Annales Mathematicae et Informaticae
63 (2026) pp. 42–54
DOI: 10.33039/ami.2026.07.002
URL: https://ami.uni-eszterhazy.hu

Student opinion mining: Automated topic extraction from student feedback*

Olivér Czimbalmos[a], Zsolt Szántó[a], Gábor Kőrösi[a], Péter Becsei[a], Beáta Udvari[b], Richárd Farkas[a]

[a]Institue of Informatics, University of Szeged
{czimbalmos,szantozs,korosig,becsei,rfarkas}@inf.u-szeged.hu

[b]Directorate of Academic Affairs, University of Szeged
udvari.beata@szte.hu

Abstract. While Student Evaluation of Teaching Work (SETW) is a cornerstone of quality assurance in higher education, the high volume and linguistic complexity of unstructured qualitative feedback often lead to its underutilization in institutional decision-making. This study addresses this "analysis gap" by developing an automated pipeline to process 34,000 unique Hungarian student responses from a major research university. To empower academic administration and faculty leadership with the ability to uncover latent thematic patterns within these responses, we propose a hybrid NLP framework that utilizes a Large Language Model (LLM) for thematic reclassification and Aspect-Based Sentiment Analysis (ABSA), combined with an unsupervised layer for fine-grained latent topic discovery using transformer-based embeddings and HDBSCAN clustering. Our proposed pipeline successfully identified new granular latent topics – such as lecture pacing, traceability, material accessibility and slides – providing a level of diagnostic detail that remains invisible to standard quantitative metrics. The results prove that modern natural language processing (NLP) techniques can effectively transform raw, unstructured student narratives into objective, actionable diagnostic tools.

Keywords: student feedback, large language models, natural language processing

*This work was supported by project 2024-1.2.3-HU-RIZONT-2024-00017, implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the 2024-1.2.3-HU-RIZONT funding scheme.

Submitted: March 31, 2026
Accepted: July 3, 2026
Published online: July 17, 2026

1. Introduction

University quality assurance mainly relies on student feedback, containing crucial information about the institution's performance. Student Evaluation of Teaching Work (SETW) provides universities with a systematic mechanism to assess pedagogical effectiveness and identify areas for improvement. While quantitative Likert-scale ratings offer a convenient numerical summary of student satisfaction, their interpretability is inherently limited. The most informative insights are often embedded in textual comments, where students articulate the underlying reasons behind their evaluations. These qualitative responses reveal nuanced perspectives on teaching practices, course organization, and learning environments that remain invisible in purely quantitative assessments. Despite their potential value, extracting actionable insights from large volumes of textual feedback remains a persistent challenge for universities. Large institutions frequently collect tens of thousands of individual responses each semester, making manual analysis both impractical and prone to subjective interpretation. As a result, a substantial portion of qualitative student feedback remains underutilized in institutional decision-making processes. This problem becomes even more pronounced in multilingual educational environments. In the Hungarian context, for example, the morphological complexity of the language significantly increases the difficulty of traditional natural language processing approaches, while institutional differences in evaluation frameworks across faculties further complicate the systematic interpretation of feedback.

Previous research has explored automated methods for analyzing student feedback, including sentiment analysis and topic modeling [5, 10, 13]. While these approaches have demonstrated promising results, they often face limitations when applied to large multilingual datasets.

To address these challenges, this study presents a large-scale case study of automated qualitative feedback analysis conducted at the University of Szeged, involving more than 34,000 unique student responses. Our research identifies previously unrecognized thematic patterns in texts, hence it transforms unstructured student narratives into structured, interpretable insights. Concurrently, this work addresses the following central research question: *How effectively can modern NLP techniques, including Large Language Models and transformer-based embeddings, automatically extract meaningful topics and sentiments from large-scale student feedback in higher education?*

To achieve this goal, we propose a hybrid Natural Language Processing (NLP) pipeline (see Figure 1) that combines large language model–based classification with modern embedding techniques and density-based clustering methods. The framework first harmonizes heterogeneous faculty-level feedback into the new institutional categories while simultaneously identifying sentiment expressed toward specific aspects of teaching. Subsequently, multilingual text embeddings are created to capture semantic similarities between responses, enabling the discovery of latent thematic clusters that reveal recurring student concerns. These clusters then recive an appropiate name based on their text such as **lecture pacing, course**

organization, or material accessibility and slides. To furthure enhance the preocessing the results are summerized in a interactive dashboard.

The primary contribution of this work is the demonstration of a scalable methodology for transforming large-scale qualitative student feedback into actionable institutional knowledge. By integrating automated categorization, sentiment analysis, and unsupervised topic discovery within a single analytical framework, the proposed approach enables universities to move beyond aggregate satisfaction scores and uncover the underlying drivers of student experience. The results illustrate how modern opinion mining techniques can support evidence-based educational governance and help institutions more effectively respond to student needs.

We also contribute to the Hungarian NLP research as we evaluate the effectiveness of open Large Language Models in processing Hungarian-language student feedback, focusing on few-shot learning for both category classification and aspect-based sentiment analysis, and conduct a systematic comparison of state-of-the-art multilingual embedding models to assess their ability to capture semantic structure in Hungarian textual data.

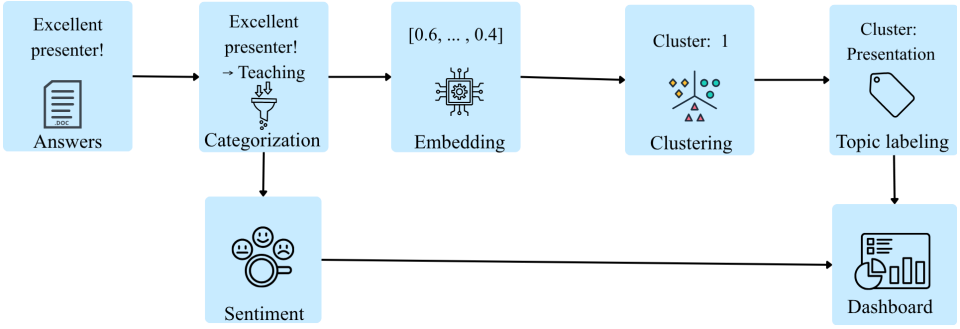

Figure 1. Schematic view of our proposed pipeline.

1.1. Related work

The automated analysis of student feedback through NLP has become an increasingly prominent field of study [1]. In the international literature, various methodologies have been proposed for thematic discovery and sentiment detection within educational datasets. A widely adopted approach is the application of Latent Dirichlet Allocation (LDA) for topic modeling. For instance, Sun, Yan (2023) utilized LDA to extract key themes related to the quality of teaching and learning from student narratives [13]. Beyond internal pedagogical assessment, student reviews from online platforms have also been utilized for institutional SWOT analyses. In this context, Srinivas, Rajendran demonstrated a framework that combines LDA-based topic modeling with rule-based sentiment detection to identify organizational strengths and weaknesses [12]. These foundational attempts at thematic discovery underscore the potential of automated qualitative analysis, which our

research seeks to refine using more contextually aware semantic models.

The rise of Massive Open Online Courses (MOOCs) has further catalyzed the development of more complex student feedback analysis. To determine the polarity of large-scale feedback, Baqach, Battou (2024) implemented a pipeline using BERT-based embeddings coupled with deep neural networks, incorporating both convolutional and dense layers for classification [3]. More recent comparative research has shifted focus toward the efficacy of different embedding strategies. Specifically, it has been observed that high-parameter foundation models and their corresponding embeddings can provide superior semantic representation in clustering tasks when compared to traditional BERT-scale architectures [6]. Our methodology builds upon this shift toward high-parameter embedder models, utilizing their deep semantic representations to enhance the granularity of student feedback clusters.

Paralleling this shift, Mersha, Kalita, et al. (2024) established an end-to-end semantic-driven topic modeling framework that utilizes Sentence-BERT representations combined with UMAP dimensionality reduction and HDBSCAN density clustering to isolate coherent thematic regions from noisy text corpora [8]. In the Hungarian higher education landscape, the quantitative results of Student Evaluation of Teaching Work (SETW) are often made publicly available by institutions such as the University of Veterinary Medicine Budapest and the University of Szeged. However, the systematic application of NLP to the qualitative, textual portions of these surveys remains less documented in the academic literature. Nevertheless, data mining and opinion detection have been applied to other Hungarian-language domains. Osváth, Yang Zijian, Kósa (2022) explored patient experience narratives, utilizing Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction and subsequent topic modeling [10]. Similar methodologies have appeared in the clinical sector as well; for example, Hinek (2024) utilized LDA to identify key thematic dimensions within Hungarian-language hotel guest reviews [5]. These studies highlight the ongoing transition from traditional statistical methods toward context-aware semantic analysis in the processing of Hungarian textual data. By situating our research within this evolving Hungarian landscape, we aim to extend these localized NLP successes into the specific domain of educational evaluation using more advanced generative architectures.

Methodologically, our research is most closely aligned with the framework presented by Osváth, Yang Zijian, Kósa (2022). Unlike their approach, we utilize Large Language Models (LLMs) for automated topic labeling, leveraging the extensive world knowledge inherent in these foundation models to generate descriptive and contextually accurate titles for the emergent clusters. Furthermore, we implement LLM-based Aspect-Based Sentiment Analysis (ABSA), allowing for a more nuanced evaluation of student feedback compared to traditional sentiment detection methods.

2. Methodology

2.1. Dataset

The dataset utilized in this study was provided by the Directorate of Academic Affairs at the University of Szeged. The corpus comprises student responses from two distinct faculties (Faculty 1 and Faculty 2) spanning three academic semesters: the full 2023/2024 academic year and the first semester of 2024/2025. All data were anonymized at the database level to ensure the privacy of both instructors and students before being released for research purposes.

The dataset is primarily composed of Hungarian textual feedback, though it also contains a small proportion of English and French responses. This multilingual characteristic necessitated the use of a multilingual embedding model in the subsequent stages of the pipeline [14]. The questionnaires used by the two faculties differed significantly in their qualitative framing:

- Faculty 1 employed multiple specific, instructor-focused questions (e.g., **"To what extent was the instructor's delivery conducive to note-taking?"**).
- Faculty 2 used a broader, more general approach, including self-reflective prompts (e.g., **"Self-reflection: What have you done to achieve the best possible performance in this course?"**).

An analysis of participation revealed a participation paradox: Faculty 2, despite offering fewer open-ended questions, generated a higher absolute volume of responses than Faculty 1. However, responses from Faculty 1 were generally longer and more detailed, with an average token count of 17.9 compared to 15.8 for Faculty 2.

The raw dataset underwent an initial preprocessing phase to ensure data quality. This included the removal of duplicate entries and responses consisting solely of non-alphanumeric characters (e.g., punctuation marks or symbols). Following these steps, the final corpus for analysis contained 34,000 unique responses with a vocabulary size of 42,809 tokens. For the statistics summary, see Table 1.

Table 1. Descriptive Statistics of the SETW Corpus.

Metric	Faculty 1	Faculty 2	Token
Number of unique responses	15,688	18,313	34,000
Mean token count	17,9	15,8	16
Median token count	11	13	12
Vocabulary size (Tokens)	27,225	26,763	42,809

From this set of answers 20 samples were selected randomly -a answer could only be picked once- from each question as a result we gathered 200 responses.

From these answers we have annotated the samples into the 6 given categories and also assigned the sentiment values for each category. There were two annotators, who used annotation guidelines provided from the Academic Affairs office[1]. The initial inter annotator agreement – averaged Cohen's kappa for each class – came out as 0.96 and for the aspect based sentiment annotation as 0.94. After each annotator finished they have discuissed the cases of disagreement and assigned the final value.

2.2. Categorization and aspect-based sentiment analysis

To address the structural heterogeneity of the faculty-level questionnaires, we implemented a standardized classification layer based on a unified framework of six core dimensions: **teachers preparation, safe study environment, requirements, teaching, teachers availability and methodology** (plus a "neutral" category). These dimensions are strategically designed to serve as the structural framework for the university's future standardized evaluation system. Given that the research corpus consists of legacy data with non-uniform qualitative framing, the necessity to reconcile historical feedback with these upcoming administrative requirements motivated the integration of this classification layer into our pipeline.

For the categorization and aspect-based sentiment analysis, we have compared three LLMs Llama-3-70B-Instruct, Qwen2-72B-Instruct, DeepSeek-llm-67b-chat [2, 11]. A critical feature of our pipeline is the integration of Aspect-Based Sentiment Analysis (ABSA). Rather than assigning a single global sentiment score to an entire response, the model identifies the specific polarity – **Positive, Neutral, or Negative** – associated with each identified category within the text.

We employed an in-context learning strategy with few-shot prompting, providing the model with category definitions and three representative examples for both classification and aspect-level sentiment detection[2]. To validate the system's performance, we established an evaluation protocol using a gold standard dataset of 200 manually annotated examples.

The vectorization, clustering steps were performed for each category separately at the review level.

2.3. Unsupervised latent topic discovery

As one of our main objectives is to discover new actionable topics, we have implemented the following steps into the pipeline.

2.3.1. Vectorization

To enable unsupervised clustering, the unstructured textual responses were transformed into high-dimensional dense vectors. In our experimental setup, we evaluated and compared two state-of-the-art multilingual transformer-based architec-

[1]Annotation guidelines: https://github.com/szegedai/StudendOpinionMining
[2]Prompt template: https://github.com/szegedai/StudendOpinionMining

tures: **Jina-v3**, drawing on the recent findings of Csányi et al. (2024), and the **Qwen3-Embedding-8B** model [4, 14].

These specific architectures were selected for their ability to handle the morphological complexity of the Hungarian language while providing a unified semantic space. This ensures a consistent and comparable representation across the multilingual (Hungarian, English, and French) corpus.

2.3.2. Topic discovery

For the discovery of latent topics, we employed Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [7]. Unlike centroid-based methods such as K-means, HDBSCAN does not require a pre-defined number of clusters and is capable of identifying groups of varying shapes and densities. Crucially, the algorithm explicitly identifies "noise" – responses that do not align with any distinct thematic cluster – thereby ensuring that the resulting topics remain semantically coherent. To optimize the clustering results, we conducted a random search for the minimum cluster size parameter, defined within the range of $[2, \sqrt{N}]$, where N represents the sample size of the given aspect. The clustering quality was validated using a density-based validity index [9], which measures the stability and separation of the identified themes.

2.3.3. Automated topic labeling

To ensure the administrative utility of the clusters, we implemented an automated labeling procedure using a LLM. The model was tasked with generating concise, descriptive titles for each cluster by analyzing a subset of representative responses. During the development of this procedure, we compared two distinct sampling strategies for providing context to the LLM. Centroid-based sampling: Selecting the ten responses closest to the cluster centroid in Euclidean distance. Random sampling: Selecting ten responses at random from within the cluster.

3. Results and discussion

3.1. Categorization

The effectiveness of the few-shot prompting strategy was validated on a manually annotated gold standard dataset of 200 examples. As summarized in Table 2, the Llama-3 model outperformed its counterparts, achieving a weighted $F1$ -score of 0.72. A notable observation during the benchmarking process was the model's sensitivity to prompt length; while few-shot examples were critical for performance, an excessive number of examples in the prompt led to a gradual degradation in accuracy, particularly for shorter student responses.

Following validation, the model was deployed to classify the full corpus of 34,000 unique responses. The distribution of labels revealed a significant class imbalance, reflecting the primary concerns of the student body. The Methodology category

Table 2. Classification performance comparison across models ($N = 200$).

Model	Weighted F1
Llama-3-70B-Instruct	**0.72**
Qwen2-72B-Instruct	0.60
DeepSeek-llm-67b-chat	0.58

emerged as the most prominent dimension, particularly within Faculty 1, accounting for nearly half of the thematic assignments.

In order to further show the result of the classification Table 3 shows each classes F1 score and their support.

Table 3. LLama F1 scores across classes and their support.

Class	$F1$ score	Support
Methodology	0.78	108
Teachers preparation	0.73	90
Save Study Environment	0.60	19
Teacher Accesibility	0.0	5
Requirements	0.76	12
Teaching	0.66	6
Neutral	0.68	37

3.2. Aspect-based sentiment analysis

The evaluation of the sentiment analysis component focused on the model's ability to assign correct emotional valence to specific classes. As with the classification task, performance was measured against the gold standard dataset of 200 manually annotated examples. The results, summarized in Table 4, show that the **Llama-3-70B-Instruct** model achieved a $F1$-score of **0.82**, outperforming both Qwen2 and DeepSeek. Table 5 summerizes the F1 scores achived with the Llama model for each sentiment level.

Table 4. Model performance comparison on aspect-based sentiment analysis ($N = 200$).

Model	Sentiment ($F1$)
Llama-3-70B-Instruct	**0.82**
Qwen2-72B-Instruct	0.79
DeepSeek-llm-67b-chat	0.72

The higher $F1$-scores for sentiment analysis compared to classification suggest that identifying emotional polarity is a less ambiguous task for LLMs than thematic categorization within a specialized administrative framework.

By employing ABSA, the pipeline successfully identified "mixed" reviews that would be lost in a global sentiment approach. For instance, in cases where a student praised an instructor's preparedness but critiqued the course requirements, the model correctly assigned a *Positive* label to the "Teachers Preparation" aspect and a *Negative* label to the "Requirements" aspect. This granularity allows university management to differentiate between interpersonal success and structural curriculum issues, providing a more balanced view of faculty performance.

Qualitative review of the ABSA results indicated that the model's "world knowledge" was instrumental in handling ambiguous feedback. Llama-3-70B proved capable of identifying the correct category even when students used non-technical language or slang. For instance the text **"I loved the long classes on Monday mornings, with the teachers soporific voice!"**, which can be interpreted as a positive feedback about the lecturers class, but if we see the question for context: **"What would you suggest to change regarding the work of the instructor?"**, the word *loved* gets a sarcastic meaning and changes the answers polarity.

Table 5. F1 scores across classes for Aspect Based Sentiment Analysis.

Sentimen	***F*1 score**
Positive	0.89
Neutral	0.43
Negative	0.86

3.3. Vectorization performance and embedder selection

The selection of the **Qwen3-embedding-8B** model was driven by a comparative evaluation against the **Jina-v3** multilingual model. To assess the models' ability to capture the semantic nuances of the Hungarian language in an academic context, we performed a similarity-based retrieval test using reference student comments. Our primary selection criterion was to identify which model consistently placed semantically similar responses in closer proximity to the reference text within the vector space.

In the first comparative example, for the reference text **"There were few practical examples"**, the Qwen model's nearest neighbor was **"Sometimes it was hard to follow the pace; it would be good to have more practical examples"**. Conversely, Jina-v3 identified **"The subject might be useful in later studies; there were few examples"** as the most similar response. While both segments of the Qwen-selected response focus on pedagogical critique regarding the lecture's delivery, Jina's selection only aligns with the reference in its second half, introducing an irrelevant positive sentiment in the first.

In the second case, using the reference text **"Communication could have been better"**, Jina-v3 retrieved **"Overall, it was a positive experience"**, which is thematically distant from the specific concern raised. In contrast, the Qwen model placed **"Maybe a bit more explanation would have helped"** in closest proximity. This result is semantically much more relevant to the original intent, as it correctly identifies "explanation" as a core component of the instructional communication mentioned in the reference.

3.4. Latent topic discovery and clustering

By applying HDBSCAN, the pipeline successfully identified new latent subtopics that were previously invisible to both the quantitative Likert scales and the high-level administrative categories. We present the detailed latent topic discovery and sentiment mapping specifically for the **Methodology** category, which emerged as the most prominent dimension in our initial classification results. As illustrated in Figure 2, the semantic space is organized into well-defined "islands", representing specific student concerns or praises.

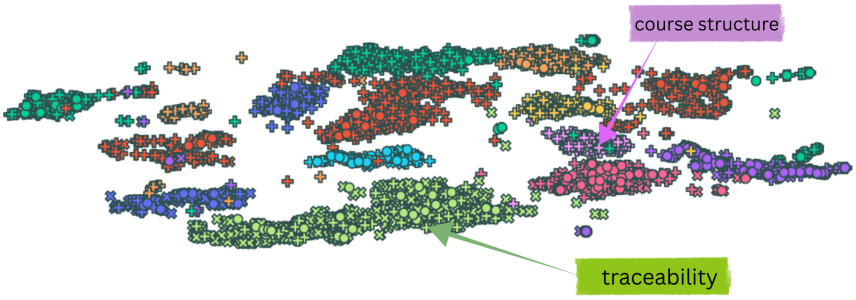

Figure 2. UMAP visualization of the identified latent student topics in the case of Methodology (Plus: positive, X: negative Circle: neutral sentiment.

The automated labeling procedure identified 11 distinct themes within the dataset, providing a comprehensive overview of the pedagogical environment. In order to get a better picture on the goodness of clustering 10 random example was picked and evaluated by hand how cohesive it was. These results are described in the following list with the clusters name and description.

- **Lecture Pace:** Feedback regarding the speed of delivery during classroom sessions. From the 10 checked answers there were only one where the text falls into another cathegory.
- **Clarity:** Comments on the intelligibility of the instructor's explanations. The same goes to this cluster as only one checked text mentioned compleately different topic.

- **Practicality:** Students' views on the hands-on applicability of the material. In this topic based on the examined answers there were three particular one that contained border line examples for the cluster. These answers were about how to improve the class, mainly with more practical examples.
- **Interactivity and Personal Opinions:** Feedback on the level of engagement and subjective student reflections. From the 10 evaluated examples there were no answer falling out of the topics scope.
- **Supplementary Materials:** The availability and quality of extra reading and study aids. The examined answere showed no outlier from the topic.
- **Traceability:** How easily students could maintain focus and follow the logic of the lecture. There were only one examined example that stated other then the one assigned. It particularly mentions lack of supporting material.
- **Course Structure:** Critiques regarding the logical organization and syllabus design. Each examined answer stated explicitly the course structure.
- **Instructor Knowledge:** Recognition of the teacher's expertise and professional command of the subject. From the 10 examined answers there were one falling out of the topics scope and could have been assigned into traceability.
- **Personal Experiences:** Anecdotal feedback regarding specific student–teacher interactions. Out of 10 answers 4 could have been assigned into practicality.
- **Material Accessibility and Slides:** Technical feedback regarding the distribution and quality of lecture slides. Each examines answers was correctly placed into the topic.
- **Real-world Examples:** The inclusion of industry-relevant or practical case studies. Out of the 10 examined answers 3 could have been assigned into practicality, 2 into Supplementary materials.

The discovery of these specific topics – such as the clear separation between "course structure" and "traceability" – demonstrates the system's ability to move beyond generic critiques. By mapping these topics to the previously analyzed sentiment scores, university management can identify not just that a course is viewed negatively, but specifically which aspect (e.g., the pace of the lecture or the lack of real-world examples) is driving the dissatisfaction.

The primary advantage of the proposed Aspect-Based Sentiment Analysis (ABSA) is the ability to map emotional valence directly onto the discovered latent topics. By projecting the sentiment scores onto the semantic manifold, we can identify not only the prevalence of specific themes but also their internal polarity distribution. As illustrated in Figure 3, the sentiment landscape reveals clear "pain points" and "success stories" within the institutional feedback.

The visualization demonstrates that certain topics, such as Course Structure is predominantly associated with positive (green) clusters, indicating it as consistent institutional strength. Conversely, topics related to Traceability show a higher density of negative (red) data points.

This granular mapping allows university management to move beyond global satisfaction metrics. For instance, while an instructor might receive an overall posi-

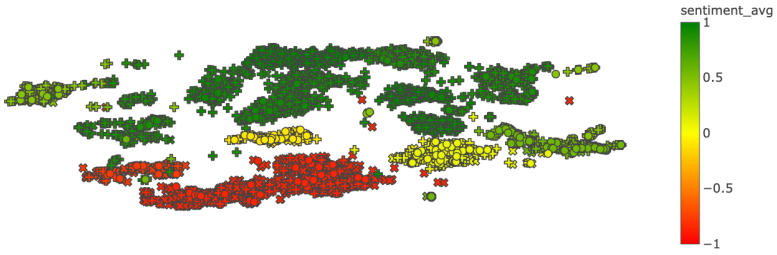

Figure 3. UMAP visualization of sentiment distribution across latent topics (within cluster sentiment average) (Green: Positive, Yellow: Neutral, Red: Negative).

tive rating, this analysis can pinpoint that the negative feedback is strictly isolated to the Traceability, rather than a lack of competence or clarity. Such precision is instrumental for targeted pedagogical interventions, as it provides a clear diagnostic tool for identifying which specific aspects of a course require immediate administrative or developmental attention.

4. Conclusion

Our study successfully established an automated pipeline for processing massive qualitative datasets within the Hungarian higher education framework. By leveraging the University of Szeged's 34,000 unique responses, the research proved that modern Large Language Models can effectively bridge the gap between vast unstructured data and precise administrative utility.

The research demonstrated the efficacy of high-parameter models like Llama-3-70B-Instruct in handling classification tasks with morphologically rich languages like Hungarian. We have also performed qualitative evaluation on the discovered topics and on their given name. Our proposed hybrid framework effectively pairs few-shot classification with unsupervised HDBSCAN clustering to capture both predefined institutional goals and emergent student concerns. The integration of Aspect-Based Sentiment Analysis allows the detection of nuanced, mixed feedback that global sentiment scores typically overlook.

The real-world application of this methodology offers immediate benefits to institutional quality assurance protocols. By automating a task that traditionally requires hundreds of manual labor hours allows university management to address pedagogical or infrastructural issues while they are still relevant to the current student cohort. Additionally, the objectivity provided by a standardized automated framework reduces the inherent bias of manual review, offering a reliable basis for faculty-level SWOT analyses and resource allocation. The unsupervised component of the pipeline is particularly valuable for discovering "hidden" systemic issues – such as specific technical failures or evolving student preferences – that remain invisible in standard quantitative Likert-scale ratings.

References

[1] A. Abdi, G. Sedrakyan, B. Veldkamp, J. van Hillegersberg, S. M. van den Berg: *Students feedback analysis model using deep learning-based method and linguistic knowledge for intelligent educational systems*, Soft Computing 27.19 (2023), pp. 14073–14094.

[2] AI@Meta: *Llama 3 Model Card* (2024), url: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[3] A. Baqach, A. Battou: *A new sentiment analysis model to classify students' reviews on MOOCs*, Education and Information Technologies 29.13 (2024), pp. 16813–16840.

[4] G. M. Csányi, D. Lakatos, I. Üveges, A. Megyeri, J. P. Vadász, D. Nagy, R. Vági: *From Fact Drafts to Operational Systems: Semantic Search in Legal Decisions Using Fact Drafts*, Big Data and Cognitive Computing 8.12 (2024), issn: 2504-2289, doi: 10.3390/bdcc8120185, url: https://www.mdpi.com/2504-2289/8/12/185.

[5] M. Hinek: *Témák és érzelmek–Szállodai vendégvélemények vizsgálata témamodellezéssel és szentimentelemzéssel* (2024).

[6] I. Keraghel, S. Morbieu, M. Nadif: *Beyond words: a comparative analysis of LLM embeddings for effective clustering*, in: International Symposium on Intelligent Data Analysis, Springer, 2024, pp. 205–216.

[7] L. McInnes, J. Healy, S. Astels, et al.: *hdbscan: Hierarchical density based clustering.* J. Open Source Softw. 2.11 (2017), p. 205.

[8] M. A. Mersha, J. Kalita, et al.: *Semantic-driven topic modeling using transformer-based embeddings and clustering algorithms*, Procedia Computer Science 244 (2024), pp. 121–132.

[9] D. Moulavi, P. A. Jaskowiak, R. J. Campello, A. Zimek, J. Sander: *Density-based clustering validation*, in: Proceedings of the 2014 SIAM international conference on data mining, SIAM, 2014, pp. 839–847.

[10] M. Osváth, G. Yang Zijian, K. Kósa: *Magyar páciensek narratív tapasztalatainak elemzése BERT témamodellezéssel és szentimentelemzéssel*, Berend Gábor–Gosztolya Gábor–Vincze Veronika (szerk.): XVIII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem TTIK Informatikai Intézet, Szeged (2022), pp. 311–324.

[11] *Qwen2 Technical Report* (2024).

[12] S. Srinivas, S. Rajendran: *Topic-based knowledge mining of online student reviews for strategic planning in universities*, Computers & Industrial Engineering 128 (2019), pp. 974–984.

[13] J. Sun, L. Yan: *Using topic modeling to understand comments in student evaluations of teaching*, Discover Education 2.1 (2023), p. 25.

[14] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, J. Zhou: *Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models*, arXiv preprint arXiv:2506.05176 (2025).