Analysis and Evaluation of Qur'an Translation Topics Using Classical, Neural, and Transformer-Based Topic Modelling

Akhmad Rinaldy Kurnia[1], Sajarwo Anggai[2*], Murni Handayani[3]
Graduate Program of Informatics Engineering, Universitas Pamulang
Jalan Raya Puspiptek No. 46, Buaran, Kecamatan Serpong, Kota Tangerang Selatan, Banten 15310
Email: akhmad.rinaldykurnia@gmail.com[1], dosen02832@unpam.ac.id[2], dosen02710@unpam.ac.id[3]

Topic modelling is an important approach for extracting latent thematic structures from text corpora, including religious texts that are characterized by dense semantics and short documents. This study aims to compare the performance of several topic modelling methods Latent Dirichlet Allocation (LDA), Biterm Topic Model (BTM), Combined Topic Model (CombinedTM), and BERTopic in extracting topics from the Indonesian translation of the Qur'an. The dataset consists of 6,236 verses, with each verse treated as a single document. Topic quality is evaluated using two main metrics: coherence score (C_v) and topic diversity. The experimental results show that CombinedTM achieves the highest coherence score, with a maximum value of approximately 0.52 at K = 10 topics, followed by BTM, which demonstrates relatively high and stable coherence scores (around 0.50) across certain topic number variations. LDA yields the highest topic diversity, exceeding 0.90, but with lower coherence scores compared to the other models, indicating its limitations in preserving semantic coherence in short texts. Meanwhile, BERTopic exhibits consistently high topic diversity (0.85–0.88) across different numbers of topics, although its bag-of-words–based coherence scores do not always increase significantly. These findings highlight that the choice of topic modelling method should be aligned with the characteristics of the corpus and the objectives of thematic analysis, particularly in the context of short-form religious texts.

Keywords: LDA, BTM, CombinedTM, BERTopic, Topic Modelling, Coherence Score, Topic Diversity, Qur'an Translation

Corresponding Author:
Sajarwo Anggai
Graduate Program of Informatics Engineering, Universitas Pamulang
Jalan Raya Puspiptek No. 46, Buaran, Kecamatan Serpong, Kota Tangerang Selatan, Banten 15310
dosen02832@unpam.ac.id

1. Introduction

The Qur'an, as a religious text, possesses distinctive linguistic and semantic characteristics. Its verses vary significantly in length, ranging from very short statements to long and complex passages, each containing dense and multilayered meanings. A single verse may encompass multiple themes, while similar themes may appear across different chapters (surahs) using varied expressions. In its Indonesian translation, this complexity increases further due to semantic shifts that naturally occur when meaning is transferred from the source language to the target language. These characteristics distinguish the Qur'an translation from general text corpora and necessitate analytical approaches that are sensitive to variations in text length and semantic interconnections among verses.

In large-scale text analysis, topic modelling is widely used as an exploratory approach to identify latent thematic structures within document collections. This approach functions as a computational aid for mapping thematic tendencies and semantic relationships across a corpus without replacing theological interpretation. In the context of Qur'an translation, topic modelling can be utilized to explore thematic distributions across verses and chapters, while maintaining its role as a supportive analytical tool rather than an interpretative authority.

Considering that the Indonesian translation of the Qur'an consists predominantly of short verses with limited word counts and dense semantic content, methodological questions arise regarding the suitability of different topic modelling approaches for this type of corpus. Various topic modelling methods have been developed based on different assumptions and mechanisms, suggesting that not all models are expected to perform equally well when applied to religious texts characterized by semantic richness and short document length.

One of the most widely used classical topic modelling methods is Latent Dirichlet Allocation (LDA), which models topics based on word distribution within documents. However, prior studies indicate that LDA performance is highly dependent on dataset size, hyperparameter configuration, and preprocessing quality. As a result, LDA often exhibits limitations when applied to short and semantically dense texts, as it relies heavily on word co-occurrence at the document level[1].

In the context of Qur'an translation, where verses are generally short and semantically complex, the Biterm Topic Model (BTM) presents a relevant alternative. Unlike LDA, which models topic distributions at the document level, BTM directly models global word-pair co-occurrences across the entire corpus. This characteristic enables BTM to produce more stable and consistent topics in short-text corpora, such as Qur'anic translations[2].

Advancements in deep learning have further contributed to the development of topic modelling approaches that leverage contextual semantic representations. The Combined Topic Model (CombinedTM) integrates bag-of-words representations, which preserve topic interpretability, with contextual embeddings that capture word meaning within sentence-level context[3]. Meanwhile, BERTopic employs transformer-based embeddings to represent documents, followed by clustering and keyword extraction processes, allowing it to generate more semantically coherent topics, particularly for short and dynamic texts[4]

The quality of topics generated by topic modelling techniques is commonly evaluated using topic coherence, which measures the semantic relatedness among top words within a topic. Topic coherence is widely used because it has been shown to correlate reasonably well with human judgment of topic interpretability[5]. In the context of religious texts such as the Qur'an translation, topic coherence is particularly important, as each topic is expected to represent a meaningful and internally consistent concept rather than fragmented semantic elements.

However, topic coherence alone does not fully capture the relationships among topics generated by a model. A model may achieve high coherence scores while producing redundant or overlapping topics if similar keywords recur across multiple topics. To address this limitation, topic diversity is employed as a complementary evaluation metric that measures the degree of lexical uniqueness among topics. Topic diversity provides insight into whether a model is capable of capturing a broad and non-redundant thematic structure within the corpus [6].

Several previous studies have demonstrated the growing application of topic modelling techniques to Qur'anic corpora and their translations. Zafar et al [7], in their study Transformer-Based Topic Modelling for Urdu Translations of the Holy Quran, applied BERTopic to multiple Urdu translations of the Qur'an and compared its performance with LDA and Non-Negative Matrix Factorization (NMF). Their findings showed that BERTopic achieved superior coherence and topic diversity scores, indicating the effectiveness of transformer-based embeddings in capturing semantic variations across translations.

In the Indonesian context, Rolliawati et al [8] conducted a study titled Text Mining Approach for Topic Modelling of Corpus Al-Qur'an in Indonesian Translation, which applied frequency-based text mining and LDA to identify dominant themes within the Indonesian Qur'an translation. Their results revealed recurring

major themes such as faith, the afterlife, and moral conduct, as well as differences in topic distribution between Makkiyah and Madaniyah verses. Although informative, this study relied primarily on traditional approaches without incorporating contextual semantic representations.

Furthermore, Syopiansyah Jaya Putra et al[9] have contributed significantly to Qur'anic text mining research in Indonesia by emphasizing the importance of language-specific preprocessing, including Indonesian stemming and normalization, to ensure semantic accuracy in computational analysis. These studies highlight that preprocessing decisions play a critical role in determining the quality of downstream topic modelling results, particularly for religious texts.

Based on the above considerations, this study positions itself to conduct a systematic comparative evaluation of classical, neural, and transformer-based topic modelling approaches, namely LDA, BTM, CombinedTM, and BERTopic, applied to the Indonesian translation of the Qur'an. The evaluation employs both topic coherence and topic diversity metrics to provide a more comprehensive assessment of topic quality. Ultimately, this research aims to identify a topic modelling approach that produces robust, interpretable, and semantically sensitive thematic representations suitable for short-form religious texts.

2. Literature Review and Problem Statement

Literature Review

Topic modelling has been extensively studied as an unsupervised approach for discovering latent thematic structures in large text corpora. Classical probabilistic models such as Latent Dirichlet Allocation (LDA) have long been adopted as baseline methods due to their interpretability and solid theoretical foundations. LDA models documents as mixtures of topics and topics as probability distributions over words. Despite its widespread adoption, several studies have reported that LDA performance is highly sensitive to document length, corpus sparsity, and hyperparameter configuration, making it less effective when applied to short and semantically dense texts[10][1].

To address the limitations of LDA on short texts, Biterm Topic Model (BTM) was introduced as an alternative that models global word co-occurrence patterns rather than document-level distributions. By learning topics from word pairs (biterms) across the entire corpus, BTM has been shown to produce more stable and coherent topics for short-text datasets such as social media posts and microblogs[2][11]. Subsequent extensions, including SparseBTM and Relational BTM, further improved topic coherence by incorporating sparsity constraints and semantic similarity through word embeddings[12].

Recent advances in neural topic modelling have introduced approaches that combine probabilistic topic modelling with distributed semantic representations. Combined Topic Model (CombinedTM) integrates bag-of-words representations with contextualized embeddings, allowing the model to retain topic interpretability while capturing deeper semantic relationships among words[3]. Empirical studies have demonstrated that CombinedTM often achieves higher topic coherence than classical models, particularly in corpora characterized by short documents and semantic complexity[13]

Transformer-based approaches represent a further evolution in topic modelling research. BERTopic, which utilizes transformer embeddings such as BERT and Sentence-BERT, performs topic extraction through document embedding, clustering, and class-based TF-IDF. This framework has been reported to outperform classical models in various domains, especially in terms of semantic coherence and interpretability[14][15]. However, transformer-based models also introduce challenges related to computational cost and the suitability of traditional coherence metrics, which are often based on bag-of-words representations[16].

In the context of religious texts, particularly the Qur'an and its translations, topic modelling has gained increasing scholarly attention. Rolliawati et al [8] applied LDA to the Indonesian translation of the Qur'an and identified dominant themes such as faith, the afterlife, and moral conduct, as well as differences between Makkiyah and Madaniyah verses. Although informative, their study relied on classical frequency-based approaches without leveraging contextual semantic representations.

More recent studies have explored transformer-based topic modelling for Qur'anic texts. Zafar et al[7] demonstrated that BERTopic outperformed LDA and NMF when applied to Urdu translations of the Qur'an, achieving higher topic coherence and topic diversity scores. Similarly, Herwinsyah [17] applied BERTopic to a combined Arabic–Indonesian Qur'anic corpus and reported semantically meaningful topics related to theological, moral, and eschatological concepts.

Indonesian scholars have also emphasized the importance of language-specific preprocessing in Qur'anic text analysis. Syopiansyah Jaya Putra et al[9] highlighted that Indonesian stemming and normalization significantly affect semantic consistency and downstream NLP tasks applied to Qur'an translations. These findings underscore the necessity of aligning preprocessing strategies with both linguistic and religious contexts.

Despite the growing body of research, most existing studies focus on a single modelling paradigm or evaluate models using limited metrics, primarily topic coherence. Comparative studies that systematically evaluate classical, neural, and transformer-based topic modelling approaches on the Indonesian translation of the Qur'an, while simultaneously considering both topic coherence and topic diversity, remain scarce.

Problem Statement

Based on the literature review, several research gaps and methodological challenges can be identified. First, there is a lack of systematic comparative evaluation of topic modelling approaches applied to the Indonesian translation of the Qur'an. Although LDA, BTM, CombinedTM, and BERTopic have each been studied in different contexts, prior research rarely compares these models within a unified experimental framework using the same dataset and evaluation criteria. Second, the majority of existing studies rely on single-metric evaluation, particularly topic coherence, to assess topic quality. While coherence provides insight into intra-topic semantic consistency, it does not adequately capture inter-topic redundancy or thematic overlap. Consequently, models with high coherence may still generate repetitive or insufficiently diverse topics. The limited adoption of topic diversity as a complementary metric restricts a comprehensive understanding of model performance[6]. Third, there is no clear consensus regarding which topic modelling approach is most semantically appropriate for Qur'anic translation texts, which are characterized by short document length, high semantic density, and religious sensitivity. Classical probabilistic models, neural topic models, and transformer-based approaches differ substantially in their assumptions, representational mechanisms, and interpretability, raising questions about their relative suitability for this domain.

Therefore, the core problem addressed in this study is the absence of empirical evidence that clearly identifies which topic modelling paradigm, classical, neural, or transformer-based, produces topics that are coherent, diverse, and semantically aligned with the characteristics of the Indonesian translation of the Qur'an. Addressing this problem is essential to support future computational analyses of religious texts and to provide methodological guidance for scholars working at the intersection of natural language processing and Islamic studies.

3. Method

This study employed a quantitative, computational research design using an unsupervised topic modelling framework to analyze thematic structures in the Indonesian translation of the Qur'an. The methodological procedure was designed to ensure comparability across different topic modelling paradigms, including classical, neural, and transformer-based approaches, while maintaining sensitivity to the linguistic and semantic characteristics of religious texts.

Data Source and Corpus Construction

The dataset used in this study consisted of the official Indonesian translation of the Qur'an published by the Ministry of Religious Affairs of the Republic of Indonesia (KEMENAG RI). The corpus contained 6,236 verses, with each verse treated as an individual document. This document-level granularity was intentionally chosen to preserve the original structural and semantic boundaries of Qur'anic verses, which are often short but semantically dense. Treating each verse as a single document aligns with prior Qur'anic topic modelling studies that emphasize verse-level analysis to avoid semantic dilution caused by document aggregation[8] [18]. The resulting corpus represents a short-text dataset characterized by high sparsity and strong contextual dependency.

Text Preprocessing

Text preprocessing was conducted to normalize the corpus while preserving essential semantic content. The preprocessing pipeline included text lowercasing, removal of punctuation and numerical characters, tokenization, stopword removal, and stemming. Indonesian stopwords were filtered using a customized list to avoid the removal of religiously meaningful terms. Stemming was applied using an Indonesian stemming approach adapted for Qur'anic translation texts, following recommendations from Putra et al. (2020), who demonstrated that language-specific stemming improves semantic consistency in Indonesian Qur'anic corpora. Preprocessing decisions were kept consistent across all models to ensure a fair comparison and to minimize confounding effects on topic quality.

Topic Modelling Approaches

To provide a comprehensive comparative evaluation, four topic modelling approaches were implemented. The Latent Dirichlet Allocation (LDA) model was used as a classical probabilistic baseline. LDA models each document as a mixture of latent topics and each topic as a probability distribution over words[19]. The model was trained using a bag-of-words representation, with multiple topic numbers evaluated to identify the optimal configuration.

The Biterm Topic Model (BTM) was applied as a short-text–oriented alternative to LDA. Unlike LDA, BTM models global word-pair co-occurrence patterns (biterms) across the entire corpus rather than document-level distributions, making it more suitable for sparse and short-text datasets[2]. This characteristic is particularly relevant for verse-level Qur'anic data. The Combined Topic Model (CombinedTM) represents a neural topic modelling approach that integrates bag-of-words representations with contextualized embeddings. This model preserves topic interpretability while capturing semantic relationships beyond surface-level word co-occurrence[3]. CombinedTM was selected to evaluate the effectiveness of hybrid representations in religious short-text corpora. Finally, BERTopic was employed as a transformer-based topic modelling framework. BERTopic generates document embeddings using transformer models, clusters documents based on semantic similarity, and extracts topic representations using class-based TF-IDF. This approach has been reported to perform well on short and semantically complex texts, including religious corpora[7].

Experimental Design

Each topic modelling approach was evaluated across multiple topic numbers to examine performance stability and sensitivity. For probabilistic models (LDA, BTM, and CombinedTM), the number of topics was varied systematically to identify optimal configurations. For BERTopic, topic reduction mechanisms were applied to control topic granularity while preserving semantic coherence. All experiments were conducted using the same preprocessed corpus and evaluation protocol to ensure methodological consistency and comparability across models.

Evaluation Metrics

To assess topic quality comprehensively, two evaluation metrics were employed: topic coherence and topic diversity. Topic coherence was measured using the C_v coherence score, which evaluates the degree of semantic relatedness among the top words within each topic. The C_v metric has been shown to correlate well with human interpretability and is widely used in topic modelling research[10][5]. Topic diversity was used as a complementary metric to measure the proportion of unique words across all topics. This metric captures inter-topic distinctiveness and helps identify redundancy among generated topics[6]. The combination of coherence and diversity metrics allows for a more balanced evaluation, addressing both intra-topic consistency and inter-topic differentiation.

Data Analysis Procedure

The analysis focused on comparing topic modelling performance across the four approaches based on their coherence and diversity scores. In addition to quantitative evaluation, qualitative inspection of dominant topics was conducted to ensure that the extracted themes were semantically meaningful and aligned with the contextual characteristics of Qur'anic translation texts. Importantly, this study did not engage in theological interpretation or exegesis of Qur'anic verses. The extracted topics were treated as computational representations intended to support thematic exploration rather than doctrinal analysis.

4. Results and Discussion

Topic Modelling Results Using LDA : Determination of the Optimal Number of Topics

Topic modelling using LDA was conducted by varying the number of topics (K) to assess the quality of the resulting topic structures. In this process, each translated Qur'anic verse was treated as a single document, enabling the model to capture latent thematic patterns at the verse level. Model evaluation was performed using two main metrics: the C_v coherence score, which measures semantic consistency among words within a topic, and topic diversity, which reflects the degree of lexical variation across topics. The results of the LDA evaluation across different numbers of topics are illustrated in Figure 1.

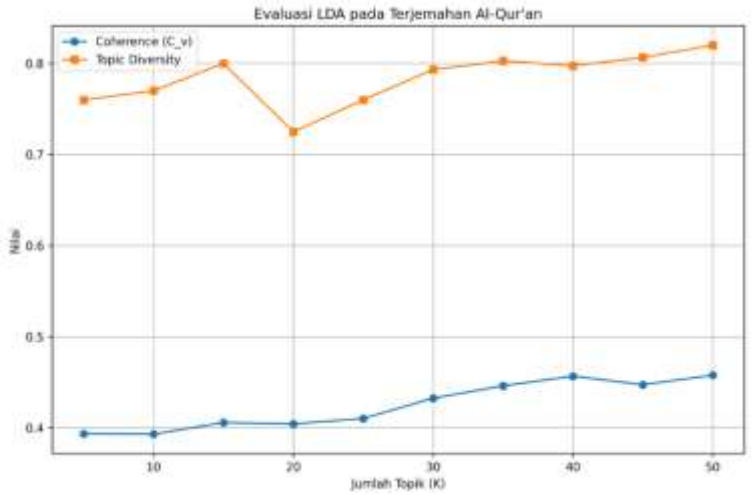


Figure 1. LDA Evaluation Graph: Coherence Score and Topic Diversity against Number of Topics

Based on Figure 1, the coherence score produced by the LDA model shows an increasing trend as the number of topics increases. The highest coherence value is achieved at K = 50, with a score of approximately 0.46, which is higher than all other topic configurations. This result indicates that at this topic number, the semantic relatedness among words within topics reaches its optimal level.

In addition, topic diversity also attains its highest value at K = 50, with a score of approximately 0.82. This finding suggests that the topics generated under this configuration exhibit a high level of lexical variation, with reduced overlap among topics. Considering both evaluation metrics, K = 50 was selected as the optimal number of topics for LDA, as it provides the highest quantitative topic quality.

Table 1. Initial Interpretation of the Top 10 Dominant Topics Generated by the LDA Model

No.	Topic ID	Dominant Keywords	Initial Interpretation
1	T07	faith, deeds, reward, sin, obedience	Faith and moral accountability
2	T12	hell, punishment, fire, torment, disbeliever	Divine punishment and retribution
3	T03	prophet, messenger, revelation, scripture, people	Prophethood and divine revelation
4	T21	Allah, Almighty, Lord, power, creation	The greatness and sovereignty of God
5	T09	world, hereafter, life, death	Worldly life and the hereafter
6	T18	command, prohibition, law, obedience	Islamic law and commandments
7	T27	people, community, destruction, denial	Narratives of past communities
8	T33	prayer, forgiveness, mercy, help	Supplication and divine mercy
9	T41	day, judgment, reckoning, recompense	The Day of Judgment
10	T05	heaven, earth, creation, signs	Creation of the universe

Using K = 50, the LDA model generated fifty distinct topics representing a wide range of themes and subthemes in the Indonesian translation of the Qur'an. A relatively large number of topics allows the model to achieve higher granularity, enabling subthemes that were previously grouped under fewer topics to be separated more explicitly. However, to maintain analytical focus and readability, not all topics are discussed in detail. Instead, the analysis concentrates on the ten most dominant topics, which are considered the most representative of the main thematic structure of the corpus. The complete list of fifty topics and their top ten keywords is provided separately in Appendix A.

The dominant topics were selected based on the highest-probability keywords within each topic and their relevance to central Qur'anic themes. This selection strategy allows the analysis to reflect the primary thematic structure while preserving the granularity achieved by using K = 50. Table 4.1 presents the ten dominant topics along with their representative keywords and initial interpretations.

Overall, the dominant LDA topics capture core Qur'anic themes such as faith and moral accountability, divine punishment, prophethood and revelation, the greatness of God, the contrast between worldly life and the hereafter, Islamic law and commandments, historical narratives of past communities, supplication and mercy, the Day of Judgment, and the creation of the universe. These themes align closely with well-established thematic categories in Qur'anic studies. Taken together, the LDA results with K = 50 demonstrate a clear separation among topics and a high degree of lexical diversity. The relatively high coherence and diversity scores indicate that the model effectively maximizes quantitative topic quality. The resulting topic granularity supports the identification of specific subthemes, while the analysis of dominant topics ensures clarity and coherence in discussion. Detailed modelling results are provided in Appendix A.

Topic Modelling Results Using BTM : Determination of the Optimal Number of Topics

Topic modelling using the Biterm Topic Model (BTM) was conducted to address the characteristics of the Qur'anic translation corpus, which consists of very short documents. Unlike LDA, which relies on document-level word distributions, BTM models global word-pair co-occurrences (biterms), making it more suitable for short-text data. The BTM model was evaluated by varying the number of topics (K) and assessed using two evaluation metrics: C_v coherence score and topic diversity. The evaluation results across different topic numbers are illustrated in Figure 2.

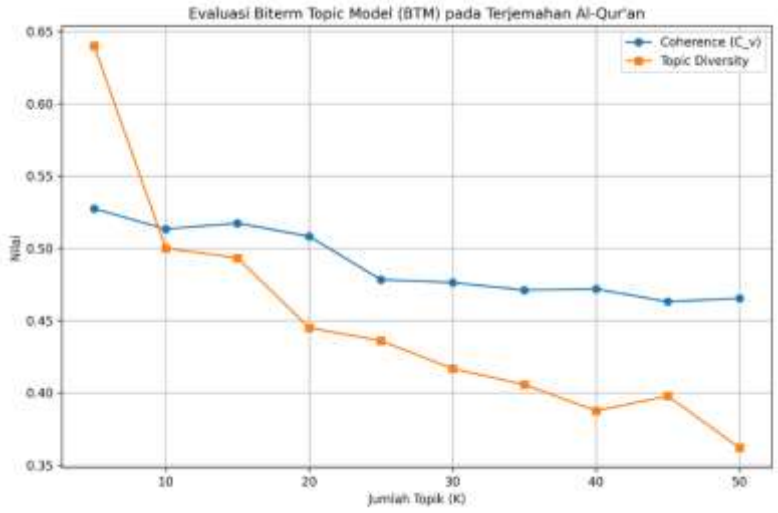


Figure 2. BTM Evaluation Graph: Coherence Score and Topic Diversity against Number of Topics

Based on Figure 2, the highest coherence score for the Biterm Topic Model (BTM) is achieved at K = 5, with a value of approximately 0.53. As the number of topics increases, the coherence score shows a decreasing trend. This pattern indicates that increasing the number of topics in BTM does not necessarily improve semantic coherence and may instead reduce the semantic consistency among words within topics. Similarly, topic diversity reaches its highest value at K = 5, with a score of approximately 0.64, before declining substantially at larger topic numbers. This finding suggests that with a smaller number of topics, BTM is able to generate topics that are both semantically coherent and relatively diverse in terms of vocabulary usage. Based on the combined evaluation of coherence and diversity metrics, K = 5 was selected as the optimal number of topics for the BTM model. This configuration provides the highest quantitative

topic quality and is well aligned with the characteristics of the Indonesian Qur'an translation corpus, which consists of short and semantically dense documents.

Using K = 5, the BTM model produced five main topics that represent broad thematic categories within the Qur'an translation. Compared to LDA, which yields more granular topic structures, BTM generates topics with wider thematic coverage while maintaining strong semantic coherence. This approach allows BTM to capture the global thematic structure of the corpus, albeit at a lower level of granularity. Given the relatively small number of topics, all extracted topics were included in the analysis. The dominant keywords for each topic serve as the basis for thematic interpretation and are presented in Table 2.

Table 2. Initial Interpretation of the Five Main Topics Generated by the BTM Model (K = 5)

No.	Topic ID	Dominant Keywords	Initial Interpretation
1	T0	Allah, earth, sky, water, Almighty	Divine power and the creation of nature
2	T1	believers, punishment, hell, disbelievers	Faith and consequences of disbelief
3	T2	Allah, Almighty, guidance, explanation, truth	Divine guidance and truth
4	T3	Allah, prophet, Muhammad, revelation	Prophethood and the transmission of revelation
5	T4	women, men, children, family, people	Social and family relations

Topic modelling using BTM with K = 5 produces relatively general but semantically cohesive topics. This result demonstrates that BTM is effective in capturing major thematic structures in short-text corpora such as the Indonesian translation of the Qur'an. Compared to LDA, which offers higher topic granularity, BTM provides a more concise and global thematic representation, making it particularly suitable for exploratory analysis of short and meaning-dense religious texts.

Topic Modelling Results Using CombinedTM: Determination of the Optimal Number of Topics

Topic modelling using Combined Topic Model (CombinedTM) was conducted to evaluate the performance of a hybrid approach that integrates probabilistic word-distribution modelling with contextual semantic representations based on embeddings. This model combines the interpretability advantages of classical topic models, such as LDA, with the ability of contextual embeddings to capture semantic meaning within context. As a result, CombinedTM is expected to generate topics that are both interpretable and semantically coherent when applied to the Indonesian translation of the Qur'an.

In this study, CombinedTM was employed to integrate bag-of-words–based distributional representations with contextual embedding representations. This hybrid approach is particularly relevant for short-text corpora, such as Qur'anic translations, where limited word counts and dense meanings often challenge purely probabilistic models. The performance of the CombinedTM model was evaluated by varying the number of topics (K) and assessed using two evaluation metrics: C_v coherence score and topic diversity. These metrics were used to examine both intra-topic semantic consistency and inter-topic distinctiveness. The evaluation results for different topic numbers are presented in Figure 3.

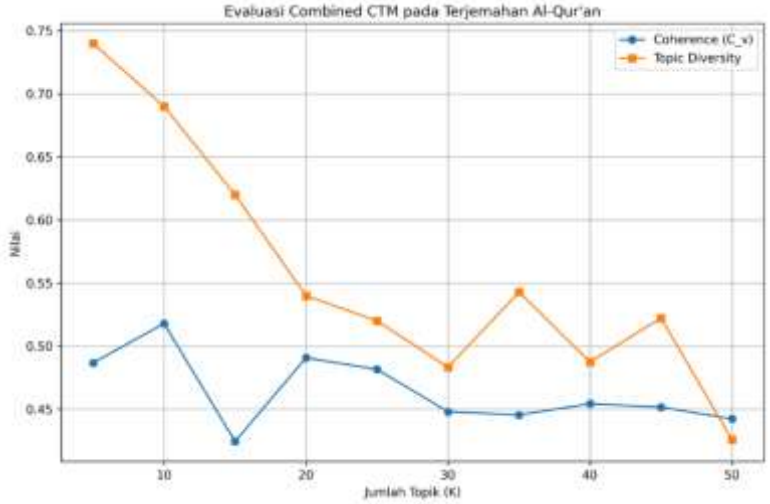


Figure 3. CombinedTM Evaluation Graph: Coherence Score and Topic Diversity against Number of Topics

Based on Figure 3, the coherence score of the CombinedTM model exhibits fluctuations across different numbers of topics, with the highest value achieved at K = 10, reaching approximately 0.52. Beyond this point, the coherence score tends to decline and does not show substantial improvement as the number of topics increases. This pattern suggests that increasing topic granularity beyond a certain level does not enhance semantic consistency within topics. From the perspective of topic diversity, the highest value is observed at K = 5 (approximately 0.74), followed by a noticeable decrease as the number of topics increases. However, at K = 10, the topic diversity score remains relatively high at 0.69, indicating that the balance between semantic coherence and topic distinctiveness is still well maintained.

Considering both evaluation metrics, K = 10 was selected as the optimal number of topics for the CombinedTM model. This configuration provides the highest semantic coherence while maintaining an adequate level of topic diversity, making it the most suitable choice for analyzing the Indonesian translation of the Qur'an. Using K = 10, the CombinedTM model generated ten main topics that represent the thematic structure of the Qur'an translation. Compared to LDA and BTM, the topics produced by CombinedTM demonstrate stronger semantic cohesion, as the model leverages contextual embeddings to capture deeper meaning relationships. The moderate number of topics allows each topic to remain interpretable without sacrificing semantic depth. Given this balance, all extracted topics were included in the analysis. The dominant keywords and initial interpretations are presented in Table 3.

Table 3. Initial Interpretation of the Ten Dominant Topics Generated by the CombinedTM Model (K = 10)

No.	Topic ID	Dominant Keywords	Initial Interpretation
1	T0	sleep, compete, eleven, one-tenth, agent	Symbolic actions and illustrative events
2	T1	prophet, Muhammad, Allah, messenger, to you	Prophethood and divine message
3	T2	punishment, hell, disbelief, hypocrite, believer	Reward, punishment, and faith attitudes
4	T3	ask, obligate, command, instruction	Commands and religious obligations
5	T4	Almighty, sky, earth, signs, your Lord	Divine power and signs of greatness
6	T5	people, believers, men, women, community	Human community and identity
7	T6	creation, loss, deeds, consequences	Creation and consequences of actions
8	T7	plan, event, decree, change	Divine decree and dynamic events
9	T8	naming, determination, designation	Naming and divine determination
10	T9	Muhammad, prophet, Qur'an, your Lord	The Qur'an and transmission of revelation

The first topic consists of narrative and symbolic terms, indicating the use of illustrative events as a means of conveying messages in the Qur'an. The second topic emphasizes prophethood and divine mission,

highlighting the role of Prophet Muhammad as the recipient and messenger of revelation. The third topic reflects moral and eschatological consequences, focusing on belief, disbelief, and accountability. The subsequent topics capture essential Qur'anic themes, including religious commands, divine power manifested through creation, social identity and community structure, the relationship between human actions and their consequences, divine decree governing events, the process of naming and determination, and the central role of the Qur'an as the ultimate source of guidance.

The CombinedTM results demonstrate strong performance in generating semantically coherent topics from short-text corpora. The highest coherence score achieved at K = 10 confirms that integrating word-distribution representations with contextual embeddings enhances topic quality. Compared to LDA and BTM, CombinedTM offers a more balanced representation, combining semantic depth with clear thematic separation, making it particularly effective for analyzing the Indonesian translation of the Qur'an.

Topic Modelling Results Using BERTopic: Evaluation of Topic Reduction in BERTopic

Topic modelling using BERTopic was conducted to evaluate the performance of a transformer-based embedding approach in extracting topics from the Indonesian translation of the Qur'an. BERTopic relies on contextual semantic representations at the sentence or document level and generates topics through density-based clustering. This approach differs fundamentally from LDA and BTM, which are based on word-distribution mechanisms, and is therefore expected to capture more contextually meaningful thematic structures. In this study, BERTopic was applied as a transformer-based topic modelling framework that integrates contextual embeddings with density-based clustering techniques. Unlike previous models, BERTopic does not require the number of topics to be specified in advance. Instead, it automatically generates an initial set of topics and provides a topic reduction mechanism to merge semantically similar topics, allowing better control over topic granularity.

The evaluation of BERTopic was carried out by comparing the C_v coherence scores between the initial topic configuration and the models produced after applying topic reduction with different numbers of reduced topics. This evaluation aims to assess how topic reduction influences semantic coherence and overall topic quality. The results of this evaluation are presented in Figure 4.

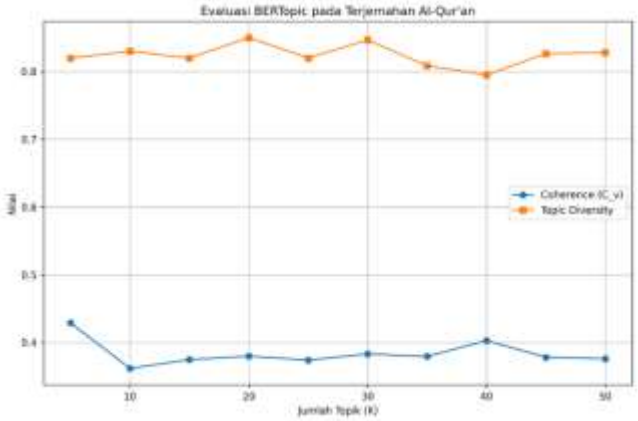


Figure 4. BERTopic Evaluation Graph: Coherence Score and Topic Diversity against Number of Topics

Based on Figure 4, the initial BERTopic model yields a coherence score of approximately 0.38. After applying the topic reduction process, the coherence score increases substantially to around 0.50 and remains relatively stable across different numbers of reduced topics. This pattern indicates that topic reduction improves semantic consistency within topics without introducing instability. In terms of topic diversity, BERTopic demonstrates consistently high values, ranging from approximately 0.79 to 0.85. The high and stable topic diversity scores indicate minimal lexical overlap among topics and a strong ability to

represent a wide range of themes within the Indonesian translation of the Qur'an. The absence of a significant decline in topic diversity as the number of topics increases further suggests that BERTopic effectively preserves topic distinctiveness.

The relationship between coherence score and topic diversity in BERTopic highlights an important characteristic of transformer-based topic modelling. Although the coherence score may not always surpass that of probabilistic models, BERTopic excels in producing topics that are semantically well separated and diverse. This outcome is largely attributed to the use of contextual embeddings, where topic quality is better reflected by semantic separation rather than by word co-occurrence frequency alone. Consequently, bag-of-words–based coherence metrics may not fully capture the semantic richness of topics generated by BERTopic. The observed pattern indicates that the topic reduction process in BERTopic does not degrade topic quality; instead, it enhances semantic coherence while maintaining high topic diversity. The stability of coherence scores across different reduced topic configurations further confirms BERTopic's robustness in producing semantically consistent topics even after topic merging. After applying topic reduction, BERTopic produces five main topics, as presented in Table 4. Each topic is represented by five dominant keywords that serve as the basis for thematic interpretation.

Table 4. Initial Interpretation of the Five Main Topics Generated by the BERTopic Model

No.	Topic ID	Dominant Keywords	Initial Interpretation
1	T-1	people, Allah, truth, Almighty, you	General background topic
2	T0	people, Allah, Almighty, truth, prophet	Prophethood and divine call
3	T1	mountain, hell, people, water, denial	Punishment and narratives of past communities
4	T2	fear God, obey, worship, Allah, to Me	Commands of piety and worship
5	T3	nature, universe, Lord, peace, praise	Divine greatness and harmony of creation

The first topic (T-1) functions as a background topic, consisting of general and frequently occurring terms such as people, Allah, and truth. This topic does not represent a specific thematic focus and primarily captures verses that are not strongly associated with other distinct topics. Therefore, it is not emphasized in the thematic analysis. The second topic (T0) relates to prophethood and divine proclamation, characterized by keywords such as prophet, Allah, and truth. This topic represents verses that emphasize the role of prophets as conveyors of divine guidance and truth to humanity. The third topic (T1) reflects punishment and narratives of past communities, indicated by terms such as hell, people, and denial. This theme highlights the consequences of rejecting divine teachings and conveys moral lessons through historical accounts of earlier communities. The fourth topic (T2) focuses on commands of piety and worship, marked by keywords such as fear God, obey, and worship. This topic captures normative calls for religious obedience and devotion as central elements of Qur'anic teaching. The fifth topic (T3) emphasizes divine greatness and the harmony of creation, represented by terms such as nature, universe, and praise. This theme underscores the order of the natural world as a sign of God's majesty and a source of spiritual reflection.

The results indicate that BERTopic is the most stable and effective model in generating diverse topics for the Qur'anic translation corpus. Its ability to maintain high topic diversity and stable semantic quality makes it particularly well suited for thematic analysis of religious texts that are short, meaning-dense, and context-rich. Consequently, BERTopic is positioned as the leading approach in this study and serves as the primary benchmark for comparison with other topic modelling methods in subsequent discussion sections.

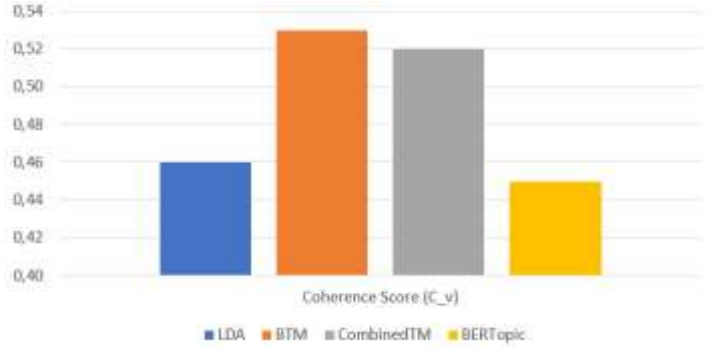
Comparative Performance Analysis Across Topic Modelling Methods

This subsection presents a comparative analysis of the performance of the four topic modelling approaches employed in this study, namely Latent Dirichlet Allocation (LDA), Biterm Topic Model (BTM), Combined Topic Model (CombinedTM), and BERTopic. The comparison is based on the evaluation results discussed in the preceding subsections and focuses on two primary metrics: coherence score and topic diversity. In addition, the analysis considers the qualitative characteristics of the topics generated by each model to provide a more comprehensive assessment of model performance.

1. Comparison Based on Coherence Score

This subsection compares the performance of the four topic modelling methods based on their C_v coherence scores. For each model, the comparison uses the best coherence score achieved under its optimal topic configuration, as identified in the previous analyses.

The coherence score reflects the degree of semantic relatedness among the dominant words within a topic and is commonly used as an indicator of topic interpretability. By comparing the highest coherence scores obtained by each model, this analysis aims to evaluate which topic modelling approach produces the most semantically cohesive topics when applied to the Indonesian translation of the Qur'an.



Fibure 5. Model Comparison Based on Coherence Score

Based on the experimental results summarized in Figure 5, the comparison of coherence scores was conducted using the best values achieved by each model under their respective optimal topic configurations. The evaluation results indicate that BTM achieves the highest coherence score, at approximately 0.53, followed closely by CombinedTM with a score of around 0.52. These values are higher than those obtained by LDA, which reaches a coherence score of approximately 0.46, and BERTopic, which records a score of around 0.45.

These findings suggest that BTM and CombinedTM are more effective in preserving semantic relatedness among words within topics when each model operates under its optimal conditions. The superior performance of BTM can be attributed to its global word-pair (biterm) modelling mechanism, which enables the model to capture stronger word associations in short-text corpora. This characteristic is reflected in BTM's highest coherence score among all evaluated models.

Meanwhile, CombinedTM achieves a relatively high coherence score due to its hybrid design, which integrates probabilistic word-distribution modelling with contextual semantic representations. However, although CombinedTM attains a coherence score of approximately 0.52, the contribution of contextual embeddings is not fully reflected in the bag-of-words–based coherence metric (C_v) used in this evaluation. In contrast, LDA produces lower coherence scores, even though its performance improves as the number of topics increases. The best coherence score achieved by LDA remains around 0.46, indicating that increasing topic numbers does not substantially enhance semantic cohesion for short and meaning-dense texts such as the Indonesian translation of the Qur'an.

BERTopic demonstrates relatively stable coherence scores across different topic reduction configurations, with values ranging from approximately 0.36 to 0.45. This stability suggests that the semantic quality of BERTopic topics is not highly sensitive to the number of reduced topics. However, the lower coherence scores compared to BTM and CombinedTM further indicate that bag-of-words–based coherence metrics may not fully capture the semantic quality of topics generated by embedding-based models.

2. Comparison Based on Topic Diversity

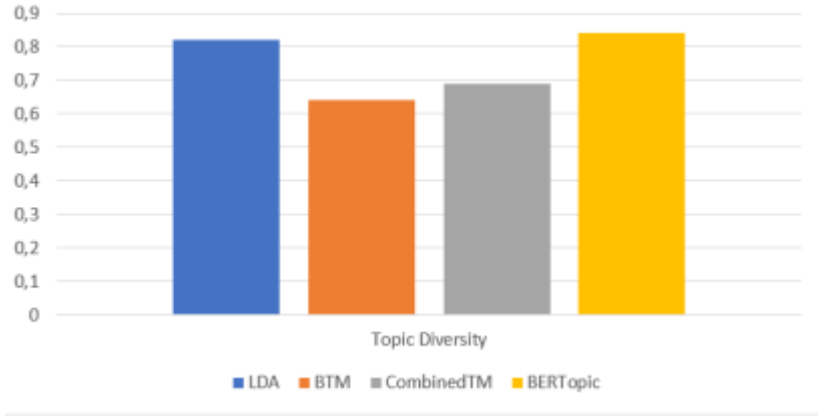


Figure 6. Comparison of Optimal Topic Diversity Across Models

Based on Figure 6, BERTopic exhibits the highest topic diversity among all evaluated models. This result indicates that BERTopic is capable of generating more diverse and semantically well-separated topics, with minimal lexical overlap. This advantage is closely related to the use of contextual embeddings and density-based clustering, which enable more flexible and context-aware separation of thematic structures.

Both LDA and CombinedTM demonstrate relatively high topic diversity scores, although their performance remains below that of BERTopic. In the case of LDA, the relatively high topic diversity reflects the model's ability to separate vocabularies across topics; however, this separation is not always accompanied by strong semantic cohesion. CombinedTM, on the other hand, occupies an intermediate position, reflecting its hybrid nature in balancing probabilistic word-distribution modelling with contextual semantic representations. In contrast, BTM produces the lowest topic diversity score among the evaluated models. This finding suggests that topics generated by BTM tend to be more concentrated and exhibit higher lexical overlap. Such behavior is a consequence of BTM's global word-pair modelling mechanism, which prioritizes focused thematic representation but limits topic diversity as the number of topics increases.

3. Trade-off Analysis and Model Suitability

The comparative results reveal a clear trade-off between coherence score and topic diversity across the four topic modelling approaches examined in this study. Probabilistic models such as LDA and BTM tend to perform well in specific aspects while exhibiting limitations in others. For instance, BTM is able to generate highly coherent topics when the number of topics is small; however, it becomes less effective in maintaining topic diversity as the number of topics increases.

CombinedTM occupies a transitional position among the evaluated models. This model improves topic stability and semantic coherence compared to LDA by integrating probabilistic modelling with contextual representations. Nevertheless, CombinedTM still shows constraints in preserving topic diversity when topic granularity is increased, indicating that the hybrid approach does not fully resolve the coherence–diversity trade-off.

In contrast, BERTopic demonstrates the most balanced performance in the context of this study. Its primary strengths lie in the stability of results and high topic diversity, allowing it to generate well-separated and contextually meaningful topics. Although BERTopic does not consistently achieve the highest bag-of-

words–based coherence scores, this limitation reflects the inadequacy of traditional coherence metrics in fully capturing the semantic quality of embedding-based topic models rather than a deficiency in the model itsel

Summary of Model Comparison

From a comparative perspective, the thematic relationships generated by LDA, BTM, CombinedTM, and BERTopic demonstrate a strong level of thematic consistency, despite being represented through different structural forms and levels of granularity. For example, the theme of faith and moral consequences appears as separate topics in LDA, specifically T07 (faith and deeds) and T12 (punishment and hell). In contrast, BTM tends to merge these aspects into a single topic (T1), linking faith directly with punishment and disbelief. A similar pattern is observed in CombinedTM through T2, which connects believers, disbelief, and recompense. In BERTopic, this relationship is modeled more compactly through T1, representing punishment, hell, and eschatological consequences as a unified thematic structure. These patterns indicate that the causal relationship between faith and recompense remains consistent across all models, although it is expressed with varying degrees of topic separation.

A comparable thematic consistency is also evident in the theme of prophethood and revelation. In LDA, this theme is represented by T03, while in BTM it appears in T3, both of which combine prophetic figures with the process of revelation. CombinedTM distributes this theme across two topics—T1 (prophethood and divine mission) and T9 (the Qur'an and the transmission of revelation)—indicating a clearer distinction between the messenger and the revealed text. In contrast, BERTopic represents prophethood and revelation more globally through T0, integrating prophetic authority, divine calling, and message delivery into a single contextual topic. This representation reflects BERTopic's tendency to consolidate closely related subthemes into broader, context-driven topics.

The theme of divine greatness and the creation of the universe also forms a coherent thematic cluster across all models. In LDA, this relationship is captured through T21 (divine greatness) and T05 (creation of nature). In BTM, both aspects are combined within T0, emphasizing God's power over the heavens and the earth. CombinedTM represents this relationship through T4 (divine greatness) and T6 (creation and consequences of human actions), highlighting the connection between divine authority and its implications for human life. In BERTopic, this thematic relationship is expressed through T3, which emphasizes divine attributes and greatness, and T4, which reflects the harmony and order of creation. These consistent patterns confirm that the creation of the universe is uniformly positioned as a manifestation of divine greatness across all models, albeit through different representational strategies.

Overall, the comparative analysis indicates that each topic modelling approach exhibits distinct characteristics and strengths in modelling the Indonesian translation of the Qur'an. LDA proves effective as a baseline method for providing an initial overview of thematic structure but performs less optimally on short and semantically dense texts. BTM excels in maintaining topic coherence when the number of topics is small, yet its topic diversity decreases as topic granularity increases, leading to greater lexical overlap. CombinedTM occupies an intermediate position, offering more stable performance than purely probabilistic models, particularly at moderate topic numbers, while still facing limitations in preserving topic diversity at higher topic configurations.

Among all evaluated models, BERTopic demonstrates the most consistent performance in generating semantically stable and diverse topics. Its ability to preserve thematic diversity and contextual coherence makes it particularly well suited to the characteristics of the Qur'anic translation corpus, which is short, context-rich, and thematically diverse. These findings suggest that no single model is universally superior

across all evaluation criteria; however, contextual embedding–based models exhibit higher overall suitability for topic modelling in the corpus examined in this study.

5. Conclusion

This study conducted topic modelling on the complete Indonesian translation of the Qur'an, comprising 6,236 verses, each treated as a short-text document. The primary objective was to evaluate and compare the performance of four topic modelling approaches—LDA, BTM, CombinedTM, and BERTopic—in extracting thematic structures from a corpus characterized by short length, high semantic density, and rich contextual meaning. The quantitative evaluation using coherence score (C_v) and topic diversity demonstrates that each model exhibits distinct strengths and limitations. LDA achieved its optimal performance at K = 50, producing a relatively high topic diversity (0.82), which indicates strong lexical separation across topics. However, its coherence score remained moderate (approximately 0.46), suggesting limitations in capturing deep semantic cohesion within short texts. BTM performed optimally at a small number of topics (K = 5), achieving the highest coherence score (approximately 0.53). This result confirms BTM's effectiveness in modelling short texts through global word-pair (biterm) representations, although its topic diversity decreased as the number of topics increased.

CombinedTM demonstrated balanced performance, achieving a coherence score of approximately 0.52 and a topic diversity of 0.69 at K = 10. The integration of probabilistic word distributions and contextual embeddings improved semantic cohesion, although the benefits of embeddings were not fully captured by bag-of-words–based coherence metrics. In contrast, BERTopic exhibited the most stable and diverse topic structures. Following topic reduction, its coherence score increased to approximately 0.50, while achieving the highest topic diversity (up to 0.85). These findings highlight BERTopic's strength in generating semantically distinct and contextually rich topics, even when traditional coherence metrics underestimate its quality. The results indicate that no single model is universally superior. Instead, model selection should be aligned with corpus characteristics and analytical objectives, with embedding-based approaches showing greater suitability for short, context-rich religious texts such as Qur'anic translations.

6. References

[1] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Front. Sociol.*, vol. 7, p. 886498, 2022.

[2] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 1445–1456.

[3] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a hot topic: Contextualized document embeddings improve topic coherence," in *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 2: short papers)*, 2021, pp. 759–766.

[4] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv Prepr. arXiv2203.05794*, 2022.

[5] S. Rahimi, "Saturation in qualitative research: An evolutionary concept analysis," *Int. J. Nurs. Stud. Adv.*, vol. 6, p. 100174, 2024.

[6] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 439–453, 2020.

[7] A. Zafar, M. Wasim, S. Zulfiqar, T. Waheed, and A. Siddique, "Transformer-based topic modeling for Urdu translations of the Holy Quran," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 23, no. 10, pp. 1–21, 2024.

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

[9] S. J. Putra, T. Mantoro, and M. N. Gunawan, "Text mining for Indonesian translation of the Quran: A systematic review," in *2017 International Conference on Computing, Engineering, and Design (ICCED)*, IEEE, 2017, pp. 1–5.

[10] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 2012, pp. 952–961.

[11] Q. Zhu, Z. Feng, and X. Li, "GraphBTM: Graph enhanced autoencoded variational inference for biterm topic model," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 4663–4672.

[12] X. Zhou, J. Ouyang, and X. Li, "Two time-efficient gibbs sampling inference algorithms for biterm topic model," *Appl. Intell.*, vol. 48, no. 3, pp. 730–754, 2018.

[13] M. H. Asnawi, A. A. Pravitasari, T. Herawan, and T. Hendrawati, "The combination of contextualized topic model and MPNet for user feedback topic modeling," *IEEE Access*, vol. 11, pp. 130272–130286, 2023.

[14] M. de Groot, M. Aliannejadi, and M. R. Haas, "Experiments on generalizability of BERTopic on multi-domain short text," *arXiv Prepr. arXiv2212.08459*, 2022.

[15] O. Babalola, B. Ojokoh, and O. Boyinbode, "Comprehensive evaluation of LDA, NMF, and BERTopic's performance on news headline topic modeling," *J. Comput. Theor. Appl.*, vol. 2, no. 2, pp. 268–289, 2024.

[16] C. Meaney, X. Wang, J. Guan, and T. A. Stukel, "Comparison of methods for tuning machine learning model hyper-parameters: with application to predicting high-need high-cost health care users," *BMC Med. Res. Methodol.*, vol. 25, no. 1, p. 134, 2025.

[17] H. Herwinsyah, "Pemodelan Topik dalam Al-Qur'an Menggunakan Library Bertopic pada Model Bahasa Bert," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 14, no. 2, pp. 319–328, 2023.

[18] K. Jamshaid, H. Farooq, and M. T. 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, 2024.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.