Topics Classification of Arabic Text in Quran by Using Matlab

Abdelkrim El Mouatasim(✉) and Jaouad Oudaani

Laboratory LabSI, Ibn Zohr University – FPO, Ouarzazate, Morocco
a.elmouatasim@uiz.ac.ma

Abstract. The book of God (Quran) referenced by more than 1.6 billion of Muslims around the world. Extracting information from the Quran is of high benefit for both specialized as well as non-specialized people in religion. The Quran language is Arabic. Since the best software of text mining like Matlab and R doesn't sport Arabic language. However, this paper proposes a technical method for using Matlab text analytic toolbox for Arabic text.

The aim of this paper is to find the approaches for analysing Arabic text of Quran and then providing statistical information which might be helpful for the people in this research area, then different text mining operations are applied like wordcloud, word embedding, clustering, topic and classification. Also in this paper the classification of verses is given by topics using LDA, SVM and neural network.

Keywords: Arabic natural language processing · Matlab · Mathematical modelling · Quran & text mining

1 Introduction

The book of God (Allah) Quran with Arabic text referenced by more than 1.6 billion of Muslims around the world. The Arabic language is a unique language, and has many special and unique features which make it suitable for it to convey; many meaning in few words, subtleties, emphasis and powerful imagery through speech alone. If Allah was to convey a message to mankind, it would be through a language which is easy to learn, and has the highest form of expressiveness.

Arabic is a language based on a system of 'roots'. In English, we often refer to the 'root' of a word to mean its origin. The Arabic root, or masdar (مصدر), refers to the core meaning of a word. This core can usually be identified by root consonants [1]. Using derivation system of roots and patterns, nouns, and verbs are derived in an almost mathematical way, leaving little room for confusion as to the desired meaning of the word. Of course the ideal model of this derivation is the Quran, and as you look through the Quran you will see these in play.

Few research studies have considered the Arabic text of Quran, mathematically-based studies [2], linear regression models [3] and text mining using R [4]. To the best of our knowledge, there is no research study that analyzed the Arabic text of the Quran using Matlab text analytic toolbox (wordcloud, word embedding, topic and

Y. Farhaoui and L. Moussaid (Eds.): ICBDSDE 2018, SBD 53, pp. 333–341, 2019.
https://doi.org/10.1007/978-3-030-12048-1_34

classification..) the way it is done in this paper. Also in this paper we use a topics finned by LDA method for classification the verses of Quran.

2 Quranic Arabic Text Mining

2.1 Preparing Quran for Analysis

The Quran has 78246 words. These words are grouped into 6236 verse (اية). A set of verses are grouped into 114 chapter (سورة).

The text of the Quran has been downloaded from Tanzil project website [5], which represents an authentic verified source of the Quran text. The downloaded file includes the whole text of the Quran without diacritic.

After that the encoding of the files have been converted by using UTF-8 because the original encoding of the files are unreadable by Matlab.

The files have then been read as a corpus and cleaned by removing stop words. Matlab does not support stop word removal for Arabic language, hence list of stop words have been created.

After remove a list of stop words and normalize the words using the Porter stemmer, the corpus have been converted into both tokenized document and bag of words.

2.2 Word Cloud

It is important for specialized as well as non-specialized people in Islamic studies to visualize the words of the Quran.

Figure 1 show the wordcloud of the raw that and clean data in Quran for the most frequent words measured using beg of words with reduction of 0.0119.

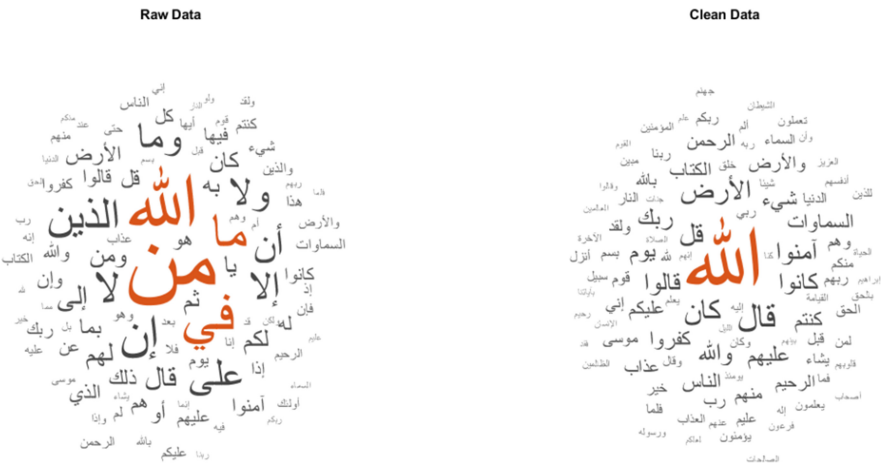

Fig. 1. Wordcloud for the most frequent words in Quran

2.3 Most Frequent Words

The bag of words has been used for calculate the frequency of the words in Quran. There are many frequent words in Quran, hence Fig. 2 depicts the 20 frequent words.

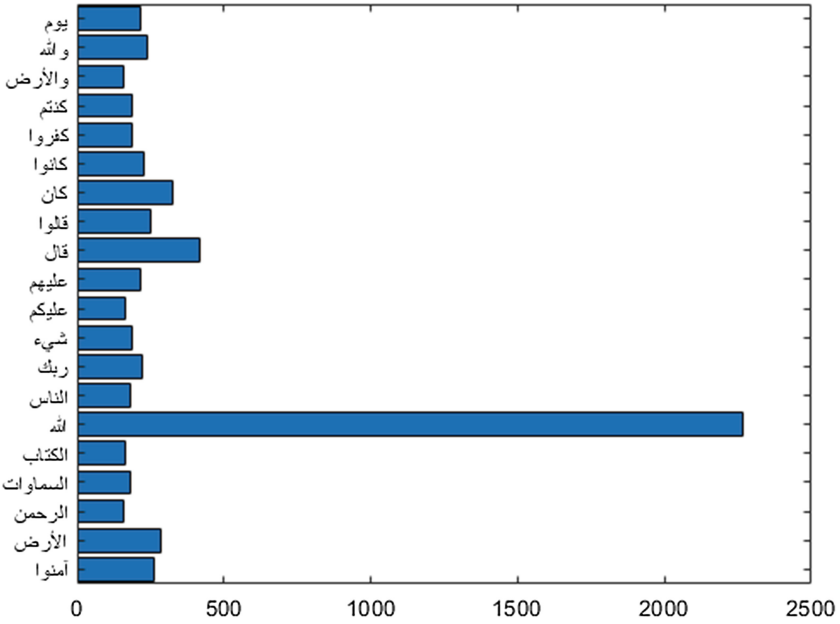

Fig. 2. Most frequent words in Quran

3 Statistical Description

3.1 Grouped Data

In this section we considered a population as words in Quran, and there statistical variables as frequency number of word in Quran (Fig. 3 and Table 1).

Table 1. Frequency distribution for frequency words in Quran

Classes of frequency word	0.5–1.5	1.5–2.5	2.5–3.5	3.5–4.5	4.5–5.5	5.5–7.5	7.5–10.5	10.5–15.5	15.5–24.5	24.5–50.5	50.5–2763
Frequency	8663	2383	1002	609	399	473	336	275	207	161	121

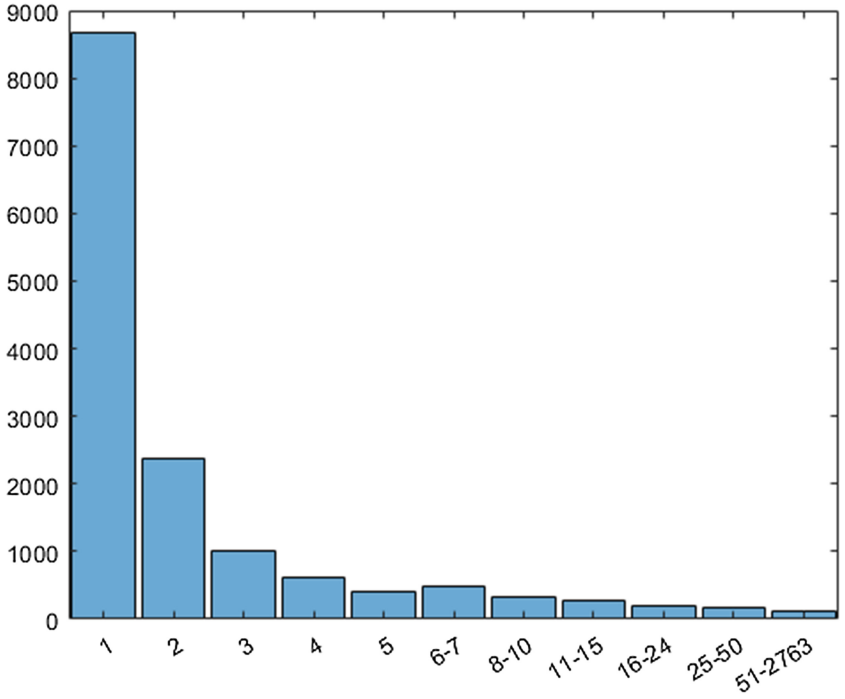

Fig. 3. Histogram of frequent words in Quran

3.2 Measuring Parameters

In this section important measurements of words in Quran are computed for ungrouped clean data as described in Table 2.

Table 2. Measurements of words in Quran

Statistic	Sum	Range	Mode	Median	Mean	Variance	Skewness	Kurtosis
Value	54440	2264	1	1	3.7214	482.8432	76.5947	7691

Table 2 shows some information about the pattern nature of the frequency words in Quran. It is clear that the mean > median = mode.

Also, the skewness is positive. This means that the data distribution are not symmetric and skewed to the right.

4 Embedding Word

4.1 Word2vec

Word embedding methods define all techniques to construct multi dimensional vector representations. Previously, we discussed the parallels between neural structure and linguistic, indicating that artificial neural networks of machine learning might be the adequate means to model natural language. Word2vec is a word embedding method that takes a corpus of words as input and produces vectors as output.

Besides two general architectures and different learning objectives, it offers several parameters which enable variations in training as well as choices between optimization techniques such as hierarchical softmax and negative sampling. Here, we propose to define word2vec method for Arabic Quran by using Matlab text analytics toolbox.

Train a word embedding using the Quran data quran-simple-clean.txt. The file contains one verse per line, with words separated by a space. Extract the text from quran-simple-clean.txt, split the text into documents at newline characters, and then tokenize the documents with Arabic stop words. The results are: Training: 100\% Loss: 2.81037 see also Fig. 4. Plot the words at the coordinates specified by XYZ in a 3-D text scatter plot, with loss: 2.80706, see Fig. 5.

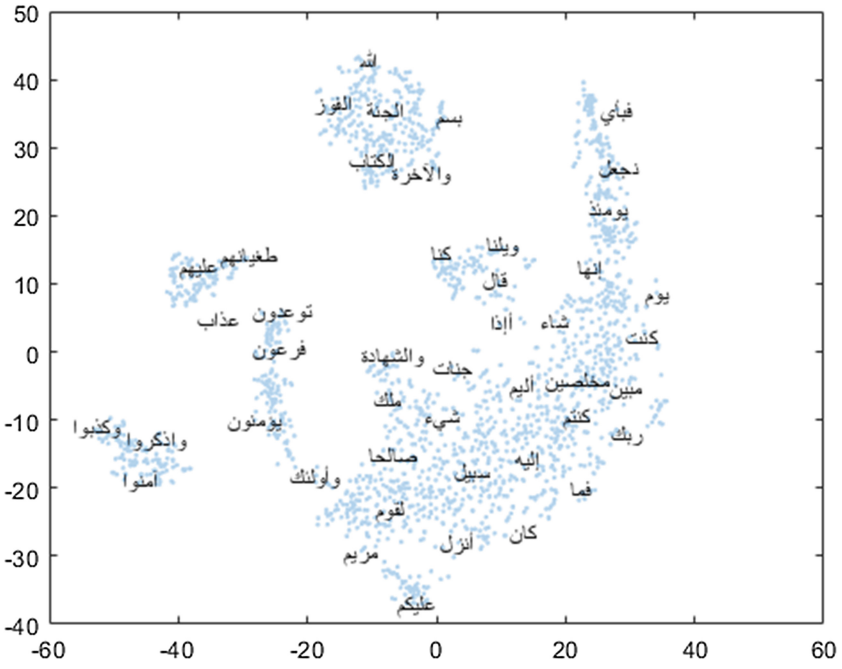

Fig. 4. Word embedding t-SNE 2D plot of Quran

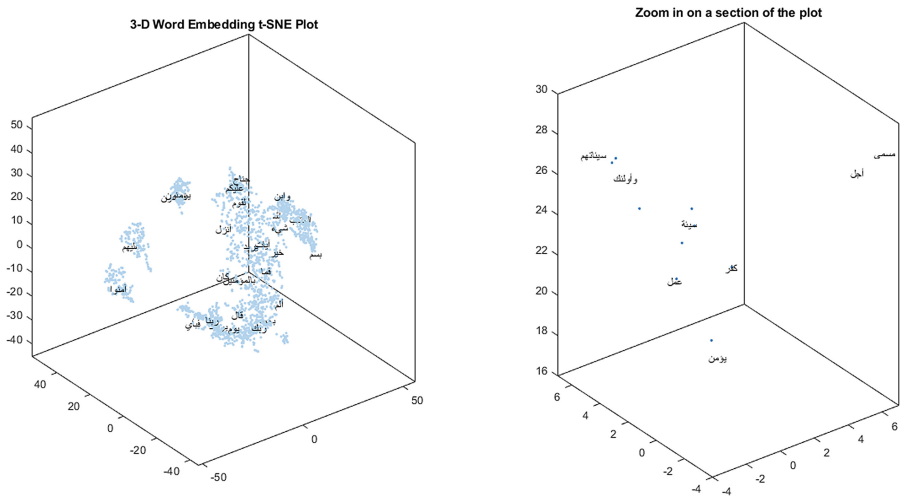

Fig. 5. 3D word embedding t-SNE plot of Quran

4.2 Clustering

We discover 7 clusters using kmeans, and visualize the clusters in a text scatter plot using the 2-D t-SNE data coordinates calculated earlier, see Fig. 6.

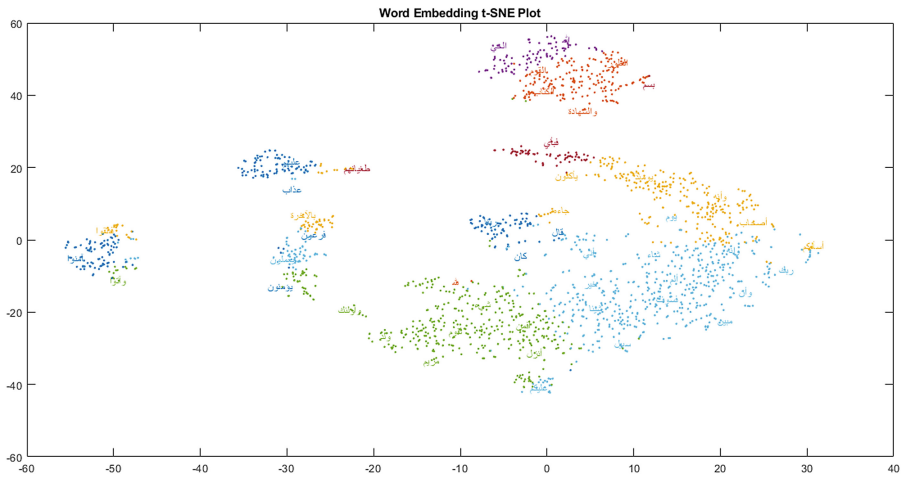

Fig. 6. Clusters of Quran

5 Topic

5.1 Latent Dirichlet Allocation

A Latent Dirichlet Allocation (LDA) model is a topic model which discovers underlying topics in a collection of documents and infers the word probabilities in topics.

To decide on a suitable number of topics, you can compare the goodness-of-fit of LDA models fit with varying numbers of topics. You can evaluate the goodness-of-fit of an LDA model by calculating the perplexity of a held-out set of documents. The perplexity indicates how well the model describes a set of documents. A lower perplexity suggests a better fit.

5.2 Choose Number of Topics

The goal is to choose a number of topics that mini the perplexity is lowest compared to other numbers of topics. This is not the only consideration: models fit with larger numbers of topics may take longer to converge. To see the effects of the tradeoff, calculate both goodness-of-fit and the fitting time. If the optimal number of topics is high, then you might want to choose a lower value to speed up the fitting process (Fig. 7).

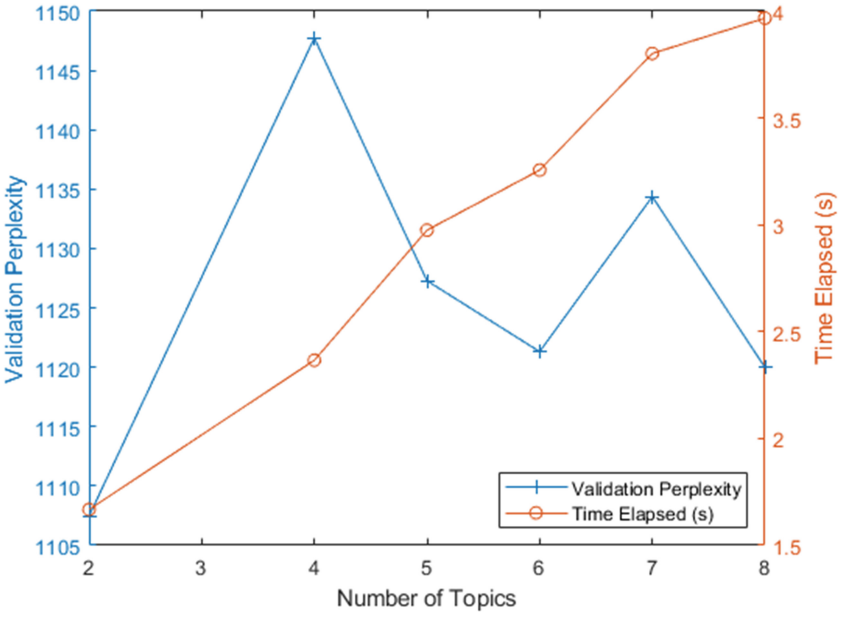

Fig. 7. Number of topic in Quran

5.3 Topic Probabilities

Use transform to transform the documents (إياك نعبد وإياك نستعين) into vectors of topic probabilities (Fig. 8).

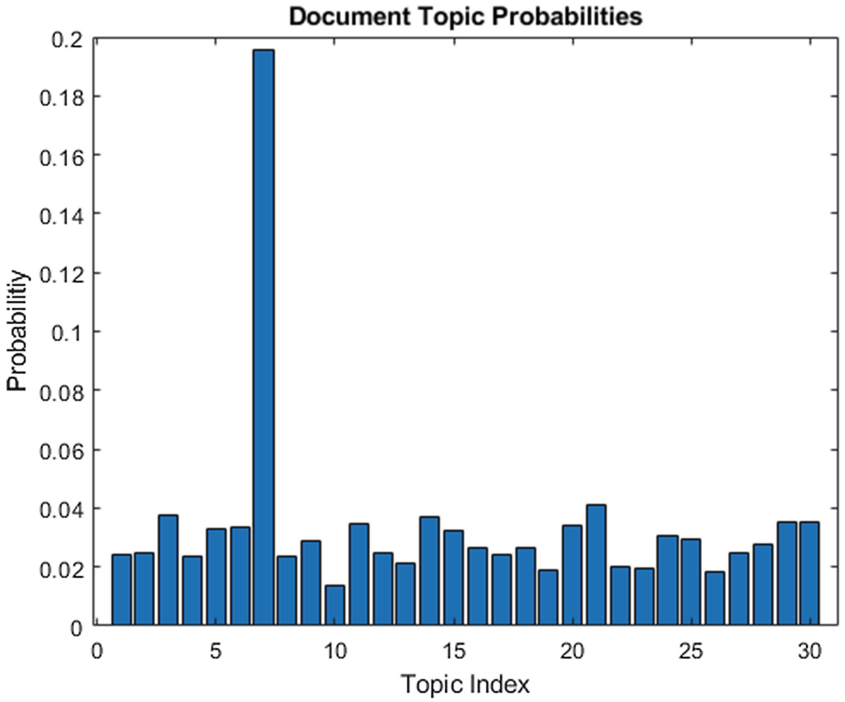

Fig. 8. Document topic of 5 verses in Alfateha chapter

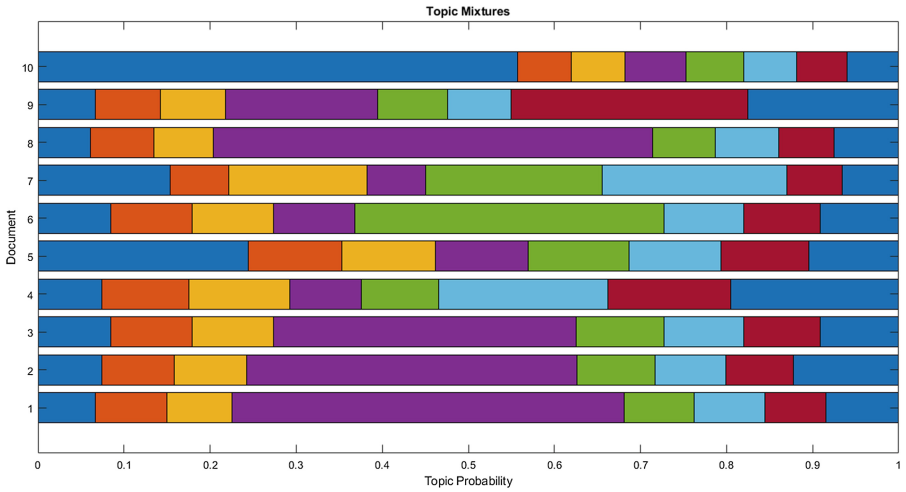

Fig. 9. Mixture topic

Visualize multiple topic mixtures using stacked bar charts. Visualize the topic mixtures of the first 10 input documents (Fig. 9).

5.4 Classification

We create a simple classification model which uses word frequency counts as predictors topicIdx = predict(ldaMdl, documents) returns the LDA topic indices with the largest probabilities for documents based on the LDA model ldaMdl.

6 Conclusions

This paper aims to layout a framework for future work that is related to the application of Arabic natural language processing, statistical and data mining to the text of Quran via Matlab.

Important a result in this paper is that a bag-of-word and tokenize a document are suggested to be used in different application like semantic search, vec2word, clustering, topic modelling and classifications.

Future work may include a text mining of Quran with optimization algorithms. If such algorithms are developed then further study on the Arabic Quranic text will be carried out to extract knowledge and important information that is useful to all humanity using machine learning techniques.

References

1. Abdul Sattar, H.: Conjugating Regular Verbs and Derived Nouns. Sacred Learning Publishers, Carrollton (2012)
2. Al-Faqih, K.M.: A mathematical phenomenon in the Quran of earth-shattering proportions: a Quranic theory based on gematria determining Quran primary statistics (words, verses, chapters) and revealing its fascinating connection with the golden ratio. J. Arts Humanit. **6**(6), 52–73 (2017)
3. El Mouatasim, A.: Simple and multiple linear regression of verbs in Quran. Am. J. Comput. Math. **8**, 68–77 (2018)
4. Alhawarat, M., Hegazi, M., Hilal, A.: Processing the text of the Holy Quran: a text mining study. Int. J. Adv. Comput. Sci. Appl. **6**(2), 262–267 (2015)
5. Tanzil.net: Tanzil Quran text download (2014). http://tanzilnet/download/