Pertanika J. Sci. & Technol. 25 (S): 133 - 142 (2017)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Diacritical Digital Quran Authentication Model

Saqib Iqbal Hakak[1]*, Amirrudin Kamsin[1], Mohd. Yamani Idna Idris[1], Abdullah Gani[1], Gulshan Amin[2] and Saber Zerdoumi[1]**

[1]*Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia*
[**]*School of Computing and Info Tech, Lakeside Campus,Taylor's University, 47500 Subang Jaya, Selangor*
[2]*Faculty of Computer Science, Shaqra University, Riyadh, Saudi Arabia*

ABSTRACT

There has been an increase of content related to Quran and Hadith on the internet over the past few years. Diacritical Digital Quran is very sensitive to tampering. Diacritics are the symbols used beneath/above Quranic verses for reading purposes of the Quran. Minor change in diacritics can alter the meaning of a particular Quranic verse. Hence, there is a need for an authentication system to differentiate between fake and original verses. In this work, a model is proposed related to automatic authentication of Digital Quran. Authentication model is divided into two phases: tokenisation and authentication. For tokenisation, regular expressions are used to split input Quranic verse into single characters. In case of authentication, existing and standard exact matching algorithm i.e. Quick search (QS) is used. On testing the proposed model by comparing popular search engines and other related existing works, our approach is 100 % accurate in terms of full verse detection.

Keywords: Diacritical verse, Exact matching, Hadith authentication, Quran authentication and integrity, Quranic verse and text authentication

ARTICLE INFO

Article history:
Received: 15 August 2016
Accepted: 18 May 2017

E-mail addresses:
saqibhakak@siswa.um.edu.my (Saqib Iqbal Hakak),
amir@um.edu.my (Amirrudin Kamsin),
yamani@um.edu.my (Mohd. Yamani Idna Idris),
abdullah@um.edu.my;
abdullah.gani@taylors.edu.my (Abdullah Gani),
amingulshan9@gmail.com (Gulshan Amin),
zerdoumisaber@gmail.com (Saber Zerdoumi)
*Corresponding Author

ISSN: 0128-7680

INTRODUCTION

There is an ongoing exchange of digital content over the internet by millions of its users. Mostly, the platform used for the exchange of information involves social media such as Facebook, Instagram, blogs and web-sites (Alsmadi & Zarour 2015). This exchange of information involving online reading and sharing has very serious consequences of copyright-protection breach, digital counterfeiting and other authenticity issues (Hakak, Kamsin, Tayan, Idris, Gani,

& Zerdoumi, 2017; Hakak, Kamsin, Tayan, Idris, Gilkar, 2017; Mohammed, Sunar, & Salam, 2015). This area of authentication with the main focus on online content has emerged as one of the most promising and challenging research domains (Mohammed et al., 2015). The issue is more serious for religious scriptures like Digital Quran (Alsmadi & Zarour, 2015; Khan & Alginahi, 2013).

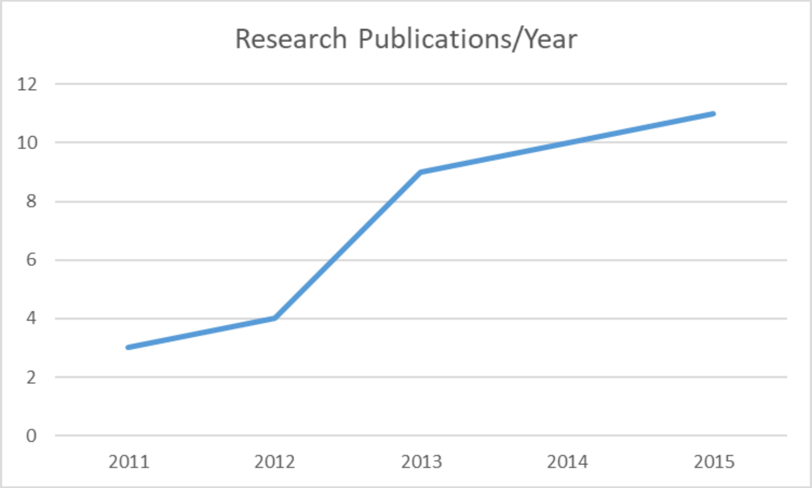


Figure 1. Research publications in the area of Quran Authentication (IEEE-Xplore, Elsevier Library)

Digital Holy Quran is the most sensitive and holiest sacred text for Muslims. It comprises 6236 verses divided into 114 chapters written in Classical Arabic. All these 114 chapters consists of variable number of verses or ayats (Ibrahim, 2010; Jamaliah Ibrahim, Yamani Idna Idris, Razak, & Naemah Abdul Rahman, 2013). It is available in different writing styles, but Uthmanic style that includes diacritics is most popular and widely used followed by simple Arabic style involving minimum diacritics (Abudena & Hameed, 2015).

In recent years, there have been numerous studies carried out with respect to authentication of the Quran as seen in Figure 1 (taken from IEEE Xplore and Elsevier library). The prime concern of these research works has been authentication of Quranic or hadith content and securing the Quranic content from being tampered through different approaches of cryptography, string matching and watermarking. In any case, there is a lot of scope to work on different research gaps related to the authentication of the Quran. One of the primary research gap is to verify Quranic verses without removal of diacritics. There are lots of symbols and diacritics used in the Quran and alteration in single diacritic changes the meaning of the entire sentence / verse. For instance, an Arabic word كتب comprising three consonants i.e. ك ت ب gives different meanings with different arrangement of diacritics [9]. Different meanings with different diacritics like كَتَبَ means wrote, كُتُب means books, كُتِبَ means written and كَتَّبَ means Make someone to write (Hakak, Kamsin, Tayan, Idris, Gilkar, 2017). Thus, it is necessary to secure and authenticate the Quranic verses available on the internet.

Accordingly, to solve the disadvantages and drawbacks of the prior art, there is an authentication model for Holy Quran. This paper is organised as follows: Recent works on this topic are discussed in Section 2. In Section 3, the proposed model is described while the results are discussed in Section 4. Conclusion is provided in Section 5.

Literature Review and Findings

Majority of internet users prefer to copy Quranic text from a specific source and paste it either on social media websites or other online blogs (Hakak, Kamsin, Tayan, Idris, & Gilkar, 2017). It is one of the most suitable and easy approach. Most non-native speakers of Arabic rely heavily on diacritics for reading and understanding Quran (Figure 2).

ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ   ٱلحمد لله رب ٱلعلمين

Figure 2. Diacritic Arabic Quran Verse

Text based authentication of digital Quran is less popular compared with image based version. This can be due to the complexity of diacritics in the text based format. Works based on the text format of Digital Quran are as follows:

Alginahi et al., (2013) proposed algorithm for verification of Arabic verses along with diacritics and other symbols. The algorithm is based on the SELECT query using MySQL that uses either linear search algorithm or binary search algorithm. However, in the pre-processing phase, all diacritics are removed for verse identification. Alsmadi et al. (2015) proposed an authentication model for Quranic verses. The authors claim that document control and digital signature are the two most widely used approaches to authenticate documents. Document control is giving permission before and after publishing the document online. In digital signature, documents should be verified by the people who signed it. The focus has been on integrity checking. The authors mention that it is challenging to read or parse Arabic diacritics correctly. Hashing approach is used in this research. Hash is calculated based on the particular verse and that hash value is compared with a hash value in database. However, there is a possibility of hash collision using this approach. Sabbah & Selamat (2013, December) proposed framework to detect and authenticate Quranic verses. The focus has been to increase detection accuracy of diacritic text. Accuracy on an average is 62%. However, this algorithm will not work with non-diacritical text and there is so much overhead associated while calculating weights and dividing the verses into two groups. The complexity of the algorithm will increase with extremely complex diacritical texts. Alshareef & El Saddik (2012) proposed a framework for Quranic verse detection. The idea is simple where the text Quranic verse is taken as input and results are displayed, whether the verse is authentic or not. There are two major components in this i.e. Quranic quote filtering and verification mechanism. In Quranic quote filtering, all Arabic diacritics and special symbols are removed. The authors claim symbols and diacritics limit the traditional search engines to provide acceptable and accurate results to the users without any valid proof or justifying the claim. Finally, after removing all symbols and diacritics, the Quranic verification mechanism is used which is based on a regular expression SQL query to verify the text. The authors have used some single verses to evaluate the authenticity of those verses. The proposed algorithm shows 89% accuracy with compared with rest of the search engines against few words. In this paper, it is assumed this accuracy will decrease if this algorithm is used on large Arabic data set due to the fact that regular expression use prefix-suffix approach for searching. Alshareef and El Saddik (2012) and Nisha, Ali, and Ali (2014) studied different search engines and their limitations. A new search engine with the name of "Truth-search-now" has been proposed. Five search engines with respect to Islamic content have been studied i.e. TheIslamic search, IntoIslam, Search-truth, IslamiCity, and Allah.pk and evaluated based on the time taken by each search engine to find a particular query. No any experimental proof is given, nor any algorithm mentioned (Nisha, Ali, & Ali, 2014).

It is observed from the recent advances that the authentication of Digital Quran is still at its initial stage and there are numerous issues such as authentication of Quran without removal of diacritics, improvement in search time of Quranic verses and other such issues. In order to address the issue of the authenticating diacritical version of the Quran, we proposed an efficient model for authentication of Quran without removal of any diacritics.

METHODS

Literature review has shown that a lot of work has been in the area of image processing while the present study's proposed model is based on binary data i.e. text as presented in Figure 3. This model can be used to detect fake Quranic verses from online sources such as social media, online blogs and web forms.

The authentication model framework consists of two phases i.e. tokenisation and searching as shown in Figure 3. The purpose of the tokenisation phase is to split the given input verse into individual characters and convert the whole string into its Unicode format. The purpose of searching phase is to verify the input from database using exact matching algorithm. A brief description of each component involved in an authentication model is discussed below:

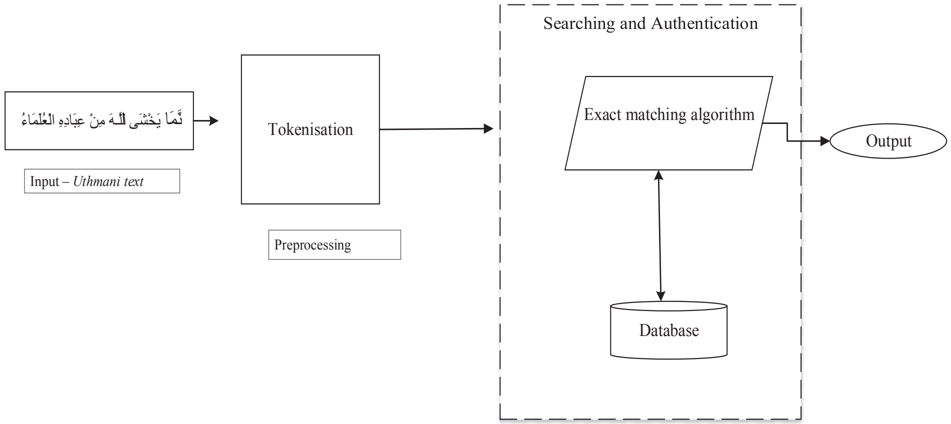


Figure 3. Proposed model for Authentication of Digital Quran

Input: In this component, the user enters the Uthmanic Quranic verse to be authenticated. After getting the Quranic verse as an input from a particular user, the output is passed to the tokenisation phase.

Tokenisation phase: In the tokenisation phase, the given input verse is tokenised based on clitics (a meaningful morphological unit of a language based on unique UTF code of Arabic characters). We use the UNICODE-16 scheme as it is a variable length encoding and suits diacritical Arabic and other complex texts. Sample Unicode representation for each Arabic character is shown in Table 1 below.

Table 1
UTF-16 Encoding for Arabic characters

Quranic Letters	UTF-16 Representation	Total Number of Verses starting with particular letter	Quranic Letters	UTF-16 Representation	Total Number of Verses starting with particular letter
ا	U+0627	1178	ض	U+0636	6
ب	U+0628	175	ط	U+0637	3
ت	U+062A	59	ظ	U+0638	1
ث	U+062B	109	ع	U+0639	42
ج	U+062C	14	غ	U+063A	2
ح	U+062D	24	ف	U+0641	698
خ	U+062E	31	ق	U+0642	530
د	U+062F	3	ك	U+0643	118
ذ	U+0630	65	ل	U+0644	262
ر	U+0631	47	م	U+0645	155
ز	U+0632	3	ن	U+0646	25
س	U+0633	48	ه	U+0647	85
ش	U+0634	4	و	U+0648	2215
ص	U+0635	5	ي	U+0649	329

Finally, UTF-16 based encoded verse passes to a search and authentication phase.

Search and Authentication phase: The role of this phase is simply to match UTF -16 based output from authenticated database. For matching it uses a modified version of Boyer-Moore String Matching Algorithm (Boyer & Moore, 1977) i.e. Quick search that finds the required verse from authenticated database. If there is a match, the verse is displayed.

The algorithm starts searching characters from right to left of the given pattern. In case of a mismatch, it can shift as many as *m* characters as shown in Figure 4.

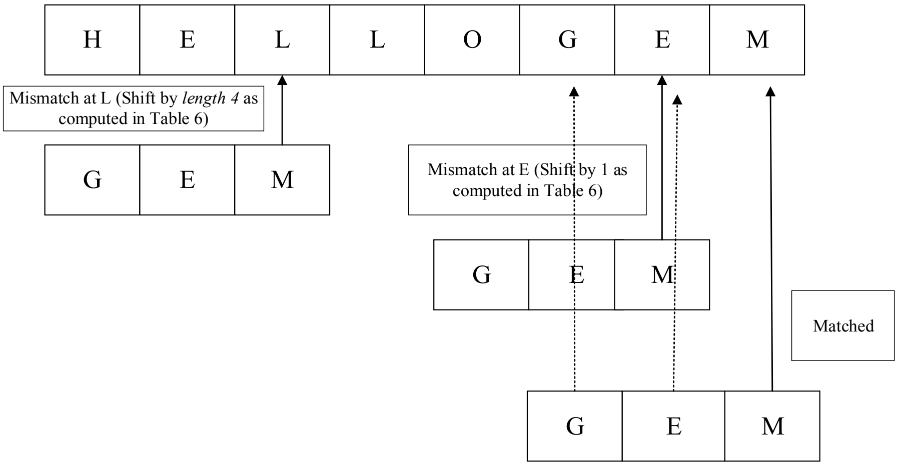

H
E
L
L
O
G
E
M
Mismatch at L (Shift by *length 4* as computed in Table 6)
G
E
M
Mismatch at E (Shift by 1 as computed in Table 6)
G
E
M
Matched
G
E
M

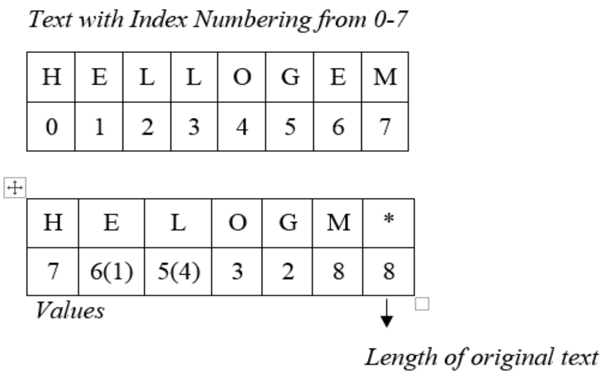
Figure 4. Searching algorithm

There are two stages in this algorithm:

1. Pre-processing
2. Searching for a given pattern from the right side of the window and using bad match table to skip characters in case of a mismatch.
    1. Pre-processing: During the pre-processing stage, a table is created which gives values on how much shift is required in case of a mismatch. This is also known as bad-match table. Thus, once a character mismatch occurs, algorithm shifts to the right of the pattern according to the value given in bad-match table.
    2. Searching starts from the tail of the pattern, i.e. from right to left of the text as compared in naive algorithm, where the search starts from left to right. The algorithm works by computing the length of search string and storing its value as default shift length. In the search string, for each character, the shift value is set as shown in Figure 4.
    3. The values can be computed using *Value = Length of pattern-1-index of character* as shown in Table 2:

Table 2
Computation of values in QS Algorithm

Length-1-Index	Value
(1, 8-1-0)	7
(1, 8-1-1)	6
(1, 8-1-2)	5
(1, 8-1-3)	4
(1, 8-1-4)	3
(1, 8-1-5)	2
(1,8-1-6)	1



Text with Index Numbering from 0-7

H	E	L	L	O	G	E	M
0	1	2	3	4	5	6	7

H	E	L	O	G	M	*
7	6(1)	5(4)	3	2	8	8

Values

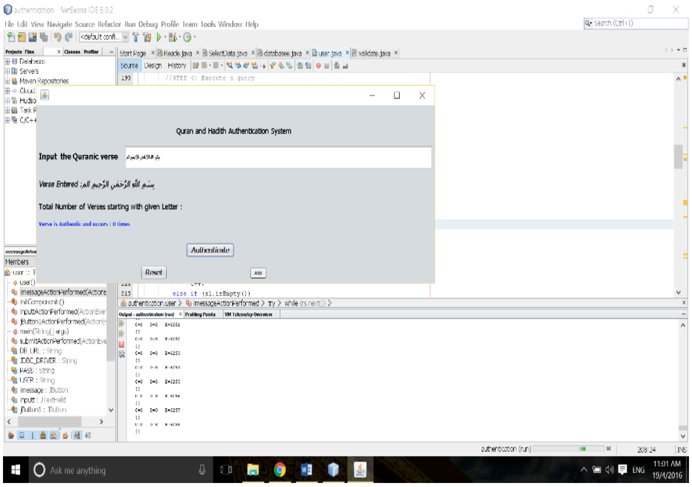
Length of original text

Figure 5. Calculated values

In Table 2, the values in brackets for character “E” and “L” in Figure 5 are the current values of shifting phase. When algorithm finds another occurrence of the same character twice, the previous value is replaced with new values. Therefore, 6 and 5 values of respective “E” and “L” are replaced by 1 and 4. For the last character, value is the length of text. This gives time complexity *O(n+m)* for the best case and in worst case *O(n*m)*. Here, m denotes length of pattern and *n* denotes length of text which is to be searched.

RESULTS

In order to test the authentication model, a prototype was developed. The prototype was implemented using Netbeans IDE environment 8.02 on i-5 Intel Processor with 4 MB cache, 4 GB RAM using Windows 10 and programming language used was Java. The initial Graphical user interface is shown in Figure 6. Random Uthmanic Quranic verses from the internet were copied and their output was analysed. For verifying authenticity, a standard Uthmani Quran dataset from http://tanzil.net/#2:1 as used. The verses were tested with two parameters of Actual number of verse to be verified in Digital Quran and Retrieved shown in Table 3.



Quran and Hadith Authentication System

Input the Quranic verse بسم الله الرحمن الرحيم الم

Verse Entered : بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ الم

Total Number of Verses starting with given Letter :

Verse is Authentic and occurs : 0 times

```
C=0    D=0    D=6251
[]
C=0    D=8    D=6252
[]
C=0    D=0    D=6253
[]
C=0    D=8    D=6254
[]
C=0    D=0    D=6255
[]
```

Figure 6. Initial prototype for Diacritical Digital Quran

In the experiment, ground truth related to tested verses was calculated manually. Each tested verse was checked manually first as how many occurrences of that verse is in digital Quran. Finally, accuracy was calculated based on the formula below:

$$Accuracy = \frac{\text{Number of particular verses Found}}{\text{Total number of particular verses}}$$

Results are shown in Table 3 and Table 4. Results were compared from existing search engines that are popular for searching Quranic verses. Search engines include http://quran.muslim-web.com/ and http://tanzil.net.

Table 3
Accuracy results

Verse No.	Quranic Verses	Actual No. of Verses	Muslim-Web	Tanzil.net	Proposed Approach	Accuracy
1	دُحُورًا وَلَهُمْ عَذَابٌ وَاصِبٌ	1	1	1	1	100 %
2	فَاتَّقُوا اللَّهَ وَأَطِيعُونِ	8	10	10	8	100 %
3	فَبِأَيِّ آلَاءِ رَبِّكُمَا تُكَذِّبَانِ	31	31	31	31	100%
4	بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ الم	8	0	135	7	100%

From Table 3, it can be observed that all existing search engines retrieved verse 1 correctly. However, in the case of verse 2 and 4, existing search engines were either not able to retrieve verses completely or showed no results indicating the limitations of detecting verses accurately. Our proposed authentication model showed promising results for all the given random verses and acquired efficiency of 100% for all the verses.

Table 4
Results with modified verses

Quranic Verses	Muslim-Web	Tanzil.net	Proposed Approach
ذلك الكتاب ل ريب فيه هدى للمتقين	No Results	Shows Results for Incorrect verse	Authenticates the verse is not authentic
يَّاكَ نَعْبُدُ وَإِيَّاكَ سْتَعِينُ ا	No results	Shows Results for Incorrect verse	Authenticates the verse is not authentic
بِمَا أُنزِلَ إِلَيْكَ وَمَا أُنزِلَ مِن قَبْلِكَ وَبِالْآ	Shows Results for Incorrect verse	No results	Authenticates the verse is not authentic
الْكِتَابَ بِالْحَقِّ مُصَدِّقًا لِّمَا بَيْنَ	Matches but displays more than one result	Matches but displays 11 results	Authenticates the verse is not authentic

Besides checking for accuracy of verse detection, we modified some Quranic verses to check whether existing search engines provide any results (see Table 4). All the existing engines gave results for incorrect verses too. For a non-native speaker of Arabic language, it will create confusion which verse is authentic, the one which is modified or the one the search engines displayed. Compared with existing search engines, the model detected fake verses accurately. The fact existing search engines give results for modified verses can be due to removal of diacritics. Most of the work in the area of Quran authentication has focused on removal of diacritics for efficient matching. Thus, if you remove any diacritic, these search engines will still show results based on character comparisons. Therefore, our model is the best in terms of providing accurate verses.

CONCLUSION

In this paper, an efficient model related to Quran authentication was proposed. The model is divided into two phases i.e. tokenisation phase and searching phase. Tokenisation is mainly responsible for the tokenisation/segmentation part while the searching phase is responsible for authenticating the verified content. Experiments conducted on the first initial prototype of the verification phase showed promising results with accuracy of up to 100%. No diacritics were removed in the whole process which is a novelty in this model. Previous models removed the diacritics for accuracy purposes. Our Future work will focus on enhancing the verification phase to improve search time and developing a complete framework by adding security phase to protect verified verses from being tampered again. The last phase will be to test the whole framework using different Quranic verses and Hadith from the internet.

ACKNOWLEDGMENT

This research was supported by the University of Malaya under FRGS grant (No: FP003- 2016), Faculty of Computer Science & Information Technology.

REFERENCES

Abudena, M. A. & Hameed, S. A. (2015). Toward a novel module for computerizing Quran's full-script writing. *International Journal of Computer Systems*.

Alginahi, Y. M., Tayan, O., & Kabir, M. N. (2013). Verification of Qur'anic quotations embedded in online arabic and islamic websites. *Int. J. Islam. Appl. Comput. Sci. Technol, 1*, 41-47.

Alshareef, A., & El Saddik, A. (2012, March). A Quranic quote verification algorithm for verses authentication. In *International Conference on Innovations in Informatioc Technology.* (pp. 339-343). IEEE.

Alsmadi, I., & Zarour, M. (2015). Online integrity and authentication checking for Quran electronic versions. *Applied Computing and Informatics,* 1-16.

Boyer, R. S., & Moore, J. S. (1977). A fast string searching algorithm. *Communications of the ACM, 20*(10), 762-772.

Hakak, S., Kamsin, A., Tayan, O., Idris, M. Y. I., & Gilkar, G. A. (2017). Approaches for preserving content integrity of sensitive online Arabic content: A survey and research challenges. *Information Processing and Management*.

Hakak, S., Kamsin, A., Tayan, O., Idris, M. Y. I., Gani, A., & Zerdoumi, S. (2017). Preserving content integrity of digital holy Quran: Survey and open challenges. *IEEE Access,* (99), 1-1.

Ibrahim, N. J. (2010). *Automated tajweed checking rules engine for Quranic verse recitation* (Doctoral dissertation), University of Malaya.

Jamaliah Ibrahim, N., Yamani Idna Idris, M., Razak, Z., & Naemah Abdul Rahman, N. (2013). Automated tajweed checking rules engine for Quranic learning. *Multicultural Education and Technology Journal, 7*(4), 275-287.

Khan, M. K., & Alginahi, Y. M. (2013). The holy Quran digitization: Challenges and concerns. *Life Science Journal, 10*(2), 156-164.

Kirchhoff, K., & Vergyri, D. (2005). Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition. *Speech Communication, 46*(1), 37-51.

Mohammed, A., Sunar, M. S., & Salam, M. S. H. (2015). Quranic verses verification using speech recognition techniques. *Jurnal Teknologi, 73*(2), 99-106.

Nisha, S., Ali, N., & Ali, A. S. (2014, November). Searching quranic verses: A keyword based query solution using. net platform. T*he 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M)* (pp. 1-5). IEEE.

Sabbah, T., & Selamat, A. (2013, December). A framework for Quranic verses authenticity detection in online forum. *Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences (32519)* (pp. 6-11). IEEE.