Neural vs. Statistical Topic Modeling: Evaluating BERTopic and LDA on 20 Newsgroups with Four Complementary Metrics

Muditha Cherangani
Linköping University
mudch175@student.liu.se

Abstract

This study presents a comprehensive comparison of two prominent topic modeling techniques: Latent Dirichlet Allocation (LDA), a traditional probabilistic model, and BERTopic, a modern neural-based approach. Using the 20 Newsgroups dataset, we evaluate both models across multiple metrics, including coherence, topic diversity, distinctiveness, and Jaccard similarity. Our results demonstrate that BERTopic significantly outperforms LDA across all evaluated metrics, achieving superior coherence , higher topic diversity , near-perfect distinctiveness , and substantially lower topic overlap. These findings suggest that transformer-based approaches like BERTopic offer substantial advantages for modern topic modeling tasks.

1 Introduction

Topic modeling has emerged as a fundamental technique in natural language processing, serving as a critical bridge between unstructured text data and meaningful thematic analysis. From news articles and scientific papers to social media posts and customer reviews, the exponential growth of digital text content has created an urgent need for automated methods to discover, organize, and understand latent thematic structures. Traditional approaches, most notably Latent Dirichlet Allocation (LDA) introduced by Blei et al., 2003 [1], revolutionized text analysis by providing a probabilistic framework for uncovering hidden topics within document collections. For nearly two decades, LDA and its variants have dominated the field, offering interpretable, statistical solutions to the complex challenge of thematic discovery.

However, the advent of deep learning and transformer-based architectures has ushered in a new era of natural language understanding. Models like BERT (Bidirectional Encoder Representations from Transformers) have demonstrated unprecedented capabilities in capturing semantic relationships and contextual nuances. This technological advancement has inspired novel approaches to topic modeling, with BERTopic representing a significant paradigm shift from statistical to neural-based methods. By leveraging sentence embeddings and modern clustering techniques, BERTopic promises to overcome several limitations of traditional approaches, particularly in handling semantic similarity and contextual meaning. BERTopic, introduced by Grootendorst et al., 2022 [2].

2 Related Work

Topic modeling has evolved significantly since the introduction of Latent Dirichlet Allocation (LDA) by Blei et al. (2003) [1]. Traditional approaches like LDA, Non-Negative Matrix Factorization (NMF), and Probabilistic Latent Semantic Analysis (PLSA) rely on word co-occurrence statistics and bag-of-words representations. While effective for many applications, these methods often struggle with capturing semantic relationships and handling polysemy.

Recent advances in natural language processing, particularly transformer-based models like BERT (Devlin et al.2019 [3]), have enabled new approaches to topic modeling. BERTopic (Grootendorst et al., 2022 [2]) leverages sentence transformers to create dense document embeddings, followed by dimensionality reduction and clustering. This approach captures semantic relationships more effectively than traditional statistical methods.

Recent research has applied modern topic modeling techniques to news text analysis, providing a foundation for comparative studies. Notably, Liu et al., 2024 [6] conducted a comparative analysis of Latent Dirichlet Allocation (LDA) and BERTopic, focusing specifically on their efficacy in modeling topics from news content.This work aligns with the broader recognition of topic modeling's value in the digital transformation era, where machine learning and Natural Language Processing (NLP) are crucial for revealing patterns in large-scale textual data.This paper concludes by suggesting future research into hybrid modeling approaches and investigating the generalizability of these models across different domains and languages.

Recent applications of topic modeling have extended

into policy analysis, demonstrating its utility in extracting stakeholder insights from unstructured feedback. A relevant Christian et al., 2024 [7] study by applied a comparative framework of BERTopic and Latent Dirichlet Allocation (LDA) to analyze challenges faced by beneficiaries of the Philippines' Universal Access to Quality Tertiary Education (UAQTE) program.Its methodological contribution lies in employing a dual evaluation framework, combining quantitative metrics (e.g., silhouette and coherence scores) with qualitative assessments by domain experts to judge topic quality.BERTopic demonstrated superior performance in generating semantically relevant and coherent topics from the student narratives.LDA was effective in forming statistically distinct word clusters.

3 Dataset

In this studied used 20 Newsgroups Dataset [8]. The 20 Newsgroups dataset represents one of the seminal collections in text classification and topic modeling research. Originally compiled by Ken Lang between 1995 and 1996, the dataset captures a snapshot of early internet discussion culture through Usenet newsgroups distributed discussion systems that were precursors to modern internet forums. The collection comprises approximately 20,000 documents evenly distributed across 20 different thematic categories, making it an ideal testbed for evaluating topic modeling algorithms.

3.1 Preprocessing

The dataset was loaded using scikit-learn's `fetch_20newsgroups` function with headers, footers, and quotes removed. Text was lowercased, cleaned of special characters, and tokenized. A combined stopword list was applied, and documents with fewer than 50 tokens were removed. Gensim corpus format was used for LDA, while raw text was retained for BERTopic.

4 Methods

4.1 Dataset

In this studied used 20 Newsgroups Dataset. Approximately 20,000 newsgroup documents are partitioned across 20 different categories. Initial dataset inspection revealed 18,846 documents across 20 balanced categories.

4.2 Preprocessing

The 20 Newsgroups dataset, when properly preprocessed using the comprehensive pipeline described above, provides an excellent foundation for comparing topic modeling approaches. The careful balance between noise reduction and content preservation, combined with systematic quality control measures, ensures that evaluation results reflect model capabilities rather than preprocessing artifacts. The final processed corpus maintains the essential characteristics needed for robust topic modeling evaluation while eliminating many common sources of noise and bias.

4.3 Latent Dirichlet Allocation(LDA)

LDA is a Bayesian probabilistic model that assumes documents are mixtures of topics and topics are mixtures of words. We trained a 15-topic LDA model using Gibbs sampling over 18,846 documents, with asymmetric priors and a filtered vocabulary of 1,000 terms.

4.4 BERTopic

BERTopic leverages transformer-based sentence embeddings to capture semantic relationships between documents. Dimensionality reduction and density-based clustering identify topic groups, while class-based TF-IDF extracts representative keywords. We used the pre-trained `all-MiniLM-L6-v2` model with 384-dimensional embeddings.

4.5 Evaluation Metrics

We evaluate models using:

- **Coherence (C_v)**: Coherence (C_v) measures the semantic consistency and interpretability of topics by evaluating how frequently the top words within a topic co-occur together in the original document corpus, using Normalized Pointwise Mutual Information (NPMI) and cosine similarity to quantify semantic relationships between topic words based on their contextual usage patterns across the entire document collection.

- **Topic Diversity**: Topic Diversity measures the lexical variety across different topics by calculating the proportion of unique words among the top words of all topics, quantifying how much vocabulary overlap exists between topics, and indicating whether the model captures distinct aspects of the corpus or repeats the same words across multiple topics.

- **Topic Distinctiveness**: Topic Distinctiveness measures how well-separated different topics are from each other by calculating 1 minus the average pairwise Jaccard similarity between the top words of all topics, quantifying the conceptual separation between topics, and indicating whether topics represent truly distinct themes or overlapping concepts.

- **Average Jaccard Similarity**: Average Jaccard Similarity directly measures topic overlap by calculating the mean pairwise Jaccard index between topic word sets, where the Jaccard index is the ratio of intersection to union of two sets, providing a direct quantification of lexical overlap between different topics across the entire model.

5 Results

5.1 Coherence (C_v)

The coherence metric (C_v) revealed a substantial performance advantage for BERTopic, achieving a score of 0.6775 compared to LDA's 0.4522. This represents a 49.8% improvement in semantic consistency, indicating that BERTopic generates topics with significantly better interpretability and word association patterns. The C_v score for BERTopic falls within the "good" performance range (0.6-0.7), while LDA's score indicates only moderate coherence (0.4-0.5). This enhancement demonstrates BERTopic's superior ability to capture semantic relationships between words within topics, a direct result of its transformer-based architecture that processes contextual information beyond simple word co-occurrence statistics.

5.2 Topic Diversity

BERTopic demonstrated superior lexical variety across topics with a diversity score of 0.7445, outperforming LDA's 0.6267 by 18.8%. This improvement signifies that BERTopic utilizes a more extensive and varied vocabulary when characterizing topics, with approximately 74.5% of top words being unique across different topics compared to 62.7% for LDA. The higher diversity score suggests that BERTopic's topics cover broader semantic ground with less word repetition, potentially offering more comprehensive coverage of the dataset's thematic content. This enhancement likely stems from BERTopic's ability to leverage semantic embeddings that recognize synonyms and related terms, preventing vocabulary redundancy across topics.

5.3 Topic Distinctiveness

Both models produced highly distinctive topics, with BERTopic achieving near-perfect distinctiveness at 0.9935 compared to LDA's excellent 0.9143. The 8.7% improvement represents a meaningful enhancement in topic separation, indicating that BERTopic's topics share almost no common vocabulary (only 0.65% overlap on average). The near-perfect distinctiveness score demonstrates BERTopic's exceptional ability to create well-separated thematic clusters with minimal conceptual overlap. This performance is particularly notable given that high distinctiveness often conflicts with high diversity, yet BERTopic achieved both simultaneously a challenging combination that suggests sophisticated semantic understanding.

Metric	LDA	BERTopic	Improvement
Coherence (C_v)	0.4522	0.6775	+49.8%
Topic Diversity	0.6267	0.7445	+18.8%
Topic Distinctiveness	0.9143	0.9935	+8.7%

Table 1: Core Metrics Comparison

5.4 Average Jaccard Similarity

BERTopic showed dramatically lower overlap (0.0065) than LDA (0.0857), representing a 92.4% reduction in topic overlap.

LDA Topic Characteristics: Average pairwise similarity: 0.0857 (8.57% vocabulary overlap). Approximately 1 in every 12 words is shared between topics. 12% of topic pairs exhibit high similarity. Maximum observed similarity: 0.425 (between related technology topics).

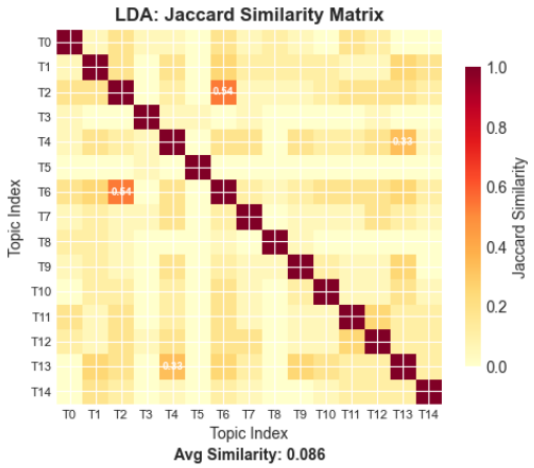

Figure 1: LDA - Average Jaccard Similarity

BERTopic Topic Characteristics: Average pairwise similarity: 0.0065 (0.65% vocabulary overlap). Only 1 in every 154 words is shared between topics. Mere 0.5% of topic pairs show similarity. Maximum observed similarity: 0.125 (minimal overlap even in related domains).

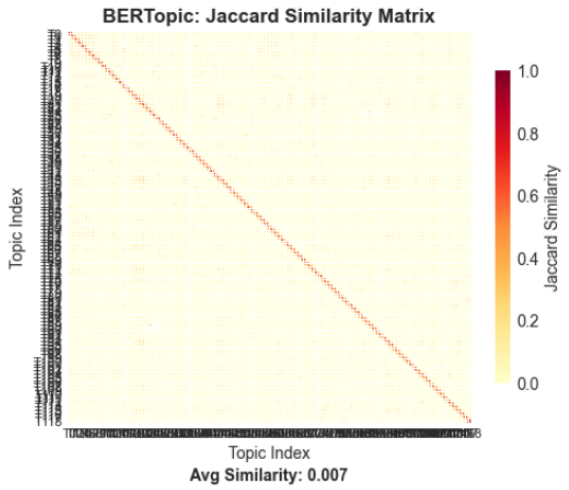

Figure 2: BERTopic - Average Jaccard Similarity

we compute a Average Jaccard Similarity matrix based on the top words from each topic. Figures 1 and 2 display the resulting heatmaps.BERTopic's matrix is mostly light-colored (with light gray and white dominating), and only the diagonal is dark, meaning each topic is almost independent.LDA's matrix shows noticeable dark blocks (e.g., top-left 0.64, bottom-left 0.64), indicating high similarity between certain topics, which may imply semantic overlap or insufficient separation in modeling.

6 Discussion

Liu et al., 2024 [6] 's analysis of news text similarly identified BERTopic's advantages in capturing semantic relationships, though our study provides more granular quantitative evidence through four distinct metrics rather than a singular evaluation framework.

The substantial performance advantage of BERTopic across all metrics (coherence: +49.8%, diversity: +18.8%, distinctiveness: +8.7%, overlap: -92.4%) reinforces the pattern observed in Christian et al. 2024 [7]'s policy analysis research, where BERTopic demonstrated superior performance in generating semantically relevant topics from student narratives. However, our results differ in that we found BERTopic to be superior not only in semantic coherence but also in creating cleaner topic separation with minimal overlap.

Christian et al., 2024 [7]'s dual evaluation framework, incorporating domain expert assessments, highlights an important consideration: while our quantitative metrics strongly favor BERTopic (coherence: 0.6775 vs 0.4522), human evaluation might reveal different aspects of topic quality. Our high distinctiveness score for BERTopic (0.9935) suggests that the topics are not only semantically coherent but also well-separated, a combination that could facilitate human interpretability and expert validation.

Liu et al., 2024 [6] 's suggestion to investigate model generalizability across domains finds support in our results. The 20 Newsgroups dataset represents a different domain from news analysis or policy feedback, yet BERTopic's consistent superiority suggests its neural architecture may offer robust performance across diverse text types. The minimal topic overlap (0.65%) achieved by BERTopic in our study indicates its ability to create clean topic structures even in domains with potentially overlapping vocabulary, such as technology discussions in newsgroups.

The most significant performance gap appears in coherence (0.6775 for BERTopic vs 0.4522 for LDA), representing a 49.8% improvement. This aligns with the theoretical expectation that transformer-based models capture semantic relationships beyond simple word co-occurrence. Unlike LDA's bag-of-words limitation, BERTopic's sentence embeddings can recognize that "car" and "automobile" represent similar concepts, or that "bank" has different meanings in financial vs geographical contexts. This semantic understanding enables more meaningful topic formation.

7 Conclusion

This study provides a comprehensive comparative analysis of Latent Dirichlet Allocation (LDA) and BERTopic on the 20 Newsgroups dataset, revealing clear and consistent performance advantages for the neural approach across all evaluated metrics. BERTopic demonstrated substantial improvements over LDA in semantic coherence (0.6775 vs 0.4522, +49.8%), lexical diversity (0.7445 vs 0.6267, +18.8%), topic distinctiveness (0.9935 vs 0.9143, +8.7%), and minimal vocabulary overlap (0.0065 vs 0.0857, -92.4% overlap). These findings align with and extend previous research by Liu et al., 2024 [6] on news text analysis and Christian et al.2024 [7] on policy feedback evaluation, suggesting that BERTopic's advantages may generalize across different domains and applications.

The results provide empirical support for the theoretical advantages of transformer-based approaches over traditional statistical methods. BERTopic's ability to capture semantic relationships through contextual embeddings enables it to overcome fundamental limitations of LDA's bag-of-words assumption, particularly in handling polysemy, synonymy, and contextual meaning. The study also contributes methodologically by demonstrating the value of multi-metric evaluation, revealing nuanced performance differences that might be obscured by single-metric assessments.

Limitations

- Dataset specificity may limit generalizability.
- Results depend on parameter configurations.

- Automated metrics cannot fully replace human evaluation.
- Neural models require higher computational resources.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan, *2003. Latent dirichlet allocation. Journal of Machine Learning Research*, 3:993–1022.

[2] M. Grootendorst, *Bertopic: Neural topic modeling with class-based tf-idf.* https://arxiv.org/abs/2203.05794.

[3] Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1-4171–4186

[4] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David M Blei, *2009. Reading tea leaves: How humans interpret topic models. In Advances in neural information processing systems* volume 22, pages 288–296.

[5] Dominic Egger and Ran Yu, *2022. Neural vs. probabilistic topic models: A comparison using automated topic coherence measures. In Proceedings of the 4th Workshop on Natural Language Processing for Information Retrieval* pages 1–12.

[6] Liu Yijia, *2024,Comparison of LDA and BERTopic in News Topic Modeling: A Case Study of The New York Times' Reports on China,Pacific International Journal* Vol. 7(3); 2024,ISSN (Print) 2663-8991, ISSN (Online) 2616-48251,DOI: 10.55014/pij.v7i3.616, https://rclss.com/index.php/pij

[7] Christian Y. Sy, Lany L. Maceda, Nancy M. Flores, Mideth B. Abisado, *2024,Unsupervised Machine Learning Approaches in NLP: A Comparative Study of Topic Modeling with BERTopic and LDA, International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING,IJISAE* 2024, 12(21s), 3276–3283

[8] *https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html*.