Student Perception in AVALIA UFPA: A Hybrid Approach Using Topic Modeling and LLM-Based Automatic Labeling

Paula Manoella M. de Brito[1], **Samara L. de Souza**[1], **Reginaldo C. dos Santos Filho**[1]

[1]Instituto de Ciências Exatas e Naturais – Universidade Federal do Pará (UFPA)
66075-110 – Belém – PA – Brasil

{paula.brito, samara.souza}@icen.ufpa.br, regicsf@ufpa.br

Abstract. *Extracting actionable insights from textual student feedback is critical for institutional improvement, but manual analysis becomes a cognitive bottleneck when scaled institution-wide semesterly. To address this, we propose an automated pipeline integrating topic modeling and Generative AI for theme extraction and semantic labeling. We compare LSA, LDA, NMF, and BERTopic, evaluating the Meta-Llama-3.1-8B-Instruct model for automated semantic labeling. Results highlight BERTopic's capacity to capture exclusive latent dimensions, alongside Llama's interpretability in generating contextually appropriate labels. This approach offers a viable pathway to convert unstructured feedback into strategic intelligence, potentially facilitating educational management.*

1. Introduction

One of the biggest challenges in academic management is converting institutional evaluation data into effective improvements. As noted by [Vieira de Sousa et al. 2024], reports frequently assume a bureaucratic nature, generating data volumes hard to interpret strategically. In Brazil, the National System of Higher Education Assessment (SINAES) governs the evaluation process based on three pillars: institutional evaluation, and course and student performance assessment. In this context, the Federal University of Pará (UFPA) conducts the AVALIA UFPA, a semesterly self-assessment of its undergraduate courses through an online questionnaire that combines closed-ended Likert scale questions with optional open-ended comments. Despite generating quantitative reports, students' textual feedback remains underutilized. This analytical gap persists not because manual reading is entirely impossible, but because it demands a cognitive and temporal effort that academic managers cannot consistently sustain semesterly across all courses and campuses of a state-wide university, creating an insurmountable managerial bottleneck.

To extract knowledge from this corpus, Natural Language Processing (NLP) emerges as a viable solution. In this context, topic modeling allows grouping documents and discovering latent structures without prior labeling, based on the co-occurrence of terms [Hammami and Faiz 2022]. However, the topic modeling algorithms output isolated keywords that often lack clear semantic interpretability, demanding highly subjective human intervention that is prone to cognitive biases and misinterpretations [Rijcken et al. 2023]. To overcome these barriers, recent studies demonstrate that Large Language Models (LLM) can effectively distill these complex word sets into concise thematic summaries [Piper and Wu 2025]. Aiming to mitigate the underutilization of the student comments, the main goal of this paper is to evaluate whether integrating topic

modeling with Generative Artificial Intelligence (GenAI)-based automatic labeling produces interpretable semantic diagnoses suitable for supporting institutional analysis.

This research presents three main contributions: (1) a hybrid pipeline integrating topic modeling with LLM-based generative labeling applied to student feedback; (2) a comparative analysis of Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and BERTopic for topic modeling, conducted through a graph-based methodology using embeddings, cosine similarity and Louvain algorithm to map exclusive thematic niches; and (3) a methodological framework designed to support the conversion of unstructured feedback into actionable strategic intelligence. Ultimately, this approach seeks to equip managers and teachers with data-driven insights, filling a regional gap in text clustering studies and valuing student perspectives in institutional self-evaluation.

2. Related Work

Recent literature on student feedback analysis highlights a transition from traditional frequency-based NLP models, which disregard word order and semantic context [Dieng et al. 2019], to contextual embeddings. While traditional algorithms often struggle with the ambiguity of short texts, the NMF model frequently emerges as a robust classical baseline, providing superior thematic recall in institutional evaluations [Neha Kardam and Makhsous 2024, Kardam and Wilson 2024]. Concurrently, transformer-based approaches have established a new scenario. Studies demonstrated that BERTopic significantly outperforms classical models, like LDA, LSA and NMF, by capturing dense, fine-grained semantic nuances in both remote learning perceptions [Hayat et al. 2024] and informal social media interactions [Zankadi et al. 2023].

While topic modeling successfully groups texts, the cognitively demanding human intervention required for semantic interpretation remains a critical bottleneck. To overcome this, recent literature has established the integration of LLMs for automatic labeling. Studies demonstrate that coupling advanced prompting techniques, such as Chain of Thought (CoT) and Persona Prompting, with algorithms like BERTopic [Khandelwal 2025, Almeida et al. 2025] and NMF [Wanna et al. 2024] produces labels that achieve high semantic alignment with human experts. Notably, open-weight architecture models have proven highly effective in these tasks, often matching or outperforming proprietary alternatives in complex contextual evaluations.

Despite these algorithmic advancements, current comparative evaluations analyze these algorithms relying on standalone metrics, such as coherence, diversity scores and traditional classification accuracy. The present study differentiates itself by introducing an alternative approach: a similarity graph-based inter-model alignment that systematically maps exclusive semantic niches uniquely captured by the algorithms. This specialized analysis is crucial to verify how each algorithm extracts thematic aspects uniquely tailored to the academic data context. Furthermore, integrating an LLM to interpret these isolated niches aims to eliminate the cognitive bottleneck of time-intensive manual interpretation by academic managers and scale unstructured feedback analysis into actionable institutional intelligence.

3. Methodology

This study employs an adapted five-stage Knowledge Discovery in Databases (KDD) methodology [Fayyad et al. 1996], as shown in Figure 1. This adaptation uses distinct preprocessing and transformation flows to suit different modeling approaches.

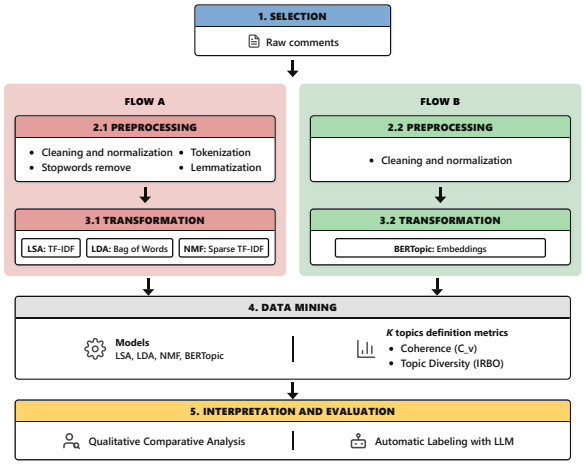


Figure 1. Flowchart of the steps in the adapted KDD methodology.

3.1. Selection

The research data was provided by the institution's Information and Communication Technology Center (CTIC) in Comma-Separated Values (CSV) format, covering the period from 2019 to 2024, except for the years 2020 and 2021, in which the questionnaire was not applied due to the pandemic and remote learning. Aggregating data spanning several years was a deliberate methodological choice to validate the pipeline's behavior across diverse temporal contexts. Nonetheless, in real-world applications, the pipeline would be executed semesterly to provide continuous institutional feedback. Because it contains sensitive information, the data is not publicly available and underwent a de-identification process before being provided for this research, safeguarding the privacy of respondents. The databases contain attributes such as course, academic unit, subject, teacher, schedule, location of classes, answers to closed questions, and open comments, totaling 86,708 records across all databases.

Initially, the database was filtered to restrict the scope to the target campus, academic unit and course, resulting in a corpus with 767 records. This restriction aims to prevent the dilution of specific logistical and pedagogical nuances that can often occur in macro-level analyses, empowering managers with precise, actionable insights. While reading these valid comments is not entirely impractical for a single course, processing them manually every semester demands significant time and effort from managers. Therefore, this dataset acts as a real representative baseline to validate an automated pipeline that can scale efficiently to the entire institution. For the analysis, only identification columns such as year of evaluation, academic unit, course, subject, teacher, and comments were retained. The comments are divided into three categories: positive comments, intended for praise; negative comments, aimed at criticism; and general comments, intended for diverse subjects such as structural and institutional aspects.

3.2. Preprocessing

The preprocessing proceeded with the data concatenation from different years into a single dataset. Subsequently, text cleaning techniques common to all models were applied, including conversion to lowercase characters and noise removal, such as numbers, punctuation, and special characters. Additionally, an anonymization operation was incorporated to remove explicit teachers' names from comments, starting from the "teacher" attribute of the dataset itself. This decision aims to ensure institutional confidentiality and eliminate noise that could bias clustering by individual identities instead of academic themes. Next, for the LSA, LDA, and NMF models, additional textual treatment was applied, including tokenization, stopword removal, and term lemmatization. Finally, null and duplicate records were excluded, resulting in a database with 619 records.

3.3. Transformation

For better data organization, the original columns for positive, negative, and general comments were unified into a single column called "comment". This strategy aimed to mitigate inconsistencies such as the insertion of criticisms in fields intended for praise and vice versa. To preserve the origin of each record after concatenation, the "label" column was created to classify the comment category: 0 for negative comment, 1 for positive, and 2 for general.

To prepare the LSA, LDA, and NMF models, a dictionary of terms was constructed and manually refined to exclude contextual stopwords. From this dictionary, vector representations were specifically tailored for each algorithm. Given its probabilistic nature, LDA utilized the Bag of Words (BoW) matrix to capture absolute term frequencies. In contrast, LSA and NMF adopted the Term Frequency-Inverse Document Frequency (TF-IDF) matrix to enhance semantic separation by weighting strictly informative terms. Additionally, the NMF implementation employed a sparse TF-IDF matrix to optimize memory and computational time during factorization by processing only non-zero values. For BERTopic, the texts were converted into embeddings using the PORTULAN/serafim-335m-portuguese-pt-sentence-encoder-ir transformer [1], chosen for its superior performance in Semantic Text Similarity (STS) and Information Retrieval (IR) in Brazilian Portuguese due to its Sentence Encoder structure, which is capable of capturing the deep context of the documents.

3.4. Data mining

The Mining stage is divided into two complementary subsections: the first one presents and justifies the algorithms selected for topic modeling, while the second one details the strategy for defining the number of topics for the models (k).

3.4.1. Topic modeling

Four algorithms were selected to ensure complementary thematic extraction under distinct premises: LSA [Deerwester et al. 1990] as a frequency-based baseline to map dominant global term co-occurrences across the corpus; LDA [Blei et al. 2003] for its probabilistic representation, suited to comments that simultaneously address multiple themes;

[1] Available on Hugging Face at PORTULAN/serafim-335m.

NMF [Lee and Seung 1999] for its additive factorization, which enforces clear semantic boundaries between themes; and BERTopic [Grootendorst 2022] as the neural counterpart, capturing deep semantic similarity through embeddings to handle the informality and syntactic complexity of student-generated texts. All topic model outputs are available in the project repository [2].

3.4.2. Defining k topics number

In topic modeling, defining the number of topics (k) to be extracted by the models determines the analysis granularity: low values generate generalist themes, while high values excessively fragment the subjects. Although related studies employ metrics that aim to mitigate the subjectivity of this choice, they do not always guarantee the quality of the topics generated, since high statistical consistency is not necessarily synonymous with high human interpretability [Chang et al. 2009].

For this research, a hybrid strategy was adopted to define k for each model. Initially, the coherence (C_v) metric was used, which evaluates semantic interpretability by measuring the degree of confirmation between the most representative words of each grouping [Röder et al. 2015]. In addition, the Inverted Rank-Biased Overlap (IRBO), an adaptation of the Rank-Biased Overlap (RBO) metric [Webber et al. 2010], was applied to assess the distinctness and non-redundance among the generated themes [Bianchi et al. 2021]. This combined analysis, along with qualitative human inspection, aimed to ensure the relevance and interpretability of the results.

3.5. Interpretation and Evaluation

3.5.1. Qualitative inter-model analysis

The qualitative comparative analysis consisted of an inter-model alignment aimed at mapping exclusive topics, which, in the context of this research, are defined as unique clusters captured by a single technique. To achieve this, we adopted a hybrid approach based on converting topic keywords into embeddings to form an all-to-all cosine similarity matrix, from which an undirected similarity graph was derived using a semantic threshold. Subsequently, by applying the Louvain algorithm for community detection, we were able to systematically filter out shared thematic structures, leaving only the exclusive niches for this analysis.

The steps of the process were as follows: **(1) Vector Representation**: the 15 keywords from each topic were converted into embeddings via the PORTULAN/serafim-335m model, calculating the topic centroid (vector mean). The limit of 15 keywords was empirically selected to balance topic representativeness with the mitigation of semantic noise, preserving the semantic core of each cluster and avoiding the inclusion of low-relevance terms. **(2) Similarity Matrix**: all-to-all cosine similarity was applied to quantify semantic proximities and latent inter-model connections based on the cosine of the angle formed between them [Salton et al. 1975]; **(3) Cutoff Threshold**: a sensitivity analysis was performed between 0.8 and 0.95, in which experiments indicated that thresholds close to 0.8 merged distinct topics, while values close to 0.95 fragmented equivalent

[2] All model outputs are available in the project repository.

themes. Thus, the threshold was set at 0.94 to ensure semantic cohesion and correct identification of analogous topics; **(4) Similarity Graph**: combining the cross-model alignment principles from [Kafiabad et al. 2025] with the network analysis methodologies proposed by [Myftaraj et al. 2025], an undirected graph was constructed, where edges connect solely nodes (topics) whose similarity exceeds the defined threshold; **(5) Community Detection**: the Louvain algorithm [Blondel et al. 2008] was applied to group densely connected nodes by iteratively optimizing the network's modularity. In the context of topic modeling, Louvain's primary advantage is its resolution parameter, which allows precise control over community granularity. As demonstrated by [Myftaraj et al. 2025], lower values merge topics into broad clusters, while higher values split them into finer subgroups. By leveraging this parameter, the algorithm functioned systematically as a filter to identify and exclude shared cross-model convergences, thereby strictly isolating the exclusive thematic niches.

3.5.2. Automatic topic labeling

To mitigate the subjectivity and cognitive load inherent in the manual interpretation of topics generated by modeling algorithms, an automatic labeling pipeline based on GenAI was implemented. For this task, the pre-trained Meta-Llama-3.1-8B-Instruct model was used. Justified by its open-weight architecture, the 8-billion-parameter model offers interpretive capabilities and viable computational costs. Specifically, the GPT-Generated Unified Format (GGUF) quantization technique was applied to reduce the precision of the model's weights, decreasing its memory footprint and enabling execution on local CPU hardware without the need for expensive dedicated GPUs. Moreover, the Instruct variant meets complex prompt constraints and offers robust training and consolidated support for the specificities of Brazilian Portuguese language.

The automated labeling process begins with the data entry stage, where keywords and representative documents are incorporated, followed by prompt engineering techniques including persona definition, system instructions, and a zero-shot approach. To ensure reproducible results, the model parameters were configured with a 4.096-token context window (n_{ctx}), random seed ($seed$=42), temperature (T=0.1), top-p (p=0.9), and a 50-token generation limit (max_tokens). The complete prompt structure is available in the project's repository on GitHub. Finally, the model processed the provided context to generate a short descriptive label.

4. Results and discussions

The presentation of the results starts with the definition of the number of topics (k), followed by the inter-model comparative analysis and concluding with automatic semantic labeling via LLM. Since the original dataset consists of texts in Portuguese, all keywords and labels in this section have been translated into English to facilitate understanding.

4.1. Topic granularity (k) definition

Initially, the search range was limited to $k \in [2, 25]$, seeking a balance between granularity and the practical usefulness of the topics. Empirically, the defined limit of $k = 25$ aims to fit the analyzed corpus, preventing an excessive number from fragmenting similar themes into multiples topics with subtle differences. This restriction aligns with the

central objective of producing interpretable semantic diagnoses to support institutional analysis, since an excessive number of groupings would increase the cognitive load on managers during the analysis, compromising the synthesis provided by the modeling.

Preliminary coherence analysis in Figure 2 showed that the LSA, LDA, and BERTopic models tended to maximize their scores at very low granularity ($k \in \{2, 3\}$). However, such general groupings, although mathematically cohesive, offer low managerial utility because they do not discriminate specific operational nuances. Given this limitation, the IRBO diversity metric was used as a criterion to explore higher granularities.

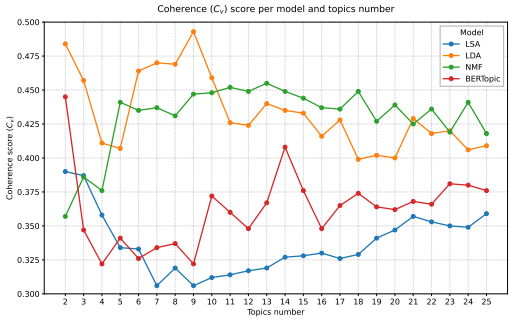


Figure 2. Coherence scores for different values of k per model.

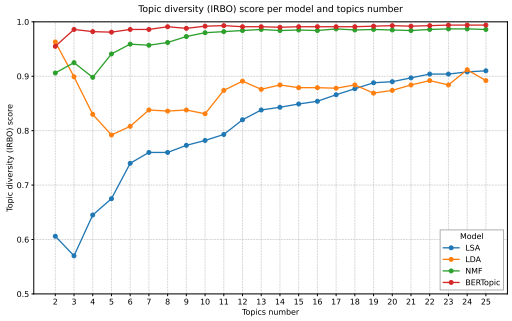


Figure 3. IRBO scores for different values of k per model.

Figure 3 shows that the four models exhibit stability in the metric at $k = 13$, indicating that the increase in the number of topics did not result in semantic redundancy detectable by the metric. Based on the scores obtained in the two metrics, the window of analytical interest was defined as the interval $13 \leq k \leq 25$. Within the established window, the value of k that maximized coherence was selected, respecting the specificities of each algorithm. The results indicated $k = 25$ as the ideal point for the LSA model, $k = 13$ for LDA and NMF, and $k = 14$ for BERTopic.

4.2. Qualitative inter-model analysis

In this research, exclusive topics are defined as clusters that did not show cosine similarity equal to or above the threshold of 0.94 with any other topic generated by the other models. The objective of this analysis is to verify if the models can extract thematic niches ignored by the other algorithms. Although they have lower recurrence in the corpus, these exclusive topics can reveal critical issues that require managerial attention and immediate action. Table 1 presents the exclusive topics found solely by NMF and BERTopic.

The critical highlight found in the analysis was BERTopic's unique ability to identify institutional risks by extracting terms such as "offensive", "unnecessary" (10), "sex" and "sexual" (12). By grouping terms by semantic proximity, the model prevented sensitive comments, infrequent in the dataset, from being overlooked. In models that rely on the frequency and co-occurrence of terms, these rare events were likely diluted into generic topics of dissatisfaction or ignored, given their low statistical representativeness.

Furthermore, BERTopic extracted niches of dissatisfaction related to the teaching profile (4) and external factors that reflect the student's daily life (2), revealing logistical and routine concerns that affect the academic experience. In topic (9), the term

Table 1. Exclusive topics extracted by NMF and BERTopic.

Model	Keywords
NMF	(8) teach, bad, weak, possess, attention, objective, rushed, incredible, point, objective, congrats, approach, terrible, patience, approach
	(12) miss, class, room, communication, experience, warn, punctual, inside, seen, of that, hardly, present, questions, personal, commitment
BERTopic	(2) schedule, times, reasons, some, time, many, too much, missed, fast, bus, night, work, many, punctual, are
	(4) didactics, little, excellent, uncompromising, pretends, didactic, disorganized, terrible, points, extremely, knows, quite, unrealistic, right, explosive
	(6) evaluation, evaluations, could, correction, clear, corrected, included, discursive, method, confusing, only, left, in, tests, forms
	(8) nothing, joined, recorded, declare, arrives, nor, occurred, early, error, social, name, moment, evaluate, did, grade
	(9) professional, ok, impeccable, qualified, fun, xxxxxxxxxxx, regular, wonderful, wonderful, present, objective, name, sir, good, horrible
	(10) comments, offensive, comment, read, comment, will be, positive, negative, unnecessary, never, other, so much, nothing, hour, “”
	(11) improve, needs, leave, didactics, experience, interest, love, attract, gain, field, way, interesting, evolve, make an effort, content-focused
	(12) sex, comments, sexual, ex, sexual, unnecessary, discussed, does, about, abuse, about, certain, approach, demanding, individual
	(13) slides, portuguese, complete, tasks, mastery, we explained, influenced, portugol, transport, portuguese, send, we would spend, delivered, correct, asked

Note: The numbers in parentheses represent the topic ID.

“xxxxxxxxxxx” is a literal student comment rather than a placeholder, while the duplicate keyword “wonderful” results from translating Portuguese gendered variants (“maravilhoso” and “maravilhosa”). Finally, the presence of empty strings in (10) is an expected behavior resulting from the exhaustion of statistically relevant terms after stopword removal combined with the implemented lexical diversity technique.

On the other hand, NMF presented two unique topics offering complementary analyses. Terms such as “teach”, “bad”, and “incredible” (8), reflect aspects related to didactic approach and teaching quality, which also mixes the polarity of feelings. This characteristic of grouping opposite polarities may be a common property in models not explicitly trained for sentiment analysis. In its second topic (12), the algorithm points to attendance and teacher communication factors, crucial for academic management.

In summary, it is noted that BERTopic presented more topics addressing rare and critical events, such as sensitive comments, which are extremely important for institutional management. Its embedding-based approach allows low-frequency terms not to be diluted among the other groupings. For management use, the embedding-based model proves suitable for highlighting comments that address sensitive topics and inappropriate professional behaviors without them being masked by the large volume of praise and criticism present in the corpus.

4.3. Automatic generative topic labeling

Previous discussions have shown that although topic modeling algorithms automate text clustering, the semantic interpretation of these groups remains dependent on human curation. Therefore, this subsection employs automatic generative labeling with LLM as a support for the interpretability of generated topics, aiming to mitigate the overload of manual inspection and provide diagnoses that support evaluation and decision-making by academic managers. For this stage, rather than conducting an inter-model comparative analysis of the generative labeling, only the topics generated by BERTopic were selected

for this task. This algorithm was chosen because it yielded the most interpretable results, effectively balancing semantic coherence and topic diversity while isolating critical institutional niches.

The Llama-3.1 labeling prompt was structured to maximize interpretive capacity by incorporating two input variables: **i) Keywords**: top 15 terms representative of the extracted topic that synthesize the central content and act as the semantic core of the grouping; and **ii) Representative Documents**: top 3 comments that, in the context of the research, best exemplify the extracted topic. The inclusion of these representative documents aims to mitigate ambiguities that isolated keywords would not resolve.

While a formal human-annotated validation was not conducted in this phase, the qualitative observation of the results in Table 2 demonstrates the face validity and semantic plausibility of the LLM-generated labels. Rather than simply outputting disjointed terms, the model synthesized the isolated niches into actionable categories, showing logical alignment with the extracted keywords. In particular, a critical highlight is that the labels for topics (10) and (12) demonstrate the model's sensitivity, inferential capacity, and fidelity to the data. Instead of hallucinating severe legal accusations (e.g., harassment or crime) without explicit context, the LLM provides safe and interpretable labels, even when sensitive terms are highlighted among the keywords. This behavior is crucial to prevent the automated classification of serious offenses without factual support, thereby ensuring an ethical and reliable institutional analysis.

Table 2. LLM-generated semantic labels for the identified topics.

LLM-Generated Label	Keywords
Excellent Professor	(0) excellent, always, great, well, good, didactics, great, good, doubts, subject, explains, content, about, super, objective
Inefficient and Disrespectful Professor With Students	(1) content, test, room, evaluation, besides, tests, passed, teaching, lack, little, only, assignments, didactics, time, times
Infrastructure Problems and Excessive Classes	(2) schedule, times, reasons, some, time, many, too much, missed, fast, bus, night, work, many, punctual, are
Subject Mastery is Well Taught	(3) mastery, subject, content, quite, day, about, well, passes, knowledge, taught, area, subjects, new, contextualized, great
Extremely Disorganized Didactics	(4) didactics, little, excellent, uncompromising, pretends, didactic, disorganized, terrible, points, extremely, knows, quite, unrealistic, right, explosive
Complicated Semester due to Lack of Communication and Strike	(5) disciplines, some, good, difficulties, despite, strike, profitable, honestly, great, subjects, pass on, after, time, general, unfortunately
Inadequate Evaluation Method	(6) evaluation, evaluations, could, correction, clear, corrected, included, discursive, method, con
Willing and Attentive Professor	(7) always, attentive, willing, solicitous, humble, hardworking, congratulations, helpful, help, well, understood, engaged, disciplined, highly, arbitrary
Dissatisfaction with Absence of Information	(8) nothing, joined, recorded, declare, arrives, nor, occurred, early, error, social, name, moment, evaluate, did, grade
Very Fun and Impeccable Class	(9) professional, ok, impeccable, qualified, fun, xxxxxxxxxxxx, regular, wonderful, wonderful, present, objective, name, sir, good, horrible
Unnecessary and Offensive Comments	(10) comments, offensive, comment, read, comment, will be, positive, negative, unnecessary, never, other, so much, nothing, hour, "
Didactics Needs Improvement	(11) improve, needs, leave, didactics, experience, interest, love, attract, gain, field, way, interesting, evolve, make an effort, content-focused
Unnecessary Sexual Comments in the Classroom	(12) sex, comments, sexual, ex, sexual, unnecessary, discussed, does, about, abuse, about, certain, approach, demanding, individual
Professor Inefficient in Didactics and Content	(13) slides, portuguese, complete, tasks, mastery, we explained, influenced, portugol, transport, portuguese, send, we would spend, delivered, correct, asked

Note: The numbers in parentheses represent the topic ID.

Furthermore, the generated labels effectively differentiate the subtle nuances of student feedback. In topic (5), the LLM captured the interaction between "difficulties", "strike", and "time", forming a situational label. Similarly, for topic (8), the model interpreted the noise of comments like "nothing to declare" captured during the modeling step. The labels also distinguish between specific behavioral issues, such as lack of respect (1), external infrastructure and logistics problems (2), pedagogical flaws like disorganized (4) or inefficient (11 and 13) teaching methods, and opposition to the evaluation method (6). Likewise, praises were logically categorized into clear axes regarding teaching methods (0), subject mastery (3), teacher profile (7), and class dynamics (9). This level of descriptive suggests that feeding the LLM with keywords alongside representative documents enables it to successfully interpret the intentionality behind the students' feedback, covering the full spectrum of evaluation. Consequently, this approach provides a robust heuristic baseline that significantly reduces the cognitive load of academic managers when analyzing unstructured text.

5. Considerations

This study explored the practical utility of integrating topic modeling with LLM-based generative labeling for the analysis of unstructured student feedback. The inter-model comparative analysis, driven by a similarity graph-based methodology, facilitated the mapping of exclusive thematic niches. The results suggest that BERTopic is particularly suited to extracting cohesive clusters and detecting critical institutional risks, such as unethical conduct and sexual comments. These rare but severe events are typically diluted by frequency-based approaches due to statistical noise and lexical overlap. Nevertheless, among the classical algorithms, NMF showed a strong tendency to capture specific qualitative and behavioral aspects of teaching methodologies.

Regarding the automatic labeling task, the Llama-3.1 model went beyond basic thematic summarization by exhibiting strict ethical and contextual sensitivity. Processing representative documents alongside topic keywords allowed the LLM to provide plausible interpretations of the intentionality behind student comments, even in the presence of noise, while avoiding premature or hallucinated legal terminology when handling sensitive subjects. From a practical standpoint, this hybrid pipeline provides a baseline that can significantly reduce the cognitive load of academic managers. Ultimately, this approach presents the potential to transform bureaucratic institutional self-evaluation into a continuous improvement engine, effectively valuing student perspectives.

Despite its contributions, this study acknowledges certain limitations. The computational cost of running neural models and the absence of a human-annotated ground truth restrict the quantitative evaluation of both the thematic clustering phase and the generated labels' accuracy. To address this, future work should include a structured validation phase where academic managers rate a sample of the generated labels for relevance and semantic alignment. Additionally, exploring Retrieval-Augmented Generation (RAG) architectures is suggested to create interactive virtual assistants, thereby democratizing access to institutional intelligence through natural language queries.

References

Almeida, T., Barbosa, K., Fernandes, S., and Aquino, A. (2025). Topic modeling in feminist debates on instagram: A generative ai approach. In *Anais do XIV Brazilian*

Workshop on Social Network Analysis and Mining, pages 173–186, Porto Alegre, RS, Brasil. SBC.

Bianchi, F., Terragni, S., and Hovy, D. (2021). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading tea leaves: how humans interpret topic models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, NIPS'09, page 288–296, Red Hook, NY, USA. Curran Associates Inc.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2019). Topic modeling in embedding spaces.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Mag.*, 17(3):37–54.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Hammami, E. and Faiz, R. (2022). Text clustering based on multi-view representations. In *CIRCLE*.

Hayat, F., Shatnawi, S., and Haig, E. (2024). Comparative analysis of topic modelling approaches on student feedback. In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR*, pages 226–233. INSTICC, SciTePress.

Kafiabad, S. T., Schiffauerova, A., and Ebadi, A. (2025). Decoding funded research: Comparative analysis of topic models and uncovering the effect of gender and geographic location.

Kardam, N. and Wilson, D. (2024). A hybrid approach to natural language processing for analyzing student feedback about faculty support. In *2024 ASEE Annual Conference & Exposition*, number 10.18260/1-2–46447, Portland, Oregon. ASEE Conferences. https://peer.asee.org/46447.

Khandelwal, T. (2025). Using llm-based approaches to enhance and automate topic labeling.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791.

Myftaraj, E., Fata, I., and Plasari, E. (2025). A topic modeling analysis of circular economy and big data research using bertopic and scibert. *International Journal of Innovative Technology and Interdisciplinary Sciences*, 8(4):1195–1243.

Neha Kardam, D. W. and Makhsous, S. (2024). A comparative analysis of natural language processing techniques for analyzing student feedback about ta support. In *2024 ASEE Annual Conference & Exposition*, number 10.18260/1-2–46425, Portland, Oregon. ASEE Conferences. https://peer.asee.org/46425.

Piper, A. and Wu, S. (2025). Evaluating large language models for narrative topic labeling. In Hämäläinen, M., Öhman, E., Bizzoni, Y., Miyagawa, S., and Alnajjar, K., editors, *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 281–291, Albuquerque, USA. Association for Computational Linguistics.

Rijcken, E., Scheepers, F., Zervanou, K., Spruit, M., Mosteiro, P., and Kaymak, U. (2023). Towards interpreting topic models with chatgpt. The 20th World Congress of the International Fuzzy Systems Association, IFSA ; Conference date: 20-08-2023 Through 24-08-2023.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Vieira de Sousa, J., Maffini Griboski, C., and Woiciechowski Batista Gumesson, D. (2024). Autoavaliação institucional no marco do sinaes: movimentos e tensões no uso dos resultados para a qualidade da educação superior : Institutional self-assessment in the sinaes framework: movements and tensions in the use of results for the quality of higher education. *Revista Cocar*, (29).

Wanna, S., Barron, R., Solovyev, N., Eren, M. E., Bhattarai, M., Rasmussen, K., and Alexandrov, B. S. (2024). Topictag: Automatic annotation of nmf topic models using chain of thought and prompt tuning with llms.

Webber, W., Moffat, A., and Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4).

Zankadi, H., Idrissi, A., Daoudi, N., and Hilal, I. (2023). Identifying learners' topical interests from social media content to enrich their course preferences in moocs using topic modeling and nlp techniques. *Education and information technologies*, 28(5):5567–5584.