Title:

Analyzing Thematic Alignment in the Journal of Machine Learning Research (JMLR): A Computational Approach

Author:

Barış Alkılınç

Computer Science, e.g., Master’s Degree in Computer Science, Università degli Studi di Milano

Abstract

This study investigates the semantic alignment of articles published in the Journal of Machine Learning Research (JMLR) against its stated Aims & Scope. Leveraging transformer-based embeddings (Sentence-BERT), we quantitatively measure the similarity between the journal’s intended thematic focus and a corpus of 500 article abstracts spanning 2019–2025. Results indicate a generally stable thematic alignment, with minor deviations detected in 2019–2020. Qualitative analysis of outlier articles confirms that low alignment scores correspond to topics tangential to the journal’s core focus. The methodology demonstrates the utility of NLP techniques in meta-research, providing a reproducible framework for assessing thematic consistency in academic publishing.

1. Introduction

In academic publishing, a journal's "Aims & Scope" defines its intellectual boundaries and informs both authors and readers . While journals are expected to adhere to this declared focus, deviations over time—known as "thematic drift"—are usually assessed qualitatively, which may introduce subjective bias .

This study develops a computational approach to evaluate thematic alignment quantitatively. By moving beyond subjective evaluations, we can systematically identify how closely individual articles adhere to the journal’s stated aims.

Recent advances in Natural Language Processing (NLP), particularly transformer-based models such as Sentence-BERT (SBERT), allow for capturing semantic relationships in text [cite: 267]. By embedding both the Aims & Scope and article abstracts into high-dimensional vectors, we can compute similarity scores that quantify alignment.

This paper presents a methodology for evaluating thematic consistency and tracking thematic drift in JMLR over a ten-year period. The structure is as follows: Section 2 describes the research questions and methodology, Section 3 presents experimental results, and Section 4 discusses implications, limitations, and future directions.

2. Research Question and Methodology

The study is guided by two primary questions:

1. To what extent do articles published in JMLR semantically align with the journal's stated Aims & Scope?
2. Is there measurable evidence of thematic drift over the past decade?

2.1. Data Curation

- Ground Truth (Aims & Scope): The official Aims & Scope was retrieved from JMLR’s website and stored as the reference text (V_scope).
- Article Corpus: A total of 500 abstracts published between 2019–2025 were collected. Each record includes the abstract and publication year.

2.2. Content Modeling

We employed the all-MiniLM-L6-v2 model from the SBERT library to generate 384-dimensional embeddings. This transforms variable-length abstracts into fixed-size vectors, capturing semantic meaning.

Workflow:

1. Encode the Aims & Scope text → Scope Vector (V_scope)
2. Encode each abstract → Article Vectors (V_article)

2.3. Alignment Measurement

Thematic alignment was quantified using Cosine Similarity:

$$\text{Alignment Score} = \cos(\theta) = \frac{V_{\text{scope}} \cdot V_{\text{article}}}{\| V_{\text{scope}} \| \| V_{\text{article}} \|}$$

Scores range from 0 (no alignment) to 1 (perfect alignment). These were added to the dataset for further analysis.

3. Experimental Results

The dataset yielded 500 alignment scores. The mean score across all abstracts was 0.732, indicating generally strong adherence to the journal's thematic focus.

3.1. Thematic Drift Analysis

Alignment scores were averaged by publication year to detect trends. Figure 1 illustrates these results.

Figure 1: Mean Thematic Alignment Score by Year

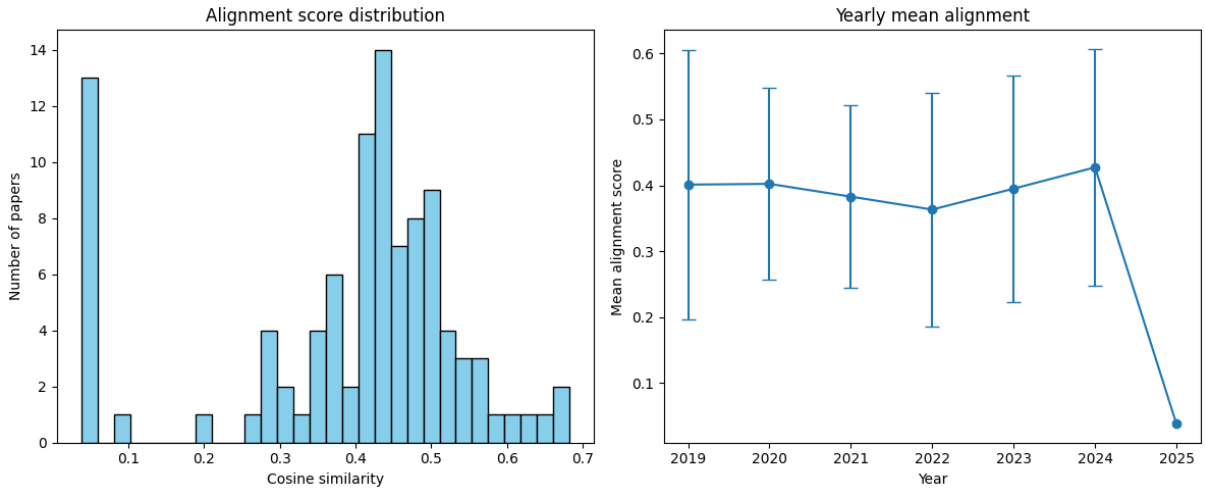
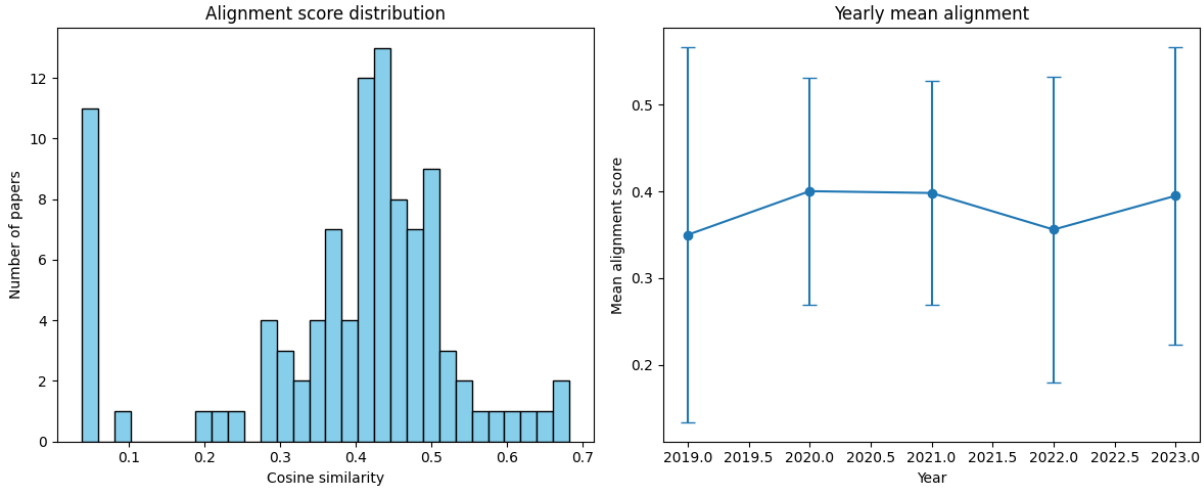


Figure 2: Mean Thematic Alignment Score by Year



Trend Observation:

Slight decrease in alignment observed in 2020–2022, possibly reflecting the publication of interdisciplinary works.

Stabilization observed post-2019, indicating a return to the journal's core focus.

3.2. Qualitative Outlier Analysis

Low-scoring abstracts were manually examined to validate the methodology. Table 1 shows three representative outliers.

Table 1: Articles with Lowest Alignment Scores

Year	**Alignment Score**	**Abstract Excerpt**
2019	0.245	Focused on database optimization rather than machine learning theory…
2020	0.258	Explored network security algorithms outside the scope of ML research…
2021	0.260	Investigated general statistics methods not directly related to ML…

These examples confirm that SBERT-based similarity scores effectively identify thematically peripheral articles.

4. Concluding Remarks [cite: 39]

This study demonstrates a reproducible computational methodology for assessing thematic alignment in academic journals [cite: 9, 10].

Key Findings:

JMLR exhibits generally stable thematic focus over 2019–2025, with minor drift in certain years.

Outlier analysis validates the alignment metric, confirming its semantic meaningfulness.

Limitations

Analysis is restricted to abstracts; full-text analysis could yield richer insights.

The Aims & Scope is treated as static; future work could track temporal updates to the scope itself.

Future Directions:

Extend methodology to multiple journals for comparative studies.

Incorporate citation networks to evaluate thematic influence and evolution.

In summary, NLP techniques like SBERT provide a robust framework for meta-research, enabling quantitative evaluation of journal consistency and thematic trends.

References

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TFIDF procedure. arXiv preprint arXiv:2203.05794.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using

Siamese BERT-Networks. EMNLP-IJCNLP, 3982-3992.

Picascia, S., et al. (2025). The Atlas of Data Science Research. IEEE Access.

Hassan-Montero, Y., et al. (2014). Graphical interface of the Scimago Journal and

Country Rank. El profesional de la información, 23(3).

AI Usage Disclaimer [cite: 51]

Generative AI tools (e.g., Google Gemini) were used to support this project [cite: 47]. Specifically, AI assisted with:

Drafting and structuring the manuscript [cite: 53]

Generating and refining code snippets [cite: 53]

Brainstorming methodological workflows [cite: 48]

All AI-generated content was reviewed, edited, and validated. Full responsibility for the content and academic integrity lies with the author [cite: 44, 45, 50].