Classifying Verses of the Quran using Doc2vec

Menwa Alshammeri[1,2] **Eric Atwell**[2] **Mohammad Ammar Alsalka**[2]
[1]Jouf University, College of Computer and Information Sciences, Sakaka, Saudi Arabia
[2]University of Leeds, School of Computing, Leeds, UK
{scmhka, E.S.Atwell, M.A.Alsalka}@leeds.ac.uk

Abstract

The Quran, as a significant religious text, bears important spiritual and linguistic values. Understanding the text and inferring the underlying meanings entails semantic similarity analysis. We classified the verses of the Quran into 15 pre-defined categories or concepts, based on the Qurany corpus, using Doc2Vec and Logistic Regression. Our classifier scored 70% accuracy, and 60% F1-score using the distributed bag-of-words architecture. We then measured how similar the documents within the same category are to each other semantically and use this information to evaluate our model. We calculated the mean difference and average similarity values for each category to indicate how well our model describes that category.

1 Introduction

The richness of the Quran and the deep layers of its meaning offer immense potential for further study and experiments. The knowledge in the Quran was presented using different approaches, mainly using the tree-structure hierarchy (Ta'a et al., 2014). As a result, determining a concept's true meaning in the Quran is difficult. We want to classify the Quran verses based on topics or meanings to assist users in identifying the religious knowledge explained in the Quran. There has been previous work on classifying textual documents and sentences in English and Arabic (Al-Kabi et al., 2013). However, only a few studies in the literature attempt to classify the verses of the Holy Quran (Al-Kabi et al., 2013; Al-Kabi et al., 2005; Ta'a et al., 2014; Akour et al., 2014).

Therefore, using NLP combined with ML, this paper presents an approach to classifying the Quran based on topics and meanings.

To do so, we need to compute the similarity in meaning between its passages. We focus on sentence/ paragraph levels. Therefore, we represent the verses of the Quran as vectors of features and compare them by measuring the distance between these features. We use Doc2ve[1] to compute features that capture the semantics of the Quranic verses. We then train a logistic regression classifier in a supervised way to learn the underlying meanings and classify the verses of the Quran into fifteen predefined classes or categories. We then use the cosine similarity measure on the vectors to examine how semantically similar the verses are in each class. We compute two metrics: average similarity and mean similarity difference to inspect the relation between the verses in the same class and other classes. This information indicates how more similar same-category documents are to each other than to documents from different categories. A higher average similarity indicates how similar the documents are in each category. A higher mean difference implies that the model can identify those documents in one class are more distinct from those in other classes. Since we are interested in a topical classification, we use the Qurany corpus to train and evaluate our model.

The rest of the paper is organized as follows: Section 2 presents studies related to the classification of the verses of the Quran. Section 3 describes our approach to classifying the Quranic verses and our experiments. Section 4 presents our evaluation and results. Finally, section 5 states our conclusions and future research directions.

[1]https://radimrehurek.com/gensim/models/doc2vec.html

Proceedings of the 18th International Conference on Natural Language Processing, pages 284–288
Silchar, India. December 16 - 19, 2021.

2 Related Work

This section briefly reviews previous work conducted on the topical classification of holy Quran verses.

Hamed and Ab Aziz (2018) proposed a Quran classification using the Neural Network classifier based on the predefined topics. The study used the English translation of the Holy Quran. They applied the classification to Al-Baqarah chapter as it contains many commands and topics. They classified the verses of Al-Baqara into two classes, Fasting, and Pilgrimage. The thesis of Al-Kabi et al. (2013) is restricted in topical classification of only two Quran chapters: Fatiha (7 verses) and Yaseen (83 verses). Another study (Al-Kabi et al., 2005), evaluated the effectiveness of four well-known classification algorithms: Decision Tree, K-Nearest Neighbor (K-NN), Support Vector Machine (SVM) and Naïve Bayes (NB), to classify Quran verses according to their topics. They used the manual topical classification of Quranic verses by (Abu Al-Khair and Kabbani , 2003) to train and evaluate the four classifiers. Three selected topics (classes) are used, and 1,227 verses were used in this study out of 6236 verses in the whole Quran. Another classification has been presented by Qurany (Abbas, 2009). This project annotates the verses of the Qur'an with a comprehensive index of nearly1100 topics; it classifies the Qur'an into fifteen main themes and subdivides the main themes into sub-themes.

In this work, we exploit the distributed representation of text to capture the semantic properties of the 6236 Arabic verses of the Quran. We transformed the verses of the Quran into a numerical form, which can be used as input to ML methods to examine the semantic similarity between the Quranic verses and classify them into topical classes.

3 Methodology

The objective of this experiment is to evaluate our model in capturing the semantic properties of verses of the Quran. Therefore, we examine our model on the following tasks:

1. Classify the verses of the Quran into 15 pre-defined categories or classes using Doc2Vec and Logistic Regression.
2. Measure how similar the verses within the same category are to each other semantically, and use this information to evaluate our model.

3.1 The Data

For the purpose of training and testing our model, we create a dataset that contains the 6236 verses of the Quran categorized into 15 main topical themes; based on Qurany[2] corpus. Table 1 shows the high – level concepts from Qurany corpus.

Main Concept	English Concept
أركان الإسلام	Pillars of Islam
الإيمان	Faith
القصص والتاريخ	The Stories and The History
القرآن الكريم	The Holy Quran
العمل	The Work
الإنسان والعلاقات الأخلاقية	Man, and The Moral Relations
الإنسان والعلاقات الاجتماعية	Man, and The Social Relations
الجهاد	Jihad
العلوم والفنون	Science and Art
الديانات	Religions
تنظيم العلاقات المالية	Organizing Financial Relationships
الدعوة إلى الله	The Call for Allah
العلاقات القضائية	Judicial Relationships
التجارة والزراعة والصناعة والصيد	Trade, Agriculture, Industry and Hunting
العلاقات السياسية والعامة	General and Political Relationships

Table 1: The high-level concepts from the Qurany corpus

Each verse is annotated with a sequence of concepts besides the main concept/theme. The verses are split into training and test sets. Table 2 shows an example of the dataset.

[2] http://quranytopics.appspot.com

Chapter	Ayah	Verse Text	
2	3	الذين يؤمنون بالغيب ويقيمون الصلاة ومما رزقناهم ينفقون	
Translation		**Class**	**Trans**
Who believe in the Unseen, and establish worship, and spend of that We have bestowed upon them;		أركان الإسلام	Pillars of Islam

Table 2: Example of the verse annotation from the dataset

3.2 Classifying the Quran Verses using Logistic Regression

Following our approach in (Alshammeri et al., 2020), we map the verses of the Quran to numerical vectors. We use ML model on our vectorized verses to classify the Quran verses into the associated concepts or classes. We set up the train/ test documents, pre-process them to be ready for training and classification. We train the Doc2vec model using the training set and generate the vectors. We then build the vector representations for the classifier; we infer vectors for the documents in the test set using the trained model. Then we train the logistic regression classifier.

3.3 Testing Category-Wise & Cross-Category Verses Similarity

Documents belonging to the same category would seem to be more similar than documents belonging to different categories, intuitively. And that's how we judge our model: a good model should generate higher similarity values for verses in the same category than for verses from different categories.

4 Evaluation and Results

4.1 Classification Results

We pre-processed the documents and transformed them into a numerical, vectorized form by training a Doc2vec model on our data. We inferred new vectors for unseen verses/documents from the test set. We tried different configurations for the hyperparameters of the Doc2vec model, we then trained the classifier with the different versions of the embeddings. Using 80% of the data set to train doc2vec classifier for the Quran verses classification, we achieved 68% accuracy, and 56% F1-score; using the distributed bag of words. We have noticed that changing the vector size did not have a big impact on the classifier performance, but with more training data, the accuracy rises to 70%, and F1- score to 60%. Table 3 shows the classifier performance results using different settings of the model: Distributed Bag-of-words (PV-DBOW) and Distributed Memory (PV-DM).

4.2 Categories Similarity Results

To inspect relationships between the verses numerically, we calculated the cosine distances between their inferred vectors from the trained Doc2vec model. We used this information to calculate the average similarity scores and the

Train set: 80% / Test set: 20%			
Doc2vec model	Vector size	Testing Accuracy	Testing F1score
PV-DBOW	50	0.68	0.56
	100	0.68	0.55
PV-DM	50	0.64	0.54
	100	0.64	0.54
Train set: 90% / Test set: 10%			
Doc2vec model	Vector size	Testing Accuracy	Testing F1score
PV-DBOW	50	0.70	0.60
	100	0.70	0.58
PV-DM	50	0.64	0.55
	100	0.64	0.54

Table 3: Classification Performance Results

mean difference for each category. We want to know how much more similar the same-category documents are to each other than to documents from other categories. Therefore, we created sets of verses pairs for all categories. More precisely, given our fifteen categories which we denoted by $C_1, \ldots, C_{15}$, where each category is a set of verses. Hence, we derived 15 average similarities per each category; one for same-category documents and 14 for cross-category documents. Finally, we calculated the mean similarity differences between the cross-category average similarities and the same-category average similarity. A higher mean difference implies that the model is able to identify documents in one category are more distinct from those in other categories. We experimented with the two architectures of Doc2vec model.

Here, we show the results of PV-DBOW architecture, with vector size of 50. We didn't include the results using vectors size of 100 as no impact were noticed. The result of the evaluation can be summarized as in Table 4.

5 Conclusion

We used NLP combined with ML to classify the verses of the Quran into 15 predefined classes. The semantics of the verses were captured using Doc2vec embeddings that were used to group similar documents. Our model achieved a classification accuracy of 68% and 56% F1 score. The results confirmed that the classifier scored higher accuracy with the distributed bags of words architecture of the Doc2vec model. Next, we evaluated our model by examining the semantic similarity of the Quranic verses. Derived classes showed relatively high average similarity for some classes using the distributed bags of words architecture. The three top classes that achieved higher average same-category similarity and mean-difference are Faith, Pillars of Islam, and Religions. The three classes scored top values consistently for both metrics with different runs (500, 700, 900, 1100 test samples). The two metrics are not relatively high for some classes. We contribute that to some classes' documents similar to those of another class. Besides, some verses in the Quran discuss more than one concept/ topic. The uniqueness and complexity of the Quran language could also be a significant reason reflected in our results. Table 5 shows an example of an instance where a verse belongs to different classes/ topics. The results confirmed that the class "Faith" has achieved the highest average similarity and mean difference.

In the future, we may incorporate subtopic chains from Qurany corpus, and we may consider creating unique classifications of Quran verses using existing knowledge resources, and test our model using them. We may consider other approaches for computing the semantic similarity, investigate their performance, and how they compare to our approach.

English Category	(Mean Difference, Same-category Avg. Similarity)
Pillars of Islam	(11%, 15%)
Faith	(21%, 26%)
Man, and The Moral Relations	(2%, 0.74%)
Stories and History	(5.4%, 5.6%)
The Holy Quran	(2.8%, 3.7%)
The work	(2.9%, 6.6%)
Man, and The Social Relations	(3.8%, 7.8%)
Jihad	(0.92%, 0.16%)
Science and Art	(7.2%, 5.3%)
Religions	(13%, 19%)
The Call for Allah	(3.2%, 6.5%)
Trade, Agriculture, Industry and Hunting	(8.6%, 8.7%)
Judicial Relationships	(4.5%, 8.2%)
Organizing Financial Relationships	(2.6%, 4.9%)
General and Political Relationships	(1.5%, 2.7%)

Table 4: Evaluation Results using PV-DBOW, Vector-size =50, # of test samples = 1100.

Verse	**Al-Nahl / سورة النحل (16, 94)**
Arabic Verse	**"وَلاَ تَتَّخِذُواْ أَيْمَانَكُمْ دَخَلاً بَيْنَكُمْ فَتَزِلَّ قَدَمٌ بَعْدَ ثُبُوتِهَا وَتَذُوقُواْ السُّوءَ بِمَا صَدَدتُّمْ عَن سَبِيلِ اللهِ وَلَكُمْ عَذَابٌ عَظِيمٌ"**
English Translation	**And make not your oaths a means of deceit between you, lest a foot should slip after its stability, and you should taste evil because you hinder (men) from Allah's way and grievous chastisement be your (lot).**
Topic	**-Judicial Relationships** **-Jihad**

Table 5: An example of an instance where a verse belongs to different classes/ topics.

Acknowledgments

The first author is supported by a PhD scholarship from the Ministry of Higher Education, Saudi Arabia. The author is grateful for the support from Jouf University for sponsoring her research.

References

Abu A. Al-Khair, and M. Kabbani. "Koran teacher intonation." The Tunisian Company for Distribution, 2003.

Azman Ta'a, Mohd Syazwan Abdullah, Abdul Bashah Mat Ali, and Muhammad Ahmad. "Themes-based classification for Al-Quran knowledge ontology." In 2014 International Conference on Information and Communication Technology Convergence (ICTC), pp. 89-94. IEEE, 2014.

Menwa Alshammeri, Eric Atwell, and Mhd Ammar Alsalka. "Quranic Topic Modelling Using Paragraph Vectors." Proceedings of SAI Intelligent Systems Conference. Springer, Cham, 2020.

Mohammed Akour, Izzat Alsmadi, and Iyad Alazzam. "MQVC: Measuring quranic verses similarity and sura classification using N-gram." 2014.

Mohammed Naji Al-Kabi, Ghassan Kanaan, Riyad Al-Shalabi, Khalid Nahar, and Basel Bani-Ismail. "Statistical classifier of the holy Quran verses (Fatiha and Yaseen chapters)." Journal of Applied Sciences 5, no. 3 (2005): 580-583.

Mohammed Naji Al-Kabi, Belal M. Abu Ata, Heider A. Wahsheh, and Izzat M. Alsmadi. "A topical classification of Quranic Arabic text." NOORIC 2013: Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, 2013.

Noorhan Hassan Abbas. "Quran'search for a concept'tool and website." PhD diss., University of Leeds (School of Computing), 2009.

Suhaib Kh. Hamed, Mohd Juzaiddin Ab Aziz. "Classification of holy quran translation using neural network technique." Journal of Engineering and Applied Sciences 13, no. 12 (2018): 4468-4475.