Topic Modeling for Hadith Corpus: A Comparison of Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), and BERTopic with AraBERT, XLM-R, MARBERT, and CAMeLBERT

Ibtisam Khalaf Alshammari[1,2,a] , Eric Atwell[1,b] and Mohammad Ammar Alsalka[1,c]
[1]University of Leeds, Leeds, United Kingdom
[2]University of Hafr Al-Batin, Hafr Al-Batin 39524, Kingdom of Saudi Arabia
[a]ML18IKFA, [b]E.S.ATWELL, [c]M.A.ALSALKA@leeds.ac.uk

ABSTRACT
The primary source of Islamic law, following the Holy Qur'an, is the collection of authentic Hadith attributed to the prophet of God, peace be upon him (PBUH). The status of the prophet's Hadith is evident in its being an explanation of the Qur'an and its abstract topics. With that, this research presents different topic modeling techniques to examine their performance on the authentic Hadith. Topic modeling is the process of clustering documents and words automatically in a textual domain. LDA and NMF are the most widely used topic modeling techniques. BERTopic is a modern technique based on BERT using pre-trained transformer-based language models for topic modeling. This study aims to apply the topic modeling approaches to the "*Matn*" part of the authentic Hadith. Then, we compare the performance of BERTopic using state-of-the-art pre-trained Arabic language models to LDA and NMF approaches. We finally evaluate the topic coherence of topic modeling methods using normalized pointwise mutual information (NPMI). The findings of this study indicate that the BERTopic model outperforms the LDA and NMF techniques in terms of overall performance.

***Keywords*: Classical Arabic Text, Hadith Corpus, Topic Modeling, BERTopic, LDA, NMF**

1. Introduction

The progress of artificial intelligence (AI), particularly in the area of natural language processing (NLP), has proven beneficial in addressing challenges across various multidisciplinary domains. One of its most prominent uses is processing religious texts to enhance comprehension of textual data, extract valuable information, and assist with knowledge discovery.

حَدَّثَنَا سُلَيْمَانُ بْنُ حَرْبٍ، حَدَّثَنَا حَمَّادٌ، عَنْ ثَابِتٍ، عَنْ أَنَسٍ، وَأَيُّوبَ، عَنْ أَبِي قِلاَبَةَ، عَنْ أَنَسٍ ـ رضى الله عنه ـ **أَنَّ النَّبِيَّ صلى الله عليه وسلم كَانَ فِي سَفَرٍ، وَكَانَ غُلاَمٌ يَحْدُو بِهِنَّ يُقَالُ لَهُ أَنْجَشَةُ، فَقَالَ النَّبِيُّ صلى الله عليه وسلم رُوَيْدَكَ يَا أَنْجَشَةُ، سَوْقَكَ بِالْقَوَارِيرِ . قَالَ أَبُو قِلاَبَةَ يَعْنِي النِّسَاءَ.**

Narrated Anas: **The Prophet (ﷺ) was on a journey and a slave named Anjasha was chanting (singing) for the camels to let them go fast (while driving). The Prophet (ﷺ) said, "O Anjasha, drive slowly (the camels) with the glass vessels!" Abu Qilaba said, "By the glass vessels' he meant the women (riding the camels).**

Figure 1: Hadith Example, Matn Part in Bold.

Hadith (the plural is ahadith) refers to the prophet Muhammad's actions, sayings, orders, or silent approval delivered through a chain of narrators. Hadith was written in Classical Arabic text, and it could be a short sentence or a long paragraph describing what the prophet said in a particular incident, a dialogue of a conversation between the prophet and someone else, or a story related by the prophet's companions to explain the prophet's acts on a particular topic such as prayers. Thus, the importance of the Hadith text lies in clarifying the abstract topics of the Holy Qur'an (Rostam and Malim, 2021). Each Hadith contains two parts, as illustrated in Figure 1. The first part is *Isnad* or *Sanad* (سند), representing the reverse chronological chain of narrators. The second essential part is the context known as the *Matn* (متن), shown in bold text, representing the actual teaching of the prophet Muhammad.

Topic modeling is an unsupervised machine-learning technique that can scientifically identify related words from a collection of documents. It clusters documents in a textual domain, where each document is represented by a topic probability distribution, thus clustering documents based on a high probability of the same topic. Topic modeling has been adapted and developed for various applications, including information retrieval (Yi and Allan, 2009), text classification (Xia et al., 2019), text summarisation (Roul et al., 2019), and search engines (Bukhari and Liu, 2018).

Popular topic modeling approaches include latent Dirichlet allocation (LDA), which was suggested in 2003, and non-negative matrix factorization (NMF). Such topic modeling methods represent a document as a mixture of latent topics without considering the semantic relationships among words within a topic. BERTopic is a topic modeling approach based on BERT (Bidirectional Encoder Representations from Transformers) that utilizes pre-trained transformer-based language models to extract document embeddings and class-based TF–IDF to generate topic representations. As the BERTopic technique showed promising results such as this study (Abuzayed and Al-Khalifa, 2021), we aim to examine the BERTopic technique using pre-trained transformer-based Arabic language models on the authentic Hadith text, "*Matn*", and compare its findings to LDA and NMF techniques. Moreover, the motivation for applying topic modeling to Hadith's text is to automatically discover hidden topics and capture the semantic relationships between words.

This paper is organized as follows: Section 2 reviews related work. Section 3 describes the detailed methodology for topic modeling techniques. Results and discussion of the experimentation have been carried out in Section 4, which is followed by the conclusion of the work in Section 5.

2. Related Work

A wide range of studies have been conducted in recent years to address the challenge of Arabic topic modeling and its applications. For example, regarding LDA, a probabilistic topic modeling algorithm to extract Qur'anic topics (Siddiqui et al., 2013; Alhawarat, 2015), Habbat et al. (2021) applied LDA and NMF to discover the topics discussed in Arabic tweets. Moreover, an experimental study evaluated the performance of LDA, NMF, and BERTopic techniques on modern standard Arabic (MSA) documents (Abuzayed and Al-Khalifa, 2021).

However, few studies have implemented AI and NLP approaches to extract topics from the Hadith corpora. This section presents an overview of previous works related to topic modeling approaches for Hadith texts.

Harrag and El-Qawasmah (2009) proposed an automatic classification of Arabic documents using the artificial neural network (ANN) algorithm and the singular value decomposition (SVD) technique to enhance the feature extraction process. They used the Hadith corpus, a collection of the nine Prophetic Encyclopaedia books. Implementing the ANN model with SVD demonstrated an adequate representation and classification of Arabic documents by achieving an F1 score of 88.33%.

Another study was presented by Al-Kabi et al. (2014) to evaluate the performance of three text classification algorithms, namely, Naïve Bayes (NB), Bagging and LogiBoost. The authors utilized the Sahih Al-Bukhari Book in their research to classify Hadith texts into one of four topics: ablution, prayer, obligatory charity (*Zakat*) and fasting. Based on their analysis, the Naive Bayes classifier categorized Hadith texts effectively.

Ramzy et al. (2023) conducted a comprehensive study on Hadith classification using a novel author-based Hadith classification dataset (ABCD). The authors' primary focus was to classify Hadith texts according to their origin of narration leveraging machine learning (ML) and deep learning (DL) approaches to *Sanad* and *Matn* separately. Consequently, they determined that ML achieved the best result on *Matn* data with a 77% F1 score, while DL obtained a 92% F1 score on *Sanad* data.

3. Experiments

This research conducts several experiments using topic modeling strategies, including BERTopic using pre-trained transformer-based Arabic language models, LDA, and NMF. The number of topics is determined based on the number of categories of the Hadith corpus. Thus, we began with an initial set of 38 topics, based on the lower number of topics in Ibn Maja's book, and subsequently increased each succeeding step by a consistent value of 10. The Gensim library was used to implement LDA, and the Sklearn library was used to implement the NMF algorithm. The methodological steps for implementing the previous topic modeling techniques in authentic Hadith texts are illustrated in Figure 2.

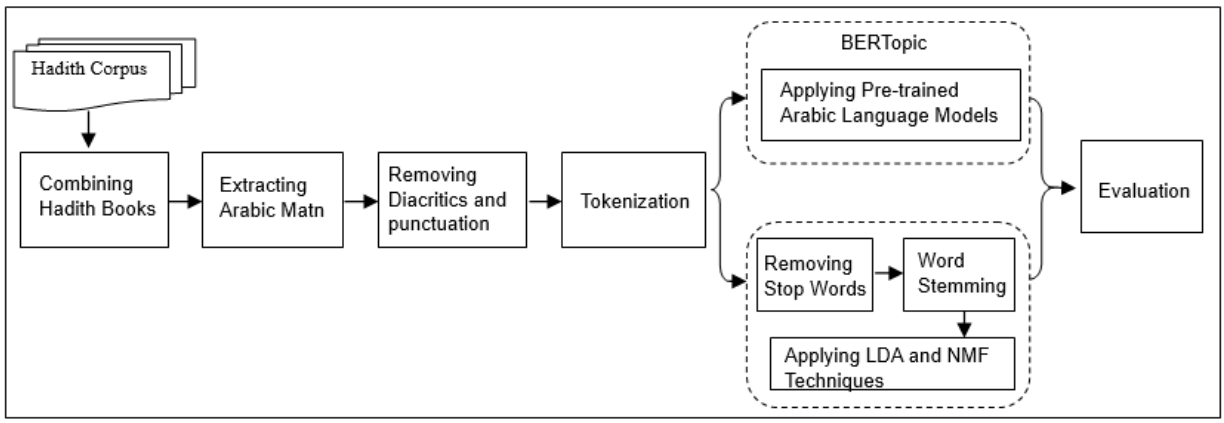

Figure 2: Hadith Topic Modeling Framework.

3.1 Data Collection and Preprocessing

The dataset used in this study is the Leeds University and King Saud University (LK) Hadith corpus presented by Altammami et al. (2020). This well-structured Arabic–English parallel Islamic Hadith corpus contains 39,038 annotated Ahadith of six canonical Hadith books. Table 1 explains the number of topics, "chapters", for each book, and the total number without duplication is 238 topics. Initially, the six Hadith books were combined into a single CSV file, extracted the "*Arabic Matn*" part from the CSV file, eliminated unnecessary data, such as diacritics[1], and then tokenized documents as a last step in the BERTopic technique. Data cleaning is essential for LDA and NMF approaches. Therefore, we removed unnecessary data such as punctuation and digits, and tokenized documents by splitting them into a list of tokens, removing stop words, and finally, word stemming.

Table 1: Number of Topics for each Book.

Book Name	Number of Topics
Sahih Bukhari	97
Sahih Muslim	56
Sunan Nesa'i	51
Sunan Tirmizi	49
Sunan Abu Daud	43
Sunan Ibn Maja	38

3.2 Latent Dirichlet Allocation

LDA is a popular method of topic modeling and a generative probabilistic model that is commonly used to analyze collections of discrete data, such as text corpora (Blei et al., 2003). The LDA model comprises a three-level hierarchical Bayesian model. Each item in the collection was represented as a finite mixture over an underlying set of topics, and each topic was represented as an infinite mixture over a collection of topic probabilities. Accordingly, LDA is considered an effective model for obtaining an explicit representation of a document.

[1] Symbols added to Arabic letters to help in pronouncing them correctly.

3.3 Non-Negative Matrix Factorisation

NMF is a linear–algebraic optimization algorithm and a non-probabilistic technique that uses matrix factorization. NMF decomposes high-dimensional vectors into two lower-dimensional representations; thus, these lower-dimensional vectors and their corresponding coefficients are non-negative. The NMF model applies the term frequency–inverse document frequency (TF–IDF) measurement to determine a word's importance to a collection of documents (Lee and Seung, 1999).

3.4 BERTopic

Grootendorst (2022) published a BERTopic technique of topic modeling that enhances the cluster embedding approach by using state-of-the-art language models and implementing a class-based TF–IDF mechanism to generate dense clusters. BERTopic employs the uniform manifold approximation and projection (UMAP) approach to minimize the dimensionality of the embeddings before the document's clustering procedure. It also generates clusters from a density-based approach using the HDBSCAN approach. To extract the document embeddings, BERTopic can use two methods. First, the Sentence–BERT (SBERT) framework can be used by two models: 'distlbert-base-nli-stsbmean-tokens' for the English language and 'xlm-rbert-base-nli-stsb-mean-tokens' for 50+ languages, one of which is the Arabic language (Reimers and Gurevych, 2019). Second, the Flair package can perform the embedding step for state-of-the-art pre-trained models (Akbik et al., 2019).

3.5 Evaluation

One of the most widely used metrics for measuring the performance of topic modeling techniques is topic coherence. The topic coherence for each topic modeling approach used in this study was evaluated using normalized pointwise mutual information (NPMI; Bouma, 2009). This coherence measure has been shown to reasonably mimic human judgment (Lau et al., 2014). The human-rated coherence score ranges between -1 and 1, where 1 refers to an ideal match. As mentioned earlier, the proposed methods were run in a range of 38 to 128 topics and were increased by 10 for each step. The NPMI score was calculated for each step of the previous models.

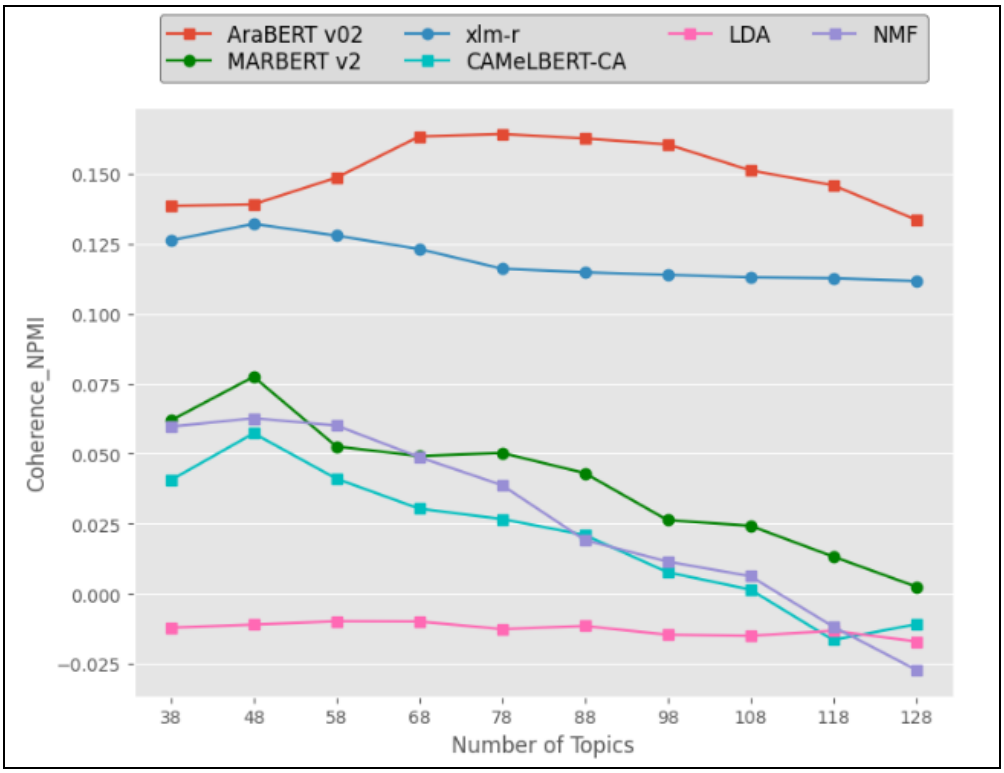

Figure 3: NPMI Scores for BERTopic with Pre-trained Models, LDA, and NMF Tools.

4. Results and Discussion

This section illustrates the results obtained from the conducted experiments. Figure 3 shows the topic models' performance in different numbers of topics. As indicated in Section 3, the initial sample size for the experiments consisted of 38 topics, which was determined based on the topics of the smallest dataset in the Hadith corpus. The experimentation was conducted until an overall total of 128 topics, as it was noticed that the coherence scores of the models began to demonstrate a decrease.

The results show that BERTopic with pre-trained Arabic language models achieved reasonably efficient performance of topic coherence compared to LDA and NMF methodologies. Notice that the LDA tool presented negative NPMI scores, in line with the study's low performance. At the same time, the NMF technique started with positive numbers and then gradually began to reduce to represent negative scores.

Figure 3 illustrates that some pre-trained models based on the Flair framework achieved high positive NPMI scores. The findings of the AraBERT V02 (Antoun et al., 2020) model represent the high behavior of BERTopic language models, while the MARBERT V2 (Abdul-Mageed et al., 2020) model obtained acceptable results. However, the outcome of CAMeLBERT for Classical Arabic text, (Obeid et al., 2020), did not meet our expectations since it started with good performance in topics between 38 and 48, but then it showed lower scores. In contrast, the performance of BERTopic using SBERT language models, XLM-R (Reimers and Gurevych, 2019), produced positive scores and was pretty similar to the outcomes of the AraBERT V02 model.

Generally, the performance of the proposed topic models was the best between topics 48 and 78 for most of the topic modeling approaches. It is worth stating that the performance of topic coherence exhibits deterioration when the number of topics increases.

5. Conclusion

This paper provides a comparison of topic modeling methods on a Classical Arabic text, the Islamic Hadith corpus. We examined the performance of the BERTopic with different pre-trained-based Arabic language models, LDA, and NMF techniques. The performance of AraBERT V02 and XLM-R were the best among the other models. Although the BERTopic technique has demonstrated considerable effectiveness, achieving high topic coherence scores remains difficult. For future work, we could improve the performance of the pre-trained-based Arabic language models by using additional hyperparameters.

Acknowledgements

The first author would like to express her deepest gratitude to the University of Hafr Al-Batin for sponsoring her PhD studies at the University of Leeds through the Saudi Cultural Bureau. She is also thankful for the anonymous comments.

References

Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2020). ARBERT & MARBERT: deep bidirectional transformers for Arabic. arXiv preprint arXiv:2101.01785.

Abuzayed, A., & Al-Khalifa, H. (2021). BERT for Arabic topic modeling: An experimental study on BERTopic technique. Procedia computer science, 189, 191-194.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019, June). FLAIR: An easy-to-use framework for state-of-the-art NLP. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations) (pp. 54-59).

Alhawarat, M. (2015). Extracting topics from the holy Quran using generative models. International Journal of Advanced Computer Science and Applications, 6(12), 288-294.

Al-Kabi, M. N., Wahsheh, H. A., Alsmadi, I. M., & Al-Akhras, A. M. A. (2015). Extended topical classification of Hadith Arabic text. Int. J. Islam. Appl. Comput. Sci. Technol, 3(3), 13-23.

Altammami, S., Atwell, E., & Alsalka, A. (2020). The Arabic-English parallel corpus of authentic Hadith. International Journal on Islamic Applications in Computer Science And Technology, 8(2).

Antoun, W., Baly, F., & Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. arXiv preprint arXiv:2003.00104.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. Proceedings of GSCL, 30, 31-40.

Bukhari, A., & Liu, X. (2018). A web service search engine for large-scale web service discovery based on the probabilistic topic modeling and clustering. Service Oriented Computing and Applications, 12, 169-182.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Habbat, N., Anoun, H., & Hassouni, L. (2021). Topic modeling and sentiment analysis with lda and nmf on moroccan tweets. In Innovations in Smart Cities Applications Volume 4: The Proceedings of the 5th International Conference on Smart City Applications (pp. 147-161). Springer International Publishing.

Harrag, F., & El-Qawasmah, E. (2009, August). Neural Network for Arabic text classification. In 2009 Second International Conference on the Applications of Digital Information and Web Technologies (pp. 778-783). IEEE.

Lau, J. H., Newman, D., & Baldwin, T. (2014, April). 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 (pp. 530-539).

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., ... & Habash, N. (2020, May). CAMeL tools: An open source python toolkit for Arabic natural language processing. In Proceedings of the 12th language resources and evaluation conference (pp. 7022–7032).

Ramzy, A., Torki, M., Abdeen, M., Saif, O., ElNainay, M., Alshanqiti, A., & Nabil, E. (2023). Hadiths Classification Using a Novel Author-Based Hadith Classification Dataset (ABCD). Big Data and Cognitive Computing, 7(3), 141.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Rostam, N. A. P., & Malim, N. H. A. H. (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.

Roul, R. K., Mehrotra, S., Pungaliya, Y., & Sahoo, J. K. (2019). A new automatic multi-document text summarization using topic modeling. In Distributed Computing and Internet Technology: 15th International Conference, ICDCIT 2019, Bhubaneswar, India, January 10–13, 2019, Proceedings 15 (pp. 212-221). Springer International Publishing.

Siddiqui, M. A., Faraz, S. M., & Sattar, S. A. (2013, December). Discovering the thematic structure of the Quran using probabilistic topic model. In 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences (pp. 234-239). IEEE.

Xia, L., Luo, D., Zhang, C., & Wu, Z. (2019, May). A survey of topic models in text classification. In 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 244-250). IEEE.

Yi, X., & Allan, J. (2009). A comparative study of utilizing topic models for information retrieval. In Advances in Information Retrieval: 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings 31 (pp. 29-41). Springer Berlin Heidelberg.