

Analyzing Thematic Alignment in Scientific Journals

Betül Gül

16 January 2026

1 Introduction

Scientific journals define their intended thematic boundaries through publisher-authored *Aims & Scope* statements, which function as an explicit contract of topical relevance for authors, reviewers, and readers. In practice, however, what journals publish can evolve over time as research agendas shift, new methods emerge, and terminology changes. This creates a non-trivial assessment problem for bibliometrics and research evaluation: do published articles remain semantically consistent with the journal’s declared scope over time, or do we observe evidence of thematic drift, localized topic changes, or atypical articles that sit far from the journal’s intended coverage?

This problem matters for several reasons. First, scholarly publishing increasingly relies on quantitative signals for monitoring and benchmarking journals, and *scope coherence* is a fundamental quality dimension: if the thematic composition drifts substantially, comparisons across years become less meaningful, and journal identity becomes harder to interpret. Second, scope mismatch may affect downstream scholarly ecosystems (peer-review routing, discoverability, indexing consistency), because metadata and venue labels are routinely assumed to imply a stable conceptual region. Third, even a small number of off-scope publications can distort topic-based or citation-based analytics if not detected, especially when alignment is evaluated over long horizons. More broadly, journal thematic stability is intertwined with how research fields reorganize over time, and measuring such stability requires methods that can adapt to semantic change rather than relying on static keywords.

A key methodological challenge is that surface word overlap is not reliable for scope adherence. *Aims & Scope* texts are short and multi-faceted, while scientific abstracts are compact, technical, and often describe similar concepts using paraphrases or evolving vocabulary. As a result, lexical representations (e.g., bag-of-words or TF-IDF) can underestimate similarity when terminology shifts, and classical topic models can be sensitive to sparse text and vocabulary drift. This motivates semantic representations that capture meaning beyond exact word matches. In particular, Sentence-BERT shows that transformer encoders can produce sentence/paragraph embeddings optimized for semantic similarity and retrieval, enabling scalable comparison between short texts through cosine similarity in embedding space [1].

In addition to paper-to-scope alignment, a complementary question is how the internal conceptual composition of a journal changes over time, even within the set of papers that remain broadly on-topic. Capturing such long-term conceptual evolution requires topic modelling methods that are robust on short technical texts and interpretable enough for reporting. BERTopic is designed for this setting by combining embedding-based document representations with clustering-based topic discovery and class-based TF-IDF topic descriptors, producing human-readable topic summaries and supporting longitudinal topic-prevalence analysis [2].

Motivated by these challenges, this study proposes a reproducible computational pipeline that (i) operationalizes the journal’s *Aims & Scope* as a thematic reference, (ii) represents pub-

lished articles using abstract text as a scalable proxy for content, and (iii) applies embedding-based alignment scoring and topic-prevalence modelling to detect distributional change, concentrated low-alignment cases, and decade-scale conceptual evolution. The pipeline is designed to be auditable: corpus construction relies on authoritative bibliographic APIs, and intermediate checkpoints and saved model artefacts ensure that results can be traced back to explicit data decisions rather than opaque dataset assembly [3, 4].

2 Methodology

This study implements a journal-centric thematic alignment pipeline whose objective is to quantitatively assess whether a journal’s published articles remain consistent with its stated *Aims & Scope* over a multi-year window, and to determine whether this alignment signal exhibits (i) distributional change over time, (ii) concentrated low-alignment outliers, or (iii) longer-term shifts in the corpus’ internal conceptual composition. The journal’s official *Aims & Scope* statement is treated as a stable, publisher-authored thematic reference, while publication content is represented through abstract text, which provides the most consistently available and scalable summary signal across articles and years. The methodological design therefore supports two complementary perspectives: semantic alignment scoring against the scope reference and topic-prevalence modelling to describe long-run conceptual evolution within the retained corpus.

The pipeline is deliberately journal-centric, evidence-driven, and auditable. Meaningful claims about drift, outliers, or conceptual evolution require (a) that the dataset truly belongs to the intended venue and (b) that intermediate processing decisions are traceable. For this reason, the methodology is structured around three principles. First, the corpus is constructed from authoritative bibliographic sources using an API-first approach rather than relying on pre-assembled datasets. Second, journal identity is validated using independent bibliographic evidence so that downstream findings can be interpreted as journal-level properties rather than artefacts of venue pollution or misindexing. Third, all modelling stages are made reproducible through checkpointing and exported evidence tables, so that the exact artefacts used for scoring and topic evolution can be reloaded and re-analysed without re-fitting or re-scraping.

Technically, the pipeline combines complementary components aligned with these objectives. Crossref is used to harvest a high-precision candidate DOI universe for the chosen journal and time range, leveraging standardized bibliographic metadata (ISSNs, container titles, publication year, and work type) as the canonical backbone for dataset construction [3]. Abstracts and minimal metadata required for text-based analysis are then retrieved via the Semantic Scholar Graph API, with resumable checkpointing to ensure robustness under rate limits and incremental processing [4]. Thematic alignment is operationalized through transformer-based semantic embeddings: a Sentence-BERT-style encoder from the Sentence-Transformers ecosystem (`all-mpnet-base-v2`) is used to represent both scope sentences and abstracts as dense vectors optimized for semantic similarity and retrieval [1, 5]. Alignment between abstracts and scope sentences is quantified using cosine similarity on L2-normalized embeddings, enabling consistent scoring at scale and supporting distributional and temporal analyses [6, 7]. Finally, to complement alignment scoring with an explicit view of thematic composition, BERTopic is applied to the same retained abstract corpus to discover latent themes and track their prevalence over time [2].

To maintain traceability, each stage exports persistent artefacts. Data curation produces intermediate snapshots (raw DOI harvest, enriched metadata, curated/scored dataset), journal-purity checks export evidence tables that can be cited directly, and topic-modelling stages persist fitted models and document-level topic assignments/probabilities. This auditability ensures that later findings—score drift, tail behaviour, outlier concentration, and topic-prevalence shifts—can be interpreted in the context of explicit filtering decisions and stable model outputs rather than as ungrounded post-hoc claims. The following subsections operationalize the methodology

through four concrete components: Data Curation, Journal Purity, Content Modeling, and Alignment Measurement, followed by evidence-based findings on alignment-score distributions, time-aware outlier diagnostics with qualitative validation, and long-term conceptual evolution within the curated journal corpus.

2.1 Data Curation

We selected the Springer journal *Machine Learning* (journal no. 10994) because it provides an official and explicit *Aims & Scope* page that can be used as a stable thematic reference, and because its bibliographic identifiers allow strong venue identity checks during dataset validation. The analysis window was fixed to 2016–2025 (inclusive), which is long enough to support longitudinal analyses. Dataset creation was designed to be API-first and reproducible: rather than relying on an existing CSV, we construct the corpus from authoritative scholarly APIs and export the curated dataset as a snapshot artefact, alongside raw logs and intermediate checkpoints.

The first step is capturing the journal’s intended thematic target. The official *Aims & Scope* page is downloaded from the publisher website, its main textual content is extracted, and a snapshot is saved to disk so the same scope reference can be reused even if the webpage changes later. This scope snapshot becomes the reference text that later stages compare against the content of published articles.

The second step is building a canonical candidate list of articles for the chosen journal and time window. We harvest DOIs from Crossref using the journal’s ISSNs and restrict results to journal articles within the target dates using cursor-based pagination [3]. This DOI universe functions as a high-precision backbone for the corpus because Crossref provides standardized bibliographic metadata and a consistent journal container title, which is more reliable for identity than free-text venue strings in aggregator datasets.

The third step is enriching the DOI list with abstracts using the Semantic Scholar Graph API [4]. Each DOI is queried to retrieve the abstract and minimal metadata required for downstream processing, while the implementation is engineered to be robust under rate limits via polite pacing, exponential backoff, and resumable checkpointing. After enrichment, the corpus is filtered using explicit quality gates: papers outside the year window are removed, and records with empty or near-empty abstracts are excluded because later content modeling (embeddings or topics) requires substantive text.

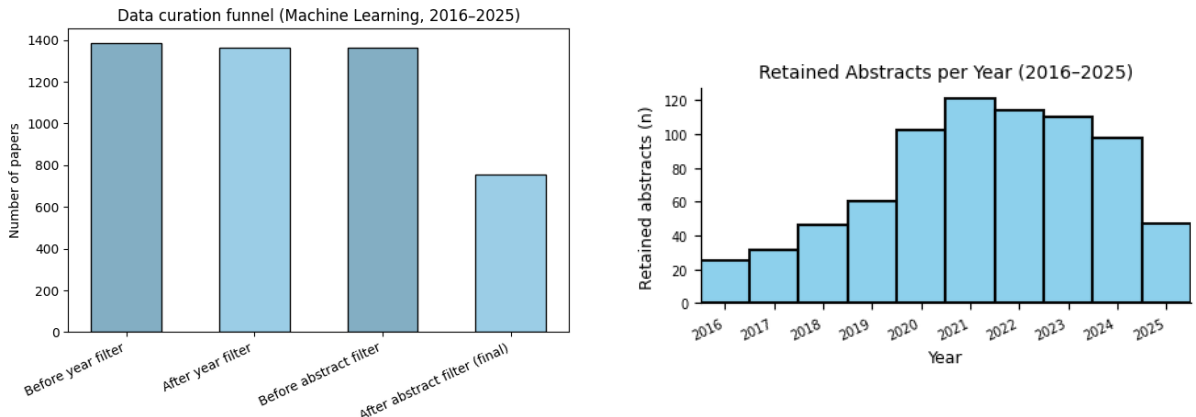


Figure 2: Retained abstracts per year.

Figure 1: Data curation funnel.

Finally, we verify that the curated dataset provides meaningful temporal coverage for longitudinal analysis and report year-wise coverage in downstream sections to account for temporal

variation in data volume.

2.2 Journal Purity

Because drift and outlier claims are only meaningful if the corpus truly belongs to the intended journal, we include an explicit journal purity validation step after curation and after scoring. Purity is assessed using independent bibliographic signals that are robust to noisy venue naming in aggregators. In particular, DOI structure (the Springer journal code prefix) provides a strong identity constraint, while Crossref container-title metadata offers a second independent confirmation at the bibliographic registry layer [3]. Taken together, these checks provide evidence that the curated and scored datasets correspond to the target journal rather than being polluted by similarly named venues or misindexed records. This validation is operationalized as exported evidence tables (overall and year-wise) that can be cited when presenting results and longitudinal analyses.

2.3 Content Modeling

Content is modeled using semantic sentence embeddings to create a structured, machine-readable representation of both the journal’s stated focus and its published articles. The core idea is to map (i) the official *Aims & Scope* text and (ii) each article abstract into a shared semantic vector space where distances reflect topical relatedness. This approach allows alignment to be assessed at the level of meaning rather than surface-level lexical overlap, making it more robust to paraphrase and terminology variation across years.

The *Aims & Scope* statement is decomposed into sentence-level units. This preserves internal granularity and enables fine-grained comparisons: an article can align strongly with one part of the scope even if it is less related to others. Article content is represented through abstract text, and each abstract is encoded into a fixed-dimensional embedding using the same encoder applied to scope sentences, ensuring that scope and article representations live in the same space and can be compared directly.

Embeddings are computed with a Sentence-BERT-style transformer encoder (SentenceTransformers, `all-mpnet-base-v2`) [1, 5]. Vectors are computed in batches and normalized so cosine similarity computations are consistent across items [6, 7]. The result of this stage is a compact set of embeddings representing the scope sentences and a corresponding embedding for each abstract, providing the foundation for alignment scoring.

2.4 Measure Alignment

Once scope sentences and article abstracts are encoded in the same semantic space, thematic overlap is quantified by computing cosine similarity between each abstract embedding and each scope-sentence embedding. When embeddings are L2-normalized, cosine similarity can be computed efficiently as a dot product without changing the underlying metric [6, 7]. For each paper i and scope sentence j , we compute a similarity value s_{ij} , forming a paper-by-scope similarity matrix.

Because the scope statement is multi-faceted, a single paper is not expected to be equally close to every scope sentence. Therefore, the alignment score is derived using two complementary aggregations over per-sentence similarities: (i) a best-match score (maximum similarity over scope sentences), and (ii) a top- (k) mean score (here $k = 3$), computed as the average of the three highest scope-sentence similarities for each paper. The top- (k) mean is treated as the primary alignment score for distributional analysis, drift estimation, and outlier detection, while the max score is retained as a supportive diagnostic.

2.5 Report Findings: Alignment score analysis (top- k mean cosine similarity)

This section analyzes alignment scores computed for the final scored set of 754 papers with valid publication year and alignment score. The analyzed period covers 2016–2025, and the alignment scores range from 0.126 to 0.627. These validation checks matter because distributional summaries, temporal comparisons, and outlier inspection all assume the same score definition is available and comparable across years, without silent missingness.

Across all years, the alignment scores exhibit a single-peaked (unimodal) distribution centered around ~ 0.35 . Dispersion is moderate (std ≈ 0.071), and the middle 50% of papers lie roughly between 0.305 and 0.403. Year-wise means suggest a mild downward tendency over time, motivating coverage-aware reporting and sensitivity checks that exclude low-coverage endpoints.

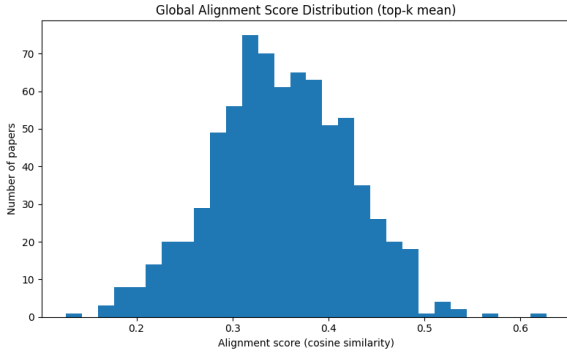


Figure 3: Global alignment score distribution.

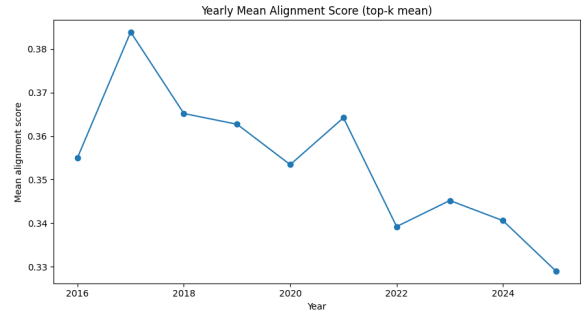


Figure 4: Yearly mean alignment score.

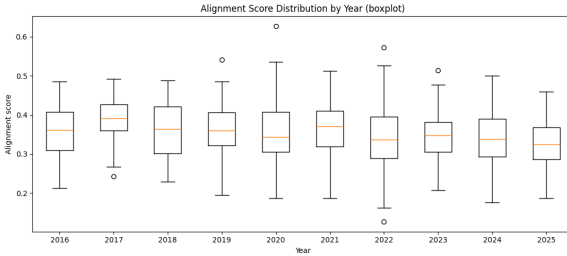


Figure 5: Alignment score distribution by year.

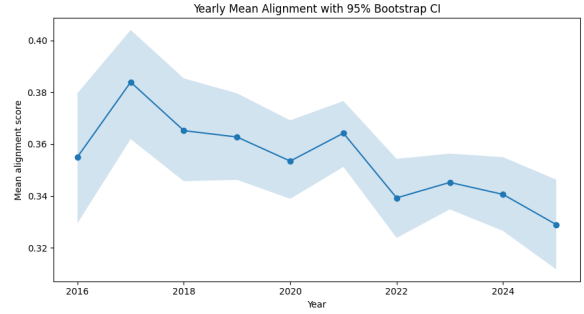


Figure 6: Yearly mean alignment score with 95% bootstrap.

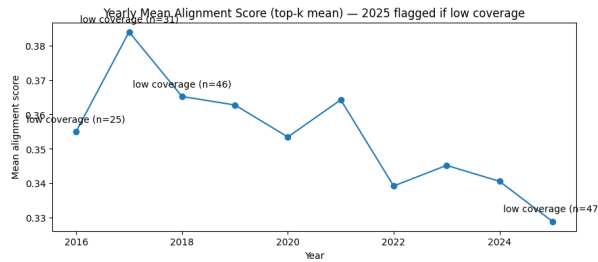


Figure 7: Coverage-aware yearly mean alignment plot.

Table 1: Global alignment score summary (top- k mean cosine similarity) for the scored dataset (2016–2025).

N	Mean	Median	Std	Min	Max	P10	P25	P75	P90
754	0.351	0.351	0.071	0.126	0.627	0.263	0.305	0.403	0.443

Because sample sizes differ by year, yearly means are reported with 95% bootstrap confidence intervals; adjacent years show substantial overlap, but early versus late years separate more clearly. A critical interpretation constraint is that scoring requires usable abstract text, so the analyzed set reflects an abstract-availability filter whose strength varies by year; this motivates coverage-aware visualization and drift-test sensitivity analyses.

Table 2: Year-wise mean alignment scores with 95% bootstrap confidence intervals.

Year	n	Mean	CI _{95%} Low	CI _{95%} High
2016	25	0.355	0.329	0.380
2017	31	0.384	0.362	0.404
2018	46	0.365	0.346	0.385
2019	60	0.363	0.346	0.380
2020	102	0.353	0.339	0.369
2021	121	0.364	0.351	0.377
2022	114	0.339	0.324	0.354
2023	110	0.345	0.335	0.356
2024	98	0.341	0.327	0.355
2025	47	0.329	0.312	0.346

Table 3: Per-year retention after requiring non-empty abstracts (raw after year filter \rightarrow curated/scored).

Year	Before (year-filtered)	After (abstract-filtered)	Retention	Dropout
2016	78	25	0.321	0.679
2017	89	31	0.348	0.652
2018	94	46	0.489	0.511
2019	100	60	0.600	0.400
2020	129	102	0.791	0.209
2021	172	121	0.703	0.297
2022	175	114	0.651	0.349
2023	163	110	0.675	0.325
2024	146	98	0.671	0.329
2025	213	47	0.221	0.779

Table 4: Drift test sensitivity: all years vs excluding the low-coverage endpoint (2016–2024). Trend slope is per-year change (negative indicates downward drift).

Subset	Years	#Years	Kruskal stat	Kruskal p	Trend slope / year
all_years	2016–2025	10	24.622	0.0034	-0.004227
years_2016_2024	2016–2024	9	19.169	0.0140	-0.003737

3 Outlier Analysis

This section examines where the lowest alignment scores come from and whether they reflect (i) genuinely weak thematic connection to the aims/scope sentences or (ii) pipeline-driven artefacts such as missing-abstract selection effects, noisy metadata, or heterogeneous text quality. The goal is not only to list outliers, but to make tail behaviour interpretable and auditable by combining a selection-bias diagnostic (retained vs. dropped pool) with a time-aware characterization of the low-score tail.

The first analysis asks whether the curated corpus is topically similar to the larger year-filtered candidate pool from which it is drawn. The workflow loads the pre-filter pool and creates a single text representation per record (title + abstract when present; title-only otherwise). Records are labelled as retained if they appear in the curated scored dataset and dropped otherwise. A global BERTopic model is fitted once on the pre-filter pool text, producing a topic assignment for each candidate record. For each year, the method compares the topic distributions among retained versus dropped records and quantifies their difference using Jensen–Shannon divergence (JSD), where higher values indicate a larger topical shift induced by the abstract filter.

The second analysis focuses directly on the low-score tail and how it changes over time. Outliers are defined globally using a percentile rule (bottom 5%) to avoid year-specific thresholds that could mask long-horizon change. A year-wise outlier rate is then computed as the fraction of curated papers in each year that fall below this global cutoff. To assess whether temporal change is concentrated in the extreme tail or present across the distribution, tail-aware statistics are computed per year (e.g., 10th percentile, median, 90th percentile). Finally, to make outlier claims auditable at the record level, the workflow exports a qualitative validation sheet for the lowest-scoring papers, including best-matching scope sentence evidence and abstract excerpts.

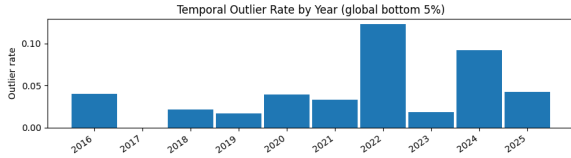


Figure 8: Temporal outlier rate by year.

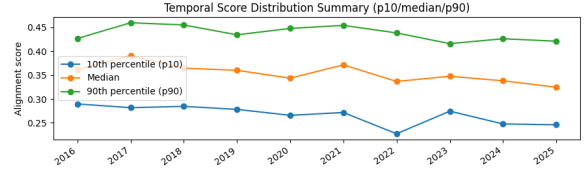


Figure 9: Temporal p10/median/p90 trends.

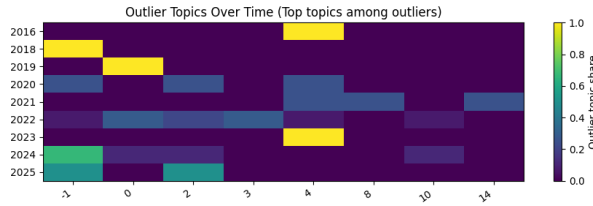


Figure 10: Outlier topics heatmap.

4 Long-Term Conceptual Evolution (Topic Evolution)

Long-term conceptual evolution is analysed by modelling how the prevalence of latent research themes changes across years within the same curated corpus of *Machine Learning* papers. The objective is not to re-check topical relevance, since alignment scoring already quantifies aims/scope consistency and its drift/outlier behaviour. Instead, the goal is to characterise

whether the internal thematic composition of the retained abstracts changes over time. To preserve comparability with the alignment-score analysis, topic modelling is applied to the exact same retained set of 754 abstracts (2016–2025).

The pipeline loads the curated scored dataset and extracts the abstract texts as the document collection for topic modelling, using a consistent year variable for temporal aggregation. A BERTopic model is fitted in a reproducible manner using checkpointing: if a trained model and document-level topic assignments already exist on disk, they are reloaded; otherwise, the model is fitted once and saved along with topic assignments and probability estimates. This design avoids repeated heavy computation and prevents stochastic re-fitting from silently changing downstream figures/tables.

BERTopic is preferred because it produces document-level topic labels that can be aggregated per year, enabling an operational definition of topic prevalence over time. The model yields discrete topic identifiers for each abstract, including the standard outlier/other bucket (−1), which captures documents that do not belong confidently to any dense semantic cluster. Topic evolution is quantified via year-wise topic share: for each year and topic, the share is the count of papers assigned to that topic divided by the total number of retained papers in that year. This normalization is essential because yearly coverage is uneven, and raw counts would largely reflect coverage rather than composition. To support interpretability and audit, the workflow exports a human-readable topic label table (top words, topic size, representative papers) and a labelled pivot table (years \times topics) for inspection.

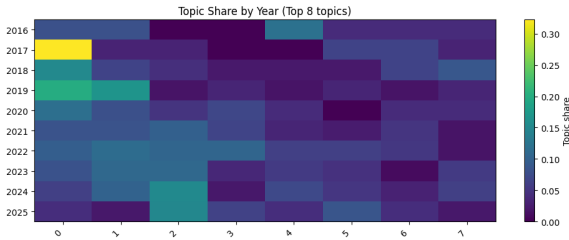


Figure 11: Topic share by year heatmap.

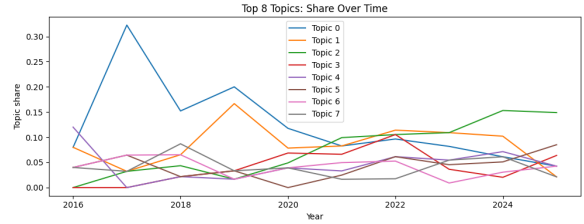


Figure 12: Top topics share over time line plot.

5 Concluding Remarks

This project aimed to build an auditable, journal-centric pipeline for assessing thematic alignment between a journal’s publisher-authored *Aims & Scope* statement and the semantic content of its published articles over time. End-to-end, the workflow (i) constructs a high-precision DOI backbone via Crossref, (ii) enriches records with abstracts via Semantic Scholar, (iii) operationalizes alignment through transformer embeddings and cosine similarity against sentence-level scope units, (iv) evaluates distributional change and tail behaviour through year-aware analyses, and (v) complements alignment scoring with BERTopic-based topic prevalence to describe long-run conceptual evolution within the curated corpus.

Critical discussion of results

The alignment-score distributions provide a compact quantitative view of scope coherence over 2016–2025, and the year-wise summaries motivate the interpretation that changes, if present, are gradual rather than abrupt. However, several constraints are important when interpreting these findings. First, the analysis is conditioned on abstract availability: requiring non-empty abstracts introduces a selection filter whose strength varies by year, which can influence both distributional shape and perceived drift. Second, the reference scope is treated as stable and

authoritative, but scope statements are short, multi-faceted, and may not fully enumerate the evolving boundaries of a research field; alignment therefore reflects similarity to the encoded scope sentences rather than a definitive ground-truth of relevance. Third, using a single embedding model and cosine similarity yields scalable and robust semantic matching, yet the score remains a proxy: it can be affected by domain-specific phrasing, abstract writing style, and the granularity of sentence segmentation in the scope. Fourth, outlier detection based on a global tail cutoff improves longitudinal comparability, but it does not distinguish between genuinely off-scope papers and edge cases driven by abstract brevity, atypical terminology, or partial topical overlap that is not well captured by the chosen similarity aggregation.

Topic evolution results improve interpretability by revealing internal thematic composition shifts among retained papers, but BERTopic assignments also depend on embedding geometry and clustering density. As a result, topic prevalence trends should be interpreted as descriptive patterns of semantic clustering rather than as definitive ontological categories, especially for small yearly sample sizes or topics with limited support.

Ideas for future work

Several extensions would strengthen both validity and explanatory power. First, incorporating full-text (or at least introduction/method sections) when available would reduce dependence on abstract-only proxies and improve robustness for terse abstracts. Second, alignment could be enhanced by adding a contrastive baseline (e.g., a negative corpus from adjacent CS fields) and reporting a contrast score alongside raw similarity, improving interpretability of “low” scores in absolute terms. Third, model robustness can be increased by triangulating multiple encoders, adding cross-encoder reranking for the most critical comparisons, and performing sensitivity analyses over k in top- k aggregation and over alternative outlier thresholds. Fourth, temporal analysis can be strengthened by explicit change-point detection, hierarchical models that account for uneven yearly coverage, and calibration against a small manually validated subset of papers. Finally, generalization beyond a single venue would enable comparative conclusions: applying the same pipeline across multiple journals would reveal whether observed drift and topic evolution patterns are venue-specific or reflect broader field-wide semantic shifts.

Acknowledgements

Use of AI tools

ChatGPT (OpenAI) was used to support drafting and language editing, and to assist with LaTeX formatting. All dataset construction, modelling, statistical analyses, and figure/table generation were carried out by the author using the described pipeline. The author reviewed, verified, and edited all AI-assisted text and assumes full responsibility for the manuscript’s content.

References

- [1] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of EMNLP-IJCNLP*, 2019.
- [2] M. Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, 2022.
- [3] Crossref. Crossref REST API Documentation (Works endpoint; filters and cursor-based pagination). <https://api.crossref.org>.

- [4] Semantic Scholar. Semantic Scholar Graph API Documentation (Paper endpoint; fields-based retrieval). <https://api.semanticscholar.org>.
- [5] Hugging Face. Model Card: `sentence-transformers/all-mpnet-base-v2`. <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>.
- [6] Sentence-Transformers Documentation. `SentenceTransformer.encode` and `normalize_embeddings`. <https://www.sbert.net>.
- [7] Sentence-Transformers Documentation. Semantic Textual Similarity: dot product on normalized embeddings as cosine-equivalent. <https://www.sbert.net>.