Integrating Large Language Models for Comprehensive Study and Sentiment Analysis of Student Feedback

Jana Kuzmanova
Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering
Skopje, N. Macedonia

Katerina Zdravkova
Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering
Skopje, N. Macedonia

Ivan Chorbev
Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering
Skopje, N. Macedonia

{jana.kuzmanova; katerina.zdravkova; ivan.chorbev}@finki.ukim.mk

Abstract

In academic year 2023/24, our university collected over 200,000 student feedback responses evaluating teaching staff and course experiences. The survey included demographic data, 10 Likert scale questions on teaching quality, a question on student attendance, and three open-ended questions about student experiences. This paper explores the integration of Large Language Models (LLM) Gemini for sentiment analysis to evaluate students' feedback quantitatively and qualitatively. We statistically analyze the Likert scale responses. To address the linguistic diversity of open-ended responses, written in both Cyrillic and Latin scripts with standard and slang expressions in several languages, we employed a preprocessing step using Gemini to standardize the input for further analyses. Sentiment analysis aims to identify various sentiment nuances, including direct answers, contradiction, multipolarity, mixed sentiment, sarcasm, irony, negation, ambiguity, understatement, and over-exaggeration. By comparing these insights with quantitative feedback, we aim to uncover deeper patterns between student perceptions and teaching performance. While the focus is on sentiment analysis, we also discuss the evaluation of the results provided by LLM. For the sentiments with less answers, the evaluation of GenAI was done manually. For the sentiments with more than 1000 entries, we suggest a semi-automated approach for sentiment categorization, to be explored in future work. This study enhances our understanding of student feedback through advanced computational methods, providing a more nuanced perspective on teaching quality and student satisfaction.

1 Introduction

In recent years, education has undergone significant transformations. Initially driven by the sudden transition to online learning due to COVID-19, these changes have been further accelerated by the emergence of generative artificial intelligence (GenAI) tools, which students have quickly embraced as essential components of their learning and assignment preparation (Zdravkova and Ilijoski, 2025). As a result, student expectations toward class delivery methods and assessments have evolved (Chan and Hu, 2023).

Student feedback has become a valuable source of insight into the quality of the educational process, encouraging teachers to improve their teaching practices, even when the sole intention of these changes is to increase their own ratings (Flodén, 2016). Well-intentioned assessments, opinions and suggestions can significantly improve teaching practices and provide guidance for institutional development and global education reforms aimed at adapting the emerging technological trends (Kastrati et al., 2021). The surveys combine assessments of overall teaching quality with sentiment analysis derived from open-ended questions (Hynninen et al., 2020).

Student surveys do not always reflect the true quality of education, as many students either skip them or respond mechanically, often omitting open-ended questions. To better understand this behavior, we asked students why they tend to fill out surveys superficially. The most common reasons included the excessive length of the questionnaires, lack of lecture attendance, which left them feeling unqualified to evaluate the teaching, the reliance on peer opinions from social networks rather than personal experience, but predominantly the impression that their responses have no impact on teaching outcomes.

These findings highlight a systematic credibility issue in the survey data. As a result, we interpreted the results with caution and used them primarily to identify general trends rather than to draw strong evaluative conclusions about specific individuals

Proceedings of Recent Advances in Natural Language Processing,pages 604–613
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-071

or courses.

After each semester, students at our university are required to complete an anonymous survey covering all faculties, all subjects, and all teaching staff. The survey for the last academic year has more than 200,000 records: 81,500 for the winter semester and 130,550 for the summer semester. Likert-scale responses were analyzed quantitatively, while sentiment analysis of open-ended answers was conducted using Gemini and its built-in training data. Manual assessment was applied only to sentiment categories with a low number of responses. Additionally, a combined analysis of numerical and textual data was performed.

The paper continues by presenting an analysis of related work. Section 3 presents a detailed description of the methodology of our experiment that includes preprocessing of open-ended questions and setting the experiment. The results are processed in section 4. The paper ends with discussions of the entire experiment, conclusions and plans for automating the process of processing the survey and evaluating the obtained results.

2 Quantitative, qualitative and AI-based approaches to survey data analysis

Surveys containing substantial amounts of numerical and textual data can provide valuable insights. Quantitative responses reveal patterns in student satisfaction, while open-ended feedback highlights specific concerns about content and teaching. However, analyzing unstructured text remains challenging due to its complexity. To address this, studies often apply scalable sentiment analysis using lexicon-based methods, machine learning, or transformer models, combined with qualitative approaches like thematic coding. This integration enables both large-scale trend detection and deeper interpretive insights that reflect contextual nuances (Kastrati et al., 2021).

2.1 Likert scale ordinal data

Methods for processing Likert-scale data include descriptive statistics, comparative analysis, and nonparametric statistical methods (Harpe, 2015).

Descriptive statistics provide initial quantitative and visual summaries of the sample (Cooksey, 2020). The numerical data in our research are ordinal, thus univariate, bivariate and multivariate statistics, as well as regression analysis and linear trend estimation, can be applied (Verhulst and Neale, 2021).

Comparative analysis allows for the examination of differences between groups. In the context of student surveys, this primarily involves comparisons across faculties, departments, and institutes (Ragin, 1998). Nonparametric tests are also suitable for such analyses. However, given that our survey collected data from twenty-two faculties and four institutes, each varying significantly in student population and educational level, we opted to exclude this aspect from our comprehensive analysis.

Nonparametric statistics encompass analyses such as contingency tables, rank tests for one or more samples, rank correlation, and basic nonparametric regression techniques (Wasserman, 2006). In our research, we primarily focused on contingency tables, which display the multivariate frequency distributions of variables.

2.2 Unstructured textual data

The sentiment analysis of the survey analyzed in our comprehensive study is designed to enhance the overall educational experience and help maintain or even improve the university already high reputation.

Sentiment analysis involves examining text to determine its polarity and identify the emotional tone of responses. Polarity refers to the general sentiment expressed in a message, typically categorized as positive, negative, or neutral. Beyond polarity, analyzing emotions offers a more nuanced understanding of feedback by recognizing linguistic patterns that reflect how respondents express their attitudes. These are typically considered direct emotional responses and are contrasted with more complex emotions such as irony and sarcasm (Filik et al., 2019; Frenda et al., 2022). In contrast, emotion classification seeks to capture more nuanced aspects of feedback, offering a deeper understanding of how respondents truly feel.

In our analysis, we extended this framework to include additional categories such as negation, multipolarity, ambiguity, exaggeration, context-dependent sentiment, incomplete understatement, contradiction, and mixed sentiment.

The most widely used NLP techniques for determining sentiment polarity include sentiment lexicons, such as VADER (Qi and Shabrina, 2023). Additionally, pattern matching, which relies on predefined rules, has proven to be highly effective for sentiment classification (Zhang and

Zhang, 2022). Finally, traditional machine learning techniques, such as the well-established Naïve Bayes classifier and Support Vector Machines (SVM), have been successful in determining sentiment polarity (Ahmad and Umar, 2023).

To detect subtle emotional nuances, several methods are employed. Emotion lexicons remain effective for classifying specific emotions (Nandwani and Verma, 2021), while rule-based detection aids in identifying sarcasm and negation (K et al., 2021). Dependency parsers, which analyze grammatical structure, help reveal word relationships (Agarwal et al., 2015). Named entity recognition (NER) also supports context-sensitive sentiment analysis, particularly when sentiment is linked to specific entities (Derczynski et al., 2015).

To carry out these tasks, we relied entirely on Gemini. First, we prompted LLM to determine the sentiment polarity of each response. Following that, we requested a classification of the emotional tone. As outlined in the introduction, we did not train or fine-tune the model ourselves. Instead, we leveraged its pretrained data and built-in classification capabilities. Throughout the process, we remained aware of the potential biases inherent in relying on an LLM and took care to interpret its outputs critically.

2.3 Integrating numerical and textual data

Surveys include numerical and textual responses. After performing quantitative and mixed analyses, meaningful patterns can be identified by comparing or integrating both data types. One common approach is the use of mixed-effects models, which establishes relationships between response variables and relevant covariates (Bates, 2010). In our study, mixed-effects models are used to compare numerical values from Likert-scale responses with sentiment derived from open-ended questions.

Correlation methods, such as Spearman's rank correlation, can also be effective tools (Sedgwick, 2014). They can compare sentiment polarity with numerical results, enabling the evaluation of similarity between sentiments and Likert-scale ratings. Multivariate analysis is another powerful technique for exploring relationships between categorical variables, i.e. those that represent different groups or categories (Wu et al., 2015).

Cluster analysis can be employed to group similar comments based on associated numerical feedback (Huang and Mitchell, 2006). For our research, we use hierarchical clustering to examine the degree of correlation between specific sentiments and corresponding Likert-scale responses.

2.4 LLM-based sentiment analysis

Since 2020, large language models have emerged as powerful tools in natural language processing, particularly for sentiment analysis. These models effectively address challenges in both polarity detection and the interpretation of subtle emotional cues in text (Zhang et al., 2024).

The best proof of the growing influence of LLMs and GenAI among the researchers is the sharp increase of scholarly articles. Namely, since 2024, more than 20,000 articles on Google Scholar have included the phrases "large language models" and "sentiment analysis", while nearly 5,000 papers reference "generative artificial intelligence" in the context of sentiment analysis. Our research leverages the large language model Gemini to enhance both the accuracy and depth of sentiment analysis, enabling a more nuanced understanding of textual data.

3 Methodological framework for LLM-based sentiment analysis

Self-evaluation is one of the strategic goals of the university, which, among other aspects, monitors and assesses the quality of higher education and scientific re-search activities. It is implemented through the university's electronic system at the end of each semester. Although students log in with their university credentials to access the survey, no usernames or personal identifiers are recorded or stored. Such authentication fully preserves anonymity, in strict compliance with GDPR (European Parliament and Council of the European Union, 2016).

In addition to demographic data, the survey included ten numerical questions rated on a 1–5 scale, one attendance question expressed as a percentage, and three open-text fields.

Students were asked to numerically evaluate the following ten aspects of each course and its teaching staff: the professors knowledge; how helpful the lectures are; whether the questions are interesting and include relevant examples; how much the student has learned during the lectures; whether there are appropriate learning materials provided; the punctuality of the teaching staff; whether the students are graded continuously with various

relevant techniques; whether the teaching staff is unbiased; as well as their availability for office hours and over email. The text fields allow students to highlight what they liked in lectures or exercises, what they disliked, and to provide any additional comments.

3.1 Preprocessing of Textual Data

The first step in the preprocessing was to remove the responses that didn't contain any textual data, only punctuation such as "/", "-", or ".", sometimes repeated multiple times. This was done through simple regular expression matching in Python, leaving only answers that contain at least one letter.

The next step was to classify the text by language used. This was done using Gemini. The requests asked for a custom structured response, which included a field for the original comment and its language. The language was an enum type, listing four languages as options: Macedonian (written with Cyrillic and Latin script), Albanian, English, and Turkish. This was done to limit the number of possible options and ensure there will be fewer misclassifications, especially with similar languages. The prompt used was "Determine the language of the following texts:"and the comments were passed in batches. Figure 1 shows the distribution by language for the textual fields.

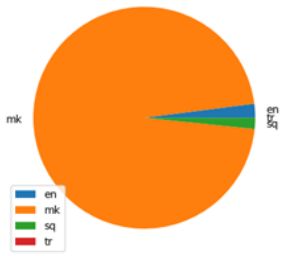

Figure 1: Distribution of comment languages

The overall accuracy of the language detection, evaluated on a random sample of 600 comments, was 96.45%. The accuracy for comments written in the Latin script was 88.69% and for Cyrillic script comments 100%. Most mistakes were on single-word comments, often with internationalisms such as ok or super, that were classified as different languages in different responses. Some responses were misclassified, such as the Macedonian word ce / se "everything", that is sometimes classified as English. However, there was also a small number of misclassified full-sentence responses. To additionally check the validity of the results, the number of responses that have different identified languages for the different textual fields were singled out. There were 679 responses that fit this description. These included the mistakes mentioned above, as well as responses where the students actually wrote different parts of the response in different languages.

After language detection, Macedonian responses written in Latin script were transliterated into Cyrillic. Using the previous step's results, Python regular expressions filtered Macedonian text to identify any Latin characters. Responses containing Latin characters were then sent to Gemini for transliteration. Similarly to language detection, this process was performed in batches, handling each column separately. The prompt used was: "Transliterate the following texts into Macedonian Cyrillic." The result was impeccable.

3.2 Experiments with Textual Data

The data was analyzed using a combination of statistical tools in Python, mostly for the numerical responses, and Gemini for analyzing the textual fields.[1]

The statistical analysis focused on the number of responses, grouped by faculty, professor, and course, as well as computing the mean, standard deviation, median, and important percentiles. The mean was computed per column, as well as by response. The mean grade across all numerical questions of each response was used to find the number of responses that have both a low mean and a non-empty text field for the positive comment, as well as for further analysis and experiments relating to grade prediction. All these statistical analyses were done using the pan-das module.

The text data was analyzed using Gemini. The gemini-1.5-flash model was used in all experiments. The data was classified by sentiment in the following categories: ambiguity, context-dependent sentiment, direct answer, humor, irony, mixed-sentiment, multipolarity, negation, over-exaggeration, sarcasm, and understatement. These were given as an enum type as part of the expected response, and comments were passed in batches.

Comments were also classified by their focus using Gemini. The categories in this case

[1]The code can be found at https://colab.research.google.com/drive/18eRXjaY-magr46ijylKI8-cKGNENfjAj?usp=drive_link

corresponded to the questions in the survey that asked for numerical answers. The possible Likert-scale categories for this classification were: Teachers' level of knowledge (abbreviated as TK), Lectures are helpful (LH) Lecture are interesting (LI), Learning outcomes (LO), Quality of lecture and study materials (QL), Punctuality (P), Continuous grading techniques (GT), Objectivity (O), Availability for consultations (AC), and Answers emails (E). Each textual response was assigned to one or more of these categories, depending on which aspect it is most likely to be commenting on.

This was later used to help evaluate Gemini's effectiveness in predicting the grade based on the textual comments. Each comment was associated with a predicted grade. Then, this grade was compared to the grade student gave to the numerical questions that correspond to the given comment. Additionally, a grade was assigned to each response by passing all textual answers in one request and asking for a predicted mean grade. This was then compared to the mean of all the numerical answers for that response.

4 Results and analysis

The 212052 responses from 4386 courses taught by 2123 teaching staff were first examined based on their demographic data and the presence of real textual answers. It was followed by the statistical analysis and comprehensive sentiment analysis.

4.1 Demographic Data

Out of all textual comments in all fields, most responses were empty. Only 35727 of the responses have at least one field that is not empty or consists only of punctuation. Figure 2 shows the number of empty, punctuation-only, and one-word responses, as well as non-empty responses that contain more than one word. There are 1871 courses, 359 teaching staff, and one institute with no non-empty textual responses. When considering the evaluation of each course as taught by a specific professor or assistant, there are a total of 5951 combinations with more than five responses, and 8628 with less than five. When considering only responses that have at least one non-empty text field, there are 1860 combinations of professors and courses with more than five responses, and 4874 with less than five.

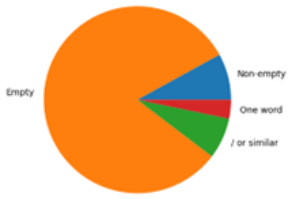

Figure 2: Distribution of empty to non-empty responses

4.2 Statistical Analysis of Numerical Answers

The numerical data was analyzed statistically using the pandas module and basic Python. Table 1 presents the mean and standard deviation of the grades given by the students for each question. The rows correspond to each of the previously mentioned questions.

	Mean	Std
PK	4.65	0.74
LH	4.53	0.86
LI	4.49	0.89
LO	4.47	0.91
QL	4.52	0.87
P	4.6	0.8
GT	4.45	0.93
O	4.56	0.85
AC	4.5	0.87
E	4.46	0.91

Table 1: Statistical analysis of numerical data

4.3 Sentiment analysis

According to the performed sentiment classification, direct expressions of sentiment were, as expected, the most frequent across all comment categories, particularly in the positive feedback section. Negation appeared significantly more often in negative comments, while mixed sentiment was most found in the additional comments field. The most common categories overall were Direct answer, Negation, Ambiguity, and Mixed sentiment, while categories such as Humor, Irony, Overexhaggeration, Understatement and Sarcasm were rare, with less than 50 comments each. The precision and recall of the sentiment categories were evaluated on a random sample of 200 comments for the larger categories, and all comments classified as belonging to the less common ones. The results are shown in Table 2. Direct answers had the highest precision among the common categories and one of the highest overall by far. However, all other categories had significantly higher recall. The Negation category captured both comments

highlighting what was missing and a substantial number of brief responses like “Nothing” or “No comment.”

	Precision (%)	Recall(%)
Ambiguity	26.47	95.23
Context-dependent	14	31.82
Direct answer	96.55	42.86
Humor	18.18	100
Mixed sentiment	66.67	93.75
Multipolarity	40	100
Negation	64.52	86.91
Overexaggeration	100	93.75
Sarcasm & Irony	80	91.67
Understatement	50	100

Table 2: Precision and recall of sentiment classification

Notable mistakes include the Humor category, as well as the Context-dependent sentiment category, which included a large number of direct answers, while a lot of comments belonging to this category were assigned to the Multipolarity category. The Humor category mostly consisted of non-answer comments like “hahaha” and only few humorous comments. The Sarcasm category also contained a mix of sarcastic comments and straightforward, mostly negative, opinions. Comments in English tend to be correctly classified as sarcastic, while there were multiple comments in Macedonian that were classified as sarcastic but weren’t. One such example is the comment Емил е цар >:(/ Emil e car >:(Emil is a king >:(, posted by a student who attended every lecture and consistently gave the young professor excellent ratings. Despite the positive verbal praise and high grades, the student included a negative emoji, illustrating mixed sentiment or multipolarity rather than sarcasm.

4.4 Joined Quantitative and Qualitative Analysis

The relationships between the rating and classified sentiment are visualized using a heatmap (Figure 3) and a boxplot (Figure 4).

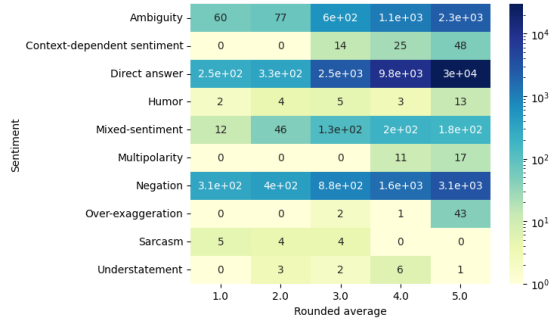

Figure 3: Comment count by sentiment and rating

Figure 3 shows the distribution of sentiment types across ratings, rounded to the nearest integer. A logarithmic scale was used due to the imbalance in frequency, since most responses are clustered around high ratings and a small number of sentiment categories.

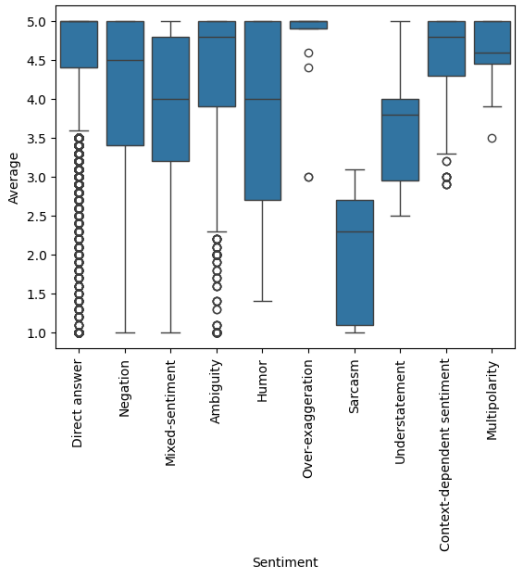

Figure 4: Rating distribution by sentiment

To more formally quantify the relationship between the rating and the sentiment, the Pearson correlation coefficient was computed for the presence of the four most common sentiment types. The results are given in Table 3. While a correlation exists between some sentiment categories and average rating, it is not very pronounced.

	Correlation coefficient
Direct answer	0.231
Negation	-0.205
Ambiguity	-0.069
Mixed sentiment	0.084

Table 3: Pearson correlation coefficient between sentiment type and average rating

Responses where students gave a low mean across all numerical answers, but still left a positive comment were also separately evaluated on whether that comment was positive. The results of the classification of these responses left in the field for positive comments are given in Figure 5. The overall accuracy of this classification was 95.54%. The precision and recall are given in Table 4.

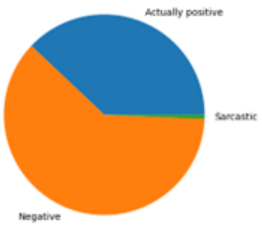

Figure 5: Types of positive comments with low average grade of response

Most of these comments were negative, generally short answers such as "Nothing", "I don't know", or similar. Other answers offered more substantial negative comments. However, some of the responses in the positive field were genuinely positive. This category contained both longer and one-word comments, while some similar or same comments were classified as negative, such as the comment Презентации Prezentacii "Presentations".

As in the general sentiment classification, sarcastic answers were rare. Most sarcastic comments were classified as negative, and many of the comments classified as sarcastic expressed straightforward negative sentiments. Additionally, there wasn't a big overlap between the comments classified as sarcastic here and in the previous classification of sentiment.

	Precision (%)	Recall(%)
Positive	99.7	92.2
Negative	93.48	99.4
Sarcastic	50	37.5

Table 4: Precision and recall of sentiment classification for positive text field in low average responses

4.5 Ratings Prediction

Ratings were predicted as integer scores on single comments and compared to the rating given for the question corresponding to that comment as described above, as well as real values predicting the mean for the entire response.

The integer rating prediction resulted in outliers with predicted grades well above the maximum allowed. These outliers lead to a large mean squared error of 280.6. Excluding these outliers, the error drops to 1.86, and only considering comments longer than three words to 1.63. The median absolute error in both of the last cases is 0, and the maximum error 4. The correlation coefficient is 0.35 and 0.49, respectively.

Table 5 shows the comparison of predicted and actual ratings, excluding outliers and only considering comments longer than three words. Actual ratings tend to be higher than predicted – in all cases of predicted ratings, there's more fives than the rating that was predicted.

pred↓/a→	1	2	3	4	5
1	711	409	568	425	820
2	87	78	158	175	303
3	89	116	316	711	2169
4	57	61	315	822	3824
5	117	128	627	1891	13006

Table 5: Predicted vs actual ratings confusion matrix

This could be related to the fact that most student ratings are all fives, but also to the lack of context. For example, the comment начинот на предавање на лекциите "the way lectures were taught" is a positive comment associated with a rating of five, but without the context that it's a positive comment, it's impossible to tell whether this comment would be associated with a positive or a negative rating. The predicted rating for this comment was 3. Additionally, many negative comments are still associated with a not too low score, but without the context of the other comments, the negative comment alone could lead to a lower predicted rating.

When considering entire responses, the results tend to be closer to the actual mean rating. The prompt for this classification included all comments from each response, labeled by whether it comes from the text field for positive, negative, or additional comments. These results are shown in Table 6. MSE stands for the mean squared error, MAE for median absolute error, MaxE is the max error, while CC is the correlation coefficient.

	MSE	MAE	MaxE	CC
all	0.66	0.5	3.78	0.39
one > 2 words	0.57	0.5	3.8	0.41
all > 2 words	0.72	0.5	3.8	0.52

Table 6: Error on predicted rating

While there are still responses with a large error, the error is smaller than when considering each comment separately. The errors were tested considering all comments, responses with at least one comment with more than two words, and responses where all textual fields were longer than two words.

Possible ways to improve the accuracy of predicted ratings include fine-tuning with a subset

of the responses, as well as including the expected grade distribution. Using an enum with the possible expected values instead of expecting a float or integer while passing the maximum possible value as part of the prompt could be a way to ensure no outliers appear in the predicted ratings.

5 Conclusions and further work

Despite the large volume of responses, only about 17% contained any meaningful textual feedback. This indicates a major limitation in student engagement with the open comment fields, which in turn limits the scope of qualitative insights.

The sentiment analysis revealed a strong dominance of direct responses across all categories, particularly in positive comments, which were often concise and unambiguous. In contrast, negative and additional fields showed significantly higher occurrences of negation, especially in low-scoring responses, indicating dissatisfaction or lack of desirable attributes (e.g., "nothing," "no comment"). Mixed sentiment was most prevalent in the additional comments, which reflects students' tendency to offer nuanced feedback when not constrained to specific positive or negative prompts.

While categories like Sarcasm and Humor were rarely observed, their classification proved challenging. Some comments in Macedonian were misclassified as sarcastic due to syntactic ambiguity or contextual misunderstanding (e.g., ironic phrasing or inconsistent emoji mistaken for sarcasm). Additionally, the Humor category was often filled with meaningless or filler content (e.g., "hahaha") rather than a genuine humorous critique. This highlights a limitation in multilingual sentiment detection, particularly because the training resources Gemini used does not have a subtler tone distinctions characteristic for Macedonian language.

There is a clear relationship between low numerical scores and the prevalence of negation in comments. Positive comments with low overall scores were often not genuinely positive, frequently consisting of "Nothing" or similarly dismissive statements. However, a small subset of such responses included legitimately positive comments, demonstrating that some students distinguish between different aspects of their experience (e.g., poor course structure but good teaching method).

The grade prediction model faced difficulties, particularly when analyzing isolated comments. The model underestimated actual scores, which is likely due to a lack of context—many neutralsounding comments (e.g., "the way lectures were taught") are associated with high scores but offer no explicit sentiment markers. When analyzing entire responses instead of individual comments, the model's predictions were more accurate, with a lower mean squared error and better correlation with actual grades. This supports the use of contextual sentiment analysis over sentence-level or phrase-level models when applying machine learning to student feedback.

The first direction of our future work will focus on integrating several additional large language models and establishing a framework for comparative analysis of the results they produce. A key objective will be to significantly enhance the accuracy of sentiment analysis performed by LLMs. Given the absence of a sentiment lexicon for the Macedonian language, one potential improvement involves machine translating the comments using the same LLM and subsequently conducting sentiment analysis using existing English-language lexicons such as VADER (Qi and Shabrina, 2023), SentiWordNet (Baccianella et al., 2010), or others (Khoo and Johnkhan, 2017).

Further refinement may be achieved by developing prototypical examples for each sentiment category and evaluating them using BERT-based sentence embeddings (Catelli et al., 2022). Although manual or crowdsourced annotation remains a valuable validation strategy, it poses logistical challenges (van Atteveldt et al., 2021).

Recognizing that it is unlikely students will significantly change their feedback behavior to reduce the high number of empty or noninformative comments, we also plan to fine-tune large language models using insights from the current study, leveraging popular retrieval-augmented generation frameworks designed for LLMs (Gao et al., 2024; Ram et al., 2023). By combining these strategies, we aim to build a more context-aware, linguistically adaptable sentiment analysis system that better captures the nuances of student feedback.

Acknowledgment

This work was supported in part by grants from the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje.

References

Basant Agarwal, Soujanya Poria, Namita Mittal, Alexander Gelbukh, and Amir Hussain. 2015. Concept-level sentiment analysis with dependency-based semantic parsing: A novel approach. *Cognitive Computation*, 7.

Hero Ahmad and Shahla Umar. 2023. Sentiment analysis of financial textual data using machine learning and deep learning models. *Informatica*, 47:153–158.

Wouter van Atteveldt, Mariken A. C. G. van der Velden, and Mark Boukes and. 2021. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2):121–140.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. volume 10.

Douglas Bates. 2010. *Lme4: Mixed-Effects Modeling With R*.

Rosario Catelli, Serena Pelosi, and Massimo Esposito. 2022. Lexicon-based vs. bert-based sentiment analysis: A comparative study in italian. *Electronics*, 11:374.

Cecilia Chan and Wenjie Hu. 2023. Students' voices on generative ai: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20.

Ray Cooksey. 2020. *Illustrating Statistical Procedures: Finding Meaning in Quantitative Data*.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council.

Ruth Filik, Alexandra Turcan, Christina Ralph-Nearman, and Alain Pitiot. 2019. What is the difference between irony and sarcasm? an fmri study. *Cortex*.

Jonas Flodén. 2016. The impact of student feedback on teaching in higher education. *Assessment & Evaluation in Higher Education*, 42:1–15.

Simona Frenda, Alessandra Teresa Cignarella, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2022. The unbearable hurtfulness of sarcasm. *Expert Syst. Appl.*, 193(C).

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Spencer Harpe. 2015. How to analyze likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, 7:836–850.

Yifen Huang and Tom Mitchell. 2006. Text clustering with extended user feedback. pages 413–420.

Timo Hynninen, Antti Knutas, and Maija Hujala. 2020. Sentiment analysis of open-ended student feedback. In *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 755–759.

Sentamilselvan K, Dr.Suresh P, Kamalam G K, S. Mahendran, and D. Aneri. 2021. Detection on sarcasm using machine learning classifiers and rule based approach. *IOP Conference Series: Materials Science and Engineering*, 1055:012105.

Zenun Kastrati, Fisnik Dalipi, Ali Shariq Imran, Krenare Pireva Nuci, and Mudasir Ahmad Wani. 2021. Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study. *Applied Sciences*, 11.

Christopher Khoo and Sathik Johnkhan. 2017. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44:016555151770351.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11.

Yuxing Qi and Zahratu Shabrina. 2023. Sentiment analysis using twitter data: a comparative application of lexicon- and machine-learning-based approach. *Social Network Analysis and Mining*, 13.

Charles C Ragin. 1998. The logic of qualitative comparative analysis. *International review of social history*, 43(S6):105–124.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Philip Sedgwick. 2014. Spearman's rank correlation coefficient. *BMJ: British Medical Journal*, 349:g7327.

Brad Verhulst and Michael Neale. 2021. Best practices for binary and ordinal data analyses. *Behavior Genetics*, 51.

Larry Wasserman. 2006. *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Wei Wu, Fan Jia, and Craig Enders. 2015. A comparison of imputation strategies for ordinal missing data on likert scale variables. *Multivariate Behavioral Research*, 50:1–20.

Katerina Zdravkova and Bojan Ilijoski. 2025. The impact of large language models on computer science student writing. *International Journal of Educational Technology in Higher Education*, 22.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Yilin Zhang and Lingling Zhang. 2022. Movie recommendation algorithm based on sentiment analysis and lda. *Procedia Computer Science*, 199:871–878.