Latent Dirichlet Allocation for Mapping Thesis Topic Saturation and Evolutionary Trends in Islamic Guidance and Counseling Studies

Khotibul Umam[1], Iswatun Hasanah[2],
Ayu Raudatul Jannah [3]
[123]Madura State Islamic University, Indonesia
khotibul.umam@uinmadura.ac.id

Abstract. This study aims to analyze the dynamics of research trends and topic saturation in undergraduate students' theses from 2023 to 2026 within the Islamic Education Guidance and Counseling (BKPI) study program at UIN Madura. Using Latent Dirichlet Allocation (LDA), 197 thesis documents were mapped to uncover their underlying thematic structure. Five topics (K = 5) proved optimal after a systematic evaluation of coherence scores, with the model peaking at C_v = 0.411. What emerges from this analysis is a clear shift in the field's research focus: conventional qualitative descriptive work is giving way to quantitative correlational designs centered on psychological variables. Topic 4 (Peer Dynamics and Bullying in Boarding Schools) illustrates this transition most starkly: its share of the corpus fell from 32.62% in 2024 to just 8.51% by 2026, a decline steep enough to signal that the topic has been largely exhausted, offering diminishing space for students to pursue further replication without meaningful variation. The trajectory of Topic 5 (Self-Concept, Social Factors, and Academic Performance) runs in the opposite direction. From a modest 7.87% in 2024, it climbed to claim 38.07% of the corpus by 2026, emerging as the field's most rapidly expanding area of inquiry. Topic 3 (Academic and Career Guidance), by contrast, remained largely unchanged, hovering between 23.07% and 23.31% across the last two years and thereby functioning as a durable, recurring core within the discipline. These patterns provide a data-driven basis for limiting replication in saturated areas and redirecting research toward emerging domains with greater potential for original contributions.

Keywords: Latent Dirichlet Allocation, Topic Modeling, Coherence Score.

1 Introduction

Systematic mapping of research trends is a fundamental component of academic quality assurance in higher education, as it enables academic programs to monitor the scholarly trajectory of their research outputs and identify emerging and underrepresented areas of study. In this context, these serve not merely as a graduation requirement but as a cumulative repository of disciplinary knowledge production. Earlier computational work on thesis corpora has demonstrated that tracking research-topic trends can ease

© The Author(s) 2026
U. Faruq et al. (eds.), *Proceedings of the 10th International Conference of Islamic Studies (ICONIS 2026) & 400th Anniversary of Shekh Yusuf Al-Makassari*, Advances in Social Science, Education and Humanities Research 1043,
https://doi.org/10.2991/978-94-6239-845-0_20

the burden of topic selection, help map out existing research areas, and surface topics that remain comparatively under-explored [1]. This function carries particular weight for the Islamic Education Guidance and Counseling (BKPI) Study Program at UIN Madura, where each graduating cohort's theses collectively trace how the program's core competencies, psychological counseling theory, Islamic educational values, and applied guidance practice are expressed in empirical research. Lacking a structured mechanism for periodic auditing of this accumulated body of work, the program has yet to fully realize the thesis repository's potential as a diagnostic tool for gauging curriculum relevance and charting future research directions.

In practice, thesis topic selection tends to cluster around a narrow set of conventional themes, a pattern driven less by any deliberate assessment of gaps in the field's knowledge base than by simple convenience: students gravitate toward replicating the variables and methodological designs found in readily accessible prior theses, largely because no systematic reference point exists to indicate which topics have already reached the point of saturation. This absence of institutional-level thesis mapping has been documented empirically in Indonesian higher education, where the lack of periodic clustering or trend analysis of accumulated thesis documents has been found to create tangible opportunities for topical overlap among student submissions [2].

This structural weakness in topic-selection guidance is not a problem unique to any single institution; a study that applied text mining to a corpus of undergraduate theses in engineering programs also identified the need for computational solutions to address the challenge of guiding students in selecting appropriate final project topics, proposing topic recommendations based on artificial neural networks as a corrective mechanism[3]. This precedent underscores a broader pattern: institutions that lack a systematic, corpus-level perspective on the outcomes of previous theses tend to leave topic selection to the individual initiative of students, thereby increasing the likelihood of unintended convergence toward themes that have already been extensively studied a risk substantial enough that several Indonesian institutions have separately developed automated title-similarity detection systems specifically to flag proposed thesis titles that overlap excessively with prior work, underscoring how widespread and consequential this gap in systematic topic oversight has become [4].

In the absence of a similar mechanism at BKPI, the risks are twofold: first, an excessive concentration on well-established themes results in an increasingly marginal contribution to the discipline. Second, genuinely under-explored intersections, such as those between the epistemology of Islamic counseling and contemporary psychosocial issues, remain untouched because there is no mechanism to surface these gaps.

Conventional approaches to auditing the topic distribution of theses, such as manual review by academic advisors, ad hoc bibliometric tabulation based on keyword frequency, or qualitative content analysis, cannot be efficiently scaled to handle the ever-increasing volume of documents, and are inherently prone to evaluator subjectivity, inconsistent categorization criteria, and limited reproducibility across evaluation cycles. Computational text mining, particularly topic modeling, overcomes these limitations by inferring latent thematic structures directly from unstructured text without requiring a predefined coding scheme. Among the various topic modeling techniques available, Latent Dirichlet Allocation (LDA) has emerged as one of the most

extensively used methods for uncovering latent structure in text mining, with applications spanning fields as varied as software engineering, political science, and both medical and linguistic research [5]. Its applicability to thesis corpora specifically has been empirically validated. The application of LDA to 666 undergraduate thesis abstracts yielded a coherence score of up to 0.448, indicating adequate topic coherence [6]. In a comparable pipeline pairing LDA with supervised classification, the method successfully identified the five most prevalent thesis topics in a corpus of 233 Computer Science theses: two on computer networks, two on software engineering, and one on multimedia [1]. In addition to identifying static topics, LDA has also demonstrated the capacity for longitudinal analysis. A large-scale bibliometric study applying LDA to 50,176 biomedical articles, stratified by publication period to track shifts in dominant themes over time [7], illustrates the method's suitability for detecting both dominant and declining topics, a capability directly analogous to the objective of identifying over-researched topics in this study. Despite this demonstrated methodological maturity, to the authors' knowledge, no prior work has extended these techniques to an Indonesian undergraduate thesis corpus specifically situated within Islamic educational guidance and counseling. It is precisely this empirical and contextual void that the present study sets out to address.

Guided by this gap, the study draws on Latent Dirichlet Allocation to mine a corpus of thesis titles and abstracts produced by BKPI students at UIN Madura between 2023 and 2026, pursuing four concrete aims: first, to surface the latent topics that most strongly characterize the program's body of thesis work; second, to pinpoint topics showing signs of saturation as a result of disproportionately heavy replication; and third, to trace how the field's research emphasis has shifted longitudinally toward emerging paradigms. and fourth, to offer an interpretive account grounded in each topic's lexical composition and the replication dynamics identified above of why these shifts have occurred. The resulting topic map is intended to serve two direct stakeholders: academic advisors, who gain evidence-based guidance to steer student proposals away from overmapped areas; and prospective students, who gain visibility into thematic gaps to formulate genuinely novel research questions, thereby contributing to a more strategically diversified and cumulatively productive research trajectory in Islamic educational guidance and counseling.

2 Method

2.1 Research Design

This study employs a quantitative, exploratory text-mining design to identify latent thematic structures in a corpus of undergraduate theses. The research process follows four sequential stages: (1) data acquisition from the institutional repository, (2) text preprocessing, (3) topic modeling using Latent Dirichlet Allocation (LDA), and (4) post-hoc analysis of topic distribution and saturation. This design is consistent with previous text-mining studies of undergraduate theses that applied an unsupervised topic-discovery pipeline to institutional repository corpora [1][6].

2.2 Data Sources and Corpus Characteristics

The corpus was assembled from the e-thesis digital repository maintained by the Islamic Education Guidance and Counseling (BKPI) program at UIN Madura, yielding 197 undergraduate thesis records across the 2023–2026 graduation cohorts. Every record links a thesis title to its Indonesian-language abstract, along with associated metadata, including the year of graduation and the name of the supervising advisor.

Although topic modeling algorithms such as Latent Dirichlet Allocation (LDA) are typically applied to large-scale corpora, using a domain-specific, highly focused corpus of 197 documents is methodologically valid for several key reasons. First, the dataset represents a total institutional census rather than a sample, encompassing all undergraduate thesis titles and abstracts produced by the Islamic Education Guidance and Counseling (BKPI) program at UIN Madura during the 2023–2026 academic period. In domain-specific text mining, corpus quality, textual density, and contextual homogeneity often outweigh sheer document volume. Second, the empirical text mining literature demonstrates that LDA can achieve stable topic extraction and adequate semantic coherence (C_v) on small-to-medium datasets when the text is bounded within a single academic discipline and subjected to rigorous, domain-specific preprocessing. Because the vocabulary in thesis abstracts from a single program is highly specialized and semantically aligned, the latent thematic structures can be effectively inferred without the noise typically found in heterogeneous, large-scale datasets.

2.3 Text Preprocessing Stages

The Indonesian-only nature of the dataset called for a preprocessing pipeline tailored to the language's morphological structure rather than a generic, off-the-shelf approach a design choice consistent with prior LDA-based topic modeling studies on Indonesian-language corpora, which have similarly emphasized that generic English-oriented pipelines are inadequate for capturing Indonesian affixation patterns and require a dedicated sequence of case folding, cleaning, tokenization, stopword removal, and stemming stages [8]. Five sequential steps make up this pipeline:

1. Case folding: Every text element was converted to lowercase, eliminating redundancy that inconsistent capitalization would otherwise introduce.
2. Text cleaning: Punctuation marks, numerical values, URLs, and other non-alphabetic characters were stripped from the corpus.
3. Tokenization: Continuous strings of text were then segmented into individual word tokens.
4. Stop Word Removal: High-frequency Indonesian stopwords were filtered out using a hybrid lexicon comprising the default Sastrawi list, augmented with domain-specific, non-informative terms recurrent in thesis abstracts, such as "*tujuan*" (objective), "*penelitian*" (research), "*hasil*" (results), and "*menunjukkan*" (shows). This kind of domain-augmented filtering is consistent with established practice in topic modeling applied to abstract-based corpora [7].

5. Stemming: Root word extraction was performed using the Sastrawi stemmer, which implements the Nazief–Adriani algorithm to handle Indonesian prefix–suffix affixation more reliably than generic stemmers, thereby normalizing affixation and morphological variations inherent in Indonesian academic vocabulary (e.g., reducing pembimbingan and dibimbing to their common root bimbing) while mitigating the over-stemming and under-stemming errors commonly reported when applying non-language-specific stemming rules to Indonesian text [9].

2.4 Topic Modeling using Latent Dirichlet Allocation

Topic modeling was performed using the Gensim library in Python. The preprocessed corpus was converted into a bag-of-words representation via a dictionary corpus mapping. Extreme frequency filtering was applied using the no_below=2 and no_above = 0.8 parameters to eliminate overly rare and excessively common words before model estimation.

The optimal number of topics (K) was determined by iteratively training the LDA model across a range of K = 2 to K = 10. Each configuration was evaluated using the C_v coherence score via Gensim's Coherence Model. The value of K that yielded the highest coherence score while exhibiting the most representatively interpretable semantics was selected as the final model configuration [6].

Following model training, each document was assigned a dominant topic label based on the highest topic probability weight. The temporal distribution was examined through a cross-tabulation of the dominant topic and the graduation year (2023, 2024, 2025, 2026) to identify initial thematic shifts, consistent with the period-stratified analysis approach demonstrated by [7].

Once the final model was established, each document was assigned to a dominant topic label based on the highest probability weight. Topic dynamics were analyzed using three distinct metrics:

6. Average Annual Proportion (Pyear): Computed by summing topic probabilities across all documents published within a given year, divided by the total document count for that year.
7. Change in Proportion (Δ): Captures the year-over-year shift by subtracting the 2025 proportion from the 2026 value (Δ = P_2026 - P_2025).
8. Relative Growth Rate (Growth: Expresses the percentage shift across the observation window from 2024 to 2026 (Growth}= [(P_2026 - P_2024) / P_2024] x 100%).

To maintain conceptual consistency, a clear distinction was made between Topic Dominance and Research Saturation. Topics maintaining an overall share above 20% (P > 0.20) without a continuous decline were categorized as Dominant or Core Topics. Conversely, Research Saturation was formally operationalized under two empirical conditions: either a topic's cumulative proportion collapsed sharply with a relative growth rate below -50% (Growth < -50%), or it followed a sustained downward trajectory, leaving its annual share below a 10% threshold (P_2026 < 0.10), a pattern interpreted as the downstream effect of excessive early-phase replication.

Topics that show a comparable declining trend but do not formally cross either threshold are reported separately as approaching saturation, rather than being classified as fully saturated, to preserve consistency between the operational definition and its subsequent application to the data.

2.5 Evaluation Metrics

Model quality in this study was assessed primarily using the C_v Coherence Score. This metric captures how semantically related the top-ranked words within each topic are to one another by combining sliding-window co-occurrence statistics with normalized pointwise mutual information and cosine-based similarity [10]. Following this rationale, the C_v score was calculated for each candidate configuration between K = 2 and K = 10 using Gensim's Coherence Model, and the configuration with the highest average score served as the primary basis for selecting the number of latent topics retained in the final model.

Perplexity was deliberately treated as a secondary rather than a primary criterion, since prior work has shown that models optimized purely for predictive likelihood do not necessarily produce topics that humans perceive as coherent. In some cases, the two measures move in opposite directions [11]. Word and topic intrusion experiments demonstrated that this gap between statistical fit and interpretability is a genuine limitation of relying solely on held-out likelihood to judge topic model quality [11]. For this reason, perplexity values were tracked over the same K range to confirm that the chosen model's probability estimates remained stable. At the same time, the actual selection of K was anchored to the coherence-maximizing configuration and cross-checked against the interpretability of the resulting keyword clusters upon manual inspection.

3 Result and Discussion

3.1 Evaluation of Coherence Score and Determination of Optimal Topic Count

The LDA modeling process began by computing the Coherence Score (C_v) as the criterion for identifying the optimal number of topics, K. This evaluation systematically swept through candidate values from K = 2 to K = 10. Among these, the value of K ultimately selected was the one that yielded the highest coherence score, an outcome taken to reflect the interpretability and semantic consistency of the keywords within each resulting topic cluster.

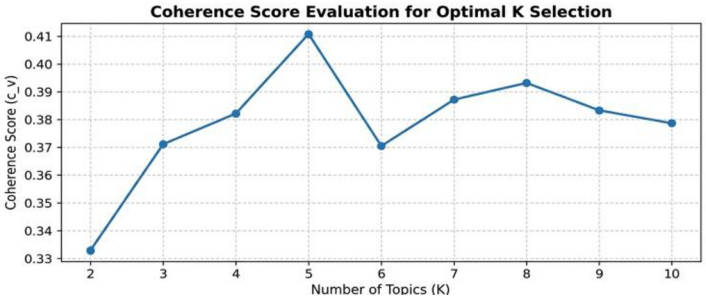


Fig. 1. Coherence Score (C_v) Evaluation across Different Topic Counts (K)

As illustrated in Figure 1, the quantitative evaluation of the LDA model across K = 2 to K = 10 shows that the Coherence Score (C_v) reaches its local maximum at K = 5 (C_v = 0.411), followed by a drop at K = 6 (C_v = 0.370) before experiencing a secondary minor peak at K = 8 (C_v = 0.393).

Based on this evaluation, K = 5 was selected as the optimal configuration, and the corresponding top words are presented in Table 1.

Table 1. Extraction of Primary Keywords (Top Words) for Each Topic from the LDA Model (K = 5)

Topic	Top Words with the Highest Probability
Topic_1	*diri* (self), *test* (test), *kelompok* (group), *konseling* (counseling), *bimbingan* (guidance), *guru* (teacher), *teknik* (technique), *layanan* (service), *remaja* (adolescent), *perilaku* (behavior)
Topic_2	*anak* (child), *orang tua* (parents), *remaja* (adolescent), *pola asuh* (parenting style), *perilaku* (behavior), *dampak* (impact), *online* (online), *game* (game)
Topic_3	*bimbingan* (guidance), *layanan* (service), *diri* (self), *guru* (teacher), *karir/karier* (career), *faktor* (factor), *akademik* (academic), *konseling* (counseling), *melalui* (through)
Topic_4	*santri* (Islamic boarding school student), *perilaku* (behavior), *bullying* (bullying), *pesantren* (Islamic boarding school), *faktor* (factor), *pondok* (boarding school/dormitory), *belajar* (learning), *teman* (peer/friend), *dampak* (impact), *bagaimana* (how)
Topic_5	*belajar* (learning), *uji* (test/examination), *sosial* (social), *hubungan* (relationship), *self* (self), *nilai* (grade/score), *sebesar* (magnitude/effect size), *diri* (self), *tinggi* (high), *prestasi* (achievement)

Table 1 presents the output of the Latent Dirichlet Allocation (LDA) topic modeling, highlighting the keywords with the highest probability within each identified cluster. Taken together, these word groupings sketch out five distinct thematic territories within the dataset:

1. Topic 1 (Group Guidance and Counseling): built around terms such as *diri* (self), *test* (test), *kelompok* (group), *konseling* (counseling), *bimbingan* (guidance), *guru*

(teacher), *teknik* (technique), *layanan* (service), *remaja* (adolescent), *perilaku* (behavior), this cluster points to research examining group-based counseling approaches that educators use to address adolescent behavioral issues.

2. Topic 2 (Parenting Styles and Online Gaming Impacts): shaped by *anak* (child), *orang tua* (parents), *remaja* (adolescent), *pola asuh* (parenting style), *perilaku* (behavior), *dampak* (impact), *online* (online), and *game* (game), this cluster captures the interplay between family dynamics, particularly parental guidance, and emerging youth behaviors tied to gaming habits.
3. Topic 3 (Academic and Career Guidance Services): drawing on *bimbingan* (guidance), *layanan* (service), *diri* (self), *guru* (teacher), *karir/karier* (career), *faktor* (factor), *akademik* (academic), *konseling* (counseling), and *melalui* (through), this cluster centers on counselor-driven interventions aimed at supporting students' academic trajectories and career planning.
4. Topic 4 (Peer Dynamics and Bullying in Islamic Boarding Schools): Defined by *santri* (Islamic boarding school student), *perilaku* (behavior), *bullying* (bullying), *pesantren* (Islamic boarding school), *faktor* (factor), *pondok* (boarding school/dormitory), *belajar* (learning), *teman* (peer/friend), *dampak* (impact), *bagaimana* (how). This thematic area examines social challenges within pesantren settings, particularly peer interactions and bullying phenomena.
5. Topic 5 (Self-Concept, Social Factors, and Academic Performance): Driven by *belajar* (learning), *uji* (test/examination), *sosial* (social), *hubungan* (relationship), *self* (self), *nilai* (grade/score), *sebesar* (magnitude/effect size), *diri* (self), *tinggi* (high), *prestasi* (achievement). The presence of analytical terms like uji (test), hubungan (relationship), and sebesar (effect size) points toward quantitative designs evaluating how psychological traits such as self-concept relate to academic outcomes.

3.2 Annual Topic Distribution (2023–2026)

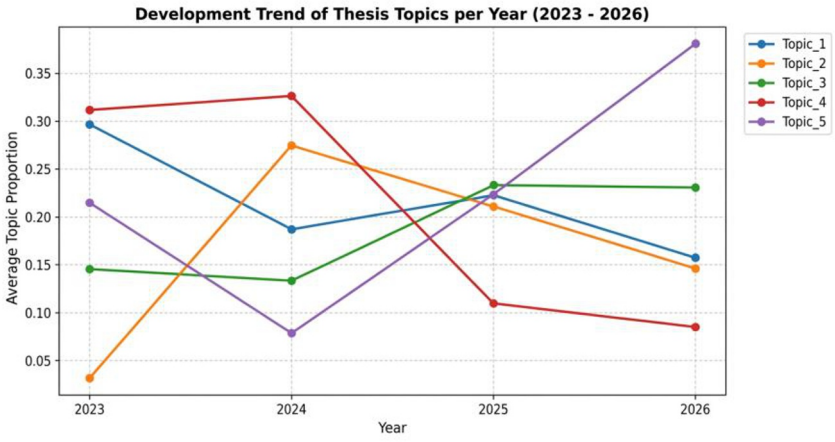


Fig. 2. Annual Trends of Thesis Topic Proportions from 2023 to 2026

Figure 2 traces how topic proportions shifted over the 2023–2026 period, revealing pronounced thematic realignment within the dataset. Topic 4 represents the most striking case: after commanding a share above 0.31 throughout 2023-2024, it fell to below 0.10 by 2026, a pattern that squarely meets the study's operational threshold for research saturation. Topic 1 and Topic 2 followed a comparable pattern of decline, each having peaked earlier before receding. Topic 5, however, moved in the opposite direction, climbing from a low of 0.08 in 2024 to overtake all others as the dominant research focus by 2026, with its mean proportion approaching 0.38. Meanwhile, Topic 3 displayed a steady growth trajectory before stabilizing around a consistent 0.23 proportion between 2025 and 2026. These empirical shifts reflect an evolving institutional research focus, transitioning from saturated classical themes toward emerging contemporary paradigms

Categorizing the 197 documents according to their computed probability distributions generated an annual average topic proportion matrix spanning 2023 to 2026. The distribution proportions for the five identified topics are outlined in Table 2.

Table 2. Average Topic Proportions per Year (2023–2026)

Year	**Topik_1**	**Topik_2**	**Topik_3**	**Topik_4**	**Topik_5**
2023	0,2965 (29,65%)	0,0318 (3,18%)	0,1454 (14,54%)	0,3116 (31,16%)	0,2147 (21,47%)
2024	0,1870 (18,70%)	0,2746 (27,46%)	0,1335 (13,35%)	0,3262 (32,62%)	0,0787 (7,87%)
2025	0,2226 (22,26%)	0,2109 (21,09%)	0,2331 (23,31%)	0,1098 (10,98%)	0,2237 (22,37%)
2026	0,1573 (15,73%)	0,1462 (14,62%)	0,2307 (23,07%)	0,0851 (8,51%)	0,3807 (38,07%)

As detailed in Table 2, the longitudinal empirical data underscore an exact proportional distribution across all five topics, documenting the precise numerical trajectory underpinning these thematic realignments from 2023 through 2026. Topic 4 held the largest share of the corpus at the outset in 2023 (31.16%), reached its peak the following year at 32.62%, then contracted sharply to 10.98% in 2025 before settling into a saturated baseline of 8.51% by 2026. Topic 5 moved in the reverse direction, bottoming out at 7.87% in 2024 before expanding rapidly to 22.37% in 2025 and ultimately claiming the largest share of the dataset in 2026 at 38.07%. Topic 3, meanwhile, grew steadily from its 2024 low of 13.35% toward a stable plateau of roughly 23.07%–23.31% across the final two years, while both Topic 1 and Topic 2 retreated from their respective historical peaks to fall below 16% by 2026.

3.3 Topic Trend Dynamics and Saturation

Trends in the trajectory of undergraduate thesis topics were evaluated using annual percentage point shifts (Δ = P2026 - P2025) and the relative growth rate (Growth =

(P2026 - P2024) / P2024 × 100%). Based on these calculations, the dynamics of the topics were grouped into three main categories:

1. Saturated Topics (Declining Clusters): Topics showing continuous proportional contraction fall under this phase. Topic 4 (Peer Dynamics & Bullying in Boarding Schools) experienced the sharpest decline; after holding over 31% in 2023–2024, it dropped to 10.98% in 2025 and settled at 8.51% by 2026 (Δ= -2.47%; Growth= -73.91). satisfying both saturation conditions simultaneously. Topic 2 (Parenting & Online Gaming), by contrast, contracted from its 2024 peak of 27.46% down to 14.62% in 2026 (Growth = -46.76%) without formally crossing either saturation threshold: its growth rate remains short of the -50% cutoff, and its 2026 share stays above the 10% floor. Given its three consecutive years of decline from that peak, it is reported here as approaching saturation rather than fully saturated, a distinction preserved throughout the remainder of this paper.
2. Emerging Topics (Rapidly Growing Clusters): Topics exhibiting substantial late-period growth represent the field's new directional momentum. Topic 5 (Self-Concept, Social Factors, & Academic Performance) drove this shift; after hitting a low of 7.87% in 2024, its proportion expanded to 22.37% in 2025 and surged to 38.07% in 2026 (Δ = +15.7%; Growth= +383.74%), making it the dominant driver of current student research.
3. Core and Stabilizing Topics: Topic 3 (Academic & Career Guidance) demonstrated strong structural stability; following initial growth from its 2024 baseline (13.35%), it stabilized into a consistent core, maintaining 23.31% in 2025 and 23.07% in 2026 (Δ = -0.24%). Meanwhile, Topic 1 (Group Guidance) retreated from its 2023 peak (29.65%) to 15.73% in 2026 (Δ= -6.53%), reflecting a shift from high popularity toward thematic consolidation.

Collectively, these empirical realignments reflect evolving institutional priorities, shifting away from saturated descriptive themes toward contemporary quantitative paradigms. The persistent drop in Topics 4 and 2 signals that further replication risks diminishing scientific novelty, whereas the dominance of Topic 5 (38.07%) and the stability of Topic 3 (23.07%) highlight where meaningful future contributions remain most viable.

3.4 Discussion

The topic modeling analysis reveals a substantial paradigm shift in undergraduate thesis research output within the study program during the 2023–2026 period. This shift reflects not only dynamic changes in topic preferences but also a methodological evolution from qualitative-descriptive studies toward more rigorous quantitative-correlational analyses. In 2023–2024, student focus was heavily concentrated on local institutional studies, particularly Topic 4 (Peer Dynamics & Bullying in Boarding Schools at 31.16%–32.62%) and school counseling services (Topic 1 at 29.65%). By 2026, however, a massive shift in focus emerged toward investigating internal student psychological variables using parametric quantitative approaches (Topic 5 at 38.07%). This observed thematic shift aligns with findings by [7], which demonstrate that longitudinal

analysis of literature output effectively maps the transition of research attention from conventional themes to emerging topics that are more responsive to empirical analysis needs.

While the topic-modeling procedure applied here is inherently descriptive, surfacing what shifted rather than directly measuring why, the internal composition of the topics themselves offers interpretable evidence toward an explanation. The lexical profile of the declining Topic 4 *(santri, perilaku, bullying, pesantren, bagaimana*) is dominated by descriptive and interrogative markers characteristic of qualitative case-study framing. In contrast, the ascending Topic 5 (*uji, sebesar, hubungan, nilai, tinggi*) is dominated by statistical-inferential vocabulary terms directly associated with hypothesis testing and effect estimation. This lexical divergence is itself evidence that the observed proportional shift is not merely thematic but methodological: BKPI thesis production is migrating from qualitative, single-site ethnographic designs toward quantitative, correlational designs.

A plausible mechanism for this migration follows directly from the replication dynamic established earlier in this paper (Section 1). In the absence of a systematic reference point for topic saturation, students preferentially adopt methodological templates modeled on readily accessible prior theses [2][4]. Under this account, Topic 4's decline and Topic 5's ascent are two faces of the same process rather than independent trends. Once early cohorts (2023–2024) had produced a critical mass of qualitative *pesantren*-behavior theses, that thematic space offered diminishing differentiation for subsequent students, who instead gravitated toward the correlational self-concept design as it became increasingly visible, accessible, and plausibly perceived as methodologically more rigorous and more readily publishable, given the wider disciplinary emphasis on quantitative, measurable outcomes in counseling and educational research. Topic 2 (Parenting & Online Gaming) appears to have followed a comparable, though less advanced, trajectory: its qualitative-behavioral framing has been losing momentum for reasons structurally similar to Topic 4, even though it has not yet crossed the formal saturation threshold and its substantive focus differs.

It is important to note that the present design, a longitudinal but purely observational topic-modeling analysis of thesis metadata, cannot on its own establish these mechanisms as causal. The explanations offered above are data-consistent interpretations rather than direct measurements of institutional facts. Confirming them would require complementary qualitative inquiry, for instance semi-structured interviews with thesis advisors regarding shifts in supervision practice, or a review of BKPI's methodology-course curriculum across the 2023-2026 period to determine whether formal training in quantitative methods expanded during this window. We flag this as a concrete direction for future research in Section (Limitations/Future Work), rather than a limitation that undermines the present findings. The topic-saturation map itself remains a valid and actionable diagnostic tool independent of the causal mechanism behind it.

The sharp decline in Topic 4 (Boarding School Peer Dynamics) from 32.62% in 2024 to 8.51% in 2026 (Growth = -73.91%) provides empirical evidence of research saturation. This saturation was triggered by the high frequency of the replication of homogeneous qualitative-descriptive designs without updating variables or analytical frameworks. A similar downward pattern observed in Topic 2 (Parenting & Online

Gaming), dropping from 27.46% in 2024 to 14.62% in 2026 (Growth = -46.76%), indicates that examining conventional behavior themes without integrating digital media or modern instruments has begun to lose its academic appeal. As highlighted by [1], research dominance in heavily mapped themes risks diminishing the novelty and scientific contribution of final projects unless novel methodological perspectives are introduced.

The surge in Topic 5 (Self-Concept, Social Factors, & Academic Performance), expanding to represent more than one-third of the total proportion in 2026 (38.07%), was driven by students' methodological need to generate statistically measurable causal or correlational evidence. Conversely, the stability of Topic 3 (Academic & Career Guidance), which consistently hovered between 23.07% and 23.31% during 2025–2026, demonstrates that this theme serves as a sustainable core topic. Career and academic development remain fundamental socio-psychological determinants unaffected by shifting methodological trends.

This topic map offers department administrators and thesis advisors an empirical basis for two strategic decisions:

1. Regarding replication control, new title proposals centered on Topic 3 or Topic 4 should face tighter scrutiny and be accepted only when they offer a genuinely novel methodological approach or an unexplored combination of variables.
2. For cross-topic collaboration to drive scientific novelty, students could be steered toward integrating emerging themes (Topic 5) with the core.

4 Conclusion

Based on Latent Dirichlet Allocation (LDA) topic modeling of 197 undergraduate thesis documents from the Islamic Educational Guidance and Counseling (BKPI) program at UIN Madura spanning 2023 to 2026, the optimal topic configuration was determined to be K = 5. The findings point to a genuine paradigm shift: the field has moved away from its early reliance on local, institution-specific qualitative-descriptive work and toward quantitative-correlational research centered on internal psychological variables.

Across the four-year observation window, the dynamics of topic evolution fell into three structural categories:

1. Saturated Topics: Topic 4 (Boarding School Peer Dynamics & Bullying) met both saturation criteria simultaneously, dropping from 32.62% in 2024 to 8.51% by 2026 (Growth = -73.91%) as a consequence of extensive qualitative replication. Topic 2 (Parenting Styles & Online Gaming) traced a comparable downward path after its 2024 peak (Growth = -46.76%; 2026 share = 14.62%) and is flagged here as approaching, though not formally meeting, the study's saturation threshold.
2. Emerging Topics: Topic 5 (Self-Concept, Social Factors, & Academic Performance) expanded rapidly, rising from 7.87% in 2024 to command 38.07% of the corpus by 2026.

3. Core Topics: Topic 3 (Academic & Career Guidance) remained a stable anchor, holding steady between 23.07% and 23.31% during 2025–2026, while Topic 1 (Group Guidance) shifted into a phase of consolidation.

These empirical trends offer actionable insights for department leadership and advisors to manage topic saturation. Recommended next steps include cross-topic integration, such as pairing emerging themes (Topic 5) with core pillars (Topic 3), or revitalizing saturated areas using modern digital tools to safeguard research novelty.

This study is descriptive and correlational by design: LDA identifies patterns of thematic change but does not, on its own, adjudicate between competing causal explanations for those patterns. Future work should complement this topic-saturation map with qualitative validation, for example, through interviews with thesis advisors to probe shifts in supervision emphasis, or through a documentary review of curriculum and methodology course offerings across the observation period to more directly establish which institutional factors, if any, drive the transitions identified here.

References

[1] H. Hairani, M. Janhasmadja, A. Tholib, J. Ximenes Guterres, and Y. Ariyanto, "Thesis Topic Modeling Study: Latent Dirichlet Allocation (LDA) and Machine Learning Approach," *Int. J. Eng. Comput. Sci. Appl.*, vol. 3, no. 2, pp. 51–60, 2024, doi: 10.30812/ijecsa.v3i2.4375.

[2] L. Cahyani and M. Arif, "Text Mining untuk Pengelompokan Skripsi di Prodi Pendidikan Informatika Universitas Trunojoyo Madura," *J. Ilm. Edutic Pendidik. dan Inform.*, vol. 8, no. 2, pp. 97–108, 2022, doi: 10.21107/edutic.v8i2.13020.

[3] X. Li and M. F. Rosas, "Graduation Thesis Topic Recommendation Based on Neural Network," pp. 409–414, 2023, doi: 10.2991/978-94-6463-040-4_62.

[4] H. Haruna, R. Y. Bakti, and T. Wahyuni, "Deteksi Tingkat Kemiripan Judul Skripsi Prodi Informatika Menggunakan Metode Support Vector Machine dan Natural Languange Processing," *Arus J. Sains dan Teknol.*, vol. 2, no. 2, pp. 447–452, 2024, doi: 10.57250/ajst.v2i2.655.

[5] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimed. Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, 2019, doi: 10.1007/s11042-018-6894-4.

[6] B. Sugiantoro, A. I. Humam, N. L. Fitriyani, G. Alfian, M. R. Maarif, and M. Syafrudin, "Utilizing Latent Dirichlet Allocation for Analyzing Topics in Undergraduate Theses," in *2023 10th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2023*, 2023, pp. 121–126. doi: 10.1109/ICITACEE58587.2023.10277046.

[7] J. Chen, M. Williams, Y. Huang, and S. Si, "Identifying Topics and Evolutionary Trends of Literature on Brain Metastases Using Latent Dirichlet Allocation," *Front. Mol. Biosci.*, vol. 9, no. June, pp. 1–9, 2022, doi: 10.3389/fmolb.2022.858577.

[8] C. Naury, D. H. Fudholi, and A. F. Hidayatullah, "Topic Modelling pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia Menggunakan LDA dan LSTM," *J. MEDIA Inform. BUDIDARMA*, vol. 5, no. 1 SE-Articles, pp. 24–33, Jan. 2021, doi: 10.30865/mib.v5i1.2556.

[9] M. U. Albab, Y. K. P., and M. N. Fawaiq, "Optimization of the Stemming Technique on Text Preprocessing President 3 Periods Topic," *J. Transform.*, vol. 20, no. 2, pp. 1–12, 2023, doi: 10.26623/transformatika.v20i2.5374.

[10] M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," *WSDM 2015 - Proc. 8th ACM Int. Conf. Web Search Data Min.*, pp. 399–408, Feb. 2015, doi: 10.1145/2684822.2685324.

[11] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei, *Reading Tea Leaves: How Humans Interpret Topic Models*, vol. 32. 2009.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

