


Fine-Tuning Topic Modelling: A Coherence-Focused Analysis of Correlated Topic Models

Syahrial Syahrial¹ and Razief Perucha Fauzie Afidh^{2,*}

¹ MANN Research Center, Banda Aceh, Aceh, Indonesia; arial.van@gmail.com (S.S.)

² Department of Informatics, Faculty of Mathematic and Natural Science, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; razief@usk.ac.id (R.P.F.A.)

* Correspondence: razief@usk.ac.id

Article History	Abstract
<p>Received 10 September 2024 Revised 13 November 2024 Accepted 22 November 2024 Available Online 30 November 2024</p> <p>Keywords: Topic modelling Correlated Topic Model Topic coherence Optimization Learning rate</p>	<p>The Correlated Topic Model (CTM) is a widely used approach for topic modelling that accounts for correlations among topics. This study investigates the effects of hyperparameter tuning on the model's ability to extract meaningful themes from a corpus of unstructured text. Key hyperparameters examined include learning rates (0.1, 0.01, 0.001), the number of topics (3, 5, 7, 10), and the number of top words (10, 20, 30, 40, 50, 80, 100). The Adam optimizer was used for model training, and performance was evaluated using the coherence score (c_v), a metric that assesses the interpretability and coherence of the generated topics. The dataset comprised 100 articles, and results were visualized using line plots and heatmaps to highlight performance trends. The highest coherence score of 0.803 was achieved with three topics and 10 top words. The findings demonstrate that fine-tuning hyperparameters significantly improves the model's ability to generate coherent and interpretable topics, resulting in more accurate and insightful outcomes. This research underscores the importance of parameter optimization in enhancing the effectiveness of CTM for topic modelling applications.</p>
	<p>Copyright: © 2024 by the authors. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License. (https://creativecommons.org/licenses/by-nc/4.0/)</p>

1. Introduction

Topic modeling is a powerful technique in natural language processing that enables the automatic identification of hidden thematic structures within large collections of text documents [1]. It involves using statistical methods to uncover groups of words that frequently occur together, thereby representing underlying topics in the data [2]. By organizing unstructured textual data into interpretable topics, topic modelling facilitates a deeper understanding of complex datasets such as news articles, social media posts, and research papers [3]. This approach has become an essential tool for exploring, summarizing, and analyzing text data across various domains, enabling researchers

and practitioners to uncover patterns, trends, and relationships that might otherwise remain hidden.

Among the various methods used for topic modeling, the Correlated Topic Model (CTM) has gained prominence due to its ability to capture relationships between topics [4]. Unlike conventional methods such as Latent Dirichlet Allocation (LDA), CTM employs a logistic normal distribution to model these relationships, providing a more nuanced and interconnected view of the hidden themes within textual data. This enhanced capability makes CTM particularly effective for analyzing large document collections, such as those found in news archives and social media, offering a more comprehensive understanding of the data's thematic structure [5].

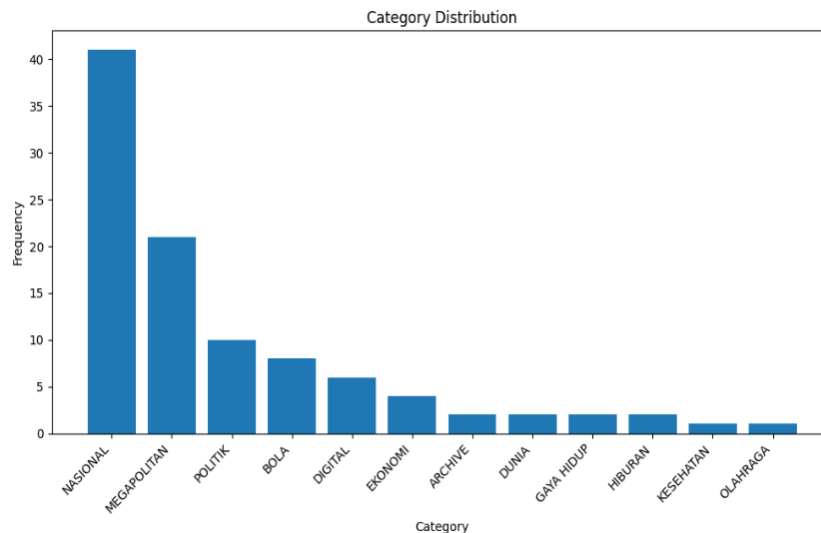


Figure 1. Document category distribution.

Parameter tuning is a critical aspect of optimizing the performance of topic models, including CTM. Key parameters such as the learning rate, the number of topics, and the number of top words significantly influence the model's ability to generate coherent and interpretable topics. Previous studies have highlighted the importance of hyper-parameter selection in topic modeling, noting that parameter variations can lead to substantial differences in model output and coherence scores [6]. For instance, Ford et al. emphasized the use of heuristic approaches based on examining the rate of perplexity change (RPC) to select the optimal number of topics in their analysis [7]. Furthermore, the influence of regularization techniques on model performance has been explored, underscoring the need for careful parameter tuning to enhance the quality of topic representations [6].

The evaluation of coherence scores serves as a vital metric for assessing the effectiveness of different parameter configurations in CTM. Coherence measures provide insights into the interpretability and semantic similarity of the topics generated by the model. By systematically tuning parameters and evaluating coherence scores, researchers can gain valuable insights into the relationships between topics and their relevance to the underlying data. This iterative process of tuning and evaluation not only improves model performance but also contributes to a deeper understanding of the data being analyzed [8].

Despite the widespread use of CTM, challenges remain in achieving optimal performance and interpretability. One significant issue is the sensitivity of CTM to parameter settings, which can greatly influence the quality of the generated topics. Inadequate parameter tuning may lead

to incoherent or irrelevant topics, limiting the model's utility in capturing meaningful patterns within the data. Moreover, the complexity of CTM, stemming from its ability to model topic correlations, adds to the difficulty of finding the best parameter configurations. This problem is further compounded by the lack of standardized approaches for systematically evaluating and fine-tuning model parameters, such as the number of topics or the learning rate. These challenges highlight the need for focused research on parameter optimization strategies and their impact on model coherence and interpretability.

This study aims to explore the application of the CTM in topic modelling, with a focus on the impact of various parameter tuning strategies on coherence scores. It aims to synthesize insights from existing literature and conduct empirical evaluations to provide a comprehensive understanding of how parameter adjustments can enhance the effectiveness of CTM in uncovering meaningful patterns within textual data.

2. Materials and Methods

2.1 Dataset

The dataset used in this study comprises 100 samples sourced from *BeritaSatu.com*, containing articles with various attributes such as title, author, and content. For the purposes of this research, only the article content was utilized, as it serves as the primary textual data required for topic modelling. The dataset is originally categorized into distinct classes, reflecting its initial thematic segmentation. The distribution of these categories is visually depicted in Figure 1, offering a clear overview of the dataset's composition.

Table 1. Model train scenario.

Learning Rate	Number of Topics	Number of Top-Words
0.1	3, 5, 7, 10	10, 20, 30, 40, 50, 80, 100
0.01	3, 5, 7, 10	10, 20, 30, 40, 50, 80, 100
0.001	3, 5, 7, 10	10, 20, 30, 40, 50, 80, 100

While the original categories are noted for descriptive purposes, they do not influence the topic modelling process or its outcomes. This selection of a limited dataset was made to enable a focused examination of parameter optimization while ensuring computational efficiency. Notably, each document in the dataset consists of multiple paragraphs, providing rich and diverse textual data for the model to analyze. This design supports the scalability of preliminary findings to larger datasets in subsequent stages of research.

2.2 Data Pre-processing

The data pre-processing stage involved several steps to prepare the text for topic modelling. Common Indonesian stop words, such as "yang," "dan," and "di," were removed to eliminate frequently used but non-informative words. Additional custom stop words were incorporated to exclude specific terms irrelevant to the analysis. Stemming was applied to reduce words to their root forms (e.g., "mempelajari" becomes "ajar"), helping to group similar words and reduce the overall vocabulary size. The text was tokenized into individual words, converted to lowercase for consistency, and non-alphanumeric characters were removed to retain only meaningful content. Documents with fewer than six words were excluded to ensure the inclusion of content-rich texts, enhancing the quality and relevance of the topic modelling process.

2.3 Correlated Topic Model (CTM)

We used CTM for the topic model. Unlike simpler models such as Latent Dirichlet Allocation (LDA), which assumes that topics are independent, CTM employs a logistic normal distribution to model topic proportions, allowing for the discovery of complex interdependencies among topics. This capability allows for a more detailed analysis of thematic patterns, making CTM especially suitable for datasets with interconnected or overlapping topics.

The Bag-of-Words (BoW) representation was generated using the CountVectorizer, which encodes each document as a vector based on the frequency of word occurrences. To convert the raw word counts into probabilities, L1 normalization was applied, ensuring that the sum of the word counts for each document equals one, thereby transforming the word counts into a probability distribution. The model employs the Adam (Adaptive Moment Estimation) optimizer for parameter

updates, leveraging its adaptive learning rate capabilities to enhance convergence [9–11]. The Evidence Lower Bound (ELBO) is utilized as the loss function, guiding the optimization process by maximizing the likelihood of the observed data while minimizing model complexity [12, 13]. The number of epochs is set to 200, indicating that the training process will iterate over the entire dataset 200 times. This iterative training approach enables the model to repeatedly refine its understanding of the data, improving its ability to capture underlying patterns and relationships.

2.4 Hyperparameter Tuning

Hyperparameter tuning is a crucial process for optimizing the performance of the CTM [14]. In this study, we systematically explored key hyperparameters to determine their impact on model effectiveness. The learning rate, which controls the step size during optimization, was adjusted across multiple values to balance convergence speed and stability. The number of topics was varied to examine how different granularities influence the model's ability to uncover meaningful patterns. Additionally, the number of top words per topic was fine-tuned to identify the optimal level of detail for interpreting topic content. Detail scenarios for hyperparameters shown in Table 1.

2.5 Performance Evaluation

The model's performance was evaluated using coherence scores, specifically the c_v metric, a widely recognized measure for assessing the quality of topics generated in topic modelling. Coherence scores evaluate the semantic similarity and interpretability of the words within each topic, providing a quantitative basis for determining the optimal number of topics [15–17]. The c_v score is computed by combining statistical co-occurrence measures with semantic similarity, effectively balancing data-driven and interpretability aspects. The calculation of the coherence score is represented as shown in Equation 1:

$$c_v = \frac{1}{|T|} \sum_{t \in T} \frac{1}{|t|^2} \sum_{w_i, w_j \in t, i \neq j} NPMI(w_i, w_j) \quad (1)$$

Where T denotes the set of topics, t represents an individual topic composed of a set of words, and w_i and w_j are words within a topic. The $NPMI$ (Normalized Pointwise Mutual Information) metric measures the

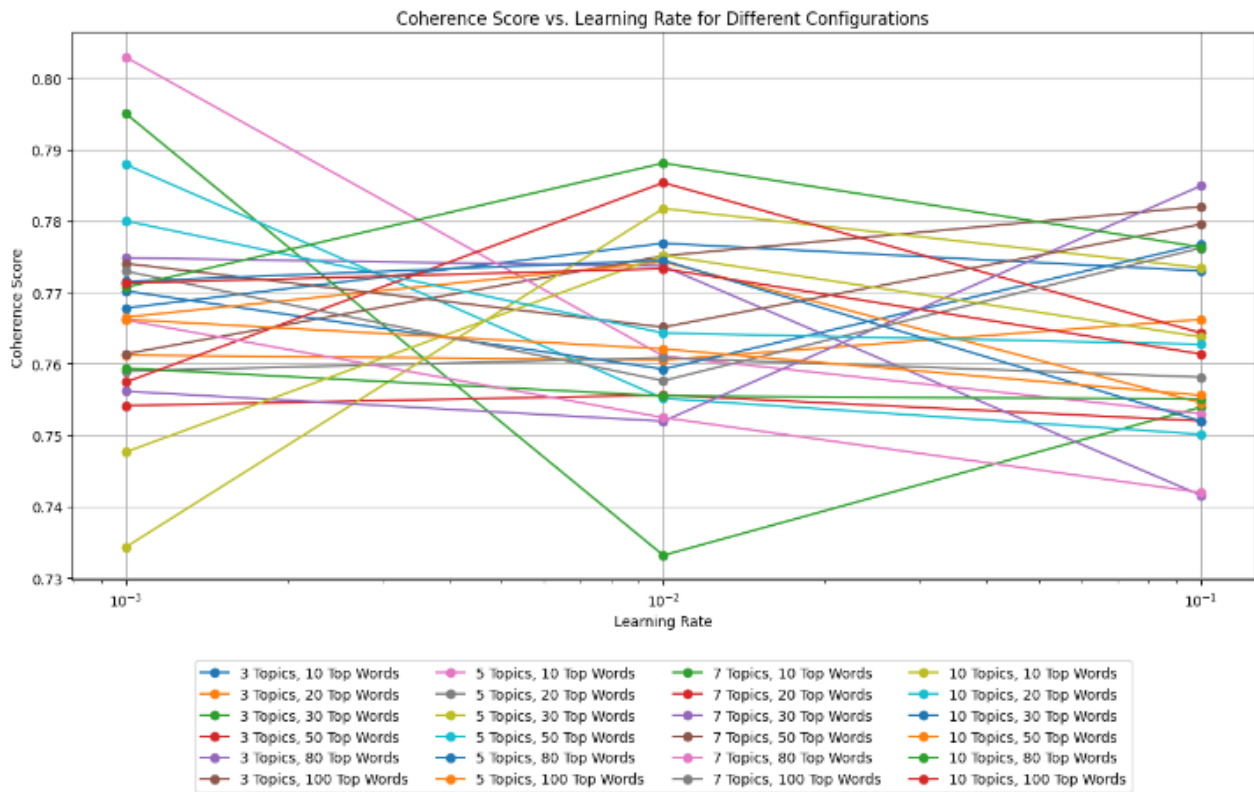


Figure 2. Comparison of coherence score vs different learning rate.

semantic association between word pairs. This approach aims to maximize c_v , ensuring that the topics generated are both coherent and interpretable while maintaining computational efficiency. By optimizing c_v , the model strikes a balance between accuracy and interpretability, providing a robust foundation for extracting meaningful insights from the textual data [18].

3. Results and Discussion

Based on the scenarios outlined in Table 1, the study's findings are illustrated in Figure 2, which presents the coherence scores for different combinations of learning rates, topic numbers, and top words. For a learning rate of 0.1, coherence scores ranged between 0.73 and 0.80. The lowest coherence score of 0.73 was observed with five topics and 30 top words, while the highest score of 0.80 was achieved with five topics and 10 top words. Similarly, for a learning rate of 0.01, coherence scores ranged from 0.73 to 0.79. The lowest coherence score of 0.73 was observed with three topics and 30 top words, whereas the highest score of 0.79 occurred with 10 topics and 80 top words. When the learning rate was set to 0.05, coherence scores ranged between 0.74 and 0.78. The lowest score of 0.74 was observed with three topics and 80 top words, while the highest score of 0.78 was achieved with seven topics and 30 top words.

As shown in Figure 3, the combination of five topics and 10 top words produced the highest coherence score of

0.803. In contrast, the combination of three topics and 50 top words yielded the lowest coherence score. These results suggest that the number of topics and the number of top words significantly influence the coherence and interpretability of the generated topics.

The findings highlight the importance of parameter tuning in optimizing the performance of the CTM. The results demonstrate that both the number of topics and the number of top words play a critical role in determining the coherence of the topics generated. Specifically, the combination of five topics and 10 top words provided the most coherent and interpretable results, indicating an optimal balance between granularity and interpretability. This finding aligns with previous research, which underscores the importance of hyperparameter selection in improving model quality and coherence scores.

The implications of this study are significant for applications of topic modelling in various domains, such as social media analysis, news classification, and customer feedback interpretation. By identifying the optimal parameter settings, researchers and practitioners can generate more meaningful and actionable insights from textual data. Moreover, the ability of CTM to capture topic correlations adds value to datasets with overlapping or interconnected themes, making it a valuable tool for analyzing complex datasets.

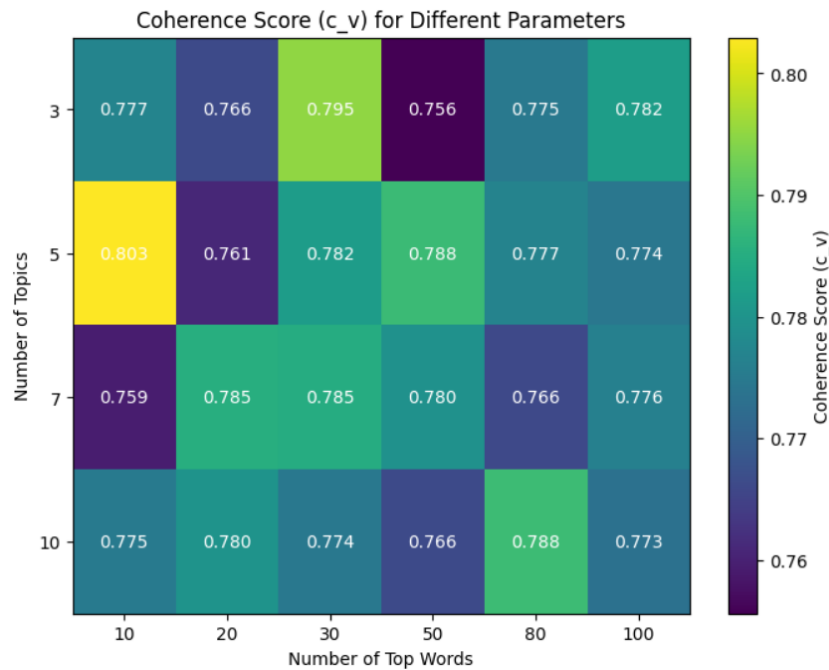


Figure 3. Heatmap of Coherence Score (c_v) for different parameters.

However, this study also has several limitations. First, the dataset used consisted of only 100 samples, which, while sufficient for preliminary analysis, may limit the generalizability of the findings. Future studies should validate the results on larger and more diverse datasets to ensure broader applicability. Second, the scope of the study was restricted to three learning rates and a fixed number of iterations (200 epochs). Exploring a wider range of learning rates and alternative optimization strategies could provide deeper insights into the model's performance. Lastly, while coherence scores are a widely accepted metric for topic evaluation, incorporating qualitative assessments of topic interpretability could further enhance the robustness of the results.

4. Conclusions

The results highlight the importance of parameter tuning in improving the performance of topic models. Key parameters like the learning rate, the number of topics, and the number of top words significantly affect the model's ability to generate meaningful and coherent topics. In this study, the best performance was achieved with five topics and 10 top words, resulting in the highest coherence score of 0.803. Other factors, such as the optimizer used, the number of training epochs, and the dataset size, also influenced the results, showing that a balanced approach to parameter selection is essential. Future research should explore how these factors interact and test on larger datasets to further enhance model performance and reliability.

Author Contributions: Conceptualization, S.S. and R.P.F.A.; methodology, S.S. and R.P.F.A.; software, S.S. and R.P.F.A.; validation, S.S. and R.P.F.A. and Z.Z.; formal analysis, S.S. and R.P.F.A.; investigation, S.S. and R.P.F.A.; resources, S.S. and R.P.F.A.; data curation, R.P.F.A.; writing—original draft preparation, S.S.; writing—review and editing, R.P.F.A.; visualization, S.S.; supervision, R.P.F.A.; project administration, R.P.F.A.; funding acquisition, R.P.F.A. All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting this study's findings are available upon request from the corresponding author.

Acknowledgments: The authors would like to thank their institution for supporting this study.

Conflicts of Interest: All the authors declare no conflicts of interest.

References

1. Kherwa, P., and Bansal, P. (2018). Topic Modeling: A Comprehensive Review, *ICST Transactions on Scalable Information Systems*, 159623. doi:10.4108/eai.13-7-2018.159623.
2. Vayansky, I., and Kumar, S. A. P. (2020). A Review of Topic Modeling Methods, *Information Systems*, Vol. 94, 101582. doi:10.1016/j.is.2020.101582.
3. Qiang, J., Qian, Z., Li, Y., Yuan, Y., and Wu, X. (2022). Short Text Topic Modeling Techniques, Applications, and Performance: A Survey, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 34, No. 3, 1427–1445. doi:10.1109/TKDE.2020.2992485.

4. Xun, G., Li, Y., Zhao, W. X., Gao, J., and Zhang, A. (2017). A Correlated Topic Model Using Word Embeddings, *IJCAI* (Vol. 17), 4207–4213.
5. Mol, M. J., Belfi, B., and Bakk, Z. (2024). Unravelling the Skills of Data Scientists: A Text Mining Analysis of Dutch University Master Programs in Data Science and Artificial Intelligence, *PLoS ONE*, Vol. 19, No. 2 February, 1–14. doi:[10.1371/journal.pone.0299327](https://doi.org/10.1371/journal.pone.0299327).
6. Koltcov, S., Ignatenko, V., Boukhers, Z., and Staab, S. (2020). Analyzing the Influence of Hyper-Parameters and Regularizers of Topic Modeling in Terms of Renyi Entropy, *Entropy*, Vol. 22, No. 4. doi:[10.3390/E22040394](https://doi.org/10.3390/E22040394).
7. Ford, J. D., Elhai, J. D., Marengo, D., Almquist, Z., Olff, M., Spiro, E. S., and Armour, C. (2022). Temporal Trends in Health Worker Social Media Communication during the COVID – 19 Pandemic, No. August, 1–16. doi:[10.1002/nur.22266](https://doi.org/10.1002/nur.22266).
8. Koltcov, S., Ignatenko, V., Terpilovskii, M., and Rosso, P. (2021). Analysis and Tuning of Hierarchical Topic Models Based on Renyi Entropy Approach, *PeerJ Computer Science*, Vol. 7, 1–35. doi:[10.7717/PEERJ-CS.608](https://doi.org/10.7717/PEERJ-CS.608).
9. Shao, Y., Wang, J., Sun, H., Yu, H., Xing, L., Zhao, Q., and Zhang, L. (2024). An Improved BGE-Adam Optimization Algorithm Based on Entropy Weighting and Adaptive Gradient Strategy, *Symmetry*, Vol. 16, No. 5, 1–16. doi:[10.3390/sym16050623](https://doi.org/10.3390/sym16050623).
10. Sun, H., Yu, H., Shao, Y., Wang, J., Xing, L., Zhang, L., and Zhao, Q. (2024). An Improved Adam's Algorithm for Stomach Image Classification, *Algorithms*, Vol. 17, No. 7, 1–13. doi:[10.3390/a17070272](https://doi.org/10.3390/a17070272).
11. Shao, Y., Yang, J., Zhou, W., Sun, H., Xing, L., Zhao, Q., and Zhang, L. (2024). An Improvement of Adam Based on a Cyclic Exponential Decay Learning Rate and Gradient Norm Constraints.
12. Wang, A., Liu, W., and Liu, Z. (2022). A Two-Sample Robust Bayesian Mendelian Randomization Method Accounting for Linkage Disequilibrium and Idiosyncratic Pleiotropy with Applications to the COVID-19 Outcomes, *Genetic Epidemiology*, Vol. 46, Nos. 3–4, 159–169. doi:[10.1002/gepi.22445](https://doi.org/10.1002/gepi.22445).
13. Chérif-Abdellatif, B. E. (2018). Consistency of ELBO Maximization for Model Selection, *Proceedings of Machine Learning Research*, Vol. 96, No. 1974, 11–31.
14. Wijanto, M. C., Widiastuti, I., and Yong, H.-S. (2024). Topic Modeling for Scientific Articles: Exploring Optimal Hyperparameter Tuning in BERT., *International Journal on Advanced Science, Engineering & Information Technology*, Vol. 14, No. 3.
15. Szigeti, Á., Frank, R., and Kiss, T. (2024). Contribution to the Harm Assessment of Darknet Markets: Topic Modelling Drug Reviews on Dark0de Reborn, *Crime Science*, Vol. 13, No. 1, 1–10. doi:[10.1186/s40163-024-00211-z](https://doi.org/10.1186/s40163-024-00211-z).
16. Nguyen, H., and Hovy, D. (2019). Hey Siri. OK Google. Alexa: A Topic Modeling of User Reviews for Smart Speakers, *W-NUT@EMNLP 2019 - 5th Workshop on Noisy User-Generated Text, Proceedings*, 76–83. doi:[10.18653/v1/d19-5510](https://doi.org/10.18653/v1/d19-5510).
17. Chandra, R., and Ranjan, M. (2022). Artificial Intelligence for Topic Modelling in Hindu Philosophy: Mapping Themes between the Upanishads and the Bhagavad Gita, *PLoS ONE*, Vol. 17. doi:[10.1371/journal.pone.0273476](https://doi.org/10.1371/journal.pone.0273476).
18. Tijare, P., and Jhansi Rani, P. (2020). Exploring Popular Topic Models, *Journal of Physics: Conference Series*, Vol. 1706, No. 1, 012171. doi:[10.1088/1742-6596/1706/1/012171](https://doi.org/10.1088/1742-6596/1706/1/012171).