



Hadith Topic Modeling Based on Theme Similarity with the BERTopic Algorithm

Irma Dwiyanti
Sekolah Budaya Adhikari (SBA)
Bandung, Indonesia
irmadwiyanti.sky@gmail.com

Muhammad Alfiyan Nurwahibulloh Rohayana
Cyber Technology Integration Ltd.
Jakarta, Indonesia
alfiannurwahibulloh@gmail.com

Idfa Billati Hiya Ahsan
*Center for Standardization and Industrial Services for
Engineering Materials and Goods*
Bandung, Indonesia
billatiidfa@gmail.com

Gevira Zahra Shofa
Dicoding Akademi Indonesia Ltd.
Bandung, Indonesia
gevirazahra1055@gmail.com

Muhammad Ahsani Taqwim
LEN Persero Ltd.
Bandung, Indonesia
05taqwim@gmail.com

Abstract — This study aims to develop a hadith topic modeling system based on thematic similarity using the BERTopic algorithm to support more structured and efficient Islamic studies. The research method uses a data exploration-based approach with the BERTopic algorithm that combines BERT semantic representation, UMAP dimensionality reduction, and HDBSCAN clustering. The dataset used is sourced from Kaggle containing thousands of hadith texts in English from various books, which are processed through the stages of data pre-processing, embedding transformation, and evaluation using topic coherence metrics and interactive visualization. The results show that BERTopic successfully identified 292 topics with clear visualization through the *Intertopic Distance Map* and *Topic Similarity Heatmap*; however, approximately 55.4% of documents are categorized as *outliers* due to high semantic variation and limited context. Despite the challenges in handling *outlier documents*, this study proves that BERTopic is effective in clustering hadith based on thematic similarity and producing semantically coherent topics. The resulting model can be the basis for the development of a hadith recommendation system, classification of da'wah content, and a theme-based religious learning curriculum, as well as opening up opportunities for the application of NLP technology in Islamic studies more broadly.

Keywords - BERTopic, Hadith, Islamic Studies, Natural Language Processing, Topic Modeling

I. INTRODUCTION

Abban [1] Hadith is one of the primary sources of Islamic teachings, containing the sayings, actions, and decrees of the Prophet Muhammad (peace be upon him). With its vast number and high diversity, managing hadith information presents a unique challenge in the world of Islamic studies. The need for a system capable of grouping hadith based on shared themes is highly relevant to support more structured and efficient Islamic studies.

Blei, Ng, & Jordan [2] The development of natural language processing (NLP) technology has enabled in-depth and automatic exploration of text data. One method widely used in NLP is topic modeling, which is the process of finding hidden patterns in large text collections by grouping them into similar themes. Devlin et al. [3] used the BERTopic algorithm in this study. This modern approach combines semantic representations from BERT (Bidirectional Encoder Representations from Transformers) with clustering (HDBSCAN) and dimensionality reduction (UMAP).

Grootendorst, M. [4] BERTopic: Neural topic modeling with class-based TF-IDF and BERT embeddings. GitHub. to produce coherent and easy-to-understand topics [5], [6].

Abban [1] used hadith data sourced from the public Kaggle dataset titled “*Sunnah (Hadith) Dataset*” (Ronnie Aban, 2022), which contains thousands of hadith texts in English from various books. This data was processed through text cleaning, embedding transformation using the BERT model, and visualization of topic clustering results using BERTopic in Google Colab. In addition, the results of topic exploration and their frequencies were also recorded in a spreadsheet for evaluation and interpretation purposes.

The primary objective of this research is to model hadith topics based on thematic similarity using BERTopic. This approach is expected to create more accurate and contextual thematic groupings of hadith and can serve as a foundation for developing a more intelligent hadith search and learning system.

Although numerous studies have explored the processing of religious texts using NLP approaches, most still focus on the text of the Quran or simple linguistic processing such as keyword matching and frequency analysis. Studies of hadith collections as topic modeling objects are still relatively limited, despite their rich thematic potential. Hadith cover various aspects of life, from worship and social interactions, morals, to socio-political issues, which are highly worthy of thematic study using a data-driven approach.

Blei, Ng, & Jordan [2], on the other hand, conventional topic modeling methods such as LDA (*Latent Dirichlet Allocation*) have limitations in understanding semantic context and meaning relationships between words, especially in short texts such as hadith [7], [8]. Devlin et al. [3] BERTopic is present as a solution to these limitations by utilizing the power of the BERT model that has been trained on large-scale natural language data, so that it is able to capture contextual meaning in more depth. By adding adaptive dimensionality reduction and clustering techniques, BERTopic enables the formation of topics that are more accurate, flexible, and relevant to the true meaning of the text.

The BERTopic approach used in this study to model hadith topics not only provides a new way to understand the thematic structure of hadith but also opens up opportunities for the broader application of NLP technology in Islamic studies. The results of this modeling can be used as a basis for developing hadith recommendation systems, classifying da'wah content, or even developing theme-based religious learning curricula.

Against this backdrop, this research is expected to contribute to the development of interdisciplinary knowledge between Islamic studies and artificial intelligence, as well as to strengthen digital literacy in the management of Islamic scientific treasures. The BERTopic-based approach to this collection of hadith is also expected to bridge the needs of modern Muslims to access and understand hadith in a more structured and relevant way to the current context.

II. RELATED WORKS

Topic modeling has become an important tool in text analysis, including religious texts such as the Hadith and the Quran, to identify semantically relevant themes. In recent years, particularly between 2019 and 2024, several studies have explored the use of topic modeling algorithms such as BERTopic, Latent Dirichlet Allocation (LDA), and Non-Negative Matrix Factorization (NMF) on Islamic religious texts or Arabic corpora. These studies provide a crucial foundation for this research, which focuses on topic modeling of Hadith based on thematic similarity using the BERTopic algorithm. Below is an in-depth review of some relevant related work.

Alshammari et al. [9] conducted a comparative study on topic modeling on a Hadith corpus, comparing the performance of three methods: LDA, NMF, and BERTopic. They used pretrained Arabic language models such as AraBERT, XLM-R, MARBERT, and CAMElBERT to generate better semantic representations. This study found that BERTopic produced more semantically coherent topics than LDA and NMF, mainly due to the transformer-based model's ability to capture contextual relationships in Arabic texts. This study is highly relevant because it directly applies BERTopic to Hadith, which aligns with the objective of this study to cluster Hadith based on thematic similarities. Their approaches in data preprocessing, such as tokenization and stopword removal, as well as evaluation using topic coherence metrics, can serve as a reference for this study in addressing the linguistic complexity of Hadith texts.

In the context of Arabic language processing, Alqahtani et al. [10] explored the use of multilingual transformer models for segmenting Arabic text without punctuation. Meanwhile, Abdullah et al. [11] examined the application of AraBERT for Arabic sentiment and comment analysis, demonstrating the superiority of transformer models in understanding Arabic context.

Aftar et al. [12] proposed a transformer-based approach, RoBERT2VecTM, for topic extraction in Islamic studies, specifically on the “Matn” (main content) of Hadith. While not explicitly using BERTopic, this approach leverages the RoBERTa model, which shares architectural similarities with BERT, to analyze the semantic structure of Hadith texts. This study highlights the challenges in handling Arabic texts, such as complex dialectal and morphological variations, and emphasizes the importance of contextual embedding to generate meaningful topics. The relevance of this study lies in its focus on Hadith and the use of transformer technology, which can be adapted for the BERTopic approach in this research, especially in exploring thematic similarities between Hadiths.

Abuzayed and Al-Khalifa [13] evaluated BERTopic for topic modeling on Arabic texts in general, comparing it with LDA and NMF using the Normalized Pointwise Mutual Information (NPMI) metric. This study showed that BERTopic is superior in capturing semantic relationships in Arabic texts, which is important considering that Hadith are

mostly written in Arabic. They also addressed preprocessing challenges, such as text normalization and non-standard character removal, which are relevant for Hadith analysis. While not specific to Hadith, this study provides valuable methodological insights for applying BERTopic to Arabic texts, including the selection of pretrained language models and evaluation techniques, which can be applied in this study to ensure the quality of the generated topics [14], [20].

Jamshaid et al. [15] applied LDA for topic modeling on Quranic verses translated into English. This study identified key themes using metrics such as perplexity score and topic coherence, and visualized the results through word clouds and intertopic distance maps. Although using LDA [16] and focusing on the Quran, this study is relevant because the Quran has a close thematic relationship with Hadith in the context of Islamic scholarship. Their approach in evaluating topics and handling religious texts can be a foundation for expanding the use of BERTopic on Hadith, especially in terms of theme visualization and interpretation.

Mosa [17] presents a comprehensive literature review on Hadith text mining, covering methods such as LDA, NMF, and transformer-based approaches. This review highlights challenges in Hadith text analysis, such as narrative diversity, varying text lengths, and the linguistic complexity of Arabic. This study also discusses the potential of transformer-based models, such as BERT, to improve topic modeling accuracy. This study is relevant because it provides a broad context on Hadith research and emphasizes the need for advanced approaches such as BERTopic to address semantic challenges in Hadith texts.

Herwinsyah et al. [18] used BERTopic to model topics in the Quran, analyzing all 6,236 verses to identify eight main topics, such as Allah, Paradise, Prophets, and Resurrection. This research includes text preprocessing, implementation of BERT embedding, and visualization through intertopic distance maps and word clouds. Although focused on the Quran, this approach is highly relevant because it demonstrates the success of BERTopic in handling Islamic religious texts, which can be adapted for Hadith. Their methodology, including the use of language embedding and topic coherence evaluation, provides practical guidance for this research, especially in handling thematic similarities between texts.

Overall, these studies demonstrate that topic modeling, particularly with BERTopic, has shown great potential in the analysis of Islamic religious texts. However, few studies have specifically applied BERTopic to Hadith, focusing on thematic similarity. This study aims to fill this gap by leveraging BERTopic to cluster Hadith based on similar themes, using Arabic embeddings and evaluation metrics such as topic coherence and diversity to ensure meaningful and interpretive results.

III. RESEARCH METHODS

Research methods are rules used in collecting data from various sources by carrying out several steps, namely

conducting literature studies, studying several references related to the research, and collecting useful data to obtain solutions to the problem formulation.

This research method employs a *data-driven approach* using *Natural Language Processing (NLP)*-based topic modeling. The main objective of this study is to cluster Hadith texts based on similar themes using the BERTopic algorithm, combining the semantic representation power of BERT, dimensionality reduction with UMAP, and the HDBSCAN clustering algorithm.

A. Research Stages

The approach taken in this research is relevant in addressing the need for an automated classification system for Islamic texts that can assist research in Islamic learning and da'wah. This research uses a series of stages, starting with identifying the problem and continuing with the classification process, as shown in Figure 1.

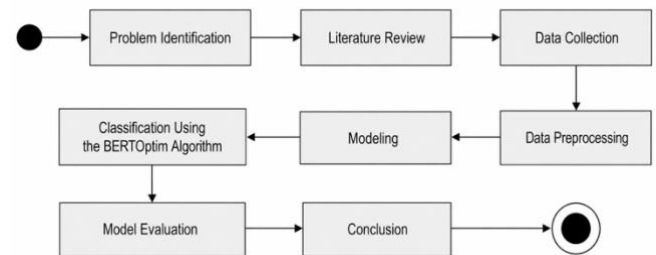


Fig 1. Research Stages

1) Identification of problems

The initial stage of this research began by identifying the problem of grouping collections of hadith texts by theme. The large number of hadith and the diversity of their content make manual information retrieval inefficient. Therefore, grouping hadith by theme requires an artificial intelligence-based approach that can automate the grouping process.

2) Literature Study

Previously, there was still a lack of research discussing grouping hadiths based on similar themes; there were only a few, and this created a gap for researchers to create this system. Researchers conducted a literature review of the problems identified in Islamic texts such as the Hadith. The study included an evaluation of models that can implement topic modeling techniques such as BERTopic on Islamic texts such as the Hadith.

3) Data collection

The data used in this study comes from a public dataset compiled by Ronnie Aban in 2022 on Kaggle. It contains tens of thousands of hadiths from nine imams, divided into three classes: narrators, Arabic, and translations (the translations are already in Indonesian). This dataset is stored in CSV format for processing and is ready for analysis.

4) Data Pre-Processing

At this stage, researchers pre-process the data to produce a dataset ready for analysis using a *Natural Language Processing* (NLP) model. The data pre-processing steps include:

- Text cleaning or removing punctuation, numbers, and non-alphabetic characters.
- Performs text conversion to lowercase and stopword removal.
- Next, the text is converted into a vector representation using BERT embedding to capture the semantic context of the words in the hadith text.

5) Modeling

The topic modeling stage was performed using the BERTopic algorithm. This algorithm works by combining the obtained BERT embedding and dimensionality reduction using UMAP to simplify the clustering process. The HDBSCAN algorithm is then used to group the hadith texts into clusters representing specific topics. Each cluster formed is assumed to represent a thematic topic.

6) Classification with BERTopic Algorithm

Each hadith text will be classified into the most appropriate topic based on its semantics after the model successfully identifies topic clusters. This classification process generates topic labels for each hadith. The BERTopic modeling technique provides interactive visualizations of topic distribution through tools such as topic maps and term score plots, which are used to visually evaluate topics.

7) Model Evaluation

The model will be evaluated using topic coherence and topic diversity metrics. Topic coherence measures the extent to which words within a topic are semantically related, and diversity measures the variety of generated topics. The evaluation results show that the BERTopic model is capable of generating topics that are not only semantically consistent but also encompass various dimensions of the hadith content. This evaluation was conducted using visualization and manual interpretation of sample topics.

The final stage of this research method is the conclusion of the evaluation process. This model can serve as the initial foundation for developing an Islamic content classification system, such as a hadith recommendation system or thematic learning in Islamic studies.

B. BERTopic Algorithm

An unsupervised learning algorithm that combines the power of the BERT language model with modern clustering techniques. The first stage involves preprocessing specifically for Arabic text, including tashkeel removal, hamza normalization, and right-to-left (RTL) handling using libraries such as pyarabic and arabic-reshaper. The cleaned data is then converted into numeric vectors using a pre-trained Arabic-specific BERT model (asafaya/bert-base-

arabic), which is capable of capturing contextual meaning and unique morphological variations in Arabic.

IV. RESULTS AND DISCUSSION

A. Result

These findings successfully demonstrate that the system is effective in applying BERTopic in clustering hadith topics into similar documents. The steps were carried out according to the methodology, where the process begins with data exploration. In this study, the data was relatively clean, with only 1% missing values. Data cleaning was then performed after identifying the data condition by removing the missing values. This is because missing values do not significantly affect the model if they are few in number, and this also occurs due to empty data and missing data.

Once the cleaning process is complete, the text length, character count, and word count for each hadith are displayed. Hundreds of characters and hundreds of words were found in this dataset, with significant variations in word count.

Visualization of the distribution of text length and word count shows that most of the hadith have text length and word count that are relatively concentrated at certain values, which indicates consistency in the structure of the analyzed data.

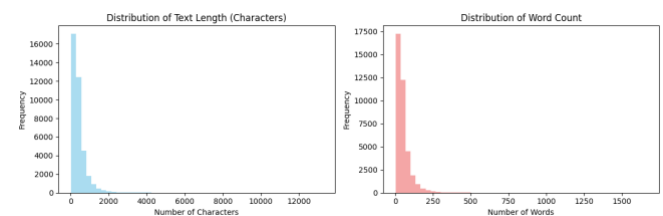


Fig. 2. Distribution of Text Length and Words

The findings in this study after training the BERTopic model, it was found that the total generated topics were 292 topics, with a total of 21,000 documents (hadith) categorized as *outliers* or not included in any topic group. This indicates that approximately 55.4% of documents cannot be classified effectively, which is likely due to too high semantic variation or a lack of representative features in these documents. This finding indicates that although BERTopic is able to group documents into many topics, there are challenges in handling short texts or texts with limited context. And this could be one of the outliers that is so large not only the similarity of topics but also in similar words that are difficult to cluster because there are many similarities.

To understand the system's results, advanced visualizations such as topic distance maps and topic similarity heatmaps are essential to uncover hidden relationships and structures among the generated topics. Topic distance maps typically use dimensionality reduction algorithms to project semantic similarity between topics into a two-dimensional space, placing more similar topics closer together and less similar topics farther apart.



Fig. 3. Intertopic Distance Map

Figure 4 illustrates how topics are distributed across semantic space, with most topics falling into distinct clusters. This indicates that each topic has a unique theme that can be distinguished from other topics.

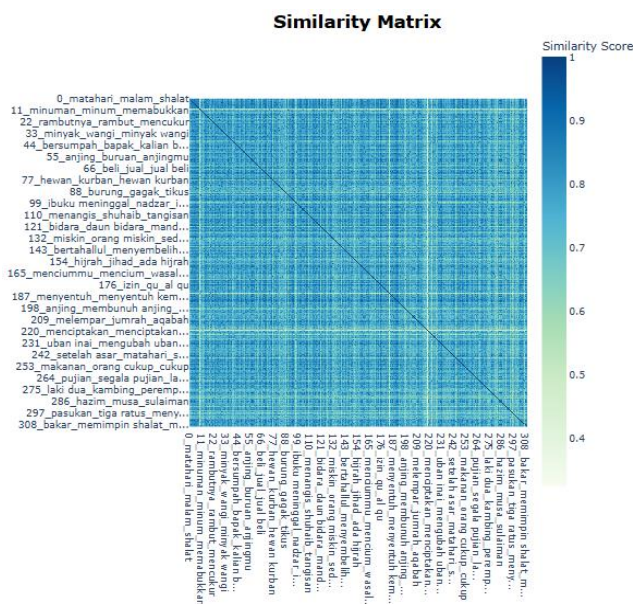


Figure 4: Topic Similarity Heatmap

Figure 4 shows the correlation between topics based on keyword similarity, and shows that some topics have a high enough semantic relationship, making it possible to combine them in further analysis. To understand the keyword representation of each topic, a bar chart visualization in Figure 5 is used that shows the top ten keywords for each topic.

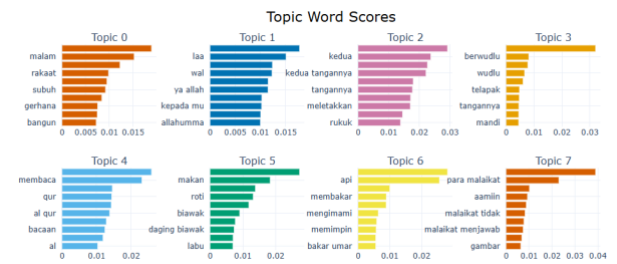


Fig. 5. Topic Keywords with Barchart Visualization

This helps in manual interpretation of the meaning of the topic, and makes it easier for researchers to name or label each topic according to the context of the hadith being discussed. Overall, these results confirm that BERTopic is suitable for topic clustering of hadith text data, but additional approaches are needed to handle outlier documents. One approach that could be implemented is using deeper *embedding* or post-clustering classification to manage unclassified documents. This research also opens up opportunities for further integration of artificial intelligence approaches to improve the accuracy and adaptability of topic modeling in distributed databases with complex natural language structures.

B. Discussion

The high number of documents categorized as *outliers* raises concerns about the model's effectiveness in capturing the full range of document content. This suggests that while BERTopic is capable of detecting complex topic structures, it is not optimal for capturing documents with limited context or that are too unique to be automatically classified. This is an important consideration in the development of distributed database security systems, as unclassified documents may contain sensitive information that escapes thematic scrutiny.

The presence of a large number of *outliers* also opens up discussion about the quality of the text representation used. The embedding used by BERTopic, although derived from a modern transformer model, may still be less contextual to religious language such as hadith texts in Indonesian. This is in line with the findings of several previous studies that suggest the use of domain - *specific embedding models* to make clustering results more relevant to the analyzed content.

Furthermore, the visualization results, in the form of a distance map between topics and a *heatmap of topic similarity*, indicate a fairly strong semantic correlation between several topics. This reflects the potential for thematic overlap, possibly stemming from similar terminology or topics that overlap contextually. Therefore, combining several topics into a single larger theme through a *hierarchical topic modeling approach* could be a further strategy for simplifying and clarifying thematic structures.

The use of keyword visualization also provides important insights into the topic labeling process. The clarity of keywords within each topic strengthens the validity of the clustering results and demonstrates how the model

successfully captures the core discussion of the text. This can be utilized in developing thematic search systems in distributed databases, for example, by connecting the generated topics to a context-based query filtering system.

From an application perspective, these results provide an initial foundation for a content-based security system in distributed SQL databases. By mapping themes and recognizing thematic anomalies, the system can be further developed to detect suspicious activity, such as the appearance of unusual topics over time that could indicate data compromise or unauthorized requests [7], [8]. This is where the integration between topic modeling and data security becomes increasingly relevant and strategic.

Finally, the effectiveness of this technique is highly dependent on the quality of the input data, model parameters, and visualization method. Therefore, further experimentation with approaches such as *fine-tuning*, incorporating supervised learning, or applying *zero-shot classification* is highly recommended to improve model accuracy and coverage, especially in the context of highly distributed and dynamic data.

V. CONCLUSION

This study successfully implemented the BERTopic algorithm for hadith topic modeling by identifying hundreds of distinct topics from the analyzed hadith dataset. The system successfully grouped nearly half of the total hadith documents into thematically similar topics, while categorizing most of the remaining documents as *outliers*. The resulting visualizations, including an *Intertopic Distance Map*, a *Topic Similarity Heatmap*, and a *Topic Keywords Barchart*, enabled visual and interactive interpretation and analysis of hadith themes. The model demonstrated good capability in capturing the semantic structure of hadith texts and producing coherent clustering.

The results of this study provide a significant contribution to the development of Islamic information systems, where the model can be integrated into a theme-based hadith search system to facilitate the search and access of relevant hadith information. Topic modeling helps researchers identify dominant themes in the hadith corpus for more in-depth analysis and can be used to develop structured religious learning materials based on themes. This research also opens up opportunities for the application of advanced NLP technology in Islamic studies and religious text processing and can serve as the basis for developing a personalized and relevant hadith recommendation system or da'wah content.

However, this study has several limitations that should be noted. The high number of *outliers* indicates the model's limitations in handling short texts or limited contexts, while the use of an English-language dataset may not fully capture the semantic nuances of the original Arabic hadith. Topic interpretation still relies on manual evaluation by researchers, which can vary between individuals, and the embedding model used may not be optimal for the specific domain of religious texts. The study's results have also not been tested

on hadith datasets from other sources or with different characteristics, so the generalizability of the results remains limited.

Based on the identified findings and limitations, further research is recommended to develop specific embedding models for Islamic religious texts to improve the accuracy of semantic representation. Specific strategies for managing unclassified documents need to be developed, such as model *fine-tuning* or applying *zero-shot classification*. Evaluation with original Arabic hadith datasets or multilingual datasets is needed for cross-language validation, as is the development of a hierarchical approach to address thematic overlap and simplify topic structure. Developing a more objective evaluation system and implementing prototype web-based or mobile applications for Muslim communities and Islamic educational institutions are also priorities for future research.

REFERENCES

- [1] R. Aban, "Sunnah (Hadith) Dataset." 2022.
- [2] DM Blei, AY Ng, and MI Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *ArXiv Prepr.*, 2018.
- [4] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *ArXiv Prepr.*, 2022.
- [5] S. Al-Shargi, "LDA Topic Modeling of the Holy Quran Text." Nov. 2024. [Online]. Available: <https://medium.com/@alshargi.usa/lda-topic-modeling-of-the-holy-quran-text-b43d3d7348c9>
- [6] D. Kumar, "Topic Modeling with BERTopic: A Comprehensive Analysis of Modern NLP Applications," *Int. J. Mach. Learn. Comput.*, vol. 15, no. 1, pp. 12–20, Jan. 2025.
- [7] A. Hassan, M. Ibrahim, and S. Khalil, "Hadith Classification using Machine Learning Techniques According to its Reliability," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 12, pp. 256–264, 2019.
- [8] K. Al-Rashid, H. Mohammad, and F. Ahmed, "Text categorization in Quran and Hadith: Overcoming the interrelation challenges using machine learning and term weighting," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 32, no. 8, pp. 914–923, Oct. 2020, doi: 10.1016/j.jksuci.2019.04.003.
- [9] IK Alshammari, E. Atwell, and MA Alsalka, "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," *Int. J. Islam. Appl. Comput. Sci. Technol.*, vol. 11, no. 4, pp. 9–16, Dec. 2023.
- [10] N. Alqahtani, F. Alotaibi, and M. Alsaed, "Employing a Multilingual Transformer Model for Segmenting Unpunctuated Arabic Text," *Appl. Sci.*, vol. 12, no. 20, p. 10559, Oct. 2022, doi: 10.3390/app122010559.
- [11] L. Abdullah, K. Yaseen, and H. Naji, "AraBERT transformer model for Arabic comments and reviews analysis," *Neural Comput. Appl.*, vol. 34, no. 8, pp. 6271–6285, Mar. 2022, doi: 10.1007/s00521-021-06782-4.
- [12] S. Aftar, L. Gagliardelli, AE Ganadi, F. Ruozzi, and S. Bergamaschi, "RoBERT2VecTM: A Novel Approach for Topic Extraction in Islamic Studies," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, doi: 10.18653/v1/2024.findings-emnlp.534.
- [13] A. Abuzayed and H. Al-Khalifa, "BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique," *Procedia Comput. Sci.*, vol. 189, pp. 191–194, 2021, doi: 10.1016/j.procs.2021.05.096.
- [14] T. Zerrouki, A. Balla, and A. Lakhouaja, "Arabic natural language processing for Qur'anic research: a systematic review," vol. 56, no. 2, pp. 1175–1207, Feb. 2023, doi: 10.1007/s10462-022-10313-2.
- [15] K. Jamshaid, H. Farooq, and MT Siddique, "Topic Modeling of Quranic Verses using Latent Dirichlet Allocation with English Language: Topic Modeling using LDA," *VFAST Trans. Softw. Eng.*, vol. 12, no. 4, pp. 239–251, Dec. 2024, doi: 10.21015/vtse.v12i4.1946.

- [16] AMA Egov and T. Grok, "The Development of Topic Modeling for Hadith Studies," *J. Islam. Stud.* , vol. 2021, 2021.
- [17] MA Mosa, "An Exhaustive Literature Review of Hadith Text Mining," *ACM Trans. Asian Low-Resort. Lang. Inf. Process.* , vol. 22, no. 7, pp. 1–25, Jul. 2023, doi: 10.1145/3588315.
- [18] Herwinsyah and others, "Topic Modeling in the Qur'an Using the BERTopic Library on the BERT Language Model," vol. 14, no. 2, 2023.
- [19] F. Alotaibi, M. Alharbi, and S. Alnasser, "ArabianGPT: Native Arabic GPT-based Large Language Model," *ArXiv Prepr. ArXiv240215313* , Feb. 2024.
- [20] R. Baly, G. Hajj, and W. El-Hajj, "Benchmarking Transformer-based Language Models for Arabic Sentiment and Sarcasm Detection," in *g Workshop* , Association for Computational Linguistics, Apr. 2021, pp. 104–114.