Jurnal Teknologi

Full Paper

A NOVEL DATASET FOR QURANIC WORDS IDENTIFICATION AND AUTHENTICATION

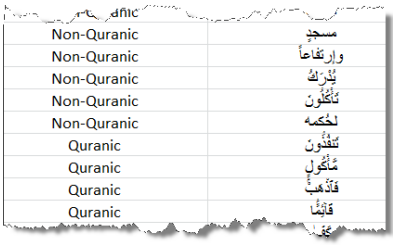
Thabit Sabbah, Ali Selamat*

Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

Article history
Received
3 December 2013
Received in revised form
2 July 2014
Accepted
25 November 2014

*Corresponding author
aselamat@utm.my

Graphical abstract



Abstract

Quran is the holy book for Muslims around the world. For the past fourteen centuries after its revelation, ithas been preserved in all possible ways from any distortions. The huge increase in Internet usage and the spread of digital media lead to the development of many websites, services, and applications related to Quran. These efforts include the conversion of Quranic verses, translations, explanations,tafseer and other Quranic sciences into digital formats. Some of these efforts are foundless authentic. The authentication dependson correct identification of Quranic words in the text. In this paper, we introduce a novel dataset for Quranic words identification and authentication. The proposed dataset contains more than 93,000 samples with64 features for each extracted in numerical form.The validation tests of the proposed dataset resulted high accuracy average.

Keywords: Quranic words, identification, authentication, dataset, Arabic, diacritic words

1.0 INTRODUCTION

Quranic words identification can be defined as determining which words belong to and are written exactly as these are written in the holy Quran. The consecutive well-ordered Quranic words form a verse. Normally, the Quranic verses are written with diacritics. According to the Arabic linguistics, the diacritics are used for distinguishing the vocal pronunciation of words and representing the vowel sounds [1].In ordinary Arabic writing, few diacritics are commonly used. These areas listed in Table 1 with their symbol and code points. However, in Quranic text writing there are many other diacritics and special symbols used to give the reader other sorts of reading guidance. We will reference all of these diacritics and special symbols as diacritics for abbreviation.

Table 1 Common diacritics used in Arabic writing

Name	Symbol	Code Point	Name	Symbol	Code Point
Arabic Fatha	◌َ	U+064E	Arabic Fathatan	◌ً	U+064B
Arabic Dammah	◌ُ	U+064F	Arabic Dammatan	◌ٌ	U+064C
Arabic Kasra	◌ِ	U+0650	Arabic Kasratan	◌ٍ	U+064D
Arabic Sukun	◌ْ	U+0652	Arabic Shaddah	◌ّ	U+0651

Muslims around the world use verses in their daily life for many reasons, such as support decision making, deduce solutions for social and religious problems, or to analyze various issues. Most of the Muslim authors quote verses as an evidence to support their conclusions and their analysis of events [2]. Quotation of verses is used not only in written publications, but it is also common in speeches, conversations, dialogs and discussions. This scientific method is popular in

Islamic societies in general and Arabs in particular [3].According to this method, in written works, the quoted verses are distinguished from the text by many ways such as surrounded by brackets, written in standard Quranic style (known as Uthmani) with full diacritics, etc. On the other hand, Quranic verses are distinguished in oral works by reciting these verses according to the standard Quranic recitation rules or by adding the common phrases used to indicate the starting and the ending of the quoted verse, and various other vocal techniques used to distinguish verses in speech. In Islamic multimedia works various advanced graphics and sound effects techniques are utilized to distinguish verses among other multimedia content. However, in most of the less authentic works especially online recourses (forums, social networks, blogs, personal websites) writers occasionally use the scientific method of quoting and citing Quranic verses. This makes it difficult for an ordinary reader to distinguish the Quranic verses (words), Moreover the automatic text processing and natural language processing techniques (Information Retrieval Systems and Knowledge Management Systems) will work less efficiently on Arabic text in general and religious content in particular.

In this paper, we propose a novel dataset for Quranic words identification and authentication; the proposed dataset contains 93,161 samples with 64 features for each. Samples are categorized into two categories; "Quranic" and "non-Quranic". The "Quranic" labeled samples of the dataset are collected from highly trusted resource, while the "non-Quranic" labeled samples are collected from thousands of Arabic posts downloaded from one of the biggest Arabic religious forum on the web (Muslim forum). In the rest of this paper, section 2 highlights the motivation by focusing on previous works related to Quranic verse authentication. Section 3 explains our framework in generating our dataset. Dataset features, specifications and statistics are shown in section 4. Section 5 illustrates the validation of our dataset, and finally this work is concluded in section 6.

2.0 RELATED WORK AND MOTIVATION

Less works are found for Quranic words identification and authentication. These studies did not consider the use of a standard dataset for the authentication purposes. Some of the works focus on authentication of Quranic quotes such as [2] in which Quranic quotes are predefined. Other previous works discussed various issues related to digital Quranic script from different aspects other than detection. A statistical study [4] aimed to protect the digital form of the Quran from corruption. The study considered the structure of the Quran in terms of number of characters in the verses and the number of verses in the chapters by representing Quranic text as a codified inference in the form of an AI natural language. Other aspects discussed by the existing studies are text information retrieval [5], semantic search [6, 7], knowledge modeling, and retrieval [8-10]. However, in all the discussed works none has used a standard dataset. The reason is that there does not exist any digital Quranic dataset encompassing all the necessary linguistic requirements

3.0 RESULTS AND DISCUSSION

In this paper, we propose a novel dataset for Quranic words identification and authentication. Figure 1 Building dataset framework shows the research approach considered to generate the dataset and the data resources and processes.

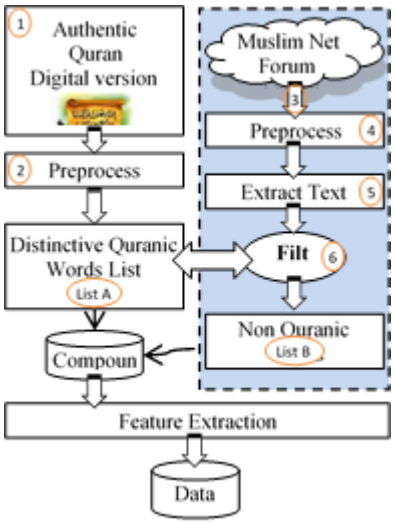


Figure 1 Building dataset framework

In the used approach, we started with Quranic text provided by Tanzil.net project (step labeled 1), which is the most reliable and precise digital Quran text available on the web[11]. In Tanzil.net project digital version, Quranic text is provided in XML format, the root node of XML file is "quran" then "sura" nodes to represent chapters, and then "aya" nodes to represent verses.

Figure 2 Quranic text as in Tanzil.net XML format shows a snapshot of the Quranic text in Tanzil.net XML file.

```
- <quran>
  - <sura name="الفاتحة" index="1">
      <aya index="1" text="بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ"/>
      <aya index="2" text="ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ"/>
      <aya index="3" text="ٱلرَّحْمَٰنِ ٱلرَّحِيمِ"/>
      <aya index="4" text="مَٰلِكِ يَوْمِ ٱلدِّينِ"/>
      <aya index="5" text="إِيَّاكَ نَعْبُدُ وَإِيَّاكَ نَسْتَعِينُ"/>
      <aya index="6" text="ٱهْدِنَا ٱلصِّرَٰطَ ٱلْمُسْتَقِيمَ"/>
      <aya index="7" text="صِرَٰطَ ٱلَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْرِ ٱلْمَغْضُوبِ عَلَيْهِمْ وَلَا ٱلضَّآلِّينَ"/>
    </sura>
  - <sura name="البقرة" index="2">
      <aya index="1" text="الٓمٓ" bismillah="بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ"/>
      <aya index="2" text="ذَٰلِكَ ٱلْكِتَٰبُ لَا رَيْبَ فِيهِ هُدًى لِّلْمُتَّقِينَ"/>
      <aya index="3" text="ٱلَّذِينَ يُؤْمِنُونَ بِٱلْغَيْبِ وَيُقِيمُونَ ٱلصَّلَوٰةَ وَمِمَّا رَزَقْنَٰهُمْ يُنفِقُونَ"/>
```

Figure 2 Quranic text as in Tanzil.net XML format

As seen in the snapshot above, the root node of XML file is "quran", then the first level of nodes ("sura" nodes) represent the chapter names of holy Quran, and the sub-nodes ("aya" nodes) represent verses. The "sura" nodes have two attributes; "name" and "index" while the "aya" nodes contain two attributes; first is the index of the verse in chapter, and second is the text of verse.

For preprocessing step (labeled 2 in our approach), the algorithm shown in Figure 3 Extracting distinctive Quranic words algorithm is developed to extract the distinctive Quranic words list (List A) from the provided Quranic text format.

Algorithm 1:
Extracting Distinctive Quranic Words List

Input: Quranic Text XML File provided by Tanzil.net

Start:
Verse Words ← ∅
Distinctive Quranic Words DQW{} ← ∅
For (Chapter ∈ Document**)**
For (Verse ∈Chapter**)**
Verse Words {} ← Verse
Loop
If Word not in Distinctive List
Distinctive Quranic Words {} ← word
End IF
Verse Words {} ← Verse Words {} - word
While Verse Words {} ≠ ∅
End // Verse in Chapter
End // Chapter in Document

Output: List of Distinctive Quranic Words

Figure 3 Extracting distinctive Quranic words algorithm

In the algorithm (Figure 3), we started with an empty set of Distinctive Quranic Words (DQW), then we access the text of each verse in each chapter; tokenizing the text and adding the tokens to the DQW set, with the consideration that the duplication of tokens in DQW is not allowed.

After that we labeled all the words in the list (List A) generated by the above algorithm with "Quranic" label. The number of distinctive diacritic Quranic words in the list is 18,994 words.

In generating Quranic words list no stemming, filtering, cleaning, letters replacement or consolidation processes were performed because of saving the Quranic words from any kind of distortion.
To generate a non Quranic words list, the standard Arabic datasets were investigated; First, two famous news datasets are downloaded namely Kaleej-2004 [12] and Watan-2004 [13]. These Arabic corpus are composed of Arabic texts for text categorization. The corpus Khaleej-2004 contains 5,690 documents. It is divided into 4 topics (categories). The corpus Watan-2004 contains 20,291 documents organized in 6 topics (categories). However, the words extracted from these datasets are useless for the problem in hand because datasets' texts are preprocessed and many considerable features are already removed. Other less famous Arabic datasets available on Internet are also found to be suffering from various problems. For example, the Arabic dataset offered by Abuaiadh [14] have only 6,726 non-Quranic diacritic words. This number of non-Quranic diacritic words is small in comparison with the number of Quranic words that is 18,994. The main consideration of the proposed dataset is to make the Quranic words distinguished from a crowd of non-Quranic words, so that a huge number of non-Quranic diacritic words is required. Moreover, the researchers who conducted Arabic language related research, for example, Arabic language Identification [15-18] used their own homemade datasets by collecting Arabic web pages from the Internet. Because the existing Arabic datasets are insufficient to meet the goal, we collected 114,398 Arabic forum posts (step 3 in Figure 2) from Muslim Net forum. Muslim Net forum is one of the biggest Islamic forums on the web containing more than 2,600,000 post sin 351,429 threads, with more than 66,100 active users. The threads in Muslim Net forum are distributed into 21 categories in various life aspects such as religion, politics, news, multimedia, computer hardware, software, design, education, and other categories. The categories are not in the consideration for this dataset. However, the diversity of topics reflects the diversity of non-Quranic words as the Quranic words have approximately the same diversity.

To extract non-Quranic words from the downloaded files we conducted three steps (steps labeled 4, 5, and 6 in framework).The preprocessing step (step 4) includes removing HTML tags, removing non-Arabic letters, punctuations, special characters, and numbers. Then, the Arabic text was extracted and to kenized. The number of Arabic words extracted from these pages is148,631 distinctive words. Filtration process was performed (step 5) against the Quranic words list. Finally, after removing the extremely diacritics words and the words containingtoo many letters (mostly incorrect words).A list of non-Quranic words isbuilt and labeled as "non-Quranic"(the output of steps 3 to 5 is (List B) a list of non-Quranic diacritic words).The generated list contains 148,333 non-Quranic diacritic words. For this dataset and its evaluation process half of these words are considered.

Next step is the merging of list A and list B to construct the compound list of words. From this compound list the features will be extracted as discussed in the subsequent section.

4.0 DATASET FEATURES, SPESIFICATIONS AND STATISTICS

In the proposed dataset, the Arabic letters, Arabic diacritics, and special symbols that appear in Quranic text are considered as the main features. The Figure 4 Example of words in compound list shows an example

of words in the generated compound list and their labels.

Non-Quranic	مسجدِ
Non-Quranic	وإرتفاعاً
Non-Quranic	يُدْرَكُ
Non-Quranic	تَأكُلُونَ
Non-Quranic	لحُكمه
Quranic	تَتقُونَ
Quranic	سَأُكُولِ
Quranic	فَاذْهَبْ
Quranic	قائِمًا

Figure 4 Example of words in compound list

From these main features, other four features are calculated including; number of letters; number of diacritics; percentage of letters; and the percentage of diacritics for each sample. Figure 5 Dataset features and their order shows the features and their order in the dataset.

	0	1	2	3	4	5	6	7	8	9
0		◌ۣ	◌ۢ	◌۬	ے	ۥ	◌۫	◌۪	◌ۨ	◌ۭ
1	◌۬	ٱ	◌ٰ	◌ً	◌ٌ	◌ٍ	◌َ	◌ُ	◌ّ	◌ِ
2	◌ٓ	◌ْ	◌ۜ	◌ٔ	◌۬	-	و	ى	ي	ف
3	ق	ك	ل	م	ن	ه	خ	د	ج	ح
4	ت	ث	ب	ة	ئ	ا	ؤ	إ	أ	ء
5	غ	ع	ظ	ط	ض	ص	ش	س	ز	ر
6	ذ	Dc	Lc	Dr	Lr					

Figure 5 Dataset features and their order

As seen in Figure 5 Dataset features and their order, there are 64 features divided in 3 categories as follows:

- First is the "Diacritics" category (the features surrounded by the bold border in the figure) consisting of 20 Arabic diacritics and other special Quranic symbols.
- Second is the "Letters" category consisting of 35 Arabic letters' shapes. Standard Arabic alphabet sethas 28 letters. However, some of these letters are written indifferent forms according to the position of the letter in the word and other considerations. For example, the first letter in Arabic alphabet which is "Alif" can be written in many forms such " أ ", " إ ", " ا ", and " آ ".In most of the Arabic text mining, classification, and language identification approaches, these forms are consolidated in one form[2].However, in Quranic studies, especially in identification and authentication, it is not proper to change or replace any of the original Quranic scripture properties. The letters category features are the features with shaded background in Figure 5 Dataset features and their order.
- Third is the "Statistical" features of each sample. It consists of four features including; Dc, Lc, Dr and Lr that represent Diacritics count, Letters count, Diacritics ratio, and Letters ratio, respectively.

These features are ordered in the dataset as f1, f2, …, f63, f64.To interpret the order of each feature from Figure 5 Dataset features and their order, consider the numbers in the most left column are tens, and the numbers in the upper row are ones. The order of the feature is the sum of its row's tens and column's ones. For example, the order of feature " ث "is 41, since it is in the row labeled 4 (i.e., 40) and column labeled 1 (i.e. 40+1 = 41), and so on.

Based on these features categories, Table 2 shows the main specification of the proposed dataset.

Table 2 Dataset specifications

Statistics	Features Category	Class		
		Non-Quranic	Quranic	All
# of samples in Dataset		74167	18994	93161
Percentage in Dataset		79.61	20.39	100.00
Mean Length of word		7.4191	9.946	7.934
Mean	Diacritics Count	5.08	4.98	5.06
	Letters Count	2.34	4.97	2.88
	Diacritics Ratio	0.700	0.498	0.659
	Letters Ratio	0.300	0.503	0.342
Maximum	Diacritics Count	11	11	11
	Letters Count	11	10	11
	Diacritics Ratio	0.92	0.80	0.92
	Letters Ratio	0.89	0.75	0.89
Minimum	Diacritics Count	1	1	1
	Letters Count	1	1	1
	Diacritics Ratio	0.11	0.25	0.11
	Letters Ratio	0.08	0.20	0.08

As seen in Table 2, the dataset contains non-Quranic and Quranic words of about 80% and 20%, respectively. The total number of samples in both classes is 93,161 samples. The average lengths of words are 7.4191 and 9.946 for both classes in same order, while the maximum and minimum values of diacritics count (Dc) and letters count (Lc) for both classes are approximately the same due to the filtration process conducted earlier during non-Quranic words generation. Figure 6 shows dataset's words length distribution.

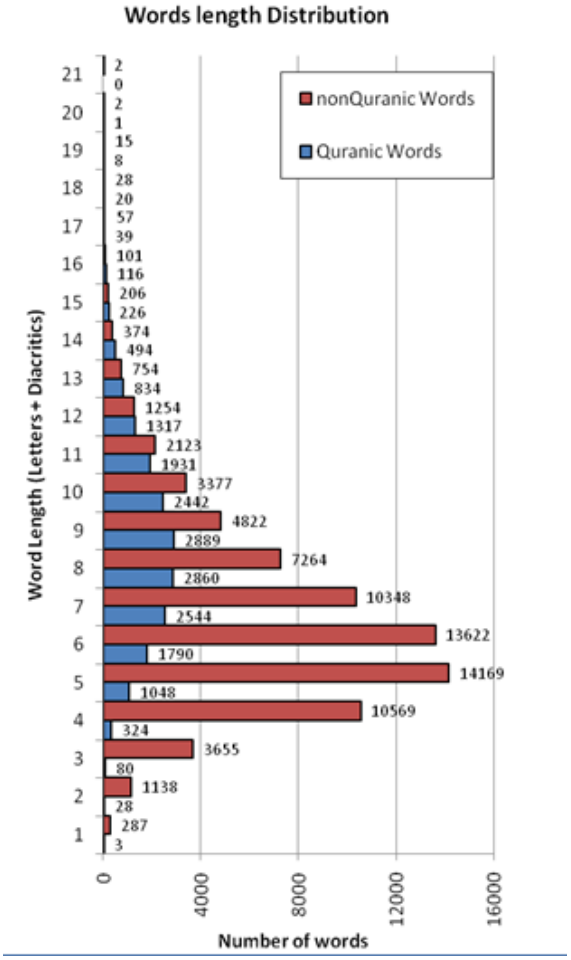


Figure 6 Class based Words length distribution

As can be seen from Figure 6 above, words lengths (length is the sum of Dc and Dl features) for both categories (Quranic and non-Quranic) is between 1 and 22. For Quranic category, highest frequency is on words with length of 9 characters, while the highest frequency is for words of length 5 characters in non-Quranic category. Figure 7 shows the word frequencies based on diacritics and letters count.

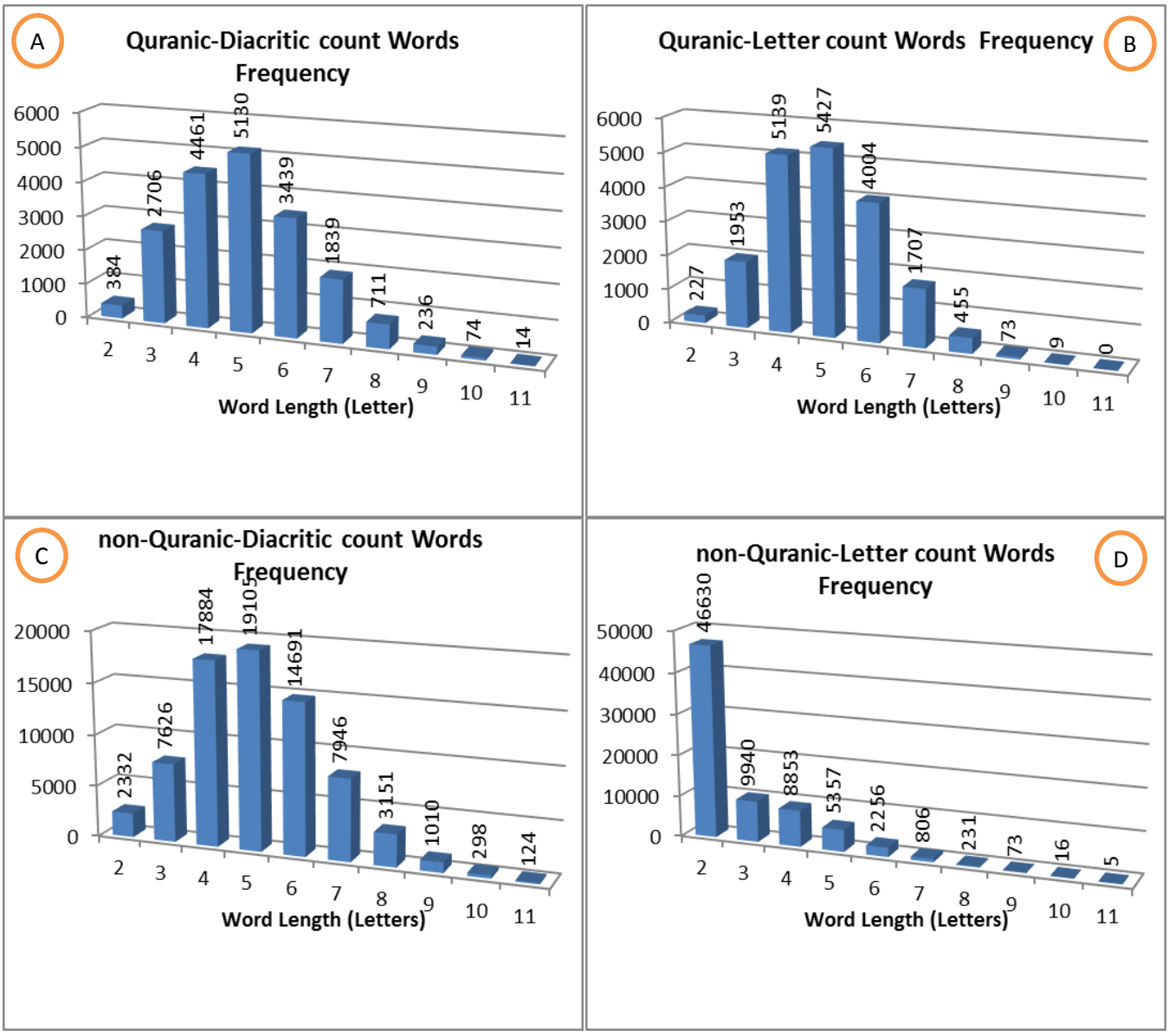


Figure 7 Diacritics and Letters count Frequencies

5.0 VALIDATION

To validate the proposed dataset, various tests based on Naïve Bayes (NB) classification algorithm are performed with ten-fold cross-validation. Table 3 shows the results of these tests in terms of accuracy, precision, and recall.

Table 3 Accuracy, Precision, and Recall results of Naïve Bayes Classification with Cross-Validation

Table 3 Accuracy, precision, and recall results of naïve bayes classification with cross-validation

Sample Size*	Accuracy	Quranic		Non-Quranic	
		Precision	Recall	Precision	Recall
5%	88.45	99.25	41.98	87.46	99.92
10%	87.84	100.00	40.99	86.71	100.00
20%	87.95	100.00	41.21	86.84	100.00
Mean	**88.08**	**99.75**	**41.39**	**87.00**	**99.97**

As seen in Table 3, the accuracy in the three experiments is between 87.84% and 88.45%, the average of accuracy is 88.08%, while the average precision is between 99.75% and 87.00% for Quranic and non-Quranic classes, respectively. The relatively high accuracy achieved compared to the limited percentage of samples used indicates the possibility of achieving higher accuracy in the case of using full dataset samples for training the classifier.

6.0 CONCLUSION AND FUTURE WORK

This paper has introduced a novel data set for Quranic words identification and authentication. The proposed dataset is created to overcome the problem of nonexistence of standard dataset to be used in the statistical machine learning Quranic studies. The dataset consists of all Quranic diacritic words (20% of the dataset), in addition to more than 74,000 diacritic non-Quranic words. Arabic letters, common Arabic diacritics, special Quranic symbols, and some statistical properties are extracted as the samples' features. Validation of the dataset shows an average accuracy of about 88% using the Naïve Bayes Classifier based on a balanced sample consisting of at most 20% of the dataset. The research community can freely access the proposed dataset by contacting the first author of this paper through http://thabitsabbah.webs.com/publications.htm. Our future work will focus on extracting more features and adding more non-Quranic diacritic samples for better classification results.

Acknowledgement

The Universiti Teknologi Malaysia (UTM) and Ministry of Higher Education (MOHE) Malaysia, under research grant R.J130000.7828.4F087 are acknowledged for some of the facilities utilized during the course of this research.

References

[1] Aabed, M. A., *et al.* 2007. Arabic Diacritics based Steganography. Signal Processing and Communications, 2007. ICSPC 2007. IEEE International Conference on. 2007.

[2] Alshareef, A. and A. E. Saddik. 2012. A Quranic Quote Verification Algorithm for Verses Authentication. Innovations in Information Technology (IIT), 2012 International Conference on. 2012.

[3] Alsulamy, E. 1999. Fundamentalists Used Quran and Sunni to Extract the Rules of Fundamentalism. Riyadh: Al Rushed library.

[4] Shamsudin, A. F. and A. Farooq. 2000. AI Natural Language in Meta-Synthetics of Al-Qur'an. TENCON 2000.

[5] Noordin, M. F. and R. Othman. 2006. An Information Retrieval System for Quranic Texts: A Proposed System Design. 2nd Information and Communication Technologies, 2006. ICTTA '06.

[6] Al-Khalifa, H. S., *et al.* 2009. SemQ: A Proposed Framework for Representing Semantic Opposition in the Holy Quran using Semantic Web technologies. International Conference on the in Current Trends in Information Technology (CTIT), 2009.

[7] Shoaib, M., *et al.* 2009. Relational WordNet model for semantic search in Holy Quran. International Conference on Emerging Technologies, 2009. ICET 2009.

[8] Baqai, S., *et al.* 2009. Leveraging Semantic Web Technologies for Standardized Knowledge Modeling and Retrieval from the Holy Qur'an and Religious Texts. Proceedings of the 7th International Conference on Frontiers of Information Technology 2009, ACM. Abbottabad, Pakistan. 1-6.

[9] Yauri, A. R., *et al.* 2012. Quranic-based Concepts: Verse Relations Extraction using Manchester OWL syntax. International Conference on Information Retrieval & Knowledge Management (CAMP), 2012.

[10] Mukhtar, T., H. Afzal, and A. Majeed. 2012. Vocabulary of Quranic Concepts: A semi-automatically created terminology of Holy Quran. 15th International in Multitopic Conference (INMIC), 2012.

[11] Tanzil.net. 2013. Who is using Tanzil?. [Online]. From: http://tanzil.net/wiki/Who_is_using_Tanzil%3F. [Accessed on 16 May 2013].

[12] Abbas, M. and K. Smaili. 2005. Comparison of Topic Identification Methods for Arabic Language. in RANLP05: Recent Advances in Natural Language Processing 2005. Borovets, Bulgary. 14-17.

[13] Abbas, M., K. Smaili, and D. Berkani. 2011. Evaluation of Topic Identification Methods on Arabic Corpora. *Journal Of Digital Information Management.* 9(5): 8.

[14] Abuaiadh, D. 2013. Dataset for Arabic document classification. 2013. [Online]. From: http://diab.edublogs.org/dataset-for-arabic-document-classification/. [Accessed on 26 June 2013].

[15] Zaidan, O. F. and C. Callison-Burch. 2013. Arabic Dialect Identification. *Computational Linguistics.*

[16] Zaidan, O. F. and C. Callison-Burch. 2011. The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content. In 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies 2011. Portland, Oregon, USA: Association for Computational Linguistics.

[17] Selamat, A. 2011. Improved N-grams Approach for Web Page Language Identification, in Transactions on Computational Collective Intelligence V, N. Nguyen, Editor. 2011, Springer Berlin Heidelberg. 1-26.

[18] Selamat, A. and C.C. Ng 2011. Arabic Script Web Page Language Identifications Using Decision Tree Neural Networks. *Pattern Recognition.* 44(1): 133-144.