiment uses 53 verses from the domain of Hajj and Umrah that are mentioned in the Quran. All three experiments must do pre-processing steps, such as the conversion from an Arabic word to an ordinary word, replacing certain capital pronouns that refer to Allah, and subsequently, the text was input into GATE before the application of OL techniques. The output of these experiments may include terms that represent concepts, relations, and instances. The experiments of these techniques are discussed in the subsequent subsection.

#### A. Experiment 1: Named Entity Recognition (NER) Technique

The NER technique identifies concepts in the translated Qur'an by recognizing uppercase letters and distinguishing specific nouns. This technique used the GATE tool to perform the extraction. Typical GATE's system consists of ANNIE processing resources that will go through a sentence splitter, tokenizer, POS Tagger, gazetteer, and JAPE transducer [22]. The JAPE transducer used the default named entities transducer in ANNIE with four predefined classes, i.e., Person, Organization, Location, and Unknown. It also excluded unrelated categories like Date and Money, which are not appropriate for Qur'an translation. In general, the mapping for concepts and instances is illustrated in Fig. 2. It will find the appropriate class mapping for the concepts and instances based on a predefined category. The outcome of this experiment

comprises the concepts and instances based on names that align with the predefined category.

```
Given: Predefined Category T= {T1, T2, Tk} and Concepts and  
Instances C= {C1, C2, Cn}  
Find a class K: C → T, namely, K(c)  
K(c), identifies the category of concepts and instances c for each c in  
C.  
For example,  
C= {Messenger, Prophet, Majesty, Lord, a raging Fire, a flaming Fire,  
Mecca, Kaabah, Satan, Jin} and  
T= {person, location, organization, unknown}
```

Fig. 2. Classification of concepts.

The NER technique involves two steps of evaluation. The first evaluation is based on the extracted names, while the second evaluation focuses on the classified extracted names. NER classifies the identified names based on predefined entity types or categories.

### B. Experiment 2: Statistical Measurement -tf, tfidf, Ridf of Hajj Documents

In contrast with the first experiment, the second experiment evaluates the performance of extracting single-term words using a statistical-based technique. For this experiment, a test collection of 53 verses in the Quran that related to pilgrimage, i.e., Hajj and Umrah, has been selected. The algorithm is depicted in Fig. 3.

---

#### Algorithm: Extraction of concepts and instances

---

```
Preprocessing Task  
Input corpus  
Sentence splitting, Tokenizing,  
for each token  
if (tokenize contain hyphen| punctuation symbol)  
Remove the hyphen and punctuation symbol  
Replace certain words with Allah  
end for  
//Find the frequency of terms using the Statistical method  
Create empty termArray; initialize countArray  
For each token T in Terms do  
Search for T in termArray  
If found  
Increment countArray[i];  
else  
Create new record  
termArray[j]=T;  
countArray[j]=1  
end if  
end for  
for each token T in termArray[i] do  
calculate each the frequency using statistical measurement  
end for  
End.
```

---

Fig. 3. The Algorithm for statistical techniques.

The statistical-based technique uses variants of measurement. Measurements, such as term frequency (tf), serve as simple yet significant statistical metrics for identifying concepts. The tf can also be normalized using inverse document frequency (idf) to produce the term frequency-inverse document

frequency (tfidf) method. Another variant of tf is Residual idf (Ridf). In particular, the following definitions apply to an extracted term from the Hajj verses: term frequency-inverse document frequency (tf-idf), term frequency (tf), and residual idf (Ridf).

$$tf(t,d) = ft,d \quad (1)$$

where t is the term, and ft,d is the frequency of term t in document d.

$$tf-idf(t,d,D) = tf(t,d) \times idf(t,D) \quad (2)$$

where idf(t,D) is given by:

$$idf(t,D) = \log \left( \frac{|D|}{df(t)} \right)$$

with:

- N = total number of documents in the corpus
- D = number of documents where the term t appears

$$-\log \left( \frac{|D|}{df(t)} \right) + \log \left( 1 - \exp \left( -\frac{tf(t,d)}{df(t)} \right) \right) \quad (3)$$

### C. Experiment 3: Qur'an Structure Pattern Based on Pattern [19]

The aim of this experiment is to identify relations whether they are taxonomy relations or non-taxonomy relations, present in the Qur'an text. It employs the existing structure of Quranic patterns proposed by Saad [19]. The chosen patterns have been previously applied in Quranic ontology learning research. It has been widely utilized for extracting semantic relations from the Quran due to the structured nature of its text and the recurrence of specific linguistic patterns. These patterns provide a linguistically informed approach to identifying relations between concepts and entities within the Quranic text. Furthermore, the patterns are inspired by Lexico-syntactic patterns by Hearst [18], a widely accepted technique for extracting taxonomic (hierarchical) relationships in computational linguistics. Hearst patterns have been successfully used in various OL tasks to automatically extract hyponym-hypernym (subclass-superclass) relationships from natural language text. However, the direct application of Hearst patterns to Quranic text presents challenges due to the unique linguistic characteristics of Quranic Arabic and the translated English versions. To overcome these challenges, the patterns were extended and modified to better suit the syntactic and semantic structure of Quranic translations.

The Quranic pattern is based on rules that came from NLP tagging for each word. For the extraction task, three patterns, as illustrated in Fig. 4, have been employed. Two considerations were considered based on the translation: 1) the formatting of the Quranic text structure, and 2) the linguistic patterns of the Quranic text. These considerations are crucial, as different translations may yield distinct outputs in terms of text structure, and patterns.

As mentioned earlier, the outcomes of this experiment are relations between concepts. Pattern 1 and Pattern 2 are used to extract taxonomy relations, specifically "part-of" relations, while Pattern 3 is used to extract non-taxonomy relations i.e. "definition" or "synonym" relations.

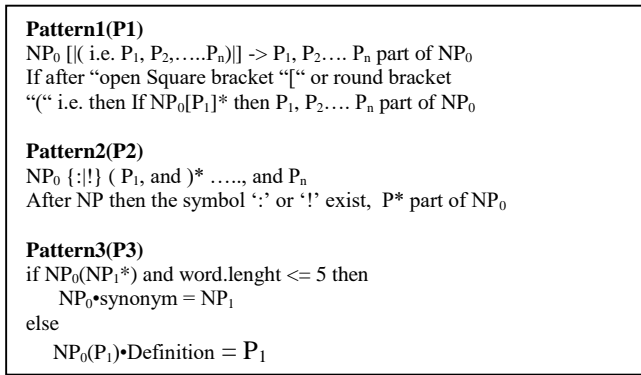


Fig. 4. The Qur’an structure pattern [19].

Each experiment in this study serves a distinct purpose in OL, contributing to the extraction of concepts, instances, and relations that form the foundation of the final ontology. By integrating the outputs from all three experiments, a structured and enriched ontology of the Quranic domain can be developed. The ontology could integrate the outputs from all three experiments to create a structured and enriched knowledge representation. Concepts and instances extracted from Experiments 1 and 2 could form the ontology classes and entities, ensuring comprehensive domain coverage. Additionally, Experiment 3 contributes hierarchical structures and definitions, refining the taxonomy and semantic clarity of the ontology. Together, these elements could help create a well-organized ontology that enhances knowledge retrieval and semantic interpretation in the Quranic domain.

#### IV. RESULTS

##### A. Result of Experiment 1: NER Technique

As mentioned earlier, the outcomes of this experiment are done using *precision*, *recall*, and *f-measure*, and the result of the extraction can be shown in Table I.

TABLE I. RESULT OF EXTRACTED CONCEPTS AND INSTANCES FOR CHAPTERS USING NER

	Al-Jin	Al-Muzammil	Al-Maarij	Al-Muddathir	Average
<b>Prec</b>	0.78	0.58	0.55	0.55	0.62
<b>Rec</b>	0.33	0.36	0.21	0.59	0.37
<b>F-M</b>	0.46	0.44	0.30	0.57	0.44

Meanwhile, the second classification evaluation reveals that only the Person and Unknown categories are suitable for entity selection as shown in Fig. 5.

##### B. Result of Experiment 2: Statistical Technique-tf, tfidf, Ridf

This experiment aims to identify the single-word terms using statistical measurements such as *tf*, *tfidf*, and *Ridf*. The results yielded 614 single terms out of a total of 3018 terms. The top 30 ranked terms are shown in Table II.

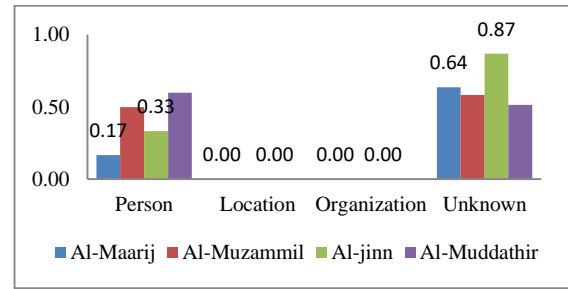


Fig. 5. The classification of extracted names.

TABLE II. SAMPLE OF SINGLE-WORD TERM CANDIDATES FROM 614 TERMS

No.	Statistical Measurement					
		<i>tf</i>		<i>Tf-idf</i>		<i>R-idf</i>
1	Allah	131	Hajj	11.96	having	0.59
2	Makkah	34	th	9.96	th	0.52
3	Hajj	23	Makkah	9.96	month	0.46
4	pilgrimage	14	Kabah	9.42	God	0.29
5	House	13	SAW	9.24	Ilah	0.29
..	..	..	..	..	..	..
..	..	..	..	..	..	..
30	Ihram	6	Verily	6.57	animals	0.28

The *tf* precision, recall, and F-measure have been measured with the performance shown in Table III.

TABLE III. THE PRECISION, RECALL AND F-MEASURE PERFORMANCE FOR TF

Precision	Recall	F-Measure
0.451	0.584	0.509

##### C. Result of Experiment 1: Qur’an Structure Pattern

This experiment focuses on the Quranic structure patterns [14]. These patterns utilize a combination of POS tagging and regex applied to the text to find “part of” relations, as well as “synonym” or “definition” relations present in the text. Table IV shows the result of extracted relations to show whether the extracted terms are correct or wrong using patterns.

TABLE IV. RESULT OF EXTRACTED RELATIONS USING QURANIC PATTERN

No.	Statistical Measurement				
	Al-Maarij Correct/Wrong	Nuh	Al-Muzammil	Al-Muddathir	Total correct
Pattern 1	3/0	3/0	2/1	2/0	10
Pattern 2	1/ 2	0/0	1/0	0/1	1
Pattern 3	6/3	3/4	6/6	9/7	24
ALL					35

Meanwhile, Table V shows the precision based on the patterns for each chapter.

TABLE V. PRECISION BASED ON QURANIC PATTERN

No.	Chapter	Precision	Average Precision
1	Al-Maarij	0.60	
2	Nuh	0.60	
3	Al-Muzammil	0.56	
4	Al-Muddathir	0.58	
			0.58

#### D. Performance of All Techniques Applied

Based on the results of running three different types of techniques, the precision and average precision of each method can be depicted in Table VI.

TABLE VI. PRECISION AND AVERAGE PRECISION OF NER, STATISTICAL-BASED AND QURANIC PATTERN STRUCTURE

No.	Technique	Average Precision
1	NER	0.62
2	Statistical	0.45
3	Pattern	0.58

## V. DISCUSSION

### A. Experiment 1: NER Technique

The Named Entity Recognition (NER) achieved an average F-Measure of 40%, with average precision above 60% and an average recall under 40%. This suggests that NER is useful for extracting relevant terms, but many relevant terms are missed. The chapter Al-Jinn had the highest precision at 0.78, while Al-Maarij had the lowest recall at 0.21. The NER shows that certain chapters miss more relevant terms where important terms like "fasting" and "water" or other significant words like "criminal" and "sinner" are not captured. This is due to the NER that may not recognize these words as entities because they aren't typical "named" entities, like places or people that we often see in everyday language. Additionally, phrases like "a weighty Word" or "the Fire of Hell" carry weight in a Quranic context. It refers to entities of "Quran" and "Hell" but this phrase is not commonly recognized as entities outside of it. These gaps suggest a need for additional NLP analysis within ANNIE module, particularly to capture more nouns and compound nouns. It was also found that the ANNIE module sometimes tagged parts of speech incorrectly, which led to extraction errors. For example, terms like "O" and "Verily" were wrongly tagged as nouns, which decreased the accuracy by misidentifying uppercase letters as meaningful nouns. To improve the recall, the NER model could train the model to Quranic texts and other religious literature to better understand contextually significant terms. The training on domain-specific text allows the model to build a specialized understanding of contextually important words and phrases, making it better at identifying them accurately within that subject area.

In Fig. 5, the Organization and Location categories show no correct classifications, with terms like "the angel" wrongly classified as an organization. The highest correct classification is for the Unknown category in Al-Jinn at 0.87, indicating more concepts like "the heaven" and "the gardens" need

identification. Al-Muddathir has the highest Person classification at 0.60. The term "House" was retrieved but misclassified as an organization; in the Qur'an, it refers to "Kaaba." Overall, the extraction process needs improvement to capture more relevant terms. In summary, even if imperfectly classified, the NER could identify terms relevant to the domain, underscoring its potential utility in extracting meaningful concepts and instances.

### B. Experiment 2: Statistical Technique-tf,tfidf, Ridf

Based on Table II, term frequency (tf) outperformed *tf-idf* and *Ridf* in extracting relevant terms as concepts or instances when considering the top 30 terms. Terms like "Allah," "Hajj," and "Makkah" were identified as more meaningful concepts compared to less significant terms like "th."

In Table III, the experiment can conclude that statistical measurements such as tf, tf-idf, and Ridf are effective for extracting single-word terms as concepts or instances. The tf performed better compared to tfidf and Ridf. It still only retrieves about 50% of relevant single-word terms, suggesting that it is not fully effective on its own. They fall short in identifying multi-word terms, which are prevalent in texts like the Qur'an. Many significant concepts are missed, which is particularly problematic for texts where important terms are often multi-word, like the Quran. Terms are in multi-word phrases, such as "Bounty of Allah," "raging Fire," "remember Allah," and "ways of Prophet Muhammad." On the other hand, the recall shows that 58% of the terms that were retrieved were relevant and might be considered as possible concepts or instances. Therefore, the lack of NER can be covered by statistical measurement to retrieve more relevant terms.

### C. Experiment 1: Qur'an Structure Pattern

Table IV shows that Pattern 1 and Pattern 2 can extract "part of" relations between concepts. Pattern 1 performs better at 28.57%, while Pattern 2 has only 2.86% accuracy. Pattern 2's low performance is due to incorrect POS tagging, where terms like "Verily" and "Nay" are both annotated as Proper Nouns (NNP), despite not being related.

Mostly, the exclamation mark in Pattern 2 is used at the end of a sentence and not at the middle sentence. In Pattern 1, the false extraction exists when it uses brackets to actually explain or elaborate more on the sentence. The extracted noun found is not "part of" relations for another noun. Pattern 3 outperformed Pattern1 and Pattern 2 with a 68.57% correct match rate, successfully extracting synonym and definition relations. However, only 54.55% of these extractions were accurate, with synonyms extracted at 81.81% but only 61.11% correct. Errors were due to incorrect POS tagging, such as misinterpreting bracketed terms, e.g., "garments (Prophet Muhammad SAW)" where "Prophet Muhammad SAW" is not a synonym for "garments."

The average precision is 0.58 in Table V reflects the limitations and challenges associated with the pattern-based approach for extracting taxonomy and non-taxonomy relations in Quranic text. Several factors contributed to this relatively moderate precision: The experiment found inconsistent use of rounded brackets "()" and square brackets "[]," which sometimes indicate synonym relation and sometimes can be a definition

relation or explanatory notes. Square brackets are often used when rounded brackets are present, as in "Messenger [Musa (Moses)]." Some words were misclassified during Part-of-Speech (POS) tagging, causing incorrect identification of relations and synonyms.

The observation from this experiment shows that the formatting of the Quran text structure and the patterns of the Quran text style can be used to extract ontological elements. But it needs to be further refinement to improve accuracy, particularly in handling text variations, multi-word terms, and contextual relationships. In fact, based on the results of running three different patterns, the average precision of 0.58 shows that half of the concepts or instances are not yet retrieved.

#### D. All Techniques

Table VI shows that the precision based on the three methods is still low, with only around 50% of concepts, instances, or relations being retrieved from the Qur'an Text. It means only half of the terms and relations can be retrieved using the techniques. The improvement is still needed to retrieve more relevant terms. On the other hand, the techniques are able to identify terms relevant to the domain, even if incorrectly classified, highlighting its potential utility in extracting meaningful concepts and instances.

Each OL technique demonstrated strengths and limitations in different scenarios. The NER approach achieved higher precision due to its reliance on predefined entity categories, ensuring accurate identification of named concepts. However, its lower recall indicates that many relevant terms were missed, particularly non-named entities. In contrast, the statistical approach effectively identified frequent terms, expanding the concept pool beyond named entities, but it struggled with multi-word expressions, leading to incomplete representations of certain Quranic terms. The pattern-based approach, while useful for extracting taxonomy and semantic relations, was limited by variations in text formatting and syntactic inconsistencies, affecting its accuracy. By integrating these techniques, the final ontology balances precision, recall, and relational depth, improving the overall quality of extracted knowledge.

This study advances OL by tailoring approaches specifically to the unique characteristics and challenges of the Qur'an, in ways that general, non-religious (ordinary text) OL research may overlook. By addressing the Quran's complex language, thematic concepts, and context-sensitive relationships, it provides a more comprehensive and accurate ontology model compared to standard OL approaches. This domain-specific refinement allows for a deeper, more authentic representation of religious knowledge, particularly in areas like theology, ritual, and ethics. In contrast, ordinary text deals with straightforward, clear language making it easier to identify entities and relationships.

## VI. CONCLUSION

In conclusion, the analysis of the table reveals that the precision levels achieved through the three methods for retrieving concepts, instances, or relations from the Qur'an Text remain comparatively low, hovering around 50%. This indicates that only half of the terms can be successfully retrieved using the

employed techniques. The findings underscore the necessity for further improvements in the existing methods to enhance precision and broaden the scope of relevant term retrieval.

Future research on ontology learning using Large Language Models (LLMs) for Quranic text will focus on refining semantic extraction methods, improving multilingual capabilities, and enhancing domain-specific training. Given the complexity of Quranic language and its deep semantic structures, fine-tuning LLMs such as AraBERT or GPT-based models on Quranic corpora will be essential to capture intricate relationships between concepts [7]. One potential approach involves integrating structured datasets with LLM-generated embeddings to improve the contextual accuracy of Ontology Learning[40]. Another promising direction is leveraging Retrieval-Augmented Generation (RAG) frameworks to enhance the extraction of non-taxonomic relationships, allowing for a deeper understanding of Quranic themes and their interconnections[9]. These advancements will contribute to more sophisticated, AI-driven Quranic knowledge representation, benefiting applications in education, comparative religious studies, and digital humanities.

## ACKNOWLEDGMENT

This project is funded partially by the Centre for Research Excellence, Incubation Management Centre (CREIM), UniSZA.

## REFERENCES

- [1] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," *Int. J. Hum. - Comput. Stud.*, vol. 43, no. 5-6, pp. 907-928, 1995.
- [2] A. Mirarab, F. S. T. Amiri, S. Dehghanianij, and N. HosseinKhalili, "Development of Qur'anic Ontologies: A Domain Review Study," *Int. J. Inf. Sci. Manag.*, vol. 21, no. 3, pp. 229-241, 2023.
- [3] F. S. Utomo, N. Suryana, and M. S. Azmi, "Question Answering Systems on Holy Quran: A Review of Existing Frameworks, Approaches, Algorithms and Research Issues," *J. Phys. Conf. Ser.*, vol. 1501, no. 1, 2020, doi: 10.1088/1742-6596/1501/1/012022.
- [4] A. Maedche and S. Staab, "Ontology Learning for the Semantic Web," *IEEE Intell. Syst.*, vol. 16, no. 2, pp. 72-79, 2001, doi: 10.1109/5254.920602.
- [5] R. Du, H. An, K. Wang, and W. Liu, "A Short Review for Ontology Learning from Text: Stride from Shallow Learning, Deep Learning to Large Language Models Trend," *arXiv Prepr. arXiv2404.14991*, 2024, [Online]. Available: <http://arxiv.org/abs/2404.14991>
- [6] R. I. Ahmed, M. H. Sayed, and T. M. Wahbi, "Quran Ontology: Review on Recent Research Issues," *Researchgate.Net*, vol. 11, no. 12, pp. 189-197, 2022, doi: 10.21275/SR221201170653.
- [7] M. M. Taye, R. Abulail, and M. Al-Oudat, "An Ontology Learning Framework for unstructured Arabic Text," in *ISAS 2023 - 7th International Symposium on Innovative Approaches in Smart Technologies, Proceedings*, 2023, pp. 1-12. doi: 10.1109/ISAS60782.2023.10391548.
- [8] R. Y. Al-Salhi and A. M. Abdullah, "Building Quranic stories ontology using MappingMaster domain-specific language," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 1, pp. 684-693, 2022, doi: 10.11591/ijece.v12i1.pp684-693.
- [9] M. Sanaei, F. Azizi, and H. B. Giglou, "Phoenixes at LLMs4OL 2024 Tasks A , B , and C : Retrieval Augmented Generation for Ontology Learning," in *Open ConfProc 4 (2024) "LLMs4OL 2024: The 1st Large Language Models for Ontology Learning Challenge at the 23rd ISWC"*, 2024, pp. 39-47.
- [10] T. Zengeya and J. Vincent Fonou-Dombeu, "A Review of State of the Art Deep Learning Models for Ontology Construction," *IEEE Access*, vol. 12, no. April, pp. 82354-82383, 2024, doi: 10.1109/ACCESS.2024.3406426.



- [11] G. Li, C. Tang, L. Chen, D. Deguchi, T. Yamashita, and A. Shimada, "LLM-Driven Ontology Learning to Augment Student Performance Analysis in Higher Education BT - Knowledge Science, Engineering and Management," C. Cao, H. Chen, L. Zhao, J. Arshad, T. Asyhari, and Y. Wang, Eds., Singapore: Springer Nature Singapore, 2024, pp. 57–68.
- [12] M. Yin, L. Tang, C. Webster, X. Yi, H. Ying, and Y. Wen, "A deep natural language processing-based method for ontology learning of project-specific properties from building information models," *Comput. Civ. Infrastruct. Eng.*, vol. 39, no. 1, pp. 20–45, 2024, doi: 10.1111/mice.13013.
- [13] A. Balali, M. Asadpour, and S. H. Jafari, "Cofee: A Comprehensive Ontology for Event Extraction from Text," *SSRN Electron. J.*, 2022, doi: 10.2139/ssrn.4117538.
- [14] A. B. Kamran, B. Abro, and A. Basharat, "SemanticHadith: An ontology-driven knowledge graph for the hadith corpus," *J. Web Semant.*, vol. 78, 2023, doi: 10.1016/j.websem.2023.100797.
- [15] Y. M. Saber, H. Abdel-Galil, and M. A. El-Fatah Belal, "Arabic ontology extraction model from unstructured text," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 8, Part B, pp. 6066–6076, 2022, doi: <https://doi.org/10.1016/j.jksuci.2022.02.007>.
- [16] J. Jung, W. Zhou, and A. D. Smith, "From Textual Data to Theoretical Insights: Introducing and Applying the Word-Text-Topic Extraction Approach," *Organ. Res. Methods*, Jan. 2024, doi: 10.1177/10944281241228186.
- [17] J. Hua, L. Luo, W. Ping, Y. Liao, and C. Tao, "Rules still work for Open Information Extraction," *ArXiv: 2403.10758*, pp. 1–29, 2024, [Online]. Available: <https://arxiv.org/abs/2403.10758>
- [18] M. A. Heart, "Automatic Acquisition of Hyponyms from Large Text Corpora Lexico-Syntactic for Hyponymy Patterns," *Proc. 14th Int. Conf. Comput. Linguist.*, vol. 2, pp. 539–545, 1992.
- [19] S. Saad, "Ontology Learning and Population Techniques for English Extended Quranic Translation Text (Doctoral dissertation)," *Universiti Teknologi Malaysia, Skudai, Malaysia*, 2013.
- [20] V. T. Phi, H. Teranishi, Y. Matsumoto, H. Oka, and M. Ishii, "PolyNERE: A Novel Ontology and Corpus for Named Entity Recognition and Relation Extraction in Polymer Science Domain," 2024 *Jt. Int. Conf. Comput. Linguist. Lang. Resour. Eval. Lr. 2024 - Main Conf. Proc.*, pp. 12856–12866, 2024.
- [21] K. Detroja, C. K. Bhensdadia, and B. S. Bhatt, "A survey on Relation Extraction," *Intell. Syst. with Appl.*, vol. 19, no. June, p. 200244, 2023, doi: 10.1016/j.iswa.2023.200244.
- [22] D. Thakker, T. Osman, and P. Lakin, "GATE JAPE Grammar Tutorial (Version 1.0)." Accessed: Jan. 29, 2024. [Online]. Available: [http://gate.ac.uk/sale/thakker-jape-tutorial/GATE\\_JAPE\\_manual.pdf](http://gate.ac.uk/sale/thakker-jape-tutorial/GATE_JAPE_manual.pdf)
- [23] Kais Dukes, "Leed University." Accessed: Nov. 20, 2023. [Online]. Available: <https://corpus.quran.com/concept.jsp?id=hajj>
- [24] A. C. Khadir, H. Aliane, and A. Guessoum, "Ontology learning: Grand tour and challenges," *Comput. Sci. Rev.*, vol. 39, p. 100339, 2021, doi: 10.1016/j.cosrev.2020.100339.
- [25] V. Lytvyn, Y. Burov, V. Vysotska, and O. Brodyak, "Approach to Automatic Construction of Interpretation Functions during Ontology Learning," *Int. Sci. Tech. Conf. Comput. Sci. Inf. Technol.*, vol. 1, pp. 267–271, 2020, doi: 10.1109/CSIT49958.2020.9321920.
- [26] M. Jackermeier, J. Chen, and I. Horrocks, Dual Box Embeddings for the Description Logic EL++, vol. 1, no. 1. Association for Computing Machinery, 2024. doi: 10.1145/3589334.3645648.
- [27] M. Schaeffer, M. Sesboué, J. P. Kotowicz, N. Delestre, and C. Zanni-Merk, "OLAF: An Ontology Learning Applied Framework," *Procedia Comput. Sci.*, vol. 225, pp. 2106–2115, 2023, doi: 10.1016/j.procs.2023.10.201.
- [28] P. Cimiano and J. Völker, "Text2Onto: A Framework for Ontology Learning and Data-Driven Change Discovery," *Nat. Lang. Process. Inf. Syst.*, pp. 227–238, 2005, doi: 10.1007/11428817\_21.
- [29] H. Dong, J. Chen, Y. He, Y. Gao, and I. Horrocks, "A Language Model Based Framework for New Concept Placement in Ontologies," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 14664 LNCS, pp. 79–99, 2024, doi: 10.1007/978-3-031-60626-7\_5.
- [30] M. Alshammeri, E. Atwell, and M. ammar Alsalka, "Detecting Semantic-based Similarity Between Verses of The Quran with Doc2vec," *Procedia Comput. Sci.*, vol. 189, pp. 351–358, 2021, doi: <https://doi.org/10.1016/j.procs.2021.05.104>.
- [31] F. Beirade, H. Azzoune, and D. E. Zegour, "Semantic query for Quranic ontology," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 33, no. 6, pp. 753–760, 2021, doi: 10.1016/j.jksuci.2019.04.005.
- [32] S. Zouaoui and K. Rezeg, "A Novel Quranic Search Engine Using an Ontology-Based Semantic Indexing," *Arab. J. Sci. Eng.*, vol. 46, no. 4, pp. 3653–3674, 2021, doi: 10.1007/s13369-020-05082-5.
- [33] N. A. P. Rostam and N. H. A. H. Malim, "Text categorisation in Quran and Hadith: Overcoming the interrelation challenges using machine learning and term weighting," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 33, no. 6, pp. 658–667, 2021, doi: <https://doi.org/10.1016/j.jksuci.2019.03.007>.
- [34] A. Mirarab, F. Sadat Tabatabai Amiri, and S. Dehghanisanij, "Quranic Ontologies: A Scoping Review of the Applications," *Libr. Inf. Sci.*, vol. 26, no. 1, 2023, [Online]. Available: [https://lis.aqr-libjournal.ir/article\\_166718.html%0Ahttps://lis.aqr-libjournal.ir/article\\_166718\\_ffe55a4d0814fd30bd2254e145cbe5ec.pdf](https://lis.aqr-libjournal.ir/article_166718.html%0Ahttps://lis.aqr-libjournal.ir/article_166718_ffe55a4d0814fd30bd2254e145cbe5ec.pdf)
- [35] R. Ahmad, F. Z. Khan, and M. A. Khan, "Ontology Based Knowledge Retrieval and Semantic Modelling of Qur'an with Contextual Information," *Int. J. Islam. Appl. Comput. Sci. Technol.*, vol. 9, no. 1, pp. 10–25, 2021.
- [36] S. . D. Nawal Masoud, "Ontology Application For The Hajj," *University Utara Malaysia*, 2009.
- [37] K. Rizwan, N. Mahmood, A. Nadeem, and A. M. G. A. Lzahrani, "Spatio-Temporal Database Modeling And Applications For Assistance Of Huge Spatio-Temporal Database Modeling And Applications For Assistance Of Huge Crowd In Hajj," *J. Eng. Sci. Comput.*, vol. I, no. May, 2019.
- [38] Youssef, Fatima Y. and Z. I. Othman, "The Hierarchical Classification for The Rituals of Hajj Using Ontology," *J. Qadisiyah Comput. Sci. Math.*, vol. Vol. 15, no. Issue 1, p. p1–13. 13p., 2023.
- [39] N. M. Sharef, M. A. Murad, A. Mustapha, and S. Shishechi, "Semantic question answering of umrah pilgrims to enable self-guided education," *Int. Conf. Intell. Syst. Des. Appl. ISDA*, pp. 141–146, 2014, doi: 10.1109/ISDA.2013.6920724.
- [40] I. M. AlAgha and M. G. Al-Masri, "An Ontology Based Approach to Enhance Information Retrieval from Al-Shamelah Digital Librar," *IUG J. Nat. Eng. Stud. Peer-reviewed J. Islam. Univ. ISSN*, vol. 24, no. 1, pp. 39–53, 2016.