RoBERT2VecTM: A Novel Approach for Topic Extraction in Islamic Studies

Sania Aftar[1*], Luca Gagliardelli[1], Amina El Ganadi[1,2], Federico Ruozzi[1]
Sonia Bergamaschi[1]
[1] University of Modena and Reggio Emilia, Modena, Italy
[2] University of Palermo, Palermo, Italy
{name.surname}@unimore.it

Abstract

Investigating "Hadith" texts, crucial for theological studies and Islamic jurisprudence, presents challenges due to the linguistic complexity of Arabic, such as its complex morphology. In this paper, we propose an innovative approach to address the challenges of topic modeling in Hadith studies by utilizing the Contextualized Topic Model (CTM). Our study introduces RoBERT2VecTM, a novel neural-based approach that combines the RoBERTa transformer model with Doc2Vec, specifically targeting the semantic analysis of "Matn" (the actual content). The methodology outperforms many traditional state-of-the-art NLP models by generating more coherent and diverse Arabic topics. The diversity of the generated topics allows for further categorization, deepening the understanding of discussed concepts. Notably, our research highlights the critical impact of lemmatization and stopwords in enhancing topic modeling. This breakthrough marks a significant stride in applying NLP to non-Latin languages and opens new avenues for the nuanced analysis of complex religious texts.

1 Introduction

In Islamic tradition, the Hadith, which includes the recorded sayings and actions of Prophet Muhammad as documented by his companions, is fundamental not only to religious practices but also garners significant interest in linguistic and anthropological studies. These records, known collectively as "Ahadith" (the plural form of Hadith), are vital to Islamic jurisprudence and serve as an important supplementary source to the Qur'an.

Islamic scholarship traditionally emphasizes the authenticity of Hadith transmission chains, often neglecting contents (Matn) scrutiny for anachronisms or inconsistencies, characterizing the conventional methodology in Hadith studies (Schacht, 1967).

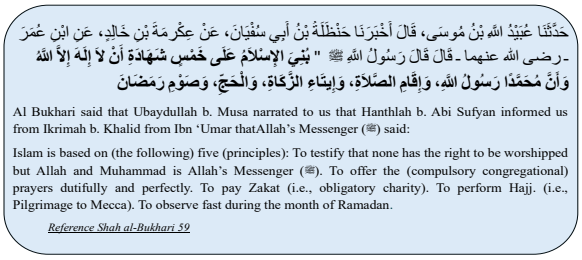

Figure 1: A Hadith with its English translation. Narrators listed first, with the bold section indicating the Matn.

With a primary focus on Religious Studies and Classical Arabic, the Hadith dataset has become a crucial asset in our research. Linguistically, the analysis of Ahadith not only highlights the evolution and usage of Classical Arabic but also uncovers subtle linguistic nuances and traces of historical development. The content of Ahadith, along with the study of their Isnad (the chain of narrators) and Matn, reveals patterns of social interaction, oral traditions, and knowledge transmission across generations (see Figure 1). The Isnad traces oral lineage and social ties of early Muslim communities, while the bolded Matn offers direct insights into their practices, beliefs, and customs. The study of Ahadith, blending religious, linguistic, and anthropological aspects, highlights their multidimensional role in Islamic studies and the importance of preserving religious values while understanding their historical and cultural contexts.

Recently, researchers from fields such as computer science and digital humanities have shown interest in Islamic research. Researchers use computational methods, including NLP, to address challenges in Islamic studies, such as identifying narrator relationships, resolving ambiguities, and answering domain-specific queries (Azmi et al., 2019). These techniques also aid in the semantic identification of Matn correspondence to other

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 9148–9158
November 12-16, 2024

Hadith, known as content identification (Mghari et al., 2022; Najeeb, 2021; Abdelrazek et al., 2023). In this context, topic modeling is an ensemble of approaches used to identify patterns and topics in unstructured text.

However, working with non-Latin languages is rigorous due to the supremacy of models and libraries built for English, which has simpler word structures and grammar. Furthermore, non-Latin languages like Arabic have complex morphology and grammar, which complicate these tasks. This study focuses on Arabic, aiming to develop a method to handle its unique aspects.

Recent developments in topic modeling introduced the Contextual Topic Model (CTM) in 2020 (Bianchi et al., 2020a). This innovation improves accuracy by using contextual embeddings instead of traditional ways (Bianchi et al., 2020a).

In this study, we develop a novel CTM-Hybrid embedding model to analyze topics within Classical Arabic texts (Hadith) and compare its performance with existing models. The use of these embeddings allow the model to handle complex Arabic morphology and syntax, generating rich semantic and contextual nuances. We compare our model with the conventional CTM using SBERT and several advanced models under various testing scenarios. Our goal is to offer a sophisticated research that provides detailed insights and deepens the understanding of Hadith topics for developers, researchers, and users to understand religious studies.

This work is part of the Digital Maktaba research project (Martoglia et al., 2023; Bergamaschi et al., 2022; Aftar et al., 2024) as a part of the ITSERR[1] one and aims to develop workflows for automatic extraction of information and metadata from non-Latin scripts, such as Arabic.

Overall contributions are listed as follows:

- We present a method for Hadith topic modeling by combining CTM with hybrid embeddings, enhancing language understanding for precise, context-sensitive topic extraction.
- We proposed framework that benefits Hadith studies, helping researchers and analysts explore new dimensions in the field.
- We proposed a specialized embedding method that accommodates Arabic language.

[1] https://www.itserr.it

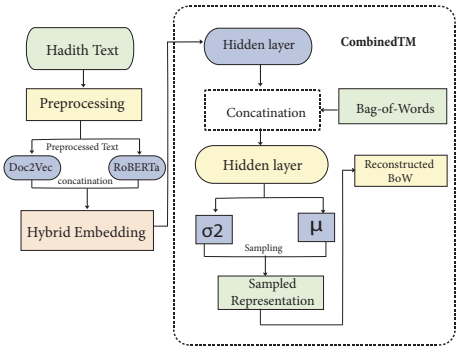

Figure 2: *RoBERT2VecTM model.*

- We extensively tested a large corpus of Ahadith (Matn), achieving superior coherence scores compared to other topic modeling approaches.

The paper is organized as follows: Section 2 details our solution; Section 3 presents experimental evaluation; Section 4 reviews related work; and Section 5 presents conclusion and future directions.

2 RoBERT2VecTM Topic Modeling

This section presents our semantic-based topic modeling for Hadith text, as illustrated in Figure 2.

2.1 Hadith Collection

Our study is based on the authoritative texts of Sahih Al-Bukhari[2] and Sahih Muslim[3], encompassing a total of more than 14,940 Ahadith across various themes. To aid computational analysis, we have adopted Unicode versions of these texts, ensuring the original sources remain unaltered.

2.2 Preprocessing

The preparatory stage uses preprocessing steps to refine the dataset for semantic similarity and topic identification (Alhawarat et al., 2021; Alhaj et al., 2019). A detailed list of adopted preprocessing steps is provided below.

Separating Isnad From Matn. This study aims to identify topics in Hadith texts by focusing on the Matn, rather than the Isnad. The Isnad, which primarily consists of narrators' names, does not contribute significantly to semantic analysis. Therefore, we center our efforts on the Matn, where the Hadith's main content resides. To achieve accurate

[2] http://sunnah.com/muslim
[3] http://www.qaalarasulallah.com

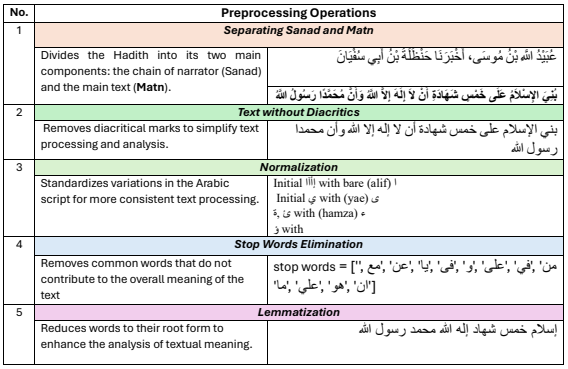

Figure 3: Data Preparation steps employed in our framework.

topic extraction and semantic interpretation, we use custom regular expressions specifically designed to identify and isolate the Matn. These regular expressions recognize the structural patterns of the Isnad and Matn, allowing us to precisely extract the relevant content for analysis. After extraction, the Matn is subjected to preprocessing for further analysis. An example can be seen in Figure 3–Step 1.

Removal of Diacritics. Removing diacritics, marks that show linguistic traits, simplifies text for more effective topic modeling. This step can be illustrated in Figure 3–Step 2.

Normalization. Normalization improves text data for analysis by standardizing Unicode characters for model compatibility and reducing word repetition to unify word forms. This step is exemplified in Figure 3–Step 3.

Stop words Removal. Stop-word removal enhances topic model precision by focusing on significant words. We used Python's *nltk.corpus* module[4]. A demonstration of this step can be seen in Figure 3–Step 4.

Lemmatization. This simplifies Arabic words to their roots. We use Farasa lemmatizer[5], recognized as one of the standardized and freely accessible tools in comparison to other lemmatizers (Mubarak, 2017). For reference, an example of this can be observed in Figure 3–Step 5.

2.3 Contextualized Model for Hadith

After preprocessing, we use our model to identify and categorize Hadith contents. This begins with a comprehensive review of CTM and hybrid embedding techniques, followed by their integration, highlighting key benefits and demonstrating improved topic identification.

Hybrid Embedding. Traditional text analysis techniques like BoW and TF-IDF are not effective at capturing the underlying meanings of words within a document as semantic based vector representation is a challenging task (Mishra and Panchal, 2022). To address this, we use a hybrid embedding strategy for the Hadith dataset. This approach combines Doc2Vec and RoBERTa to capture its semantic in two different ways: (*i*) Doc2Vec generates embeddings covering the entire content based on the similarity of the words producing an embedding of a 128-size; whereas RoBERTa provides nuanced embeddings for each word in its surrounding context. Given a text, RoBERTa first converts it into tokens, then returns an embedding of 768-size for each of them, resulting in a matrix of (#tokens x 768)-size. To effectively combine the embeddings from both models, it is crucial that they have the same dimensionality. UMAP (McInnes et al., 2018) is therefore applied in this study to reduce RoBERTa's embeddings to 128 dimensions, ensuring the preservation of the original semantics. The resulting embeddings are then utilized in the proposed model to enhance performance.

To further refine the data and achieve dimensionality reduction, our model employs an autoencoder with a 64-unit bottleneck, using ReLU activation to introduce non-linearity and capture complex patterns in the data. The model is trained for 200 epochs, optimized using the Adam optimizer, and minimizes reconstruction error through Mean Squared Error (MSE). This approach produces a compact yet expressive representation of the data, preserving essential features for subsequent analysis.

Combined Topic Model. Contextualized Topic Model introduced by Bianchi et al., suggests two different methods: *CombinedTM (Combined topic model)* (Bianchi et al., 2020a) and *ZeroShotTM* (Bianchi et al., 2020b). Both techniques utilize contextualized document embeddings from pre-trained language models, yet they differ in their motivations and methods. In our research, we employ CombinedTM, which integrates both contextualized embeddings SBERT and BoW representations as input. This CTM model is based on ProdLDA, which employs variational autoencoders. As part of the exponential-family PCA class, ProdLDA shares traits with exponential-family harmoniums but stands out due to its non-Gaussian priors (As-

[4]from nltk.corpus import stopwords
[5]farasa-api.qcri.org

nawi et al., 2023). The ProdLDA model is represented by the following equation:

$$p(v_i|\phi,\alpha) \propto \prod_{j=1}^{J} p(v_i|z_i=j,\alpha)\phi_j \quad (1)$$

In the aforementioned equation, v_i denotes the word at position i, ϕ is the vector that shows the proportion of each topic within a document, and α is the matrix that describes how words are distributed across various topics. z_i represents the latent topic assignment for the word v_i. In the CombinedTM framework, SentenceBERT embeddings are combined with BoW representations and fed into the inference network of ProdLDA. This same strategy is adopted in our study, based on the inference network of ProdLDA model. In our approach, we replaced SBERT with our hybrid embeddings and fine-tuned the model parameters to improve coherence scores (see Section 3.1). Incorporating CombinedTM with Hybrid embedding involves multiple steps. Initially, a document is inputted and separate embeddings are generated each using Doc2Vec and RoBERTa. These embeddings are then processed through a series of steps to create contextualized embedding that encapsulates both meaning and context of the document. This contextualized embedding is then reduced to a lower-dimensional space, with each dimension aligned to a specific word in vocabulary. In the mean time, same document is use to generate BoW vectors. After that reduced Hybrid embedding and the Bag-of-words vector are then combined and passed through a hidden layer, creating a latent representation of the document. To approximate the posterior distribution of the latent representation, which follows a Gaussian distribution, **variational inference techniques** are applied. This process yields μ (mean) and σ^2 (variance), which parameterize the distribution. During inference, a latent representation is generated by sampling from this Gaussian distribution. Since directly sampling from μ and σ^2 would not be differentiable, the **reparameterization trick** is employed. Instead of sampling the latent variable z directly from $\mathcal{N}(\mu,\sigma^2)$, an auxiliary variable ϵ is sampled from a standard normal distribution $\mathcal{N}(0,I)$. The latent variable z is then computed using the transformation:

$$z = \mu + \sigma \odot \epsilon$$

This trick ensures that the sampling operation is differentiable, allowing gradients to flow through

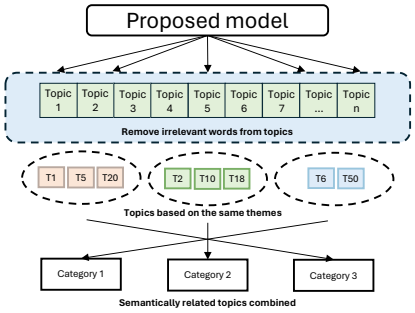

Figure 4: Mapping topics to predefined categories.

μ and σ during backpropagation. Once the latent variable z is obtained, it is passed through a **product-of-experts decoder**, which reconstructs the **Bag-of-Words (BoW)** vectors from this latent representation. The decoder's goal is to minimize the **reconstruction loss** during the model's training, ensuring that the output closely resembles the input document. After this procedure, the model is capable of generating semantically related topics from the latent space.

Furthermore, to improve the quality and coherence of these generated topics, we refined them by removing duplicates and short words. This refinement process allowed us to categorize the topics into eleven selected themes, enhancing their interpretability and context understanding. For example, we labeled Islamic obligations as **'Prayer and Worship'** and moral considerations as **'Islamic Ethics and Morality'** to clearly present our corpus key themes. The objective of this procedure is to provide an additional coherent and detailed representation of the essential themes in our corpus.

3 Experimental Evaluation

To evaluate our proposed model, we compared it to various models, including probabilistic and neural network models. All models were trained on the same pre-processed dataset for fairness.

LDA with TF-IDF. Latent Dirichlet Allocation (LDA) (Gupta et al., 2022) effectively distributes topics across large corpora. We enhanced LDA by replacing BoW with Term Frequency-Inverse Document Frequency (TF-IDF) to improve accuracy, as discussed in (Gupta et al., 2022).

Non-negative Matrix Factorization. NMF (Arora et al., 2012) is widely used to identify latent topics in text. We used NMF with TF-IDF for text analysis and fine-tuned the model, using it as a

baseline.

Topic Modeling in Embedding Spaces. ETM, a generative model combining topic models and word embeddings (Dieng et al., 2020), was trained on the Hadith dataset using Word2Vec. It groups words into topics, revealing the dataset's themes.

Product of Latent Dirichlet Allocation. ProdLDA (Srivastava and Sutton, 2017), a neural variation of LDA, produces more distinct and less overlapping topics. We trained it with our Hadith dataset.

Neural Variational Document Model. NVDM, an unsupervised topic modeling approach (Miao et al., 2016), uses an MLP encoder to learn semantic latent variables from documents' bag-of-words representations. We adapted NVDM for Arabic with AraBERT embeddings to enhance document modeling and semantic analysis of the Hadith dataset, enabling performance comparison with our proposed model.

BERTopic. BERTopic employs pre-trained transformer models along with a class-based TF-IDF method to enhance the coherence and interpretability of identified topics. We have adopt this model as our baseline to compare with our proposed model on an Arabic dataset, utilizing the *xlm-r-bert-base-nli-stsb-meantokens* (Grootendorst, 2022) for its supports multiple languages. Even though BERTopic can automatically determine the number of topics, we set it to 20-70 for comparison with other models.

CluWords. CluWords (Viegas et al., 2019) creates meta-words from pre-trained embeddings, improving document representation. We adapted this for our Ahadith corpus using FastText embeddings due to their effectiveness with Arabic (Grave et al., 2018). The remaining steps are similar to those discussed above.

RoBERT2VecTM settings. Before initiating the embedding process, we conducted a comprehensive analysis of our dataset to identify and compile a list of Arabic terms deemed irrelevant for our analysis[6]. This carefully selected list consists of Arabic terms that were considered non-contributory to our analysis. The strategic removal of these words from each document is a crucial step that significantly enhances the quality of our subsequent topic modeling efforts. By reducing the noise in the data, we facilitate a more refined analysis, enabling our models to capture the essential themes and patterns with greater accuracy.

To generate *Hybrid Embeddings*, we configured Doc2Vec with 200-dimensional embeddings and set a minimum word count of 5 to filter out common terms. A context window of 10 was used to enhance our understanding of relationship between words. The training was conducted over 50 epochs to enhance the capture of semantic nuances, and a sampling rate of 1e-5 was used for optimized learning. For RoBERTa, we used a version fine-tuned on Arabic[7], along with its corresponding tokenizer (learning rate: 5e-05, train batch size: 10). Finally, we trigger CTM to generate a set of topics on a given dataset. For the best performance of CTM we have selected four parameters including activation function, number of neurons, optimization function and number of layers details is shown in Table 1. After spotting several topics, we further redefine them to eliminate infrequently used ones and combine duplicates. Moreover, we have systematically extracted the top 50 words for each of them for a more precise understanding. During their interpretation, we examined closely related semantic themes, leading us to combine multiple topics into singular overarching categories.

Hyperparameter	Value
Activation Function	ReLU
Optimization Function	Adam
Layer Number	1
Neuron Number	128

Table 1: RoBERT2VecTM Settings.

3.1 Evaluation Metrics

To evaluate the topic model, we used three coherence metrics: *NPMI*, *Coherence Score*, and *Topic Diversity*. These metrics ensure accurate interpretation and demonstrate the model's effectiveness.

Coherence CV. (Röder et al., 2015) introduced Coherence CV. In CV, each topic word is compared to the entire topic set using a sliding window to detect word co-occurrences. A *"word vector"* of size N is constructed for the N most probable words, with each cell containing the *Normalized Pointwise Mutual Information (NPMI)* between the word and word i. These vectors are aggregated into a single

[6]The list of identified terms is available on our GitHub repository: https://github.com/saniaaftar/RoBERTa2VecTM.

[7]https://huggingface.co/Davlan/xlm-roberta-base-finetuned-arabic

topic vector. The CV score is calculated by averaging cosine similarities between each topic word and its topic vector.

Coherence NPMI. Coherence NPMI, introduced by Aletras and Stevenson (Aletras and Stevenson, 2013), uses normalized pointwise mutual information (NPMI) to assess word co-occurrence within a topic. It normalizes scores from -1 to 1, providing a detailed evaluation of topic coherence by averaging NPMI scores of word pairs within a topic. The formula for Normalized Pointwise Mutual Information (NPMI) is given by:

$$\text{NPMI}(x, y) = \frac{\log\left(\frac{P(x,y)}{P(x)P(y)}\right)}{-\log(P(x, y))} \quad (2)$$

Topic Diversity. Topic diversity evaluates the uniqueness of words across different topics. It calculates the ratio of unique words to the total number of top words across topics, measuring the percentage of unique words among all top words (Wu et al., 2017). The formula for Topic Diversity is expressed as:

$$\text{Topic Diversity} = \frac{|x|}{tw \times |\text{topics}|} \quad (3)$$

In this formula, $|x|$ represents the count of unique words in the topic-word distributions, tw is the number of top words per topic, and $|\text{topics}|$ is the total number of topics in the model.

3.2 Topic Selection

For selecting the best number of topics we have tested topic ranges from 20 to 70, using the Coherence CV score as the main performance metric. Consequently, 50 topics were identified as the optimal number, as shown in Figure 5.

3.3 Benchmark Comparison

Table 3 presents the results obtained from the different models after applying all the preprocessing steps on the input dataset.

RoBERT2VecTM obtains the best CV Score, Topic Diversity, and NPMI, outperforming all the baselines on all the metrics.

In terms of CV Score and NPMI, the best baseline performer is BERTopic, obtaining a value of 0.55 and 0.053 respectively. RoBERT2-VecTM improves these values by 10.0% (0.61) and 13.2% (0.053). While, for Topic Diversity, NMF is the second best performer (0.52), and RoBERT2VecTM outperforms it by 1.9% (0.53).

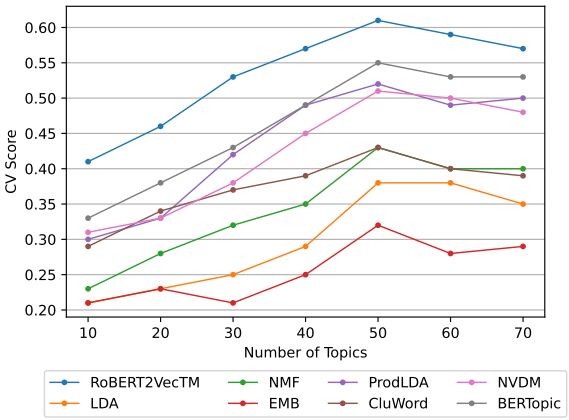

Figure 5: Identifying the Best Number of Topics.

The robustness of RoBERT2VecTM is further evidenced by its performance in topic diversity. Our model achieves a diversity score of 0.53, which indicates a strong capability to differentiate between various themes in the data.

Additionally, Table 2 shows a sample of randomly selected Ahadith with recommended categories and model-assigned probabilities. Although the model was originally tested using Ahadith in Arabic, translated versions are used in this paper to ensure clarity and accessibility for readers who do not understand Arabic.

3.4 Ablation Study

To investigate the impact of lemmatization and stopwords removal, we conducted a comparative analysis of our model against baselines with and without them. The goal of this experiment is twofold: *(i)* analyze the impact of lemmatization and stopwords removal; *(ii)* discover the robustness of the models against dirty data. The obtained results are presented in Table 3, while Figure 6 shows the improvement in percentage when lemmatization is used and stopwords removed.

Our findings underscore the critical role of lemmatization in boosting the effectiveness of the models, leading to significant improvements in all the metrics. On average, the use of lemmatization significantly enhances model outputs, with an increase in CV Score by 22.0%, topic diversity by 8.5%, and NPMI by 19.6%, highlighting its crucial role in boosting topic quality.

However, as seen in Figure 6 (a), not all models benefit equally from lemmatization. For instance, the EMB model does not show improvement in CV Score, and its NPMI actually decreases by 8%, suggesting that lemmatization does not significantly

Category	Probability	Hadith
Prayer and Worship	0.86	When Ramadan begins, the gates of Paradise are opened.
Belief and Theology	0.81	Allah will not be merciful to those who are not merciful to mankind.
Islamic Law and Society	0.79	One of the evil deeds with bad consequence from which there is no escape for the one who is involved in it is to kill someone unlawfully.

Table 2: Random Hadith and Corresponding Categories.

enhance its performance. Similarly, the CluWord model shows negligible changes. In contrast, models like RoBERTa2VecTM, BERTopic, and NMF consistently exhibit improvements, especially in coherence scores, indicating their effective adaptation to lemmatization.

Stopwords removal is intended to filter out unnecessary words, potentially clarifying and focusing on the meaningful ones. Each model behaves differently regarding the removal of stopwords, as clearly seen in Figure 6 (b) particularly those with high initial NPMI and CV scores such as RoBERTa2VecTM and BERTopic, handle the absence of stopwords effectively, enhance their ability to extract meaningful topics. These models exhibit relatively balanced and moderate performance, suggesting a robustness that allows them to benefit from stopwords removal without significant detriment. On the other hand, the models like LDA and EMB displays no substantial change from this process.

Further insights can be observe from Table 3 it has been clearly seen that RoBERT2VecTM demonstrates good performance across all metrics, surpassing competing models. It delivers consistently balanced results, showcasing its robustness and versatility across diverse evaluations. It compares favorably against most models in term of lemmatization as well as stopwords. EMB model have difficulty in identifying meaningful topics, possibly due to embedding strategy that does not effectively capture the thematic structure in Arabic.

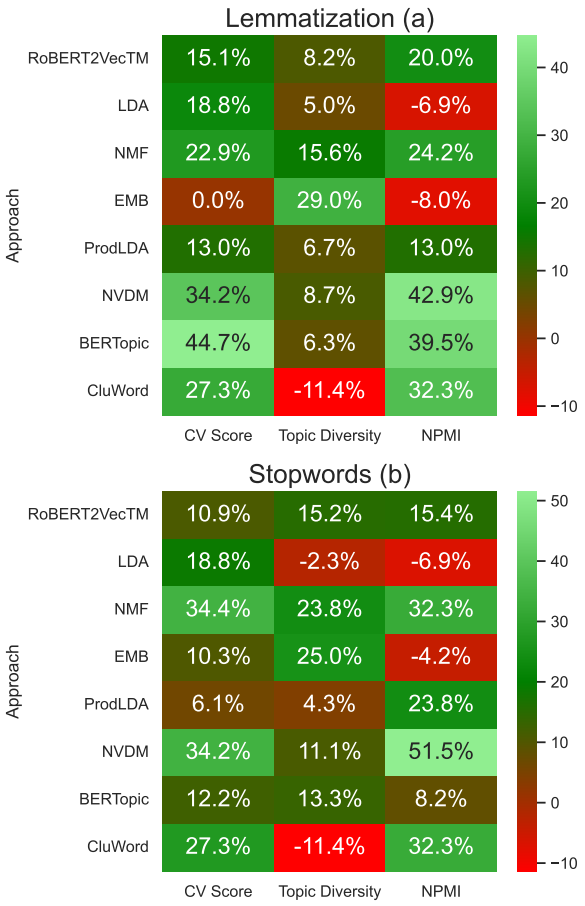

Figure 6: Percentage performance improvements with lemmatization and stopword removal.

3.5 Impact of Hybrid Embedding

To validate the effectiveness of our proposed model, RoBERT2VecTM, which incorporates a hybrid embedding strategy, we systematically compared its performance against several well-established configurations. These included individual implementations of RoBERTa and Doc2Vec, as well as the original Contextualized Topic Model (CTM) that utilizes SBERT embeddings. This comparative analysis was designed to rigorously evaluate the influence of different embedding strategies on the overall performance and robustness of topic modeling. As detailed in Table 4, our integrated model, RoBERT2VecTM, which combines the strengths of Doc2Vec and RoBERTa embeddings, achieved a remarkable CV-Score of 0.61. This score not only exceeds the performance of the standalone RoBERTa and Doc2Vec models but also surpasses the CTM with SBERT embeddings by significant margins: 13.0% (0.54) higher than RoBERTa alone, 74.3% (0.35) better than Doc2Vec alone, and 10.9%

Approach	CV Score	TD	NPMI
(a) Lemmatized and Stopword-Free			
RoBERT2VecTM	**0.61**	**0.53**	**0.060**
LDA	0.38	0.42	0.027
NMF	0.43	0.52	0.041
EMB	0.32	0.40	0.023
ProdLDA	0.52	0.48	0.052
NVDM	0.51	0.50	0.050
BERTopic	0.55	0.51	0.053
CluWord	0.42	0.31	0.041
(b) Without Lemmatization			
RoBERT2VecTM	**0.53**	**0.49**	**0.050**
LDA	0.32	0.40	0.029
NMF	0.35	0.45	0.033
EMB	0.32	0.31	0.025
ProdLDA	0.46	0.45	0.046
NVDM	0.38	0.46	0.035
BERTopic	0.38	0.48	0.038
CluWord	0.33	0.35	0.031
(c) Considering Stopwords			
RoBERT2VecTM	**0.55**	**0.46**	**0.052**
LDA	0.32	0.43	0.029
NMF	0.32	0.42	0.031
EMB	0.29	0.32	0.024
ProdLDA	0.49	0.46	0.042
NVDM	0.38	0.45	0.033
BERTopic	0.49	0.45	0.049
CluWord	0.33	0.35	0.031

Table 3: Comparative Analysis of Topic Modeling Approaches using 50 topics.

Approach	Embedding	CV-Score
RoBERT2VecTM	*Doc2Vec + RoBERTa*	**0.61**
CombinedTM	*RoBERTa*	0.54
CombinedTM	*Doc2Vec*	0.35
CombinedTM	*SBERT*	0.55

Table 4: Comparing RoBERT2VecTM Performance with Original Model.

(0.55) greater than the SBERT-enhanced CTM.

These results shows the superior capability of the hybrid embedding approach in RoBERT2VecTM, outperforming the original CTM configurations across various metrics. The enhanced performance and robustness provided by RoBERT2VecTM because of its ability to combine both contextual information and structural details. This enables the model to uncover latent themes within the dataset more comprehensively .

The success of RoBERT2VecTM underscores the advantages of hybrid models in topic modeling, particularly in integrating diverse linguistic and contextual information for better insights and more accurate analysis.

Topics Representation. Figure 7 displays word clouds for different categories from our model, representing key thematic clusters identified in the

Figure 7: Word cloud for selected categories.

Hadith dataset. Meanwhile, the top words for some prominent topics can be examined using Figure 8. Note that only a selection of topics is displayed, as visually representing all 50 topics within this paper is not feasible.

Topics	Topic Top Keywords
1	ركعتين (Two units of prayer), صلاة (Prayer), مسجد (Mosque) -, تهجد (Night prayer) -, تراويح (Special nightly prayers during Ramadan) -, وتر (Odd-numbered prayer) -, قيام (Standing in prayer) -, دعاء (Supplication) -, سجود (Prostration) -, تكبير (God is the greatest)
2	إيمان (Faith) -, صدقة (Charitable giving) -, احسان (Beneficence) -, أخلاق (Morals) -, أمانة (Trustworthiness) -, حياء (Modesty) -, عفو (- Forgiveness), كرم (Generosity) -, صبر (Patience) -, شكر (Gratitude)
3	تفكر (Reflection) -, تدبر (Deep contemplation) -, قلب (Heart) -, روح (Soul) -, نفس (Self) -, عبادة (Worship) -, يقين (Certainty) -, معرفة النفس (Self-knowledge) -, تصفية القلب (Purification of the heart)

Figure 8: Prominent Words for Each Topic.

4 Related Work

In this section, we present related work, specifically focusing on topic modeling, machine learning (ML), and natural language processing (NLP) as applied to Hadith studies.

Topic modeling. Researchers in topic modeling have developed specialized techniques to address challenges. Neural models enhance traditional approaches by using various embeddings for improved topic construction (Larochelle and Lauly, 2012; Zhao et al., 2021; Terragni et al., 2021; Bhat et al., 2020). Recent topic modeling developments focus on embedding models over traditional techniques highlighting the efficacy of embedding-based techniques, as shown in studies by (Bianchi et al., 2020b) , (Dieng et al., 2020). The Contextualized Topic Model shows that pre-trained language models improve topic modeling (Bianchi et al., 2020a). (Wu et al., 2017) Introduced a topic modeling-based method for summarizing long texts by identifying key sentences. (Xie et al., 2020) used a multilingual BERT-enhanced LDA for multilingual text topic trends. (HABBAT et al., 2021) combined ProdLDA with AraBERT, surpassing traditional LDA in term of topic coherence. CTM,

based on ProdLDA, represents a significant advancement in topic modeling techniques (Asnawi et al., 2023).

Content-Based Study of Hadith. Different NLP based studies have been conducted to categorize Sahih Al-Bukhari into different topics (Jbara et al., 2009), (Naji Al-Kabi et al., 2005), (Abdelaal et al., 2019). (Rostam and Malim, 2021) proposed an alternative text categorization method by analyzing relationships between resources, integrating the Quran and Hadith. (Nohuddin et al., 2016) explored keyword interrelationships in Hadith chapters using text mining and cluster analysis to identify keyword frequencies and similarities across chapters. (SR et al., 2017) used different classifiers to recognize and assess the performance of Malay-translated Hadith based on sanad.

Narrator Based Study. The Shamela library [8] is a key resource for Hadith study, offering detailed narrator analysis but lacking automatic distinction between 'Sahih' and 'Da'ief' Hadith. While most research emphasizes the Isnad, several studies propose different models to classify Hadith based on narrators (Najeeb, 2021, 2020, 2016; Yotenka et al., 2022; Gaanoun and Alsuhaibani, 2022).

5 Conclusion

This study address the challenge of analyzing and extracting meaningful information from the Hadith. We developed a contextualized topic model with hybrid embeddings to identify significant topics in complex Arabic and Islamic contexts. Experiments show that our model, outperforms many state-of-the-art models in understanding thematic nuances in the Hadith dataset. This advancement enables precise topic identification and categorization into Islamic domains like *"Prayer and Worship"* and "Islamic Law and Society". The model's proficiency with key Islamic terms has been rigorously validated through comprehensive testing.

6 Limitations

Our study primarily focuses on addressing the challenge of topic modeling in non-Latin languages, specifically targeting Classical Arabic for the experiments. Extending our model to multiple non-Latin languages could present unique challenges and may not be straightforward due to the distinct complexities and nuances in the morphology and syntax of each language. Moreover, all experiments were conducted using the Hadith dataset. As a result, the model's effectiveness and generalizability might be limited when applied to other datasets with different linguistic characteristics.

Future research will expand the model to multiple languages to assess its cross-lingual performance and adaptability, demonstrating its robustness and effectiveness across diverse cultures.

[8] http://shamela.ws (Accessed on 16 Jul 2023)

Acknowledgments

This work was conducted within the PNRR project ITSERR - Italian Strengthening of the ESFRI RI RESILIENCE" (Avviso MUR 3264/2022) funded by EU – NextGenerationEU - Grant No IR0000014.

References

Hammam M. Abdelaal, Berihan R. Elemary, and Hassan A. Youness. 2019. Classification of hadith according to its content based on supervised learning algorithms. *IEEE Access*, 7:152379–152387.

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

Sania Aftar, Luca Gagliardelli, Amina El Ganadi, Federico Ruozzi, and Sonia Bergamaschi. 2024. A novel methodology for topic identification in hadith. In *Proceedings of the 20th Conference on Information and Research science Connecting to Digital and Library science (formerly the Italian Research Conference on Digital Libraries), Bressanone, Brixen, Italy - 22-23 February 2024*, volume 3643 of *CEUR Workshop Proceedings*, pages 117–125. CEUR-WS.org.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th international conference on computational semantics (IWCS 2013)–Long Papers*, pages 13–22.

Yousif A Alhaj, Jianwen Xiang, Dongdong Zhao, Mohammed AA Al-Qaness, Mohamed Abd Elaziz, and Abdelghani Dahou. 2019. A study of the effects of stemming strategies on arabic document classification. *IEEE access*, 7:32664–32671.

Mohammad O Alhawarat, Hikmat Abdeljaber, and Anwer Hilal. 2021. Effect of stemming on text similarity for arabic language at sentence level. *PeerJ Computer Science*, 7:e530.

Sanjeev Arora, Rong Ge, and Ankur Moitra. 2012. Learning topic models–going beyond svd. In *2012 IEEE 53rd annual symposium on foundations of computer science*, pages 1–10. IEEE.

Mohammad Hamid Asnawi, Anindya Apriliyanti Pravitasari, Tutut Herawan, and Triyani Hendrawati. 2023. The combination of contextualized topic model and mpnet for user feedback topic modeling. *IEEE Access*, 11:130272–130286.

Aqil M Azmi, Abdulaziz O Al-Qabbany, and Amir Hussain. 2019. Computational and natural language processing based studies of hadith literature: a survey. *Artificial Intelligence Review*, 52:1369–1414.

Sonia Bergamaschi, Stefania De Nardis, Riccardo Martoglia, Federico Ruozzi, Luca Sala, Matteo Vanzini, and Riccardo Amerigo Vigliermo. 2022. Novel perspectives for the management of multilingual and multialphabetic heritages through automatic knowledge extraction: The digitalmaktaba approach. *Sensors*, 22(11):3995.

Muzafar Rasool Bhat, Majid A Kundroo, Tanveer A Tarray, and Basant Agarwal. 2020. Deep lda: A new way to topic model. *Journal of Information and Optimization Sciences*, 41(3):823–834.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2020b. Cross-lingual contextualized topic models with zero-shot learning. *arXiv preprint arXiv:2004.07737*.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Kamel Gaanoun and Mohammed Alsuhaibani. 2022. Fabricated hadith detection: A novel matn-based approach with transformer language models. *IEEE Access*, 10:113330–113342.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Maarten R. Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *ArXiv*, abs/2203.05794.

Rahul Kumar Gupta, Ritu Agarwalla, Bukya Hemanth Naik, Joythish Reddy Evuri, Apil Thapa, and Thoudam Doren Singh. 2022. Prediction of research trends using lda based topic modeling. *Global Transitions Proceedings*, 3(1):298–304.

Nassera HABBAT, Houda ANOUN, and Larbi HASSOUNI. 2021. Arabertopic: A neural topic modeling approach for news extraction from arabic facebook pages using pre-trained bert transformer model. *International Journal Of Computing and Digital System*, 14.

Khitam Mahmoud Abdalla Jbara, Azzam T Sleit, and Bassam H Hammo. 2009. *Knowledge discovery in Al-Hadith using text classification algorithm*. University of Jordan.

Hugo Larochelle and Stanislas Lauly. 2012. A neural autoregressive topic model. *Advances in Neural Information Processing Systems*, 25.

Riccardo Martoglia, Sonia Bergamaschi, Federico Ruozzi, Matteo Vanzini, Luca Sala, and Riccardo Amerigo Vigliermo. 2023. Knowledge extraction, management and long-term preservation of non-latin cultural heritages - digital maktaba project presentation. In *Proceedings of the 19th The Conference on Information and Research science Connecting to Digital and Library science, IRCDL 2023, Bari, Italy, February 23-24, 2023*, volume 3365 of *CEUR Workshop Proceedings*, pages 153–161. CEUR-WS.org.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mohammed Mghari, Omar Bouras, and Abdelaaziz El Hibaoui. 2022. Sanadset 650k: Data on hadith narrators. *Data in Brief*, 44:108540.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736. PMLR.

Asha Rani Mishra and VK Panchal. 2022. A novel approach to capture the similarity in summarized text using embedded model. *International Journal on Smart Sensing and Intelligent Systems*, 15(1):1–20.

Hamdy Mubarak. 2017. Build fast and accurate lemmatization for arabic. *arXiv preprint arXiv:1710.06700*.

Moath Mustafa Ahmad Najeeb. 2016. Xml database for hadith and narrators. *American Journal of Applied Sciences*, 13(1):55–63.

Moath Mustafa Ahmad Najeeb. 2020. A novel hadith processing approach based on genetic algorithms. *IEEE Access*, 8:20233–20244.

Moath Mustafa Ahmad Najeeb. 2021. Towards a deep leaning-based approach for hadith classification. *European Journal of Engineering and Technology Research*, 6(3):9–15.

Mohammed Naji Al-Kabi, Ghassan Kanaan, Riyad Al-Shalabi, Saja I Al-Sinjilawi, and Ronza S Al-Mustafa. 2005. Al-hadith text classifier. *Journal of Applied Sciences*, 5(3):584–587.

PNE Nohuddin, Z Zainol, KF Chao, M Tarhamizwan, S Marzukhi, and A Nordin. 2016. Keyword based clustering technique for collections of hadith chapters. *International Journal on Islamic Applications in Computer Science And Technologies–IJASAT*, 4(3):11–18.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Nur Aqilah Paskhal Rostam and Nurul Hashimah Ahamed Hassain Malim. 2021. Text categorisation in quran and hadith: Overcoming the interrelation challenges using machine learning and term weighting. *Journal of King Saud University-Computer and Information Sciences*, 33(6):658–667.

Joseph Schacht. 1967. *The origins of Muhammadan jurisprudence*. Oxford University Press.

Mohammad Najib SR, Nurazzah Abd Rahman, N Alias, MN Alias, et al. 2017. Comparative study of machine learning approach on malay translated hadith text classification based on sanad. In *MATEC Web of Conferences*. EDP Sciences.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. Octis: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270.

Felipe Viegas, Sérgio Canuto, Christian Gomes, Washington Luiz, Thierson Rosa, Sabir Ribas, Leonardo Rocha, and Marcos André Gonçalves. 2019. Clusterwords: exploiting semantic word clustering representation for enhanced topic modeling. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 753–761.

Zongda Wu, Li Lei, Guiling Li, Hui Huang, Chengren Zheng, Enhong Chen, and Guandong Xu. 2017. A topic modeling based approach to novel document automatic summarization. *Expert Systems with Applications*, 84:12–23.

Qing Xie, Xinyuan Zhang, Ying Ding, and Min Song. 2020. Monolingual and multilingual topic analysis using lda and bert embeddings. *Journal of Informetrics*, 14(3):101055.

Rahmadi Yotenka, Sekti Kartika Dini, Achmad Fauzan, and Atina Ahdika. 2022. Exploring the relationship between hadith narrators in book of bukhari through spade algorithm. *MethodsX*, 9:101850.

Xiaowei Zhao, Deqing Wang, Zhengyang Zhao, Wei Liu, Chenwei Lu, and Fuzhen Zhuang. 2021. A neural topic model with word vectors and entity vectors for short texts. *Information Processing & Management*, 58(2):102455.