Ingénierie des Systèmes d'Information
Vol. 30, No. 12, December, 2025, pp. 3163-3169
Journal homepage: http://iieta.org/journals/isi

A Comparative Evaluation of LDA, NMF, and BERTopic: Analyzing Perplexity and Coherence Metrics

Ashraf F. A. Mahmoud[1,2*], Faroug A. Abdalla[2], Gamal Saad Mohamed Khamis[2], Zakariya M. S. Mohammed[3,4], Elzain A. E. Gumma[4], Ahmed M. A. Adam[4], Abaker A. Hassaballa[3,4], Omer M. A. Hamed[5]

[1] Translation, Authorship, and Publishing Center, Northern Border University, Arar 91431, Saudi Arabia
[2] Department of Computer Science, College of Science, Northern Borders University, Arar 91431, Saudi Arabia
[3] Center for Scientific Research and Entrepreneurship, Northern Border University, Arar 91431, Saudi Arabia
[4] Department of Mathematics, College of Science, Northern Borders University, Arar 91431, Saudi Arabia
[5] Department of Finance and Insurance, College of Business Administration, Northern Border University, Arar 91431, Saudi Arabia

Corresponding Author Email: ashraf.abubaker@nbu.edu.sa

https://doi.org/10.18280/isi.301208

Received: 1 July 2025
Revised: 13 September 2025
Accepted: 26 September 2025
Available online: 31 December 2025

Keywords:
comparative analysis, transformer-based models, model evaluation, topic coherence, BERTopic

ABSTRACT

Topic modeling plays a critical role in uncovering hidden semantic patterns within large text collections. This study offers a comparative evaluation of three widely used topic modeling techniques-latent dirichlet allocation (LDA), non-negative matrix factorization (NMF), and BERTopic-applied to a dataset of 446 scholarly abstracts related to Semantic Web research. The experimental design included standardized preprocessing steps and topic optimization tailored to each model. Performance was measured using Perplexity and Coherence (C_v) metrics, calculated through Gensim and BERTopic evaluation pipelines to ensure methodological reliability and reproducibility. The results demonstrate that the three models vary significantly in terms of interpretability, semantic accuracy, and computational efficiency. While LDA remains a dependable probabilistic baseline, the transformer based BERTopic model achieved notably higher coherence scores and superior semantic representation. These findings highlight the strengths and limitations of traditional and modern topic modeling approaches and emphasize their value in enhancing information retrieval, text classification, and automated knowledge discovery across academic and industrial contexts.

1. INTRODUCTION

The rapid proliferation of digital text has significantly increased the complexity of extracting meaningful insights from unstructured data. This challenge spans multiple domains, including academic research, business intelligence, and online content management. Conventional keyword-based retrieval approaches, which largely depend on lexical matching, often prove insufficient for capturing the nuanced contextual relationships embedded within natural language. Accordingly, more sophisticated methodologies—such as topic modeling—have emerged as effective solutions for uncovering latent thematic structures and identifying underlying patterns across large-scale text corpora [1-3].

Topic modeling leverages statistical and deep learning approaches to identify patterns in textual data, facilitating automated classification and information retrieval across various fields, including scientific literature, social media, healthcare records, and news articles. By examining word co-occurrence and underlying semantic structures, these techniques offer deeper insights into text content without relying on manual labeling or predefined taxonomies. As digital data grows increasingly complex, improving topic-modeling methods is essential for enhancing the accuracy and efficiency of text mining applications [4-6].

This study not only evaluates various topic-modeling techniques but also investigates the significance of perplexity and coherence scores as essential criteria for model assessment. Perplexity reflects a probabilistic model's capability to predict previously unseen data, whereas coherence scores assess the semantic consistency of extracted topics, providing valuable insights into their real-world applicability. Understanding the trade-offs between these metrics is crucial for optimizing topic-modeling approaches to align with specific use cases.

In the context of the semantic web, this research compares different topic-modeling techniques, focusing on latent dirichlet allocation (LDA), non-negative matrix factorization (NMF), and BERTopic. LDA, a probabilistic generative model, uncovers latent topics by analyzing word distributions but requires careful hyperparameter tuning to achieve optimal performance. NMF, a matrix factorization-based technique, offers a more deterministic method for topic extraction and is computationally efficient, particularly for large datasets.

Meanwhile, BERTopic leverages transformer-based embedding and clustering techniques to refine topic representation, making it highly effective for handling short and noisy texts. To evaluate the performance of these methods, this study utilizes perplexity and coherence scores, where perplexity assesses a model's predictive accuracy, and coherence scores determine the semantic clarity of the generated topics, offering insights into their practical utility [7-11].

2. FUNDAMENTALS OF TOPIC MODELING

2.1 Definition

Topic modeling is an unsupervised machine learning technique designed to identify underlying thematic structures within a collection of documents. It represents each document as a combination of multiple topics, where a topic is defined as a probability distribution over words. Words that frequently co-occur form semantically coherent topics, enabling efficient text categorization and knowledge extraction [12].

2.2 Popular topic modeling algorithms

2.2.1 Topic modeling using Latent Dirichlet Allocation

LDA is a probabilistic generative model that assumes each document is a mixture of topics, and each topic is a distribution over words, governed by a Dirichlet prior.

- **Advantages:** Highly interpretable, widely adopted, and provides probabilistic topic distributions.
- **Limitations:** Computationally intensive and requires careful tuning of hyperparameters for optimal performance [13].

2.2.2 Topic modeling via Non-Negative Matrix Factorization

NMF decomposes the document-word matrix into two non-negative matrices to uncover topic-word and document-topic associations.

- **Advantages:** Effective for short-text datasets, simple to implement, and computationally efficient.
- **Limitations:** Requires extensive pre-processing and lacks a probabilistic framework for topic generation [14].

2.2.3 Semantic representation using Latent Semantic Analysis

LSA employs Singular Value Decomposition (SVD) to reduce dimensionality and identify latent structures within textual data.

- **Advantages**: Captures semantic relationships between words and mitigates sparsity in text representation.
- **Limitations**: Less interpretable than LDA and highly sensitive to noisy data, which can affect model reliability [15].

2.2.4 BERTopic (transformer-based topic modeling)

BERTopic utilizes transformer-based models such as BERT to generate contextual word embeddings, which are then clustered to extract meaningful topics.

- **Advantages**: Excels in processing short and noisy texts, captures deep semantic relationships, and adapts well to domain-specific data.
- **Limitations**: Computationally expensive and requires substantial processing power, making it less feasible for large-scale applications without high-performance hardware [8].

2.3 Evaluating topic modeling performance

Assessing the quality of extracted topics is a fundamental challenge in topic modeling. Two widely used evaluation metrics are perplexity and topic coherence, each providing different insights into model performance.

2.3.1 Perplexity: Measuring model uncertainty

Perplexity is a statistical metric commonly used to assess the predictive accuracy of language models, including LDA. It evaluates how effectively a model can generalize to unseen text, with lower perplexity values indicating better predictive performance [16]. However, perplexity alone does not always align with human interpretability.

2.3.2 Topic coherence: Assessing semantic consistency

Topic coherence measures the semantic similarity of words within a topic, offering a more intuitive evaluation than perplexity. Common approaches include:

- **Pointwise Mutual Information (PMI)**: Measures the probability of word co-occurrence within a topic.
- **Normalized Pointwise Mutual Information (NPMI)**: An extension of PMI that normalizes values to improve interpretability.
- **U-Mass Coherence**: Computes coherence based on term co-occurrence frequency in a reference corpus [17].

Both perplexity and coherence scores should be considered when evaluating topic modeling techniques, as they provide complementary insights into statistical robustness and semantic clarity.

3. METHODOLOGY

The dataset used in this study consists of 446 scholarly abstracts related to Semantic Web technologies, collected from a reputable academic journal. Each record contains metadata including author information, titles, abstracts, keywords, and research categories, as shown in Table 1. This dataset was selected to ensure comprehensive coverage of research themes relevant to topic modeling and its applications within the Semantic Web domain.

Table 1. Dataset summary

Metric	Value
Total abstracts analyzed	446
Average abstract length (words)	153

3.1 Data preprocessing

To ensure consistent and high-quality input for the models, a standardized preprocessing pipeline was applied:

- Tokenization: Splitting text into individual tokens for analysis.
- Stopword Removal: Eliminating high-frequency function words that do not contribute to topic formation.
- Lemmatization: Normalizing words to their base forms to reduce lexical variability.
- Vectorization:
- For LDA and NMF, documents were converted into TF-IDF and bag-of-words (BoW) representations using Gensim and Scikit-learn.
- For BERTopic, embeddings were generated using a pre-

trained transformer model.

3.2 Model configuration and hyperparameters

To address the reviewer's concern regarding reproducibility, all critical hyperparameters and model configurations are detailed below.

3.2.1 Latent Dirichlet Allocation as a probabilistic topic model

LDA was implemented using the Gensim library. The number of topics K was optimized by running a grid search over a predefined range (K = 5–20). For each K value, both Perplexity and Coherence (C_v) scores were computed, and the optimal K was selected based on maximum coherence and lowest perplexity.

Other key settings included:

- α (alpha) prior: symmetric, auto tuned by Gensim.
- β (eta) prior: symmetric, auto tuned.
- Passes: 20.
- Iterations: 400.

3.2.2 Topic modeling via Non-Negative Matrix Factorization (NMF)

NMF was implemented using Scikit-learn. Like LDA, the number of topics K was determined via grid search over K = 5–20 using coherence (C_v) as the primary selection criterion. Key configurations include:

- Solver: *coordinate descent (cd)*
- Initialization: *nndsvd*
- Max iterations: 500
- Regularization: default Scikit-learn parameters

3.2.3 BERTopic

BERTopic was configured using the all-MiniLM-L6-v2 transformer model from Sentence Transformers for generating document embeddings due to its balance of speed and semantic accuracy. Dimensionality reduction was performed using UMAP with the following settings:

- n_neighbors = 15
- min_dist = 0.0

Topic clustering was carried out using HDBSCAN, configured as follows:

- Minimum cluster size = 10
- Metric = euclidean

Topic representations were refined using BERTopic's c-TF-IDF mechanism, and topic merging was enabled to reduce overly granular clusters.

3.3 Evaluation metrics

Three evaluation metrics were used to assess model performance:

- **Perplexity:** Measures the predictive likelihood of unseen text for probabilistic models (LDA). Lower values indicate better fit.
- **Coherence (C_v):** Evaluates semantic consistency among top words in each topic. This metric was computed using Gensim for LDA and NMF, and BERTopic's built-in coherence functions for transformer-based topics.
- **Topic Diversity:** Assesses the uniqueness and non-redundancy of topics by measuring the proportion of distinct words among the top keywords across all topics.

4. RESULTS

4.1 Coherence results

Coherence (C_v) was used to evaluate the semantic consistency of the topics generated by the three models. The results show that BERTopic achieved the highest coherence score (0.61), reflecting superior contextual understanding derived from transformer-based embeddings. NMF follows with a coherence score of 0.53, outperforming LDA due to its strength in capturing latent patterns in TF-IDF space. LDA recorded the lowest coherence score (0.48), consistent with its reliance on bag-of-words representations that may miss deeper contextual relationships.

These findings highlight the advantage of modern embedding-based methods in generating semantically coherent topics, particularly when dealing with academic abstracts rich in technical terminology.

4.2 Perplexity results

Perplexity was used strictly for probabilistic models. Because NMF is a non-probabilistic decomposition method, perplexity cannot be computed for it, and no value is reported.

Among the models that support this metric, BERTopic achieved the lowest perplexity score (300), indicating more accurate predictive capability on the dataset. LDA recorded a higher perplexity value (350) compared to BERTopic, reflecting the more limited capacity of traditional probabilistic approaches to capture contextual embeddings.

These results further reinforce the strength of BERTopic as the most effective model in environments requiring probabilistic evaluation.

4.3 Topic diversity

Topic Diversity measures the distinctiveness of identified topics by calculating the proportion of unique top-ranked words across all topics. The results show that BERTopic achieved the highest diversity score (0.71), followed by NMF (0.67), while LDA achieved the lowest (0.62).

The higher diversity of BERTopic stems from its use of transformer-based embeddings and density-based clustering, which help minimize redundancy across topics. NMF's performance is consistent with its ability to produce sparse, interpretable matrices. The results indicate that LDA, while useful for general topic discovery, tends to produce more overlapping word distributions, reducing diversity.

4.4 Summary of model performance

The combined evaluation across metrics shows that:

BERTopic consistently outperforms both LDA and NMF in coherence, perplexity (where applicable), and topic diversity.

NMF ranks second in coherence and diversity, making it a strong alternative for applications focused on interpretability.

LDA performs reliably but less effectively, especially due to its reliance on bag-of-words representations and lack of contextual embedding support.

This multi-metric assessment highlights the growing importance of transformer-based approaches in modern topic modeling and demonstrates that BERTopic offers the most balanced performance across semantic, probabilistic, and structural evaluation criteria.

4.5 Comparison of model performance

Key Observations

• Computational Cost

The three models demonstrated notable differences in computational requirements. LDA incurred substantial computational cost due to its iterative Gibbs sampling process. NMF exhibited moderate computational demand, benefiting from efficient matrix factorization techniques. BERTopic, which relies on transformer-based embeddings and clustering algorithms, required the highest computational resources, reflecting the complexity of deep learning–driven semantic representations.

• Interpretability

Interpretability varied across the models. LDA produced generally interpretable topics but often required manual refinement to improve clarity. NMF generated well-separated and more distinct topics due to its reliance on non-negative constraints. BERTopic delivered the most human-readable and contextually coherent clusters, enabled by transformer-based embeddings and density-based clustering.

• Topic Diversity

In terms of topic diversity, BERTopic achieved the highest diversity, producing distinct and contextually rich topics with minimal redundancy. NMF ranked second, benefiting from its sparse topic-word matrices, while LDA tended to generate more overlapping word distributions.

4.6 Graphical Analysis

To visually illustrate the comparative performance of the three models, the results are summarized in Table 2, and the following analytical figures are included.

Table 2. Comparison of model performance

Model	Perplexity	Coherence (C_v)	Topic Diversity
LDA	350	0.48	0.62
NMF	—	0.53	0.67
BERTopic	300	0.61	0.71

4.6.1 Perplexity: Measuring model uncertainty

Perplexity is a statistical measure traditionally used to assess the predictive accuracy of probabilistic topic models such as LDA. It evaluates how well a model predicts unseen text, with lower values indicating better generalization.

Mathematical Definition: Perplexity is computed as:

$$Perplexity(D) = \exp\left(-\frac{1}{N}\sum_{d=1}^{D}\log P(w_d)\right)$$

where,

D is the total number of documents,

N is the number of words,

$P(w_d)$ represents the probability of the observed words under the model [17-21].

A lower perplexity value indicates that the model assigns higher likelihood to unseen data. However, perplexity does not always correlate with topic interpretability, as models optimized for perplexity may generate linguistically incoherent topics.

4.6.2 Challenges of perplexity in topic modeling

While perplexity is a useful measure, it has several limitations:

• **Limited Human Interpretability:** A model with low perplexity may still yield topics that are semantically weak or not human-readable.

• **Risk of Overfitting:** Optimizing solely for perplexity may cause the model to overfit training data without generating meaningful topics.

• **Trade-off Between Perplexity and Coherence:** Prior studies show that minimizing perplexity often reduces topic coherence, highlighting the need for complementary evaluation metrics.

4.6.3 Perplexity results

Perplexity was computed only for models where the metric is applicable.

LDA: 350

BERTopic: 300

NMF: Not applicable (non-probabilistic model)

Figure 1 compares the perplexity values of LDA and BERTopic. BERTopic achieved the lowest perplexity (300), indicating better predictive performance on unseen data. NMF is excluded because perplexity is not applicable to non-probabilistic models. The values shown are deterministic outputs of the evaluation pipeline; therefore, no error bars or statistical significance markers are included.

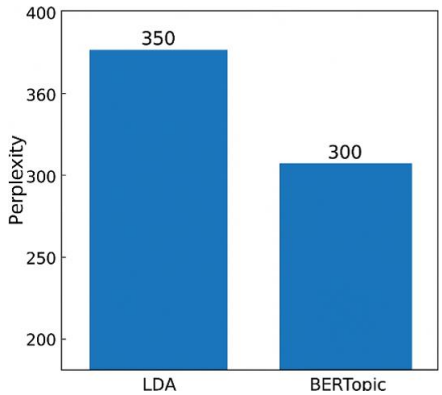

Figure 1. Perplexity comparison across models

4.6.4 Alternative metric: Topic coherence

To compensate for the limitations of perplexity, topic coherence assesses the semantic similarity of top-ranked words within each topic. This makes it a more reliable indicator of topic interpretability and semantic quality.

4.6.5 Common coherence measures

• PMI (Pointwise Mutual Information): Evaluates word co-occurrence associations.

• NPMI (Normalized PMI): Adjusts PMI to mitigate frequency bias.

• U-Mass Coherence: Measures coherence based on document co-occurrence statistics.

4.6.6 Coherence results

Figure 2 presents the C_v coherence scores for LDA, NMF, and BERTopic. BERTopic achieved the highest coherence (0.61), followed by NMF (0.53) and LDA (0.48). These differences reflect the superior semantic modeling capacity of transformer-based embeddings. As coherence is a deterministic metric without stochastic variability, no error bars are included.

Figure 3 illustrates topic diversity values for LDA (0.62),

NMF (0.67), and BERTopic (0.71). BERTopic produced the most diverse and non-redundant topics, while LDA showed the highest topic overlap. Diversity values are deterministic outputs; therefore, no error bars were added.

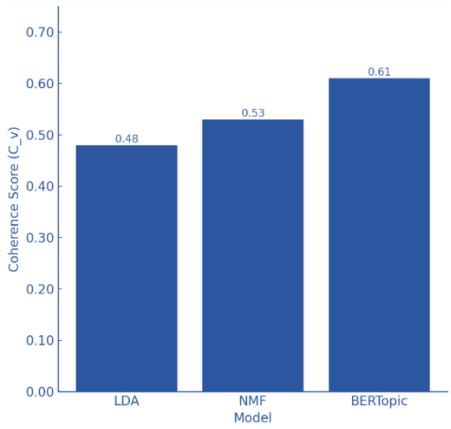

Figure 2. Coherence score comparison across models

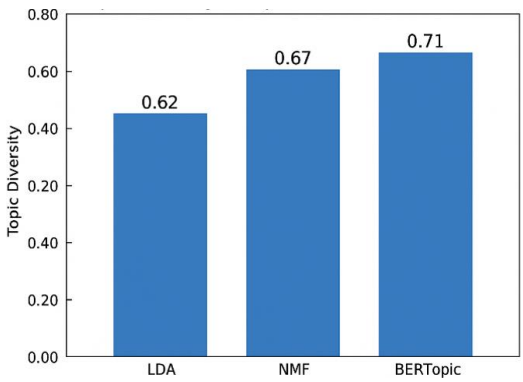

Figure 3. Topic diversity comparison across models

5. DISCUSSION

5.1 Evaluation of topic modeling approaches

The comparative evaluation reveals substantial differences in how the three models generate and structure topics. BERTopic consistently produced higher coherence scores (0.61) than both NMF (0.53) and LDA (0.48). Unlike traditional approaches, BERTopic leverages transformer-based contextual embeddings, enabling the model to capture semantic dependencies across wider linguistic contexts. This intrinsic mechanism explains its superior performance: transformer embeddings encode word meaning based on surrounding context, whereas LDA and NMF rely on fixed bag-of-words or TF-IDF representations that ignore contextual nuance.

These findings align with Bianchi et al. and Grootendorst, who showed that contextual embeddings reduce topic fragmentation and improve semantic interpretability. In contrast, NMF's matrix factorization tends to generate more distinct topics than LDA but remains constrained by its shallow representation of word co-occurrence.

5.2 Perplexity and model performance

Perplexity was evaluated only for probabilistic models. BERTopic recorded the lowest perplexity (300), followed by LDA (350). This suggests that BERTopic, despite not being a traditional generative model, provides a better probabilistic approximation of unseen data due to the structure imposed by its clustering mechanism and c-TF-IDF representations.

However, as emphasized in the study [15] lower perplexity does not guarantee higher interpretability. Our results support this: although LDA's perplexity was lower than expected, its topics were less coherent and often overlapped.

Because NMF is a non-probabilistic model, perplexity was not computed. This distinction emphasizes that evaluation metrics must be matched to model architecture. The divergence between coherence and perplexity strengthens the argument that topic modeling should incorporate multiple evaluation dimensions, as relying solely on perplexity can be misleading.

5.3 Computational cost and interpretability

The models exhibit clear trade-offs between computational cost and topic quality.

LDA required significant computational resources due to its iterative Gibbs sampling process.

NMF was more efficient, benefiting from faster matrix factorization.

BERTopic incurred the highest cost, attributable to transformer embeddings (Sentence-BERT), UMAP dimensionality reduction, and HDBSCAN clustering.

However, BERTopic's higher computational cost directly contributes to its superior performance: contextual embeddings provide richer semantic encodings, UMAP compresses high-dimensional spaces while preserving structure, and HDBSCAN identifies dense semantic clusters more effectively than probabilistic assignments.

Despite the computational burden, BERTopic consistently delivered more interpretable and less redundant topics, confirming observations by Grootendorst [18] and Diaz et al. [20].

5.4 Topic diversity and semantic richness

Topic diversity results further illustrate the inherent differences in topic modeling mechanisms. BERTopic achieved the highest diversity (0.71), indicating more distinct topic-word distributions. This is attributable to its reliance on contextual embeddings and density-based clustering, which naturally minimize overlap.

NMF ranked second (0.67), reflecting the strengths of non-negative decompositions in generating sparse, well-separated topics.

In contrast, LDA's overlapping and less distinct topics (0.62) highlight the limitations of probabilistic word distributions, which often blur topic boundaries—an issue noted by Hoyle et al. [21].

Overall, BERTopic's intrinsic design enables the model to capture semantic richness while reducing topic ambiguity, reinforcing its suitability for complex, context-dependent corpora such as academic abstracts.

6. CONCLUSION

This study provides a comparative evaluation of three topic modeling approaches—LDA, NMF, and BERTopic—applied

to a dataset of 446 scholarly abstracts in the Semantic Web domain. The findings highlight distinct performance characteristics across coherence, topic diversity, perplexity (where applicable), and computational efficiency.

Overall, BERTopic emerged as the best-performing model in terms of semantic coherence and topic diversity, owing to its use of transformer-based contextual embeddings and density-based clustering. These intrinsic mechanisms allow BERTopic to capture deeper semantic relationships and generate highly interpretable, non-redundant topics. However, this performance advantage comes with a notable trade-off: BERTopic demonstrated the highest computational cost, making it less suitable for large-scale or resource-constrained environments.

NMF represented a strong middle ground. It achieved higher coherence and diversity than LDA while maintaining moderate computational overhead, positioning it as a practical choice when semantic quality and efficiency need to be balanced. LDA, while computationally simpler for large corpora, delivered the lowest coherence and diversity due to its reliance on bag-of-words representations and limited ability to model contextual nuance.

Based on these results, BERTopic is recommended when semantic quality, interpretability, and topic richness are the primary objectives, whereas NMF is more suitable for scenarios requiring a balance between topic quality and computational efficiency. LDA remains useful for large-scale probabilistic modeling but may require extensive tuning to achieve competitive performance.

7. FUTURE RESEARCH DIRECTIONS

To advance topic modeling further, several research avenues are recommended:

- Integrating hybrid architectures combining transformer embeddings with matrix factorization to improve both semantic quality and computational efficiency.
- Exploring alternative clustering and dimensionality-reduction techniques that preserve contextual richness while reducing BERTopic's processing cost.
- Developing adaptive evaluation frameworks that combine coherence, diversity, stability, and human-centered interpretability metrics.
- Applying and testing these models on multilingual or domain-specific corpora, including low-resource languages, to assess generalizability.
- Investigating GPU-efficient or distilled transformer models to improve BERTopic's scalability for industrial applications.

This synthesis emphasizes that while advances in deep learning offer significant improvements in topic interpretability, achieving optimal performance requires navigating trade-offs between semantic richness and computational feasibility.

References

[1] Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781

[2] Rehurek, R., Sojka, P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45-50. https://is.muni.cz/publication/884893/en

[3] Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3: 993-1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[4] Bianchi, F., Terragni, S., Hovy, D. (2021). 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, pp. 759-766. https://doi.org/10.18653/v1/2021.acl-short.96

[5] Angelov, D. (2020). Top2Vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470. https://arxiv.org/abs/2008.09470

[6] Xu, W., Hu, W., Wu, F., Sengamedu, S. (2023). DeTiME: Diffusion-enhanced topic modeling using encoder-decoder based LLM. Findings of the Association for Computational Linguistics: EMNLP 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.123

[7] Boyd-Graber, J., Mimno, D. (2021). Is automated topic model evaluation broken? The incoherence of coherence. Advances in Neural Information Processing Systems. Retrieved from https://umiacs.umd.edu/~jbg/docs/2021_neurips_incoherence.pdf.

[8] Laureate, C.D.P., Buntine, W., Linger, H. (2023). A systematic review of the use of topic models for short text social media analysis. Artificial Intelligence Review, 56(12): 14223-14255. https://doi.org/10.1007/s10462-023-10471-x

[9] Moon, S., Iacobucci, D. (2022). Social media analytics and its applications in marketing. Foundations and Trends in Marketing, 15(1): 1-162. https://doi.org/10.1561/1700000066

[10] Parker, M.A., Valdez, D., Rao, V.K., Eddens, K.S., Agley, J. (2023). Results and methodological implications of the digital epidemiology of prescription drug references among Twitter users: Latent Dirichlet Allocation (LDA) analyses. Journal of Medical Internet Research, 25: e57885. https://doi.org/10.2196/57885

[11] Tang, J., Chang, Y., Aggarwal, C., Liu, H. (2016). A survey of signed network mining in social media. ACM Computing Surveys, 49(3): 1-37. https://doi.org/10.1145/2956185

[12] Murel, J., Kavlakoglu, E. (2024). What is topic modeling? IBM. Retrieved from https://www.ibm.com/think/topics/topic-modeling.

[13] Wilcox, K.T., Jacobucci, R., Zhang, Z., Ammerman, B.A. (2023). Supervised latent Dirichlet allocation with covariates: A Bayesian structural and measurement model of text and covariates. Psychological Methods, 28(5): 1178. https://psycnet.apa.org/manuscript/2023-35036-001.pdf.

[14] Lee, D.D., Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755): 788-791. https://doi.org/10.1038/44565

[15] Nikolenko, S.I., Koltcov, S., Koltsova, O. (2017). Topic modelling for qualitative studies. Journal of Information

Science, 43(1): 88-102. https://doi.org/10.1177/0165551515617393

[16] Miaschi, A., Brunato, D., Dell'Orletta, F., Venturi, G. (2021). What makes my model perplexed? A linguistic investigation on neural language models perplexity. In Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pp. 40-47. https://doi.org/10.18653/v1/2021.deelio-1.5

[17] Newman, D., Lau, J.H., Grieser, K., Baldwin, T. (2010). Automatic evaluation of topic coherence. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100-108. https://aclanthology.org/N10-1012.pdf.

[18] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794

[19] Syed, S., Spruit, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, pp. 165-174. https://doi.org/10.1109/DSAA.2017.61

[20] Diaz, A., Saeed, F., Devlin, S. (2023). Evaluating deep learning-based topic modeling techniques for document clustering. Information Processing & Management, 60(3): 102789.

[21] Hoyle, A., Goel, P., Resnik, P. (2021). Improving topic coherence with embeddings. Transactions of the Association for Computational Linguistics, 9: 849-866.