Digital Quran Computing Algorithms and Applications

Amro Ali Badawy
Department of Computer Science
Faculty of Computers and Informatics
Zagazig University, Egypt

ABSTRACT
Digital Quran Contents (DQC) are growing in fast pace due to the wide spread of smartphones. The analysis of the digital Quran contents attracted significant research efforts with related discussions and studies covering a range of information technology (IT) subject domains under the generic topic of DQC. In this work, this paper will provide an updated literature review of these efforts to report on the developmental trends, challenges, and research directions in DQC. The research work in DQC include several research areas such as security, authentication, natural language processing, voice recognition, and knowledge based systems. The aim of the current work is to expose the existing research efforts in the field of DQC and explore the research opportunities and challenges in a systematic and scientific manner. In this research, the author proposed covering the research on DQC over three topics, namely, Quran authentication, Quran classification, and Quran topic analysis.

Keywords
Quran, Authentication, Classification, topic analysis, Text Mining

1. INTRODUCTION
Muslims holy book is Quran. Islamic knowledge is taken from Quran text. This led us to deal with two points. First: for Muslims, the Quran is reserved; thus, its text cannot be modified or changed because its rules from their god. Second: the text of Quran is used daily in any Muslim's life activities. Despite the Quran is preserved from distortion in Muslims' believes, the huge growth of digital media used on the internet presented new challenges to deal with possible un authenticated Quran text or contents this; a situation which is not accepted in by the believers of this religion. Thus, it is very important to address this issue [1]. Besides, it is of great demand to provide the Quran through devices such as Mobile phone and computers for Muslims, as then need to access Quran in daily activities throughout the entire day.

One of the main tasks in digital Quran that it is very important that one can determine a genuine Source of the Quran, or whether the content related of the Quran has been altered or not. There are three topics used in handling Quran content. Muslim people have a core Sharia body in charge of overseeing and determining the authenticity of the Quran. By developing and evaluating Quran authentication system [2].

The second main task is Quran classification. This assignment entails creating strategies for applying the semantic search strategy to search the Quran's knowledge. The conceptions in the Quran are categorized and structured according to a certain theme [3].

The third task is topic analysis which is about developing ways for searching and looking up specific information from the Holy Quran [4]. To manage data and information over the internet, Meta data related to each word and creating and assessing a method and application to analyses the authenticity of the language in electronic version of the Quran may be done. Readers have no way of knowing whether a verse is genuine or not. It's difficult to verify the validity of a verse due to an unintended misspelling or a deliberate conduct. The authors presented a method for systematically finding and categorizing appropriate techniques for maintaining the Digital Holy Quran's content integrity.

Online media and the internet are becoming increasingly popular. As a result, the number of individuals engaging with the Quran is growing, as is the availability of Quranic verses, scripts, translations, and other Quranic sciences through digital media.

The words from web content are classified using a classification model. Experiments with several feature categories. To achieve higher evaluation measurements, a prototype is produced utilizing machine learning and optimization techniques [5].

The Quran is the basic resource that has a wealth of patterns, themes, and facts that Muslims use to build their faultless pure knowledge. When analyzing a work of literature like the Quran, approaches that go beyond word level representation to sentence level representation are required. Deep semantic analysis and domain expertise are required to extract the inferred linkages, which necessitate learning approaches that go beyond word level representation to attain sentence level representation. With the help of a deep learning model offered in, this assignment is done [6].

2. QURAN AUTHENTICATION
Building and testing a Quran authentication system to help both the core and end users determine the validity of digital Quran apps, with the objective of utilizing it as a tool or mechanism to enhance digital Quran publication laws. The objective is to develop an innovative and accurate digital Quran authentication system capable of semantic and linguistic validation of the Quran. Accuracy is required to guarantee that the usage of original sources of references is authentic and acknowledged by Muslim scholars [7].

Watermarking has become a common solution for copyright and integrity issues in digital material. To preserve the integrity of PDF Digital Holy Quran files generated with the DCT method for feature extraction and a GEAR hash function for tamper detection, the authors developed an invisible fragile watermarking approach. To reduce the amount of time spent hashing and therefore achieve quicker performance, the watermark is produced by hashing the image's feature, which is obtained via the DCT method. The watermark is implanted using SLSB techniques, which results in less color distortion [8].

Integrity verification methods for electronic Quranic verses

are provided, with the primary job of the cryptographic hash function being to validate the integrity of the transmitted data. To create the Holy Quran's hash table, one technique employs cryptographic hash methods. The hash algorithms SHA256 and RIPEMD160 were chosen, while the other option is a single compression procedure that manipulates data in real time.

For the Arabic character set, the compression technique utilizes two bytes in Unicode UTF8. The results demonstrate that the sizes of the generated hash tables are less for a digitally encoded copy of the Holy Quran in Unicode UTF8, 84.73 percent and 90.46 percent, respectively. (6.55-fold and 10.48-fold) [9].

The Quran's 114 chapters (Suras), as well as all of its verses and phrases, have all been preserved in their entirety. As a result, the author created and tested a system and method for verifying wording integrity in Quran e-versions by producing Meta data for all words in the Quran while maintaining counts and positions. Hash algorithms are used in security to verify the integrity of a device and its data files, and every minor change in the data results in a different hash value [10].

The input image is transformed into the wavelet domain using the discrete wavelet transform, which operates block wise in the wavelet domain and pixel wise in the spatial domain, in the fragile watermarking approach.

The coefficients matrix of the wavelet transformed picture is then split into several blocks using a block-wise technique for further watermarking the authentication bits.

By replacing only one original element with +1 and inverting the modified coefficients matrix back to the spatial domain, the authentication bits are embedded into a block of 2×n components. By changing one inconsequential bit, the difference in the coefficients matrix of the changed and inverted is recorded on the pixel.

The authentication bits are encrypted using public-key cryptography. The authentication bits are calculated using a well-known hash function from the input image [11]. With the development of cyber security, attempts have been made to tamper with the Quran's contents. The purpose of this article is to systematically identify and categories viable techniques for maintaining the Digital Quran's content integrity. The emphasis has been on approaches that are only suitable for text and picture forms.

Even without diacritics, a native Arabic speaker can read Arabic text correctly. Diverse vowel sounds are represented by diacritics, As a result, the same term might have several meanings. Approaches to content protection and content verification were classified and categorized [12].

The verification and security parts of the authentication process are separated. The Boyer-Moore technique, an established and common exact matching algorithm, is employed for the verification component. During the security stage, watermarking will be used to safeguard the approved and tested verse. Furthermore, because the system is still being built, only the verification step of the proposed framework has been tested. To deal with diacritics in Quranic text, clitics segmentation utilizing UTF-16 encoding is utilized. The first prototype of the verification phase yielded encouraging results, with up to 98.6% [13].

Digital multimedia content literature has identified a variety of security challenges that need to be addressed, including digital copyright protection, proof-of-content-authenticity prevention, and content-originality verification. Such needs are clearly more prevalent in the case of specialized and religious information, similar is the Holy Quran's digital material.

Text, photos, audio, and video are the four types of digital multimedia material found on the Internet, with the difficulty being to offer safe, robust, and dependable storage and distribution for each [14]. The authors provide a digital Quran certification framework for certifying and authenticating Multimedia Quran apps in digital format using modern digital authentication and certification methodologies.

The certification procedure is scrutinized by a developed mechanisms and a religious panel, and the digital Quran application receives a digital certificate once all standards have been met. The granted digital certificate can be easily verified online by a common user of the service. The proposed framework intends to reduce the possibility of altering digital content of the holy Quran in order to protect consumers' faith and confidence [15].

For the authentication and verification step of digitized Quran images, a delicate watermarking approach is used. If there is evidence of harmful tampering, the recommended approach uses the discrete wavelet transform (DWT) to detect the tampered with pixels. The wavelet coefficients are taken into account while embedding authentication code that is encrypted with a secret key, guaranteeing great security. A block of wavelet coefficients contains the authentication binary code. The results of the experiments show that the proposed technique may successfully identify tampering and maintain image quality after watermarking while using a little amount of watermark payload [16].

The most significant issue facing Quran authentication and security is improving the accuracy and precision of text detection. It's also crucial to analyses and categories contemporary research that are relevant to preserving and validating the content integrity of the Quran. The current study is organized by format and methodology, such as the online formats in which Quranic text is available, the mechanisms used to secure Quranic material from modification, and the procedures used to verify Quranic content [17].

3. QURAN CLASSIFICATION

Understanding the Quran necessitates the use of an ontology that can capture knowledge and deliver it in a machine-readable format. Because they use traditional ways to define ideas of knowledge without tying them to a related subject of knowledge, current ontology methodologies are inappropriate and inaccurate in establishing authentic concepts of Quran knowledge. Knowledge themes are crucial in providing a true understanding and explanation of the Quran's knowledge classification. It's critical to show the evolution of the Quran's ontology as well as the way for searching for Quran knowledge utilizing a semantic search approach [18].

A question and answer system can be found in any language publication. Other algorithms have been highly significant in the process of constructing a question and answer system, which is still ongoing, and testing the Nave Bayes algorithm. The Nave Bayes technique is the first option in the

examination since it is easy to calculate. The Nave Bayes method is still relevant for application, as evidenced by its average accuracy rate of 90.5%. The use of classification is critical, especially when establishing the document's topic or theme. The Decision Tree approach, which is used to classify translations of Quran verses in the category, has been the subject of research on the classification of Quran translations in this category of science [19].

Both expert and non-specialized people in religion gain from extracting information from the Quran. Arabic is the language of the Quran. Finding methods for studying the Quran's Arabic text and then presenting statistical data could be beneficial to Muslims. Text mining operations such as word cloud, word embedding, clustering topic, and classification are implemented when various text mining operations are used to this research area [20].

In the realm of information retrieval, substantial research efforts have gone into the creation of various natural language and information retrieval systems connected to the Arabic language for various natural language and information retrieval system methods. MQVC is a method for finding verses that are the most similar to a query verse entered by the user. Information retrieval and natural language processing applications commonly use document similarity evaluation [21].

Automatic themes-based categorization is the process of automatically categorizing Quran verses into predefined categories or topics. It is a must-do activity for all Muslims and anybody interested in learning the Quran. Several natural language processing (NLP) domains, including as search engines, data mining, question–answering systems, and information retrieval applications, might benefit from Quran themes-based classification. A multi-label classification method based on themes and subjects is used to automatically identify and classify Quran verses [22].

It is quite tough to recite the Quran on a regular basis. To avoid recitation errors that could result in a mistranslation of the uttered words or sentences, the recitation must be completed according to Tajweed criteria. A processing approach and an artificial neural network are used to create Tajweed is a computerized speech-based categorization model. The dataset was generated by combining the Quran recitations of well-known reciters. Tajweed categorization makes use of a neural network. To assess the neural network's training process, three unique training methods – Gradient Descent with Momentum, Resilient Backpropagation, and Levenberg – were utilized the name Marquardt. The Levernberg Marquardt training algorithm 77.7% obtains the highest test accuracy, followed by Gradient Descent with Momentum 76.7% and Resilient Backpropagation 76.7% [23].

The surah is divided into Makkiyah and Madaniyah groups based on the point of descent. This split is based on the projected decrease of the surah or certain verses. This grouping is accomplished by identifying the data's classification and employing an algorithm. The C4.5 is an induction decision tree that is used to do classification and assess the accuracy of C4.5-based applications. The algorithm has a 95.6% success rate. The C4.5 approach is well-known in the classification process for the categorization of Suras in the Quran based on this finding [24].

The act of building software tools capable of allocating previously seen texts to predefined categories or topics is referred to as automatic text classification (ATC). This is used to automatically categories the verses of the Quran (ayat, sentences) according to Islamic scholars' categories. The classic linear classification function is used for automatic text categorization (score function). To categories the different verses in each Sura, a system (classifier) was created and deployed (chapter). In the first stage, this technique totally normalizes the verses, after which they are classified into classes based on their highest scores [25].

To detect Quran words in text created from online sources, a machine learning approach is utilized. By training the learner on the Quran Words dataset, the recommended approach for detection creates a learning model of Quran words using Support Vector Machines. The developed classification model is then utilized to classify terms found in internet material. Experiments on a variety of feature categories, including Diacritics and statistical features, have been carried out, and a prototype has been created. The recommended method provides outstanding accuracy and other evaluation criteria [26].

The main objective of identifying a Quran verse is to figure out what topic it belongs to. As a result, the current approach for labelling Quran verses is reliant on the availability of Quran scholars proficient in Arabic and tafseer. Using text classification techniques, automate the labelling of Quran verses. Three text categorization methods were employed to automate the procedure: k-Nearest Neighbor, Support Vector Machine, and Nave Bayes. Quran verses may be classified with more than 70% accuracy using text classification algorithms [27]. Table 1 summarizes the accuracy rates of different classification methods.

Table 1: Quran Classification methods

Reference	Accuracy
[9]	90.46%
[13]	98.6%
[19]	90.5%
[23]	77.7%
[24]	95.6%
[27]	70%

4. QURAN TOPIC ANALYSIS

Meaning of words is a major topic of research in Natural Language Processing that seeks to identify deeper meanings in a text (NLP). To provide a framework for searching the holy Quran using semantics, Modelling techniques are investigated. Verses are the greatest structure to utilize because it uses the least amount of energy for data. It's critical to find and retrieve valuable information using methods other than simple keyword searches. Muslims believe that the holy Quran is God's speech, and that its meanings are limitless. The text is not in the same order as human words [28].

The fundamental parts of each surah are close together, and each surah is built around a single central subject. The resemblance was calculated using natural language processing algorithms, which were based on the two approaches of word2vec and Roots' accompaniment in verses. Surah titles

and surah concepts are compared to see how similar they are. In the random mode, the conceptual similarity and the distance between chapters are computed and compared. The facts indicate that the surah's title was picked logically [29]. Ontology is a term used in computer science to express a shared area of understanding. Ontology is made up of individuals, concepts, and relationships, and it is used to formally characterize the domain of interest. Although ontology learning from text tries to automate the ontology creation process, most outputs still need to be reviewed and modified by humans before being used in applications. The results may be compared to a similar resource to automate validation and assess the quality of ontology learned [30].

When analyzing a text like the Quran, approaches that go beyond word-level representation to sentence-level representation are required. In order to learn an informative representation of Quranic verses, graph vectors are employed in a deep learning approach. Vectors may be used as both inputs and outputs for machine learning models for subject analysis. The author was able to create a document embedding space that models and explains word distribution using the paragraph vector model. The Holy Quran's spatial dimensions show the data's semantic structure, helping in the identification of important themes and concepts in the book. [31].

The Quran content has been widely translated into several languages all over the world. It's difficult to contribute to this arena because it's difficult to understand the Quran text in Arabic and other languages. The Quran has yet to depict the ease with which one can seek for a specific topic needed for a certain purpose. Because definitions and instances ideas pass over from one ayah to the next and from one surah to the next, tracing the implicit links would necessitate more in-depth research and patience to uncover the hidden concepts and patterns [32].

The author think working with Quran is very sensitive, but it’s an introduction to make fixed methods to handle similar cases that deal with reserved words. After the authentication step it becomes harder because classification of Quran requires good knowledge of Arabic language and also in Muslims religion. The most impressing for me is topic analysis, Quran is very reach, it is interesting to extract topics and apply algorithms. For Muslims, Quran is a special book that have multi subject in few words and have one word in multi meanings, using machine learning will improve the work and have a lot of advantages.

5. CONCLUSIONS

This paper discussed the recent research work related to the topic of DQC. The paper went through three different topics namely, Quran authentication, Quran classification, and Quran topic analysis. In the Quran authentication, the most notable used method is the discrete wavelet transform (DWT) to detect the tampered with pixels. In the Quran classification, the best performance was achieved by the C4.5 algorithm, 95.6%. Finally, the word2vec and ontology seem to be the most common tools by the different topic analysis methods. From the discussed methods, it is obvious that the machine learning methods are the most utilized approaches in the topics of Quran classification and topic analysis. On the other hand, the watermarking topic is the most widely explored topic in the field of Quran authentication.

6. REFERENCES

[1] A New Fragile Digital Watermarking Technique for a PDF Digital Holy Quran Mohammad A 2013

[2] Authentication and Tamper Detection of Digital Holy Quran Images (Fajri Kurniawan) 2013

[3] Statistical Classifier of the Holy Quran Verses (Fatiha and Yaseen Chapters) Khalid Nahar 2018

[4] Extracting Topics from the Holy Quran Using Generative Models (Mohammad Alhawarat)2015

[5] Tajweed Classification Using Artificial Neural Network (Fadzil Ahmad)2018

[6] Quranic Topic Modelling Using Paragraph Vector Menwa Alshammeri)2018

[7] Program for Developing the Novel Quran and Hadith Authentication System Amirrudin Kamsin , Abdullah Gani (2015)

[8] A New Fragile Digital Watermarking Technique for a PDF Digital Holy Quran Mohammad A. AlAhmad, Imad Fakhri Alshaikhli (2013)

[9] Integrity verification for digital Holy Quran verses using cryptographic hash function and compression Mishal Almazrooie a , Azman Samsudin a , Adnan Abdul-Aziz Gutub b , Muhammad Syukri Salleh c , Mohd Adib Omar a , Shahir Akram Hassan (2018)

[10] Online integrity and authentication checking for Quran electronic versions Ezzat Alssmadi , Mohamad Zaror (2015)

[11] Exploiting Digital Watermarking to Preserve Integrity of The Digital Holy Quran Images Fajri Kurniawan, Mohammed S. Khalil, Muhammad Khurram Khan and Yasser M. Alginahi(2014)

[12] Content Integrity Techniques for Digital Quran Gulshan Amin Gilkar, Saqib Hakak, Wazir Zada Khan, Hussain Hameed Alshamrani(2020)

[13] A Framework for Authentication of Digital Quran Saqib Hakak , Amirrudin Kamsin , Jhon Veri , Rajab Ritonga , and Tutut Herawan (2018)

[14] The Role of Information Security in Digital Quran Multimedia Content Dr. Omar Tayan (2014)

[15] A Secure Framework for Digital Quran Certification Muhammad Khurram Khan, Zeeshan Siddiqui , Omar Tayan, (2017)

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

[17] Preserving Content Integrity of Digital Holy Quran: Survey and Open Challenges SAQIB HAKAK, AMIRRUDIN KAMSIN, OMAR TAYAN, MOHD. YAMANI IDNA IDRIS , ABDULLAH GANI , AND SABER ZERDOUMI, (2017)

[18] AL-QURAN ONTOLOGY BASED ON KNOWLEDGE THEMESTa'a1, Q. A. Abed , and M. Ahmad (2017)

[19] Document Classification using Naïve Bayes for Indonesian Translation of the QuranSyopiansyah Jaya Putra, Yuni Sugiarti, Galuh Dimas, Muhamad Nur Gunawan,Tata Sutabri, Agung Suryatno (2019)

[20] Topics Classification of Arabic Text in Quran by using

MatlabAbdelkrim El Mouatasim and Jaouad Oudaani (2019)

[21] MQVC: Measuring Quranic Verses Similarity and Sura Classification Using N-GramAkour (2014)

[22] An Ensemble Multi label Themes Based Classification for Holy Qur'an Verses Using Word2Vec Embedding Ensaf Hussein Mohamed • Wessam H. El Behaidy (2020)

[23] Tajweed Classification Using Artificial Neural Network Fadzil Ahmad, Saiful Zaimy Yahya, Zuraidi Saad, Abdul Rahim Ahmad (2018)

[24] The Quranic Classification Uses Algorithm C4.5 Mohamad Irfan, Wisnu Uriawan1, Nur Lukman1, Opik Taupik Kurahman1, Wahyudin Darmalaksana2 (2020)

[25] Statistical Classifier of the Holy Quran Verses (Fatiha and Yaseen Chapters) Khalid Nahar (2005)

[26] Support Vector Machine based approach for Quranic Words Detection in Online Textual Content Thabit Sabbah, Ali Selamat (2014)

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

[28] Extracting Topics from the Holy Quran Using Generative Models Mohammad Alhawarat (2015)

[29] The Study on Quranic Surah's' Topic Sameness Using NLP Techniques Ehsan Khadangi, Mohammad Moein Fazeli, Amin Shahmohammadi (2018)

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

[31] Quranic Topic Modelling Using Paragraph Vectors Menwa Alshammeri , Eric Atwell , and Mhd Ammar Alsalka (2020).

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