Response to Reviewers

Manuscript ID: JIPS-D-24-00176

Title: Semantic-based Evaluation Framework for Topic Models: Integrated Deep Learning and LLM Validation

Cover Letter to Reviewers

We would like to express our sincere gratitude to the reviewers for their detailed and constructive feedback on our manuscript. Your valuable comments have significantly improved the academic quality and clarity of our paper. We have carefully reviewed all the points raised and revised our manuscript accordingly. Below, we provide a point-by-point response to each reviewer's comments, detailing the changes we have made.

Reviewer #1 Comments

1. The key contributions of the study should be more clearly outlined.

Response: We thank the reviewer for this important suggestion. We have now clearly articulated the key contributions of our study in the Introduction section (page 2, paragraph 3). Specifically, we have added four distinct contribution points:

1. **Novel Semantic-based Metrics**: We develop comprehensive evaluation metrics based on semantic analysis principles specifically tailored for neural topic models.
2. **Experimental Validation with Controlled Datasets**: We validate the effectiveness of our semantic-based metrics through experiments with three synthetic datasets designed to represent varying degrees of topic overlap.
3. **LLM-based Validation Framework**: We introduce an innovative approach using Large Language Models (LLMs) as proxy domain experts for validation.
4. **Integrated Evaluation Approach**: Our research bridges the gap between evaluation methodologies and modern topic modeling techniques.

Each contribution is now explained with specific details and quantitative improvements where applicable.

2. The related work section is comprehensive, but the connection to the proposed research needs to be strengthened.

Response: We have substantially restructured the Related Work section (pages 2-5) to strengthen the connection to our proposed research. The revisions include:

1. Added Figure 1 (page 2) that illustrates the evolution of topic model evaluation approaches from 2010 to 2024, clearly positioning our research in this trajectory.
2. Added Table 1 (page 4) that compares statistical and semantic-based evaluation methods across seven key evaluation aspects.
3. Reorganized the section into four subsections with clearer logical progression:
    - 2.1 Evolution of Topic Model Evaluation Metrics
    - 2.2 Limitations of Current Evaluation Metrics
    - 2.3 Transition from Statistical to Semantic Evaluation
    - 2.4 Recent Developments in Neural Topic Model Evaluation (2022-2024)
4. Added explicit connections between previous work and our approach at the end of each subsection, particularly in Section 2.3 (page 4, last paragraph).

3. The study categorizes the experimental datasets into Distinct, Similar, and More Similar, but the specific criteria used to differentiate these categories are not fully detailed.

Response: We have added quantitative criteria to differentiate the dataset categories in Section 3.1 (page 5, last paragraph):

"The quantitative similarity levels between these three datasets were determined using cosine similarity measures between topic embeddings. The Distinct dataset has an average inter-topic similarity of 0.21, the Similar dataset shows an average of 0.48, and the More Similar dataset demonstrates an average of 0.67, confirming the intended gradation of semantic overlap in our experimental design."

4. Some table titles and contents lack consistency.

Response: We have corrected all inconsistencies in table titles and contents. Specifically:

- Ensured consistent naming in Table 3 (page 9) by using "Coherence" instead of "Semantic Coherence" to align with the metric categories

- Added detailed explanatory notes beneath Tables 3 and 4 (pages 9 and 13) to clarify the meaning of Mean and CV (Coefficient of Variation) values
- Ensured all table references in the text match the actual table numbers throughout the manuscript

Reviewer #2 Comments

1. Abbreviations Are Not Well Defined

Response: We have now defined all abbreviations upon first use throughout the manuscript:

- CV (Coefficient of Variation) - page 9, Table 3 notes
- RBO (Rank-Biased Overlap) - page 8, Section 3.3.1
- NMF (Non-negative Matrix Factorization) - page 4, Section 2.3
- NPMI (Normalized Pointwise Mutual Information) - page 8, Section 3.3.1
- KL (Kullback-Leibler) Divergence - page 8, Section 3.3.1
- JSD (Jensen-Shannon Distance) - page 8, Section 3.3.1
- IRBO (Inverted Rank-Biased Overlap) - page 8, Section 3.3.1

2. Poor Explanation of Tables

A. Lack of Explanation for Mean Values

Response: We have added clear explanations of Mean values in the notes below Tables 3 and 4 (pages 9 and 13):

"Mean represents the average metric value across all datasets. CV (Coefficient of Variation) is calculated as (standard deviation/mean) × 100, indicating the relative variability of each metric. Lower CV values suggest greater consistency across different dataset conditions."

B. Expected Trend for Each Metric is Not Defined

Response: We have clarified the expected trends for each metric in Sections 4.2 and 4.3 (pages 8-10). For example, in Section 4.3, we have explained:

"This metric evaluates the semantic relatedness between words within a topic, with

higher scores indicating stronger semantic relationships between a topic's keywords." (for Coherence)

"This metric measures how semantically different one topic is from other topics, with higher values indicating clearer boundaries between topics." (for Distinctiveness)

"Higher diversity scores indicate a broader coverage of different concepts and more balanced distribution of topics." (for Diversity)

C. Unclear Measurement of CV (Coefficient of Variation)

Response: We have added a clear definition of CV in the notes below Tables 3 and 4 (pages 9 and 13):

" CV (Coefficient of Variation) is calculated as (standard deviation/mean) × 100, providing a standardized measure of dispersion. A CV of 3.5% indicates acceptable stability across multiple evaluation runs, which is lower than the variability observed in traditional statistical approaches."

D. Incorrect Placement of "Semantic Coherence" in a Statistical-Based Table

Response: We have corrected this inconsistency in Table 3 (page 9) by using consistent terminology for the metrics, ensuring that metric names align properly with their categorization in Section 3.3.

3. Overlapping Representations in Figures 1 and 2

Response: We have consolidated the previously overlapping figures into a single Figure 1 (page 8) that shows all three datasets (Distinct, Similar, and More Similar) with the caption:

"Figure 1. t-SNE Visualization of Topic Distributions: Distinct (left), Similar (center), and More Similar (right) datasets"

The description has been streamlined to focus on the key patterns in each dataset visualization.

4. Poor Presentation of Comparison with Previous Semantic-Based Metrics

Response: We have enhanced the comparative analysis in Section 5.1 (pages 10-11) by:

1. Adding Table 6 (page 11) that quantitatively compares the performance gaps between statistical and semantic metrics across dataset pairs

2. Providing specific improvement percentages: "36.5% improvement in discriminative power ($p < 0.001$)" (page 11)
3. Adding correlation statistics between our metrics and human judgments: "$r = 0.85$, $p < 0.001$" (page 11)
4. Discussing the statistical significance of our results throughout Section 5

5. Missing Explanation of LLM Prompts

Response: We have significantly expanded our explanation of the LLM-based evaluation method in two ways:

1. Enhanced Section 3.3.3 (page 7) to include:
    - The complete system prompt used for LLM evaluation
    - The three-component evaluation process (Individual Topic Assessment, Topic Pair Comparison, Model-Level Synthesis)
    - Inter-rater reliability measures between LLM platforms (Cohen's Kappa, $\kappa = 0.91$)
2. Added a detailed Appendix A (pages 13-15) that provides:
    - The full system prompt text
    - Metric-specific prompts for each evaluation aspect
    - Code examples for coherence, distinctiveness, diversity, and semantic integration evaluation
    - The score calculation methodology
    - Inter-rater reliability calculation method

We once again thank the reviewers for their valuable feedback. We have diligently addressed all the points raised and believe that these revisions have significantly improved the clarity, consistency, and academic rigor of our paper. We hope that the revised manuscript meets your expectations, and we welcome any additional comments or suggestions you may have.

Sincerely, The Authors