International Journal of Computers and Informatics

Journal Homepage: https://www.ijci.zu.edu.eg

Int. j. Comp. Info. Vol. 7 (2025) 79–90

Paper Type: Original Article

Digital Quran Computing Algorithms and Applications

Amro Badawy [1], Ahmed Salah [1,2,*] and Mahmoud Mahdy [1]

[1] Department of Computer Science, Faculty of Computer and Informatics, Zagazig University, Zagazig 44519, Egypt.
[2] College of Computing and Information Sciences, University of Technology and Applied Sciences, Ibri, Sultanate of Oman.
Emails: microtouch_group@yahoo.com; ahmad.salah@utas.edu.om, ahmad@zu.edu.eg; mamahdi@zu.edu.eg.

Received: 31 Jan 2025 **Revised:** 28 Mar 2025 **Accepted:** 20 Jun 2025 **Published:** 23 Jun 2025

Abstract

The computational analysis of the Quran presents unique challenges due to its linguistic complexity, thematic richness, and cultural significance. This paper explores advanced text mining techniques to uncover thematic structures in three Surahs: Al-Kahf, An-Naml, and Al-Baqarah. Two primary methodologies are employed: (1) Latent Dirichlet Allocation (LDA) for topic modeling, applied to Surahs Al-Kahf and An-Naml to extract latent themes such as faith, morality, and divine guidance, validated by Quranic scholars; and (2) a TF-IDF/UMAP/K-means clustering pipeline for Surah Al-Baqarah, which identifies semantically coherent thematic groups (e.g., "Divine Law and Guidance," "Faith and Belief") through dimensionality reduction and unsupervised learning. Key findings demonstrate the efficacy of these methods in bridging traditional exegesis (tafsir) with data-driven approaches, revealing nuanced thematic interconnections and validating known Quranic structures quantitatively (e.g., 10-cluster solution for Al-Baqarah with a silhouette score of 0.21). The paper contributes to digital Quranic studies by providing reproducible frameworks for thematic analysis, addressing challenges in Arabic text preprocessing, and highlighting the potential of hybrid computational-hermeneutic methodologies. This work advances the interdisciplinary dialogue between Islamic scholarship and modern NLP, offering tools for scalable, objective analysis while preserving theological nuance. The study leverages LDA topic modeling to analyze Surahs Al-Kahf and An-Naml, identifying five key themes per Surah, such as "Stories of Prophets" and "Divine Parables," which align with classical exegetical classifications. Preprocessing steps—including diacritic removal, stop-word filtering, and tokenization—were tailored to Quranic Arabic, ensuring linguistic fidelity. Expert validation confirmed the theological coherence of extracted topics, demonstrating LDA's utility in augmenting manual tafsir with scalable, data-driven insights.

Keywords: Quran; Authentication; Classification; Topic Analysis; Quranic Text Mining; LDA Topic Modeling; TF-IDF; UMAP; K-means Clustering; Thematic Analysis; Digital Humanities.

1 | Introduction

1.1 Background: The Quran in the Digital Age

In addition to its spiritual instruction, the Holy Quran, the main religious literature of Islam, is admired for its unparalleled linguistic, thematic, and structural richness. Because of its complex rhetorical tactics, polysemous language, and nuanced meanings, the Quran—written in classical Arabic—is very challenging to interpret. Systematic thematic analysis is difficult because of its non-linear structure, which sometimes incorporates multiple themes—such as history, ethics, theology, and law—into a single chapter or verse. Quranic interpretation has traditionally been founded on tafsir, an exegetical discipline grounded in a wealth

Corresponding Author: ahmad@zu.edu.eg

Licensee International Journal of Computers and Informatics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

of linguistic, theological, and historical knowledge. These traditional methods are subjective, time-consuming, and difficult to scale, despite their enormous usefulness. The exponential increase of digital texts and advances in computational linguistics have made it possible to analyze religious texts on a large scale. Researchers can use techniques like machine learning, text mining, and natural language processing (NLP) to explore Quranic themes in scalable and data-driven approaches. These techniques can reveal trends that manual analysis would miss. These methods promise objectivity, reproducibility, and the ability to supplement traditional exegesis with quantitative insights. [1] .

1.2 | Problem Statement

Despite the potential of computational tools, NLP techniques provide unique challenges when analyzing the Quran. Classical Arabic is rich in morphology and semantics, with many terms having many contextual meanings yet having a same origin. Furthermore, the Quran's use of diacritical marks, extensive intertextual references, and spiritual emphasis make it difficult to apply normal computational approaches without careful customization. To ensure the preservation of theological context and interpretive integrity, collaboration with subject-matter specialists is required.

This work tackles the primary issue in digital Quranic studies: How can the Quran's subject structure be identified and examined while respecting its linguistic and theological complexity using unsupervised machine learning techniques, specifically topic modeling and semantic clustering ? [2]

1.3 Computational Approaches to Religious Texts

Large-scale literary, historical, and religious text analysis is made possible by the study of digital humanities, which combines computational techniques with humanities disciplines. Religious experts may now examine holy writings including the Bible, Buddhist texts, and the Quran structurally, thematically, and emotionally thanks to computational analysis. These methods improve interpretive depth while facilitating scalability and reproducibility across vast corpora. Within Islamic studies, computational Quranic analysis has become a more popular area of study. Numerous problems, including topic modeling, information retrieval, semantic categorization, and Quranic authentication, have been investigated by researchers. The Quran's language diversity, theological relevance, and evolving digital presence have sparked research on themes, verse classification, intertextual links, and authenticity verification across a variety of digital media. [3].

1.4 Natural Language Processing (NLP) Fundamentals

Making it possible for computers to understand and process human languages is the aim of the natural language processing (NLP) subfield of artificial intelligence and linguistics. It is essential to textual analysis because it makes it possible for machines to extract, classify, and interpret textual data using statistical models and algorithms.

However, Arabic natural language processing has numerous challenges. Arabic contains several derivative forms that come from its roots, giving it a complicated morphology. The distinctions between Classical Arabic (used in the Quran), regional dialects, and Modern Standard Arabic further complicate model building. Due to its spiritual background, polysemy, historical orthography, and rhetorical tactics, Quranic Arabic presents other difficulties. Compared to English, Arabic NLP presents a unique challenge for high-quality computational analysis since it lacks annotated corpora, pretrained models, and standardized tools [4] .

1.5 Text Preprocessing Techniques

Effective text preparation is a solid foundation for computer text analysis, especially when dealing with linguistically complicated and religious texts like the Quran. Preprocessing standardizes textual representations, reduces noise, and prepares data for feature extraction and modeling .

The procedure often begins with tokenization, which separates the material into distinct words or important elements. Normalization often follows in the form of lemmatization and stemming, which reduce words to their base or root form. However, these approaches are constrained by Arabic's root-pattern morphology. Excessive stemming of the Quran can remove religiously significant affixes or root patterns. [5]. Eliminating stopwords is a crucial next step. Arabic stopwords can be found in generic lists, but lists tailored to the Quran must be carefully curated to avoid omitting terms that are important both contextually and theologically. For Quranic analysis, specifically designed stopword lists can significantly improve model performance .

Other important steps are diacritical removal to reduce orthographic variation and punctuation filtering (which helps separate important tokens). These steps are especially necessary for pipelines that use semantic clustering and LDA topic modeling. [6].

1.6 Feature Extraction and Representation

Because machine learning algorithms require numerical input, text must be transformed into quantitative representations. In this case, feature extraction methods such as TF-IDF are helpful. Term Frequency–Inverse text Frequency (TF-IDF) is a widely used vectorization technique that balances a term's frequency within a text with its frequency across the corpus. More weight is given to terms that are important in one text but rare in another. This facilitates the differentiation of semantically rich terms from common or uninformative ones. TF-IDF serves as the basis for both dimensionality reduction and cluster analysis [7].

1.7 Unsupervised Learning for Text Analysis

1.7.1 Topic Modeling

In a text corpus, a class of techniques called "topic modeling" locates hidden semantic patterns, or themes. These models aim to uncover latent patterns that explain the observed word distributions across different manuscripts .

Latent Dirichlet Allocation (LDA) is a probabilistic generative model that assumes that each document is a collection of themes, each of which is a distribution across words. The program estimates the likelihood of themes in each document by learning from word co-occurrence patterns.

Researchers evaluate the quality of topic models using coherence and perplexity scores. Perplexity assesses model fit, while coherence measures issue interpretability. Coherence in religious texts is sometimes supplemented by expert validation, which is used to ensure the theological consistency of the retrieved issues [8].

1.7.2 Dimensionality Reduction

The high-dimensional spaces in which text data is typically represented lead to the so-called curse of dimensionality, where noise and sparsity hinder model performance and visualization. Dimensionality reduction simplifies these representations while preserving their meaningful structure.

Singular Value Decomposition (SVD), particularly in its truncated form, is one matrix factorization technique utilized in Latent Semantic Analysis (LSA). It reduces dimensionality by approximating the original term-document matrix with fewer latent components. This thesis states that SVD groups Quranic verses before UMAP does.

A non-linear method called Uniform Manifold Approximation and Projection (UMAP) preserves both local and global structures when projecting high-dimensional data into lower dimensions. When it comes to displaying semantic clusters, UMAP is particularly useful.UMAP is quicker and more effective at preserving the topological integrity of clusters than alternatives like t-SNE.[9]

1.7.3 Clustering

Clustering is an unsupervised learning technique that groups related data points based on feature similarity. It is commonly used in exploratory data analysis and thematic segmentation .

The K-means method iteratively separates the data into "k" clusters by minimizing intra-cluster distance. Prior to convergence, cluster assignment and centroid updates alternate after random initialization. One of the primary disadvantages of K-means is the need to predetermine the number of clusters.

Several measures are used to evaluate cluster quality:

- The Silhouette Score measures separation and cohesion, or how near points are to one another in a cluster relative to other clusters.
- The Calinski-Harabasz Index measures the proportion of within-cluster dispersion to between-cluster dispersion .
- The average similarity between each cluster and its most similar counterpart is assessed by the Davies-Bouldin Index [10]. These metrics were used to identify the best clustering

2 | Related Work in Computational Quranic Analysis

Numerous research that have employed topic modeling, often with LDA on translations into Arabic or English, have focused on the Quran. These studies look for recurrent themes in the Surahs, with some focusing on theological, moral, and historical categories. However, the Quran's complex structure has often resulted in low coherence scores and frequent topic overlaps.

Clustering techniques have also been studied. While some employ paragraph vectors, others use hybrid clustering approaches to group verses based on semantic similarities. This demonstrates how a UMAP-K-means pipeline can be effectively used to separate themes like "Faith and Belief" or "Economic Justice." Despite promising results, formal clustering evaluation measures have been used in few studies.Beyond topic modeling and clustering, computational works have investigated semantic search, authenticity detection, and sentiment analysis within Quranic corpora. However, few efforts combine multiple NLP techniques into a unified framework or evaluate them comparatively using a single large Surah.

The Quran is the holy book of Muslims. The text of the Quran is the source of Islamic knowledge. We had to deal with two points as a result. First of all, the Quran is considered sacred by Muslims; as such, its words cannot be altered or changed since they are divine decrees. Second, every Muslim uses the Quran in their daily activities. Although Muslims believe that the Quran is not distorted, the enormous proliferation of digital media on the internet has created new difficulties in dealing with potentially unauthenticated Quranic text or materials, a condition that is unacceptable to Muslims. Therefore, it is imperative that this issue be addressed. [11]. Additionally, Muslims have a strong need for access to the Quran through gadgets like computers and mobile phones because they use them for regular tasks all day long.

Finding the authentic source of the Quran or determining whether or not its content has been altered is one of the most crucial responsibilities in the digital Quran. The subject of the Quran is handled using three topics. The primary Sharia body of Muslims is responsible for monitoring and assessing the Quran's veracity. By creating and assessing an authentication system for the Quran.[12]

Classifying the Quran is the second major effort. Developing methods for using the semantic search strategy to look up information in the Quran is the task at hand. The Quran's ideas are arranged and classified based on a certain subject .[13]

The third duty, topic analysis, focuses on creating methods for locating and examining particular content found in the Holy Quran [14]. It is possible to manage data and information online by designing and evaluating an application and method to analyze the language's validity in the electronic edition of the Quran, as well as by managing meta data associated with each word. There is no way for readers to determine if a verse is authentic. When a passage has intentional or inadvertent misspellings, it might be challenging to confirm its authenticity. The authors provided a framework for methodically identifying and classifying suitable methods for preserving the content integrity of the Digital Holy Quran .

The internet and online media are growing in popularity. Because of this, more people are studying the Quran, and Quranic passages, scripts, translations, and other Quranic disciplines are becoming accessible through digital media.

A classification model is used to categorize the words found in online content. tries out a number of feature categories. Machine learning and optimization strategies are used to develop a prototype with a higher level of assessment measures [15]. The Quran serves as the primary resource which is rich in patterns, themes, and facts that Muslims rely on to create their flawless pure knowledge. When studying a literary work such as the Quran, we should look for techniques that move beyond word level to sentence level representation. To extract the inferred linkages, deep semantic analysis and domain expertise are needed, which means learning strategies that go beyond word level representation to achieve sentence level representation. This task is done with the help of a DL model provided and proposed by [16].

2.1 Quran Authentication

With the goal of using it as a tool or method to improve digital Quran publication laws, a Quran authentication system is being developed and tested to assist core and end users in determining the legitimacy of digital-Quran applications. The goal is to create a novel and precise digital Quran authentication system that can validate the Quran both linguistically and semantically. To mak sure that the use of unique refernces of references is genuine and accepted by Islamic specialities, accuracy is necessary.[17]

A popular remedy for copyright and integrity problems in digital content is watermarking. The authors overcame a new sophisticated approach to digital content security, as a watermarking technique, that is invisible and fragile, with the intention of preserving the contents and authenticity of files that contain the Holy Quran in digital format, mainly PDF. The proposed method, which considers two computational methods, applies the Discrete Cosine Transform (DCT) method to the digital image data of the file for the process of feature extraction. The extracted feature data, captures the key visual properties of the original image. Secondly, the GEAR hash function uses this extracted information to be used as a means of reliably identifying whether the file has been altered from its original state. The watermarking working procedure does not involve simply hashing the document image, but involved implementing a hashing function, GEAR hash, to generate the corresponding hash values specifically from the image extracted features obtained from the DCT method. This is significant because only hashing the features substantially reduces the time and efficiency that is generally used for these types of procedures, which in turn leads to a faster and overall more efficient watermarking system to achieve the object of protecting the sacred text effectively. There is minimal color distortion since the watermark is inserted using SLSB techniques.[18]

For electronic Quranic verses, integrity verification techniques are offered, with the cryptographic hash function's main responsibility being to confirm the transmitted data's integrity. One method uses cryptographic hash algorithms to generate the hash table for the Holy Quran. The hash algorithms SHA256 and RIPEMD160 were selected, although a single compression process that modifies data in real time was the alternative

The compression technique just mentioned in the previous section addresses the specific needs of Arabic character sets by using a two-byte per character scheme and the popular Unicode UTF8 encoding standard. The generated results from the use of the compression method are impactful, and the savings are readily apparent. For example, the results show that when a digital copy of the Holy Quran is encoded, compressed, and stored using the Unicode UTF8 encoding standard, the generated hash table size drops dramatically compared to other methods. In fact, the results are significant, yielding a 84.73 percent and 90.46 percent hash table size reduction, as shown quantitatively. To address the compression aspect of the standard, the reduction of hash table size equates and leads to lost digital space of an equivalent amount, discovering stored space reductions of 10.48 and 5.55 respectively, as noted in section [19]. It is important to highlight that there is a salient requirement and inherent feature of the proposed method, which is the complete fidelity of the original content of the sacred Quran. This means keeping every single verse, phrase, and the overall organization of the Quranic text in its entirety, consisting of 114 complete chapters or Suras. From the foundation of data image fidelity, the author was able to create and test a total system and process. The goal of the overall system is to create a secure and automated way to check that the wording is accurate in the different e-versions of the Quran as it pertains to an unambiguous encoding, while generating specific metadata for every single word in the entire Quran. The metadata is designed to keep track of important characteristics of each word within the text, not the least of which is how many times each word appears in the Quran and the position in each verse, and chapter. This system to check the fidelity of wording, is a reliable mechanism based on the unique properties of cryptographic hash algorithms that are used universally in digital security for their unique property of reliably checking the integrity of stored digital data files [20]. The main property or characteristic of these hash functions is that they are very volatile, even the smallest change or modification to the input into the hash function leads to a completely different hash output. In addition to this because the output is not tied to the input, the hash function can be treated like an indicator that has been modified.

Moving to another aspect of digital security for image data, the fragile watermarking technique presented earlier uses a key signal processing technique, the discrete wavelet transform (DWT). The DWT is used to convert the original input digital image picture data into the wavelet domain. This transformation process works block-wise, where coefficients are created in wavelet domain transfer functions, commonly reducing coefficient matrices into a smaller block domain. It works pixel-wise in a spatial domain because it is trying to tie back to the structure of the original image. Conversion in to the wavelet domain is critical to the fragile watermarking because the watermark can be embedded in certain frequency bands that are more difficult for the naked human eye to see, while sensitive to modification.

In order to further watermark the authentication bits, the wavelet-transformed image's coefficients matrix is subsequently divided into many blocks using a block-wise approach .

The authentication bits are included into a block of 2×n components by substituting +1 for only one of the original elements and flipping the altered coefficients matrix back to the spatial domain. The difference between the inverted and altered coefficients' matrix is recorded on the pixel when one insignificant bit is changed.

The public-key cryptography is used to encrypt the authentication bits. The authentication bits are calculated using a well-known hash algorithm from the input picture [21]. There have been attempts to alter the contents of the Quran as cyber security has advanced. This article's goal is to methodically find and classify workable methods for preserving the content integrity of the Digital Quran. The focus has been on methods that are solely appropriate for text and image formats.

A native Arabic speaker may accurately read Arabic text even without the use of diacritical marks. Because diacritical markings are used to signify a variety of vowel sounds, a single phrase may have multiple interpretations. Methods for content inspection and protection were divided into different categories.[22] The authentication procedure is divided into security and verification steps. The verification component uses the Boyer-Moore approach, a well-known and widely used exact matching algorithm. Watermarking will be utilized to protect the authorized and tested verse during the security phase. Additionally, only the verification stage of the suggested framework has been tested because the system is currently under development. UTF-16 encoding is used for clitics segmentation in order to handle diacritical characters in Quranic text. Results from the verification phase's initial prototype were promising, reaching up to 98.6%.[23]

The literature on digital multimedia content has identified a number of security issues that require attention, such as content-originality verification, proof-of-content-authenticity prevention, and digital copyright protection. Similar to the digital content of the Holy Quran, such needs are obviously more common in the case of specialized and religious knowledge .

The challenge is to provide secure, reliable, and robust storage and delivery for the four categories of digital multimedia content that may be found on the Internet: text, images, audio, and video [24]. The authors offer a framework for digital Quran certification that uses contemporary digital authentication and certification techniques to certify and authenticate multimedia Quran apps in digital format .

A religious panel and devised processes examine the certification process, and if all requirements are satisfied, the digital Quran application is granted a digital certificate. A typical user of the service can readily verify the provided digital certificate online. To safeguard customers' faith and confidence, the suggested structure aims to lessen the likelihood of modifying digital content of the holy Quran.[25]

A delicate watermarking technique is employed for the digital Quran images' authentication and verification process. The suggested method detects the tampered pixels using the discrete wavelet transform (DWT) if there is proof of detrimental manipulation. To ensure high security, the wavelet coefficients are considered while embedding an authentication code that is encrypted using a secret key. The authentication binary code is included in a block of wavelet coefficients. The trials' findings indicate that, with a small quantity of watermark payload, the suggested method may be able to detect tampering and preserve image quality after watermarking.[26]

Improving the precision and accuracy of text detection is the biggest challenge facing Quran security and authentication. Analyzing and classifying recent studies that are pertinent to maintaining and confirming the Quran's content integrity is also essential. The current study is structured according to technique and format, including the online formats in which Quranic text is accessible, the safeguards against tampering with Quranic content, and the processes for confirming Quranic content.[27]

2.2 Quran classification

For the purpose of achieving a truly comprehensive and deeply nuanced computational understanding of the Quran, the implementation of an ontology becomes an absolutely necessary prerequisite. Such an ontology must possess the capability to not only meticulously store complex knowledge related to the sacred text but also to make this stored information readily available and accessible in a machine-readable fashion. This machine-readable format is crucial for enabling automated processing and analysis. However, existing or current ontology development procedures and methodological approaches are frequently found to be unsuitable and regrettably imprecise when the objective is to develop truly authentic and accurate conceptual representations specifically tailored for Quranic knowledge. This limitation often arises because these conventional methods describe concepts of knowledge using generic techniques without effectively connecting them to a specific, related subject area or domain of knowledge inherent within the Quran itself. Consequently, in order to properly comprehend, categorize, and explain the intricate classification system underpinning the Quran's vast and multifaceted knowledge base, the identification and structuring of distinct knowledge topics or themes are absolutely essential and play a fundamental role. Furthermore, demonstrating precisely how an ontology designed for the Quran has evolved over time and illustrating the practical application of employing a sophisticated semantic search strategy as the primary method for effectively discovering and retrieving specific knowledge contained within this theological text are both deemed particularly crucial steps for advancing research in this area [28].

It is a widely observed characteristic that any significant publication or body of text, regardless of the specific language in which it is composed, inherently contains or allows for a system of questions and potential answers to be derived or addressed. In the specific context of developing automated systems, particularly during the ongoing research process focused on rigorously validating the effectiveness of the Naive Bayes algorithm for specific tasks and simultaneously constructing a functional question and response system, the contributions of various other machine learning algorithms have demonstrably played a significant and indeed crucial supporting role alongside Naive Bayes. The Naive Bayes approach was initially selected as the first and primary choice for testing within this experimental framework, largely because of its inherent simplicity and computational efficiency; it is relatively simple and fast to compute results compared to many other, more complex models. The empirical results obtained from testing yielded an average accuracy rating of a notable 90.5 percent. This achieved level of performance strongly indicates and provides empirical evidence suggesting that the Naive Bayes approach retains its viability and is still demonstrably applicable and effective for tasks related to question answering or classification within this particular domain of study. The capability for accurate classification is generally considered an essential function or skill within natural language processing systems, holding particular importance and relevance when the task involves precisely determining the central topic, main subject, or underlying theme being discussed or represented within a given document or research "paper." Within this broader scientific category of textual analysis and automated classification, dedicated research specifically addressing the classification of various Quran translations has prominently featured the application of the Decision Tree technique [29]. This particular technique is utilized effectively as a systematic method to group or categorize different translations of individual Quranic verses based on their shared characteristics, predicted class labels, or thematic content.

Information from the Quran is beneficial to both religious experts and laypeople. The language of the Quran is Arabic. Muslims may benefit from the development of techniques for analyzing the Arabic text of the Quran and subsequently providing statistical data. When several text mining procedures are applied to this research area, they include word cloud, word embedding, clustering topic, and classification.[30]

Significant research has been conducted in the field of information retrieval to develop a variety of natural language and information retrieval systems that are linked to the Arabic language for different natural language and information retrieval system techniques. MQVC is a technique that finds verses that are most comparable to a user-inputted query verse. Document similarity evaluation is widely used in natural language processing and information retrieval applications.[31]

The practice of mechanically classifying Quranic verses into predetermined categories or topics is known as automatic themes-based categorization. For all Muslims and anybody else interested in learning the Quran, it is an essential exercise. Quran themes-based classification may be useful for a number of natural language processing (NLP) fields, including as search engines, data mining, Q&A systems, and information retrieval applications. Quran verses are automatically identified and categorized using a multi-label classification system based on themes and subjects.[32]

Regularly reciting the Quran is quite difficult. The recitation must be finished in accordance with Tajweed rules in order to prevent recitation errors that can lead to a mistranslation of the spoken words or sentences. A computerized speech-based categorization model called Tajweed is developed using an artificial neural network and a processing technique. The Quranic recitations of renowned reciters were combined to create the dataset. Neural networks are used in Tajweed categorization. Evaluation of the neural network's training procedure involved applying three distinct methods: Gradient Descent with Momentum, Resilient

Backpropagation, and Levenberg-Marquardt [1]. The test accuracy is highest for the Levernberg Marquardt training algorithm (70.7%), followed by Resilient Backpropagation (76.7%) and Gradient Descent with Momentum (76.7%.[33] (

According to the point of descent, the surah is separated into Makkiyah and Madaniyah groups. This division is predicated on the surah or specific verses' anticipated decline. By determining the classification of the data and using an algorithm, this grouping is achieved. Applications based on C4.5 are classified and their correctness evaluated using C4.5, an induction decision tree. The success rate of the algorithm is 95.6%. Based on this discovery, the C4.5 technique is well recognized in the classification process for the Quran's Suras.[34] .

Automatic text categorization (ATC) is the process of developing software tools that can assign previously viewed texts to certain groups or subjects. This is used to automatically classify the Quranic verses (ayat, sentences) based on the classifications made by Islamic scholars. Text categorization is done automatically using the traditional linear classification algorithm (score function). A system (classifier) was developed and implemented (chapter) to categorize the many verses in each Sura. This method completely normalizes the verses in the first step, and then they are categorized into classes according to their greatest scores.[35]

A machine learning method is used to identify Quranic terms in text produced from online sources. The suggested method for detection builds a learning model of Quran words using Support Vector Machines by training the learner on the Quran Words dataset. Terms detected in online content are then categorized using the created categorization model. A prototype has been developed and experiments have been conducted on various feature categories, such as statistical characteristics and diacritical features. Excellent accuracy and additional evaluation criteria are provided by the suggested approach.[36]

Determining the topic to which a verse in the Quran belongs is the primary goal of identification. Therefore, the present method of labeling Quranic passages depends on the availability of Quran scholars who are tafseer and Arabic-speaking. Automate Quran verse labeling using text classification techniques. The process was automated using three text categorization techniques: Nave Bayes, Support Vector Machine, and k-Nearest Neighbor. Text classification systems can classify Quranic verses with above 70% accuracy.[37]

2.3 Quran Topic Analysis

The expansive and ever-contemporary domain of Natural Language Processing (NLP) has a concentration area that revolves around a challenging exploration of word meaning, i.e., semantics. This exploration does not simply involve definitions, but rather, it pursues a deeper exploration of the layers of meaning that exist in textual data. When it is said that these computational linguistics are applied methods, it refers to the careful examination and development of more complex modeling techniques that are structured to provide a solid framework for semantic search functions on the valuable, multi-layered text of the holy Quran.

Choosing the unit of analysis is also an important consideration. Studies have suggested that, in light of efficiency in computational expenditure or data preparation, verses are the most appropriate structural unit for this kind of analysis. This is due to the fact that verses are discrete units, and the amount of processing data at the verse level, while attending to meaning, would likely cost the least energy and resources, compared to larger or smaller units of textual analysis. Reading only functional, basic keyword search would be neither sufficient nor effective in probing a complicated piece of literature like the Quran. It is imperative and critical to use complicated or modulated approaches to supersede simple word contrasts so that meaning, categorizing information, and redundancy can be achieved in a search for elements or components of significance to understanding the text.

From this critical lens, all Muslims agree that the meanings contained in the Quran as a holy text are endless and infinite; this reflects the belief and origin of it as the word of God delivered to all humanity, and the complexity contained in the text warrants a more complicated examination. The sacred Quran text has an organization unlike the typical architectures or forms and patterns one would expect to encounter in texts that discuss and describe elements of human interactions or ideas [38]. However, each surah (or chapter) in the Quran does address or anchor around a primary theme or notion. For example, all parts and fundamental pieces of the surah (which are generally regarded as verses) are deeply tied to one another and related (or related) to one another, which catalyzes a thread of relation or meaning density in a coherent way for a semantic mapping.

To computationally examine the inherent and complementary meanings and relationship, natural language processing algorithms were uniquely applied to do so. The algorithms were based on implementing and integrating two different approaches, the common word embedding (concept) analogous word2vec (which basically maps or conveys spatial relationship to conceptual attachments as defined by positions), and an approach called Roots' verse accompaniment (likely for its application in examining some of the unique literate characteristics about the organization of Quranic text in both grammatical and linguistic approaches). These abstract or concept orientations were used to compute semantic comparability (similarity) across linguistic constituent elements (word meaning). The similarities between surah titles and concepts are compared. Conceptual similarity and chapter distance are calculated and compared in the random mode. The evidence suggests that the title of the surah was chosen in a reasonable manner [39]. In computer science, the term "ontology" refers to a common field of knowledge. Ontology is used to explicitly characterize the domain of interest and is composed of people, concepts, and relationships. The majority of outputs still require human review and modification before being used in applications, despite efforts to automate the ontology building process through ontology learning from text. To automate validation and evaluate the caliber of ontology taught, the outcomes might be contrasted with a comparable resource.[40]

Approaches that go beyond word-level representation to sentence-level representation are necessary for evaluating a text such as the Quran. Graph vectors are used in a deep learning approach to learn an informative representation of Quranic verses. Machine learning models for subject analysis can use vectors as both inputs and outputs. Using the paragraph vector model, the author was able to develop a document embedding space that models and explains word distribution. The spatial dimensions of the Holy Quran reveal the semantic structure of the data, which aids in identifying key ideas and themes within the text.[41] The substance of the Quran has been extensively translated into a number of languages worldwide. Contributing to this field is challenging due to the difficulty of comprehending the text of the Quran in Arabic and other languages. The simplicity with which one can look for a certain topic required for a particular purpose is not yet depicted in the Quran. Tracing the implicit connections would require more thorough investigation and perseverance to reveal the hidden concepts and patterns because definitions and examples ideas are carried over from one surah to the next and from one ayah to the next.[42]

Though it offers a beginning to creating fixed procedures to handle similar instances that deal with reserved terms, we believe that working with the Quran is extremely sensitive. It gets more difficult after the verification stage since Quran classification necessitates a solid understanding of both Arabic and Islam. Topic analysis is what most impresses me; the Quran is incredibly comprehensive, and it's fascinating to extract topics and use algorithms. Whether or not you believe in the Quran, it is a unique book with many topics in a small number of words and a single word with multiple meanings. Using machine learning will enhance the task and offer numerous benefits.

3 | Conclusions and Significance

In the research paper, three areas were addressed. Quran authenticity, classification, and topic analysis. The methods outlined in the previous sections demonstrate very clearly that machine learning methods are the most common methods used for Quran classification, as with topic analysis. Watermarking is the area of research that is most researched when it comes to Quran authenticity [43].

This study comes to the conclusion that strong, scalable tools for Quran analysis are provided by unsupervised machine learning techniques like topic modeling and semantic clustering. When properly validated, these computational methods enable systematic thematic investigation and verse categorization, which effectively supplements conventional Tafsir. The significance of this work lies in its dual contribution: methodological and scholarly. On one hand, it offers a replicable framework for Quranic text mining using LDA and clustering algorithms; on the other, it delivers insights into the Quran's thematic composition. It highlights both the promise and the limitations of using modern data science tools in religious text analysis, emphasizing the need for interdisciplinary collaboration between computer scientists and Islamic scholars. Concluding Remarks

As digital Quranic studies continue to evolve, integrating machine learning with religious scholarship will unlock new paths for interpretation and educational tools. The future of computational Quranic analysis is rich with potential—balancing algorithmic precision with theological sensitivity will be key to advancing this field responsibly and meaningfully [43].

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

This study is based on a conceptual framework, and no empirical data were generated or analyzed.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Gaanoun, Kamel, and Mohammed Alsuhaibani. "Sentiment preservation in Quran translation with artificial intelligence approach: study in reputable English translation of the Quran." Humanities and Social Sciences Communications 12, no. 1 (2025): 1-15.

[2] Shahid, Usama, Muhammad Zunnurain Hussain, and William Sayers. "Computational Analysis of Quran Text Using Machine Learning and Large Language Models." In 2025 8th International Conference on Data Science and Machine Learning Applications (CDMA), pp. 18-24. IEEE, 2025.

[3] Nirwana, A. N. "Multimedia Tafsir: Exploring the Meaning of the Quran in the Digital Era." Available at SSRN 4785707 (2024).

[4] Rostam, Nur Aqilah Paskhal, and Nurul Hashimah Ahamed Hassain Malim. "Text categorisation in Quran and Hadith: Overcoming the interrelation challenges using machine learning and term weighting." Journal of King Saud University-Computer and Information Sciences 33, no. 6 (2021): 658-667.

[5] Safeena, Rahmath, and Abdullah Kammani. "Quranic computation: A review of research and application." In 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, pp. 203-208. IEEE, 2013.

[6] Siddiqui, Muazzam Ahmed, Syed Muhammad Faraz, and Sohail Abdul Sattar. "Discovering the thematic structure of the Quran using probabilistic topic model." In 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, pp. 234-239. IEEE, 2013.

[7] Bsoul, Qusay, Rosalina Abdul Salam, Jaffar Atwan, and Malik Jawarneh. "Arabic text clustering methods and suggested solutions for theme-based quran clustering: analysis of literature." Journal of Information Science Theory and Practice 9, no. 4 (2021): 15-34.

[8] Alhawarat, Mohammad. "Extracting topics from the holy Quran using generative models." International Journal of Advanced Computer Science and Applications 6, no. 12 (2015): 288-294.

[9] Allaoui, Mebarka, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. "Considerably improving clustering algorithms using UMAP dimensionality reduction technique: a comparative study." In International conference on image and signal processing, pp. 317-325. Cham: Springer International Publishing, 2020.

[10] Slamet, Cepy, Ali Rahman, Muhammad Ali Ramdhani, and Wahyudin Darmalaksana. "Clustering the verses of the Holy Qur'an using K-means algorithm." Asian Journal of Information Technology 15, no. 24 (2016): 5159-5162.

[11] Mohammad, A. "A New Fragile Digital Watermarking Technique for a PDF Digital Holy Quran." 2013.

[12] Kurniawan, Fajri. "Authentication and Tamper Detection of Digital Holy Quran Images." 2013.

[13] Nahar, Khalid. "Statistical Classifier of the Holy Quran Verses (Fatiha and Yaseen Chapters)." 2018.

[14] Alhawarat, Mohammad. "Extracting Topics from the Holy Quran Using Generative Models." 2015.

[15] Ahmad, Fadzil. "Tajweed Classification Using Artificial Neural Network." 2018.

[16] Alshammeri, Menwa. "Quranic Topic Modelling Using Paragraph Vector." 2018.

[17] Kamsin, Amirrudin, and Abdullah Gani. "Program for Developing the Novel Quran and Hadith Authentication System." 2015.

[18] AlAhmad, Mohammad A., and Imad Fakhri Alshaikhli. "A New Fragile Digital Watermarking Technique for a PDF Digital Holy Quran." 2013.

[19] Almazrooie, Mishal, Azman Samsudin, Adnan Abdul-Aziz Gutub, Muhammad Syukri Salleh, Mohd Adib Omar, and Shahir Akram Hassan. "Integrity Verification for Digital Holy Quran Verses Using Cryptographic Hash Function and Compression." 2018.

[20] Alssmadi, Ezzat, and Mohamad Zaror. "Online Integrity and Authentication Checking for Quran Electronic Versions." 2015.

[21] Kurniawan, Fajri, Mohammed S. Khalil, Muhammad Khurram Khan, and Yasser M. Alginahi. "Exploiting Digital Watermarking to Preserve Integrity of the Digital Holy Quran Images." 2014.

[22] Gilkar, Gulshan Amin, Saqib Hakak, Wazir Zada Khan, and Hussain Hameed Alshamrani. "Content Integrity Techniques for Digital Quran." 2020.

[23] Hakak, Saqib, Amirrudin Kamsin, Jhon Veri, Rajab Ritonga, and Tutut Herawan. "A Framework for Authentication of Digital Quran." 2018.

[24] Tayan, Omar. "The Role of Information Security in Digital Quran Multimedia Content." 2014.

[25] Khan, Muhammad Khurram, Zeeshan Siddiqui, and Omar Tayan. "A Secure Framework for Digital Quran Certification." 2017.

[26] Kurniawan, Fajri, Mohammed S. Khalil, Muhammad Khurram Khan, and Yasser M. Alginahi. "Authentication and Tamper Detection of Digital Holy Quran Images." 2013.

[27] Hakak, Saqib, Amirrudin Kamsin, Omar Tayan, Mohd. Yamani Idna Idris, Abdullah Gani, and Saber Zerdoumi. "Preserving Content Integrity of Digital Holy Quran: Survey and Open Challenges." 2017.

[28] Ta'a, Q. A. Abed, and M. Ahmad. "Al-Quran Ontology Based on Knowledge Themes." 2017.

[29] Putra, Syopiansyah Jaya, Yuni Sugiarti, Galuh Dimas, Muhamad Nur Gunawan, Tata Sutabri, and Agung Suryatno. "Document Classification Using Naïve Bayes for Indonesian Translation of the Quran." 2019.

[30] El Mouatasim, Abdelkrim, and Jaouad Oudaani. "Topics Classification of Arabic Text in Quran by Using Matlab." 2019.

[31] Akour. "MQVC: Measuring Quranic Verses Similarity and Sura Classification Using N-Gram." 2014.

[32] Mohamed, Ensaf Hussein, and Wessam H. El Behaidy. "An Ensemble Multi-Label Themes Based Classification for Holy Qur'an Verses Using Word2Vec Embedding." 2020.

[33] Ahmad, Fadzil, Saiful Zaimy Yahya, Zuraidi Saad, and Abdul Rahim Ahmad. "Tajweed Classification Using Artificial Neural Network." 2018.

[34] Irfan, Mohamad, Wisnu Uriawan, Nur Lukman, Opik Taupik Kurahman, and Wahyudin Darmalaksana. "The Quranic Classification Uses Algorithm C4.5." 2020.

[35] Nahar, Khalid. "Statistical Classifier of the Holy Quran Verses (Fatiha and Yaseen Chapters)." 2005.

[36] Sabbah, Thabit, and Ali Selamat. "Support Vector Machine Based Approach for Quranic Words Detection in Online Textual Content." 2014.

[37] Adeleke, Abdullahi O., Noor A. Samsudin, Aida Mustapha, and Nazri M. Nawi. "Comparative Analysis of Text Classification Algorithms for Automated Labelling of Quranic Verses." 2017.

[38] Alhawarat, Mohammad. "Extracting Topics from the Holy Quran Using Generative Models." 2015.

[39] Khadangi, Ehsan, Mohammad Moein Fazeli, and Amin Shahmohammadi. "The Study on Quranic Surahs' Topic Sameness Using NLP Techniques." 2018.

[40] Alrehail, Sameer Mabrouk A. "Ontology Learning from the Arabic Text of the Qur'an: Concepts Identification and Hierarchical Relationships Extraction." 2017.

[41] Alshammeri, Menwa, Eric Atwell, and Mhd Ammar Alsalka. "Quranic Topic Modelling Using Paragraph Vectors." 2020.

[42] Rolliawati, Dwi, Indri Sudanawati Rozas, Khalid, and Muhamad Ratodi. "Text Mining Approach for Topic Modeling of Corpus Al Qur'an in Indonesian Translation." 2020.

[43] Badawy, Amro. "A Survey on Digital Quran Computing (DQC)." DQC Survey, 2022.