VAWKUM Transactions on Computer Sciences
http://vfast.org/journals/index.php/VTCS@ 2024, ISSN(e): 2308-8168, ISSN(p): 2411-6335
Volume 12, Number 4, October-December 2024 pp: 239-251

Topic Modeling of Quranic Verses using Latent Dirichlet Allocation with English Language

Kashmala Jamshaid Akhter [1], **Humera Farooq** [1*], **Muhammad Tariq Siddique** [1]

[1]Department of Computer Science, Bahria University, Karachi, Pakistan

Keywords: *Quran Dataset, Natural Language Processing, Topic Modelling, Visualization.*

Journal Info:
Submitted:
October 20, 2024
Accepted:
December 20, 2024
Published:
December 31, 2024

Abstract

This study aims to assess the effectiveness of topic modeling in the English translation of the Holy Quran. Topic modeling is a popular text mining technique for uncovering latent semantic patterns in the collection of textual documents and helps to annotate the documents based on these topics. This study identifies the most significant topics in each document as well as grasping an understanding of the topic distribution throughout the document sets. Different steps are performed to acquire the dominant topics in each document and identify the distribution of topics across documents. In this context, the present research work chose to employ Latent Dirichlet Allocation as an unsupervised approach for topic modeling since there is no requirement for a training phase as hidden topics can be discovered throughout the topic modeling process. For this, the word cloud is generated to understand and interpret the results after pre-processing. A dictionary and corpus are created to extract the features from the dataset using the Bag of Words approach. The results are evaluated by calculating the perplexity and coherence score, where high coherence indicates the goodness of well-structured topic models and low perplexity score indicates the correctness of prediction made by the topic models. Lastly, the visualization step is performed.

***Correspondence author email address:** humerafarooq.bukc@bahria.edu.pk
DOI: 10.21015/vtse.v12i4.1946

1 Introduction

Natural Language Processing (NLP) is a set of theoretically computational methods used for evaluating and modeling naturally striking texts [1]. It helps to enable human-like language that proceeds at a singular or many levels of linguistic analysis handling activities and application at various levels [2]. Over the last several years, knowledge discovery has become the greatest significant research topic in the realm of NLP. To extract usable information from source documents, knowledge extraction research employs a range of methodologies, such as stopwords, chunking, Part-of-speech (POS) tagging, stemming, and word sense disambiguation [3].

The Quran is a religious text revealed to the Prophet Muhammad (Peace Be Upon Him) over 14 centuries ago. It is considered as a spiritual and comprehensive guidance for Muslims all over the world. Scholars all over the world always research to get in-depth meanings and important topics from the Holy Quran by translating it into different languages. While living in the modern era of Artificial Intelligence, researchers worldwide apply different techniques to understand the semantic meanings and build models to explore the hidden topics or thematic representations of Holy verses. However, manual research interpretation and translation is time-consuming due to the overlap of verses from chapter to chapter [4]. While working in Natural Language Processing, existing literature reported different studies to solve these issues including extracting the hidden topics from the documents [3].

Topic modeling can be defined as a form of modeling through statistical conduct used to find hidden topics that appear in various documents. It generally refers to the process of determining which topics best characterize a group of documents. Therefore, these topics will only appear as a result of the topic modeling procedure [5]. According to [6], it is an unsupervised method of identifying or extracting themes by detecting patterns and is quite similar to clustering algorithms. According to [7], the terms are depicted as a combination of the topic in the modeling enabling different topics, with each topic being a distribution of probability over the vocabulary words. The method employs the thematic framework to understand structural correlations and develop relationships among vocabulary as well as documents. Furthermore, it is significant to concentrate on the point that topic modeling has been applied in different research areas such as data mining in healthcare[8].

2 Related Studies

The existing literature shows that researchers proposed different methods to address the challenge of topic classification for the Quranic Text. A study conducted by [9] performed the topic modeling for the Indonesian translation of the Holy Quran for daily life. A theme-based classification of Quranic topics has been performed using the Ensembling technique. To extract the semantic features, Word2vec has been applied [10]. The variability in the contents and context of different Urdu translations of the Quran has been analyzed using topic modeling techniques with BERTopic [11]. The detection of textual content (Islamophobia) was performed using text mining and NLP techniques. The study applies topic modeling using LDA. For feature extraction GloVe and Word2Vec and text classification BERT and GPT have been used [12].

Keeping in view the imbalanced classification problem of the dataset for the Holy Quran different techniques have been proposed [13][14]. Ensemble learning (boosting and bagging) has been proposed to classify the Quranic topics by [13]. Whereas in [15], random over-sample (ROS), synthetic minority over-sampling technique (SMOTE), and random under-sample (RUS) methods are used to classify Quranic topics [14]. The Ensemble Method with Naïve Bayes has been used to classify the Quranic verses by considering it as a multi-label classification problem [16].

Word Centrality measurements have been used to classify Quranic verses into 10 topics. They also did a comparison of different Machine Learning (ML) techniques such as the k-Nearest Neighbor algorithm, Decision Tree, Naive Bayes, and Support Vector Machine and evaluated the performance of each classifier [17]. Similarly, another study reported applying a Genetic Algorithm to classify the Quranic topics [18]. Recently, Deep Learning (DL) models have also been applied for the classification of Quran verses [19][20]. The word embedding has been used with Naïve Bayes classifiers [21]. Machine Learning techniques are used to identify similarities and differences between ten different English translations of the Quran. For the experiments, the authors have selected two surahs (Al-Humazah and An-Nasr) [4]. Semantic similarity has been detected using the Siamese transformer-based architecture for the Quranic dataset [22].

A generative model has been proposed that is capable of modeling topics for unseen documents

[15]. The elementary concept is that text documents are depicted as subjects of latent with random mixes, where each subject is defined through word distribution. Furthermore, it is observed that Latent Dirichlet Allocation (LDA) has a significant advantage over Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA) in terms of dimensional reduction and through the potential used in more compound techniques [23]. In [24], the authors performed a comprehensive survey for topic modeling and LDA.

The presented research aims to see how the LDA technique affects the topic modeling for Quranic texts. The reason for selecting LDA as there is no requirement for a training phase as hidden topics can be discovered throughout the topic modeling process. The proposed framework serves as a guide to analyze the impact of using LDA for the topic modeling of Quranic verses and to provide a substantial contribution to the study of LDA models for the English language and its impacts, specifically on the translation of Quranic verses.

3 Proposed Methodology

The primary goal of this research study is to conduct a thorough and comprehensive investigation that would address the present gap in previous studies about LDA implementation on the dataset of English translation of the Holy Quran. The proposed framework includes the Acquisition of data, Pre-processing, and Feature Engineering using the bag representation of words. It is critical to conduct it correctly, since it turns textual data into a numerical representation, allowing the model to comprehend the data and begin training on it. After that, the probabilistic topic modeling approach LDA is utilized to develop the topic model for the translation of Quranic verses. The LDA model belongs to the category of unsupervised machine learning, accepts documents as input, and presents topics as output. The terms of the topic connected with the document can be explored using this methodology. It also aids the researcher in compiling a list of major keywords for each topic. In addition to this, the model is evaluated so that the findings of the study can be interpreted accurately. Since an LDA model has been established, the next step is to assess the formed topics and associated keywords. As a result, the interactive optimum method is used to visualize the dominant topics. Figure 1 shows the framework of the proposed work.

3.1 Pre-processing Steps

It is observed that there is no single and consistent source of the dataset which can be utilized for the research. Therefore, the investigator of the study decided to use the open-source Quran dataset from Kaggle for the English language in CSV format [25]. Since the focus of the study is on the modeling of the English dataset for Quranic verses, the surah and ayah columns were omitted during the data preparation phase. The purpose of pre-processing is to clean up the raw noisy dataset so it can be used in the subsequent steps. It is the most critical phase after data gathering. The pre-processing techniques are adopted from the study conducted by [26]. Besides this, [27] research is considered as it briefly examines the pre-processing, feature extraction, and other aspects of the English translation of Quranic verses. Moreover, the researcher decided to look into two more studies in the realm of English language text processing. The study by [28] along with [29] has been considered.

The pre-processing for topic modeling of the dataset is based on the following phases: The first phase is Text normalization, in which text is converted into lowercase after the removal of punctuations, brackets, and whitespaces (all special characters) from the verse dataset. It is critical to ensure that the English dataset does not contain any null values before removing noise from it. As a result, the sum of null values is examined. The dataset does not hold any NAN or null values. For data noise removal, the study conducted by [28] discussed removing all special and non-alphabetic letters during the punctuation and noise removal phase. This also includes the removal of whitespaces from the dataset. In this context, the study of [29] stated that the removal of text noises improves the experimental outcomes and therefore, it is usually reduced in the text processing step. In the pre-

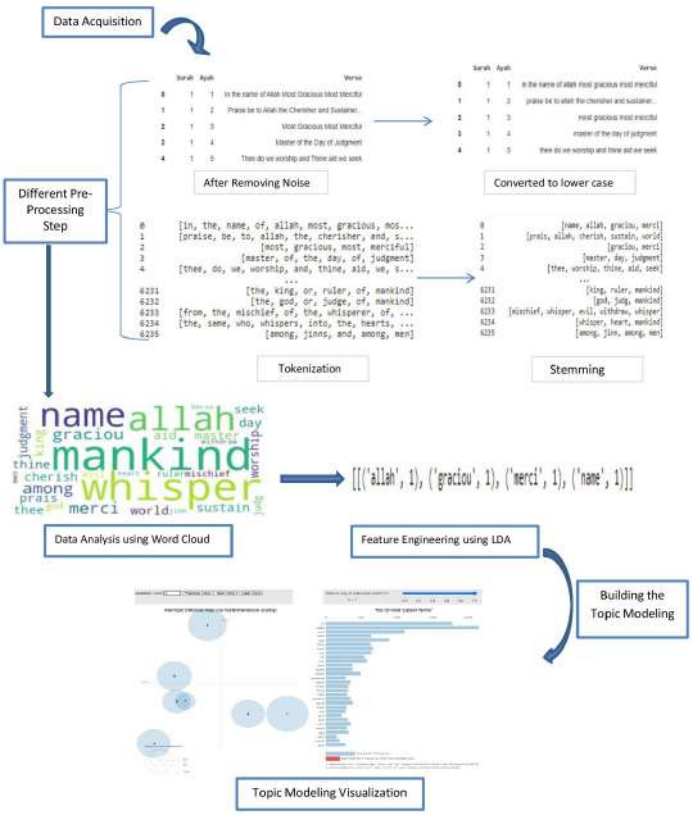

Figure 1. Workflow of Proposed Methodology

sented study we have performed text normalization by first excluding null values from the dataset. Additionally, punctuation marks, non-alphabetic letters, symbols, whitespaces, and brackets are eliminated since they are considered text noise. In addition, to avoid the same words having distinct meanings, the characters are transformed into lowercase. This process is crucial because it is required to verify the precise amount of repeated terms in the verses of the Holy Quran specifically in the translation in English language [26]. A phrase is made up of words, and tokenization divides a sentence into a list of terms that may be used to reconstruct it. In the second phase of Tokenization, each verse string sequence breaks down into single-word pieces. The third phase is to remove stop words, which are auxiliary words that are thought to have no relevance to the phrase. It has been noticed that some stopwords in a text must be stemmed by a stemmer and can no longer be filtered by specified stopwords. For example, Porter stemmer converts "was" to "wa," and when you stem first before eliminating stopwords, "wa" stays in your vector after filtering stopwords that contain "was." As a result, stopwords must be removed in their original form before running stemmer. For the purpose of this research, English stopwords are removed from the stopwords list which is set for the English language translation. Finally, the last phase is stemming, which is useful for determining the fundamental form of each word from a group of words with the same meaning [30].

3.2 Exploratory Data Analysis

To validate the pre-processing results, the data analysis step is performed. According to [31], rather than providing data in the form of text or tables, it is simpler to grasp the content of data when it is presented in a visual format. It is furthermore observed that visualization aids in the comprehension of complicated facts. Despite the vast quantity of data employed, the data pattern may be rapidly and readily comprehended.

One type of data visualization that has been seen is the word cloud. A word cloud is a graphic depiction of the frequency with which words appear in books or on the internet. The occurrence frequency of a word is determined by its font size where the larger the font size indicates the higher the word frequency. In contrast to this, the smaller the font size indicates the lower the word frequency [32][33]. As a result, the word cloud based on the pre-processed words dataset is generated. Figure 2 shows the example of a word cloud for pre-processed data.

3.3 Feature Engineering using LDA

As stated by [27], feature extraction can be performed using a Bag of Word representation after getting clean words from the processing stage. Assessing the number of occurrences of each word based on each class in the training set is being used to generate this representation. LDA, like any other algorithm, is limited to numeric values. Another study by [28], stated that the extraction of features is more important in identifying significant patterns and obtaining information. Selecting these attributes aids in dimensionality reduction, which improves text mining performance. The characteristics that were extracted and chosen are represented as vectors. The vector representation is a method of translating meaningful words into numbers, with each word being distributed with weights, embeddings based on frequency, prediction-based embedding, and non-negative matrix factorization. These are the methods commonly used for extracting words. It also includes document term frequency, count vector, and vector of co-occurrence.

According to the study of [34], a text document is often represented as a vector of term weights known as word features from a collection of words using the frequency count of each term in the document using the Bag of Word representation model. This document representation approach is also known as a Vector Space Model. Therefore, it is significant to transform cleaned and tokenized words into a bag of word representation, which you can think of as a dictionary, with the key being the word and the value being the number of times that word appears in the corpus. In this regard, it is observed that the LDA topic model has two key inputs to do this:

- Dictionary: This indicates each unique word that entails its own id.
- Corpus: The number of times a specific term appears in each document

A unique id for each word in the document is created. The produced corpus is a mapping of (word-id, word-frequency). For example, if (0, 1) implies, word id ’0’ occurs once in the first document. Likewise, for (1,1) word id ’1’ occurs 1 time and so on. This is used as the input by the LDA model. It is necessary to pass the ID as a key to the dictionary to determine what word a particular id relates to. In order to make it a more readable format, the list is further refined.

3.4 Building Topic Modling

To build the LDA model, it is necessary to provide several topics by the corpus and dictionary. Aside from that, alpha and beta are hyper parameters that influence topic sparsity. Both are defaulted to 1.0/num topics previously, according to the Gensim documentation. The number of documents utilized in each training is determined. The model parameters are adjusted and updated, and passes are the total number of training passes. Moreover, the model takes the pass parameter value, which is the number of times the model is trained on the entire corpus and is controlled by the pass parameter (set to 10). "Epochs" can be another term for passes. It is critical to have a large enough number of "passes" and "iterations". Moreover, per-word-topics is set to True. Setting this to True enables the extraction of the most likely topics for a given term. The training procedure is structured in such a manner that each word is allocated to a certain topic. Words that are not indicative will be eliminated if this is not the case. Lastly, the most important parameter in training this LDA model is alpha. The value of the hyper-parameter alpha determines how topics are dispersed among documents in a topic model. A higher alpha number indicates that a topic will be spread out more over the documents, whereas a lower alpha value indicates that a topic will be spread out more narrowly among the documents. As per the research of [35] and [36], paying attention

Figure 2. Wordcloud of Preprocessed Data

to alpha settings is critical while building a robust topic model. In this regard, various research uses topic models for diverse objectives, however, simply use the default value for the alpha hyper-parameter. The default value for alpha is symmetric [36]. This indicates that the alpha value for each topic is the same. Gensim python module employs the formula of dividing 1.0 by the number of topics in the model to determine the symmetric value for alpha. If there are 7 topics in the model, alpha will be set to 0.142. For this study, we have used alpha 'auto', 'symmetric', and 'asymmetric' values.

4 Experiments

The English language dataset file has 6237 rows having three columns with the names: surah, ayah, and verse [25]. However, surah and ayah columns were removed from the initial data preparation phase because the research focused on modeling Quranic verses.

4.1 Experimental Protocols

To evaluate topic models, the perplexity and coherence scores are used. As reported in the literature [37], a good model can create well-coordinated topics with high coherence values. Furthermore, according to [38], the perplexity score may be utilized to assess the accuracy of a trained LDA model prediction. Where a low perplexity score suggests that the test sample is correctly predicted by the topic model [6].

After conducting several experimental trials, we have determined that the optimal parameter settings for topic modeling include an alpha value of 'auto,' resulting in the identification of 7 topics". It results in a -7.30 perplexity score and a 0.36 coherence score. Furthermore, for topic visualization, we utilized LDAvis, an interactive web-based tool for visualizing topics estimated by LDA [8].

4.2 Experimental Results

Since the value of k, or the number of topics, has a direct influence on the LDA model coherence and perplexity score. Therefore, in this experiment, the researcher intends to determine the optimal value for k so that the LDA model for the English language can be trained on it. As a consequence, several iterations of k and their related coherence scores are employed with alpha = auto, symmetric, and asymmetric values. Table 1. shows the optimum values for k, as well as the coherence scores associated with them. Furthermore, past research has shown that topics with fewer overlapping bubbles across the plot are considered a good topic model through which less coherent topics are acquired. Moreover, if the number of topics is larger than the topics repeated it implies the badness

Table 1. Optimal value of k Along with Coherence Score

Number of Topics (k)	Coherence score (c-v) Auto	Symmetric	Asymmetric
2	0.28	0.28	0.27
7	0.36	0.35	0.35
12	0.38	0.38	0.36
17	0.40	0.36	0.39
22	0.42	0.38	0.39
27	0.46	0.42	0.44
32	0.47	0.44	0.44
37	0.49	0.47	0.45

of the model. For the alpha=auto, the final choice for parameter k is 7, which has a coherence value of 0.36, based on the same criteria as those given above for picking k. The perplexity score after developing the LDA model for seven topics is -7.30

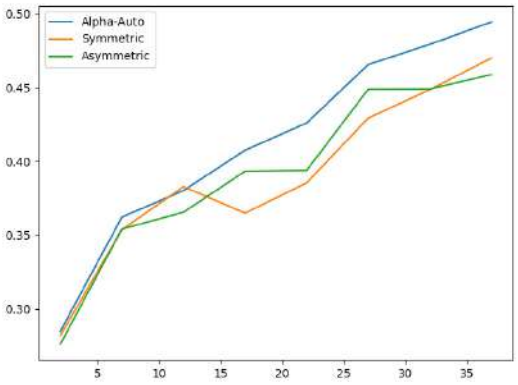

Figure 3. Number of Topics and Coherence Score Graph in LDA for Alpha Auto, Symmetric and Asymmetric values

The default 'symmetric' alpha, employs a fixed symmetric prior of 1.0 divided by the number of topics. For alpha= symmetric, the final parameter k value is 7, with a coherence score of 0.3623 and a perplexity score is -7.45. The asymmetric value for alpha has been used to normalize the asymmetric prior estimate of 1.0 / (topic index + sqrt(number of topics)). The researcher employed alpha = asymmetric with various iterations of k and their related coherence values to find the ideal value for k. Furthermore, based on the same considerations as used to select k for alpha=auto and 'symmetric', parameter k is 7, which has a coherence value of 0.35 for asymmetric value. Upon creating the LDA model for seven topics, the overall perplexity score was -7.44. The number of topics and their respective coherence scores is shown in Figure 3.

Furthermore, the relevance of topics and their relations are highlighted by providing the words that make up each topic for the English topic model. In addition, the visual representation is created after deciding on the topic for developing the LDA model for Quranic verses in English. For the demonstration, Figure 4 depicts the top 30 most relevant keywords that make up the particular topic, together with their anticipated phrase frequencies for alpha = 'Auto'. The general topic distribution is displayed on the left by circles, whilst overall word frequencies are depicted on the right with sky blue bars. When k = 7 is chosen, the LDA model yields a total of 10 keywords, as shown in Table 2.

Furthermore, Table 2 also provides the analysis of the results. The table depicts the most significant terms that are acquired after constructing the LDA model for the topic modeling of Quranic verses for the top ten keywords for each topic (0-6). The number score represents the importance of a specific term in a certain topic. As seen in the table, the representation of topic 0 is based on these keywords: allah '0.082', verili '0.036', earth '0.036', made '0.031', thing '0.030', heaven '0.021', except '0.015', doth '0.015', blaze'0.013' and path '0.013'. This indicates that the most representative keywords from topic 0 are"allah", "verili", "earth"

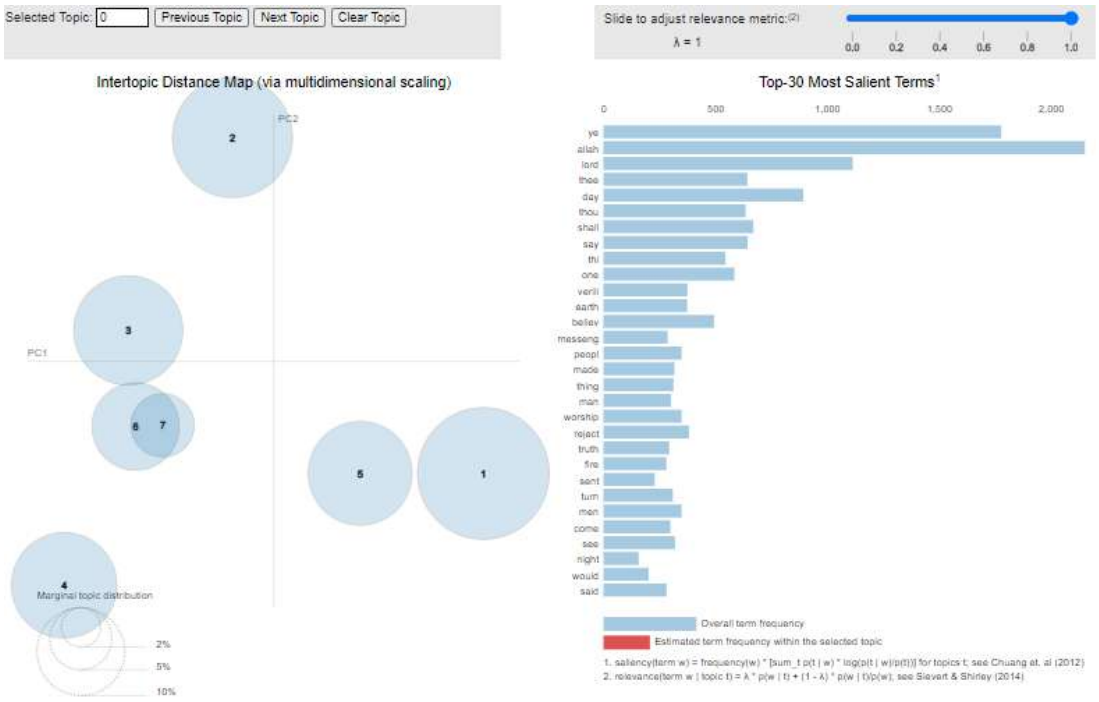

Figure 4. LDA Visualization for Alpha = 'Auto'

and so on. From the above table, it is observed that the most significant word in topic 0 is "allah" whereas the least important term is path due to its weightage. In contrast, the representation of topic 6 is based on these keywords: man '0.028', fire '0.026', give '0.022', deed '0.020', evil '0.019', garden '0.018', power '0.017', therein '0.017', good '0.016'and life '0.016'. This means that the most relevant keywords for topic 6 are "man", "fire", "give" and so on. Whereas the most significant term is "man" and the least significant term is "life" based on the keyword weightage.

5 Analysis and Discussion

As the study aims to use the LDA topic model to decide which topic would best describe a document,therefore, the investigator tries to find the topic in a document with the largest percentage of contribution. From the literature, it is concluded that the perplexity value needs to be lesser which is most probably in negative figures, whereas, the coherence score needs to be larger likely to be in positive numbers. Moreover, model perplexity is an intrinsic assessment metric that is extensively used to evaluate language models. The smaller this metric is, the better it produces results. It measures how surprised the model is by data it had never seen before. Whereas, topic coherence metrics, on the other hand, assess a topic by determining how semantically comparable high-scoring terms in the topic are.

The pre-processing procedures for the English translation dataset of Holy Quran are based on the study conducted by [26]. The best results for the topic modeling of this dataset are acquired when the number of topics is set to 7 for utilizing alpha as auto. As a result, the coherence score is 0.36 and the perplexity score is -7.30. Table 3 exhibits the dominating topic and its contribution to a specific document in this aspect. At this point, each verse is treated as a separate document. It is observed that at index 6, document 6 with the highest topic percentage contribution of around 33% is located where topic 3 contributed the most, in contrast to this, the least contributed topic percentage is 19% which is located at index 1. In this document, topic 4 contributed the most.

Additionally, we wish to identify the most represen-

Table 2. Topics based on LDA and the top ten keywords for each topic (0-6) using auto alpha parameter

Topic 0		Topic 1		Topic 2		Topic 3		Topic 4		Topic 5		Topic 6	
allah	0.08	thou	0.05	night	0.04	lord	0.08	ye	0.1	thee	0.08	man	0.02
verili	0.03	peopl	0.03	wealth	0.03	thi	0.03	allah	0.07	messeng	0.03	fire	0.02
earth	0.03	truth	0.02	high	0.02	one	0.03	day	0.05	sent	0.03	give	0.02
made	0.03	inde	0.02	return	0.02	worship	0.02	shall	0.04	would	0.02	deed	0.02
thing	0.03	make	0.02	help	0.02	turn	0.02	say	0.03	hand	0.02	evil	0.19
heaven	0.02	sign	0.01	anoth	0.02	come	0.02	believ	0.03	woe	0.02	garden	0.01
except	0.01	mischief	0.01	togeth	0.01	said	0.02	reject	0.02	book	0.19	power	0.01
doth	0.01	like	0.01	cherish	0.01	us	0.01	men	0.02	follow	0.01	therein	0.01
blaze	0.01	hath	0.01	appoint	0.01	truli	0.01	see	0.01	way	0.01	good	0.01
path	0.01	mankind	0.01	throne	0.01	right	0.01	creat	0.01	clear	0.01	life	0.01

tative document for each topic, as it may be difficult to comprehend the most representative topics if each document is simply examined for its dominant topic. As a consequence, the researcher aims to construct a list of the documents in which the topic seems to have the greatest impact. In this context, Table 4 demonstrates the topics, as well as the most representative materials for Quran verses in English translation. In addition, the verses that highlighted these topics and their contributions are presented to help in the analysis since the researcher is looking for samples of sentences that best portray a certain topic. Topic number 4 contributed the most to the verses dataset in the English language.

One of the goals of the study is to figure out how widely a topic is covered by analyzing the distribution of topics among documents. Table 5 demonstrates the topics, as well as their overall distribution throughout the documents and percentages while bearing in mind that they are verses in particular. Topic 0 which is trained on LDA auto=alpha is distributed in 443 documents out of 6236 documents. In addition, Topic 1 contributed 803 documents, while Topic 2 contributed 30 documents. Whereas, topic 3 is covered in 1472 documents, followed by topic 4 in 2596, topic 5 in 240, and finally topic 6 in 652 documents respectively. It is observed that topic 4 appears to be prevalent in the majority of documents in the topic modeling of English verses. Whereas, topic 0 contributed the most to the modeling of Arabic verses as mentioned above. This implies that both languages have different topic dominance due to the diversity of language even if the nature of the dataset is the same. Furthermore, while working with a vast number of documents, it is helpful to know the overall size of the documents. As a result, Figure 5 is produced, which shows the mean, median, quantile, and standard deviation for the number of documents as well as the document word count.

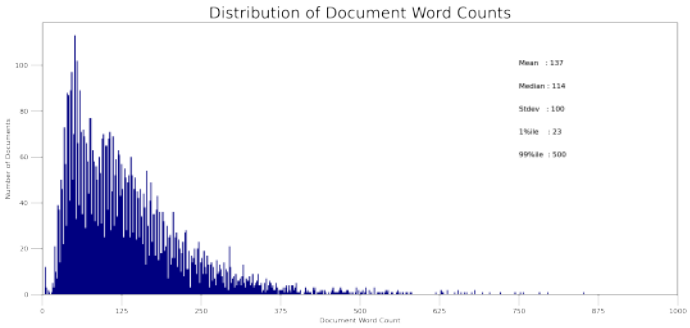

Figure 5. Distribution of Document word counts

For the comparison with other studies, due to the limitation of studies for acquiring dominant topics in the English translation of the verses of the Quran, the researcher considers other relevant studies. In this context, [39] find the dominant topics and most representative documents in the open-source articles regarding the Covid-19 dataset. They also followed the same pre-processing steps and generated features to feed the LDA model using Bag of Words models. However, they have selected 15 topics for the number of topics values k and generated keywords

Table 3. Using the Alpha Auto Parameter to Determine the Dominant Topic in Each Sentence Across the Document

Document-No	Dominant-Topic	Topic-Perc-Contrib	Verse
0	3.0	0.30	in the name of allah most gracious most merciful
1	4.0	0.19	praise be to allah the cherisher and sustainer
2	3.0	0.27	most gracious most merciful
3	4.0	0.23	master of the day of judgment
4	0.0	0.24	thee do we worship and thine aid we seek
5	5.0	0.22	Show us the straight way
6	3.0	0.33	the way of those
6	3.0	0.33	thou hast bestowed th…
7	1.0	0.20	a | m
8	4.0	0.22	this is the book in its guidance sure without
9	6.0	0.24	who believe in the unseen are steadfast in pra..

from dominant topics. The circles in a good topic model should be quite large and non-overlapping, and they should be dispersed around the chart rather than clinging to one quadrant. A topic model with several topics may lead to smaller circles and far more overlaps, concentrated into one quadrant. In [10], a thematic approach has been adopted for the semantic meanings of the topics of the Quran. They selected Chapter 2 as a "Mushaf Al-Tajweed" and performed different pre-processing steps. They have designed a large classical Arabic corpus for the training and perform multi-classification. The significance of their study is to word embedding however they have only worked at Chapter 2 of the Holy Quran. Another interesting work was reported for the topic modeling based on Islamophobia in which the author performs text classification based on the word embedding and transformer technique [12]. Recently few studies reported results based on the utilization of topic modeling for the translation of the Holy Quran in other languages such as [9] and [11].

6 Conclusion

The presented study's aim was to apply topic modeling applied to English-translated Quranic verses. The main focus was to analyze and see the impacts of the LDA model on the topic modeling of Quranic verses having an unlabeled dataset based on the English language. For this purpose, a series of rigorous trial evaluations were conducted. This study would provide a substantial contribution to the study of LDA models for the English language and its impacts specifically for Quranic verses. The results were evaluated by calculating the perplexity and coherence score, where high coherence indicates the goodness of well-structured

Table 4. Topics with Their Most Representative Document

Topic-Num		Topic-Perc-Contrib	Verse
0	0.0	0.46	allah forgiveth not that partners should be se…
1	1.0	0.46	thou causest the night to gain on the day and…
2	2.0	0.31	say i find not in the message received by me b…
3	3.0	0.50	they said beseech on our behalf thy lord to ma…
4	4.0	0.63	after this it is ye the same people who slay a …
5	5.0	0.53	we have sent thee inspiration as we sent it to…
6	6.0	0.57	but give glad tidings to those who believe and …

Table 5. Utilizing Alpha Auto to Determine Topic Distribution across Documents

Num-Documents		Perc- Documents
0.0	443	0.07
1.0	803	0.01
2.0	30	0.00
3.0	1472	0.23
4.0	2596	0.41
5.0	240	0.03
6.0	652	0.10

topic models and low perplexity score indicates the correctness of prediction made by the topic models. One significant advantage of utilizing topic modeling in the context of the Quranic corpus is its ability to facilitate the exploration of hidden themes and topics within the text. This method allows researchers to analyze vast amounts of data efficiently, uncovering connections and patterns that may not be immediately apparent through traditional reading methods. In the future, asymmetric and symmetric approaches and semi-supervised methods will be used to improve the performance of the system. Another opportunity would be to apply the proposed methodology for Hadith datasets. Moreover, or future research, it is recommended to integrate the results obtained from LDA with Generative AI to enhance the analysis of textual data. By leveraging the insights gained from LDA, Generative AI Large Language Models (LLM) agents can generate coherent and contextually relevant narratives around the identified topics, enriching the interpretation of the data. This combination has the potential to provide a more nuanced understanding of themes and trends within large datasets, facilitating deeper insights and more effective decision-making processes.

Author Contributions

Kashmala Jamshaid Akhter: Conceptualization, Data curation, Methodology, Software, Writing- Original draft preparation**Humera Farooq**: Supervision, Writing- Reviewing and Editing **Muhammad Tariq Siddique**: Visualization, Writing- Reviewing.

Compliance with Ethical Standards

It is declare that all authors don't have any conflict of interest. It is also declare that this article does not contain any studies with human participants or animals performed by any of the authors. Furthermore, informed consent was obtained from all individual participants included in the study.

References

[1] K. R. Chowdhary and K. R. Chowdhary, "Natural language processing," *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

[2] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 2, pp. 604–624, 2020.

[3] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia tools and applications*, vol. 82, no. 3, pp. 3713–3744, 2023.

[4] G. M. Al and K. Badruddin, "Assessment of performance of machine learning based similarities calculated for different english translations of holy quran," *International Journal of Computer Science and Network Security*, vol. 22, no. 4, pp. 111–118, 2022.

[5] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Information Systems*, vol. 94, p. 101582, 2020.

[6] T. R. Hannigan, R. F. Haans, K. Vakili, H. Tchalian, V. L. Glaser, M. S. Wang, S. Kaplan, and P. D. Jennings, "Topic modeling in management research: Rendering new theory from textual data," *Academy of Management Annals*, vol. 13, no. 2, pp. 586–632, 2019.

[7] C. U. Ghanshyambhai and A. Shah, "Optimizing topic coherence in the gujarati text topic modeling: A relevant words based approach," Ph.D. Thesis, 2018.

[8] T. Islam, "Ex-twit: Explainable twitter mining on health data," *arXiv preprint arXiv:1906. 02132*, 2019.

[9] D. Rolliawati, I. Rozas, and K. Khalid, "Text mining approach for topic modeling of corpus al qur'an in indonesian translation," in *Proceedings of the 2nd International Conference on Quran and Hadith Studies Information Technology and Media in Conjunction with the 1st International Conference on Islam, Science and Technology, ICONQUHAS & ICONIST, Bandung, October 2-4, 2018, Indonesia*, 2020.

[10] E. H. Mohamed and W. H. El-Behaidy, "An ensemble multi-label themes-based classification for holy qur'an verses using word2vec embedding," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 3519–3529, 2021.

[11] A. Zafar, M. Wasim, S. Zulfiqar, T. Waheed, and A. Siddique, "Transformer-based topic modeling for urdu translations of the holy quran," 2024.

[12] A. Saeed, H. U. Khan, A. Shankar, T. Imran, D. Khan, M. Kamran, and M. A. Khan, "Topic modeling based text classification regarding islamophobia using word embedding and transformers techniques," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2023.

[13] B. Arkok and A. M. Zeki, "Classification of quranic topics using ensemble learning," in *2021 8th International Conference on Computer and Communication Engineering (ICCCE)*. IEEE, 2021, pp. 244–248.

[14] B. S. Arkok and A. M. Zeki, "Classification of quranic topics based on imbalanced classification," *Indones. J. Electr. Eng. Comput. Sci*, vol. 22, no. 2, p. 678, 2021.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[16] M. R. Choirulfikri, K. M. Lhaksamana, and S. Al Faraby, "A multi-label classification of al-quran verses using ensemble method and naïve bayes," *Building of Informatics, Technology and Science (BITS)*, vol. 3, no. 4, pp. 473–479, 2022.

[17] A. S. Aiman, K. M. Lhaksmana, and Others, "Topic classification of quranic verses in english translation using word centrality measurement," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 5, pp. 803–809, 2022.

[18] B. Arkok and A. M. Zeki, "Gos: A genetic oversampling algorithm for classification of quranic verses," in *2022 13th International Conference on Information and Communication Systems (ICICS)*, 2022, pp. 289–293.

[19] A. M Alashqar, "A classification of quran verses using deep learning," *International Journal of Computing and Digital Systems*, vol. 16, no. 1, pp. 1041–1053, 2024.

[20] M. Alshammeri, E. Atwell, and M. A. Alsalka, "Quranic topic modelling using paragraph vectors," in *Intelligent Systems and Applications: Proceedings of the 2020 Intelligent Systems Conference (IntelliSys) Volume 2*. Springer, 2021, pp. 218–230.

[21] G. Mediamer, "Semantic feature analysis for multi-label text classification on topics of the al-quran verses," *Journal of Information Processing Systems*, vol. 20, no. 1, 2024.

[22] M. Alshammeri, E. Atwell, and M. A. Alsalka, "A siamese transformer-based architecture for detecting semantic similarity in the quran," in *The International Journal on Islamic Applications in Computer Science And Technology-IJASAT*, vol. 9, no. 4. Design For Scientific Renaissance, 2021.

[23] A. Rafea and N. A. GabAllah, "Topic detection approaches in identifying topics and events from arabic corpora," *Procedia computer science*, vol. 142, pp. 270–277, 2018.

[24] U. Chauhan and A. Shah, "Topic modeling using latent dirichlet allocation: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 7, pp. 1–35, 2021.

[25] K. Kaggle, vol. 2021, no. 20 Feburary, 2017. [Online]. Available: https://www.kaggle.com/zusmani/the-holy-quran

[26] N. S. Jamil, K. R. Ku-Mahamud, A. Mohamed Din, F. Ahmad, N. Che Pa, W. H. Wan Ishak, R. Din, and F. K. Ahmad, "A subject identification method based on term frequency technique," *International Journal of Advanced Computer Research*, vol. 7, no. 30, pp. 103–110, 2017.

[27] M. S. M. R. A. Pane and N. S. Huda, "A multi-lable classification on topics of quranic verses in english translation using multinomial naive bayes," pp. 481–484, 2018 2018.

[28] S. Likhitha, B. S. Harish, and H. M. K. Kumar, "A detailed survey on topic modeling for document and short text data," *International Journal of Computer Applications*, vol. 178, no. 39, pp. 1–9, 2019.

[29] S. Mifrah and E. Benlahmar, "Topic modeling coherence: A comparative study between lda and nmf models using covid'19 corpus," *International Journal of Advanced Trends in Computer Science and Engineering*, pp. 5756–5761, 2020.

[30] K. M. O. Nahar, M. Ra'ed, M. Al-Shannaq, M. Daradkeh, and R. Malkawi, "Direct text classifier for thematic arabic discourse documents," *Int. Arab J. Inf. Technol.*, vol. 17, no. 3, pp. 394–403, 2020.

[31] M. F. Bashri and R. Kusumaningrum, "Sentiment analysis using latent dirichlet allocation and topic polarity wordcloud visualization," in *5th International Conference on Information and Communication Technology (ICoIC7)*. IEEE, 2017, Conference Proceedings, pp. 1–5.

[32] F. Miley and A. Read, "Using word clouds to develop proactive learners," *Journal of the Scholarship of Teaching and Learning*, pp. 91–110, 2011.

[33] S. Jayashankar and R. Sridaran, "Superlative model using word cloud for short answers evaluation in elearning," *Education and Information Technologies*, vol. 22, no. 5, pp. 2383–2402, 2017.

[34] R. Koulali and A. Meziane, "A comparative study on text representation models for topic detection in arabic," *Computación y Sistemas*, vol. 23, no. 3, pp. 683–691, 2019.

[35] H. Wallach, D. Mimno, and A. McCallum, "Rethinking lda: Why priors matter," *Advances in neural information processing systems*, vol. 22, 2009.

[36] B. Carron-Arthur, J. Reynolds, K. Bennett, A. Bennett, and K. M. Griffiths, "What's all the talk about? topic modelling in a mental health internet support group," *BMC psychiatry*, vol. 16, no. 1, pp. 1–12, 2016.

[37] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 530–539.

[38] H. Amoualian, W. Lu, E. Gaussier, G. Balikas, M. R. Amini, and M. Clausel, "Topical coherence in lda-based models through induced segmentation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, Conference Proceedings, pp. 1799–1809.

[39] J. Le, M. Weber, and D. Wild, "Topic modeling of covid-19 open research dataset using latent dirichlet allocation," 2020.