ISSN: 2473-4901
v10 n6210

From Angry Reviews to Classroom Success: Using LLMs to Synthesize RateMyProfessors.com Data

Nicholas Caporusso
caporusson1@nku.edu

My Hami Doan
doanm4@mymail.nku.edu

Bikash Acharya
acharyab2@mymail.nku.edu

Priyanka Pandit
panditp1@mymail.nku.edu

Sushant Shrestha
shresthas11@mymail.nku.edu

Na Le
len4@mymail.nku.edu

Rajani Khatri
khatrir2@mymail.nku.edu

Will Pond
pondw1@mymail.nku.edu

Human-Computer Interaction Lab
Northern Kentucky University
Highland Heights, USA

Abstract

In recent years, online professor review platforms have become increasingly prevalent in higher education. While previous studies have examined various aspects of these platforms, such as review sentiment and content validity, their potential as a source of information for academic success has been largely unexplored. This paper investigates the use of Large Language Models to analyze anonymous professor reviews and identify common themes related to effective teaching practices, course design, and student engagement. The goal is to provide students with actionable suggestions on how to succeed in specific courses rather than focusing on elements that do not directly impact educational outcomes.

Our study analyzed reviews of nearly 40,000 computer science instructors, producing meaningful insights into course experiences. The methodology introduces a novel approach to process reviews and extracts content that contributes to student success in computer science. In addition to highlighting effective teaching strategies, this research also identifies areas for potential improvement in computer science education. Our work demonstrates how natural language processing techniques can be utilized to elicit actionable information for both students and educators. The methodology demonstrated in the paper on online reviews can be utilized to summarize Student Evaluations of Teaching.

Keywords: Large Language Models (LLMs), Computer Science Education, Student Evaluation of Teaching, Educational Data Mining, Natural Language Processing, Mixed Methods Analysis.

1. INTRODUCTION

Education, and particularly the field of computer science (CS), is continually evolving, driven by advancements in technology as well as changing student interests, backgrounds, and learning preferences (Luxton-Reilly et al., 2018). It is important for instructors and departments to understand teaching approaches and course design elements that resonate with today's learners to keep pace with these changes and provide an effective and engaging educational experience for CS students (Stephenson et al., 2018). There is a growing body of academic literature on pedagogical best practices in CS; student voices and perspectives are often missing from this discourse (Robins et al., 2003).

Student reviews are collected systematically through Student Evaluation of Teaching (SET) systems such as surveys administered by the university at the end of the semester to gather students' perceptions of their learning experiences and the effectiveness of their instructors. This, in turn, provides feedback to professors and informs personnel decisions (Coladarci & Kornfield, 2007). Although SETs are widely used in higher education institutions, the analysis of such valuable data often lacks the same systematic approach, and the utilization of data is often fragmented and inconsistent across departments and institutions due to several factors. First, SETs generate a large amount of qualitative and quantitative data, making it challenging to process and interpret the information effectively (Spooren et al., 2013). Moreover, different departments and institutions may employ varying methods to analyze SETs, leading to a lack of standardization and comparability across the board (Uttl et al., 2017). Also, SETs are often not publicly shared and are used for evaluating individual instructors' performance rather than identifying broader trends and best practices in teaching (Hornstein, 2017). This fragmented approach to SET analysis hinders the ability to derive meaningful insights and actionable recommendations for improving teaching effectiveness at a larger scale (Linse, 2017).

On the other hand, RateMyProfessors.com (RMP) enables students to anonymously and publicly rate their teachers on various quantitative criteria, including clarity, helpfulness, and easiness (Timmerman, 2008). RMP has gained immense popularity among students because of the ability to publicly share opinions about professors, primarily thanks to the possibility of getting access to reviews before making enrollment decisions. Although its validity and usefulness have been questioned by scholars and educators, RMP offers a wealth of student reviews and opinions about CS courses and instructors. However, the unstructured text format of these reviews makes it challenging to efficiently distill overarching themes and evidence-based insights, especially when trying to digest reviews of professors in different disciplines. Furthermore, many reviews contain elements unrelated to pedagogy, including personal retaliation, inappropriate comments, and swear words.

This paper proposes a novel approach to analyzing RMP reviews based on the use of Large Language Models (LLMs) and their capabilities in natural language processing tasks, including text classification and summarization. Our methodology enables the extraction of a summary of the learning experience based on key dimensions, such as teaching style and classroom environment, learning approach and course content, participation and interaction, workload and expectations, and overall experience, rather than focusing on aspects that are not related to academic success. By leveraging the power of LLMs, our method automatically analyzes professors' reviews to identify key pedagogical themes and filter out irrelevant or biased information.

2. RELATED WORK

In the past decades, several studies suggested that universities should consider making their own SET data publicly available online to provide students with more representative and comprehensive data (Coladarci & Kornfield, 2007). Nevertheless, to this date, RMP remains the largest dataset of professors' reviews. As a result, the platform has drawn considerable research attention, and several studies have explored the use of RMP data to gain insights into various aspects of higher education, overcoming the limitations of SETs in terms of public availability. Researchers have investigated the correlations between RMP ratings and traditional SETs (Coladarci & Kornfield, 2007), finding generally strong correlations, suggesting some degree of the validity of publicly available reviews as an indicator of instructor performance. Simultaneously, (Coladarci & Kornfield, 2007) found that RMP may be useful for identifying very highly rated instructors but less effective for differentiating among instructors with lower ratings and, therefore, that RMP is not a substitute for formal in-class evaluations. Other studies noted that easiness and quality ratings on RMP were positively correlated, suggesting that students tend to rate professors more favorably when they perceive the course as less challenging (Kindred & Mohammed, 2005). Several research groups conducted thematic content analyses of RMP comments and found that students often comment on both instructor competence and personal characteristics (Felton et al., 2008). Also, several studies (Kindred & Mohammed, 2005) analyzed the content of RMP reviews to identify common themes and factors that influence student ratings and found that students often mentioned professor personality, teaching style, and course difficulty as key factors in their evaluations, and they cautioned that RMP reviews should be interpreted with care, as they may not always reflect the actual quality of teaching. The authors of a study (Legg & Wilson, 2012) found that students who voluntarily rate their professors on RMP tend to provide more negative evaluations compared to formal in-class evaluations. This self-selection bias raises questions about the representativeness of RMP ratings and their ability to reflect the overall student experience accurately. Also, other potential biases in RMP ratings have been a significant concern for researchers. Studies have shown that factors such as a professor's age, gender, ethnicity, and even physical attractiveness can influence student ratings on RMP (Legg & Wilson, 2012). The latter findings suggest the presence of biases and, consequently, raise questions about the fairness and objectivity of RMP evaluations and their impact on instructors' careers. For instance, (Gordon & Alam, 2021) found that students often comment on the accents of instructors with "Asian" last names, highlighting the potential for racial and linguistic biases in these evaluations. Additionally, some authors (Rosen, 2018) observed that professors in science, technology, engineering, and mathematics (STEM) fields tend to receive lower ratings on RMP compared to those in the humanities and arts, suggesting potential disciplinary biases. Despite the concerns regarding validity and biases that have been a subject of ongoing debate, RMP remains popular among students, with millions of users relying on it to inform their course selections (Boswell & Sohr-Preston, 2020). RMP has several limitations. However, it also offers valuable insights into student perceptions and preferences. RMP can provide instructors with feedback on various aspects of their teaching, including their rapport with students, communication skills, and classroom management. By analyzing RMP data, researchers can gain a deeper understanding of the factors that students consider important in their learning experience. This information can be used to improve teaching practices and enhance student satisfaction. Studies have suggested that RMP comments and qualitative feedback can provide insights into effective teaching practices (Hartman & Hunt, 2013). However, limited research has explored its use as a tool for identifying best practices in teaching.

More recently, AI techniques have been applied to analyze educational data and provide insights into teaching practices. The authors of a study (Sutoyo et al., 2020) used Machine Learning techniques, including sentiment analysis and natural language processing (NLP) frameworks such as BERT to analyze student comments from course evaluations. They identified key themes such as course content, teaching style, and assessment methods that influenced student satisfaction and learning outcomes. Their findings highlighted the importance of engaging students, providing clear explanations, and offering timely feedback. Also, the authors of (Wang et al., 2020) found that BERT was effective at identifying themes and sentiments in the comments, outperforming traditional machine learning approaches. These studies provided insights into student perceptions and learning outcomes in CS education and demonstrated the growing interest in using LLMs to analyze SET and RMP data. However, more research is needed to fully understand the potential and limitations of LLMs in this domain. There remains a gap in leveraging the rich qualitative data available in RMP reviews to identify best practices.

3. MATERIALS AND METHODS

Recently, advancements in Machine Learning have led to the development of powerful LLMs capable of understanding and generating human-like text very accurately. One of their viable applications in education consists of analyzing large volumes of unstructured data, such as student reviews of professors, whether from SET or other sources and processing them in a way that provides instructors and students with more actionable items for improvement or selection. As the goal of our work is to enhance the assessment of professors' teaching quality to benefit instructors and students, in this study, we investigate the use of LLMs to analyze publicly available reviews of computer science professors and extract key features that can inform and improve pedagogical practices as well as guide students in succeeding in academic courses. In this regard, the massive dataset offered by RMP is an exceptional testbed to evaluate different approaches based on LLMs, their feasibility, and their performances.

Instead of focusing on quantitative ratings such as professor quality, difficulty, and whether students would take the course again, our strategy takes a qualitative approach to the analysis of textual reviews on RMP. Our methodology consists of the following steps.

1. Code the emerging themes identified from the student reviews, and we define a set of key dimensions of teaching quality that will be utilized to assess all professors.
2. Process a professor's reviews and present a coherent and concise overview of the professor's classroom experience based on the dimensions identified in the previous step. This information provides prospective students with insights into course selections, and it can be utilized by the instructor to improve their teaching.
3. Based on the insights extracted from professors' reviews, define suggestions that can help students succeed in the class.

LLMs can assist in every step of this process, from identifying key themes and their relationships to summarizing the findings from the coded themes to generating summaries that capture the essence of the expected classroom experience and a list of insights for student success based on the most salient points. Throughout this process, we leverage the LLM to abstract from aspects of the original reviews that can influence students negatively, such as the sentiment of the reviewer and their ability to articulate their opinions. Furthermore, we filter out inappropriate information, including sexist comments (Boswell & Sohr-Preston, 2020). The output of the process consists of (1) a single review that conveniently summarizes key aspects of a professor's teaching and (2) a set of actionable suggestions that students can use to succeed in the courses.

Data collection

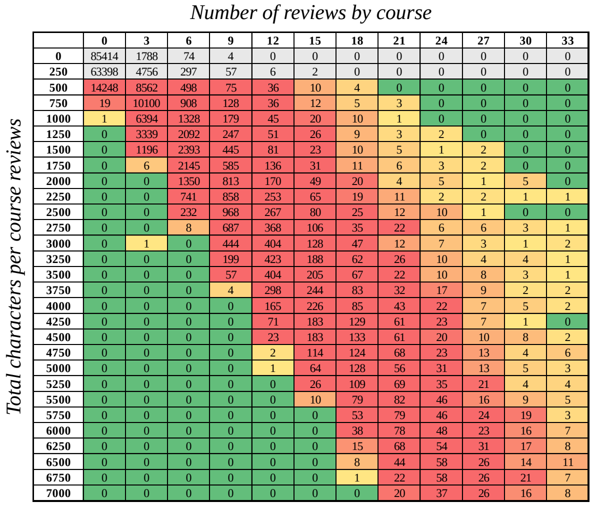
To obtain the dataset for our study, we developed software that automatically retrieved data from RMP using GraphQL, a query language for Application Programming Interfaces (APIs). GraphQL enabled us to query RMP's server and specify the exact data fields required for our analysis. This approach allowed us to efficiently collect complete information about schools, professors, and their associated ratings. The initial dataset consisted of a total of 9,244 schools, 2,050,784 professors, and over 23,311,429 ratings. After retrieving the initial dataset, we applied a filtering process to narrow the scope of our study to professors specifically within CS departments. We focused on a single academic field to extract more targeted information and insights and actionable insights that are directly relevant to CS education. Therefore, we limited our dataset to 727,315 reviews from 227,687 individual CS courses taught by 49,147 professors at 3,502 schools.

Pre-processing

Subsequently, we pre-processed our data to filter out irrelevant reviews. In fact, many students' comments involve just a few characters or a single word, or reviews such as "no comment'', lacking useful information. Therefore, we removed a total of 12,099 professors whose reviews accounted for a total of less than 500 characters, regardless of the number of reviews, as shown in the first two lines of Figure 1. By doing this, we avoided analyzing reviews that, in addition to providing very little insight into the course experience, would cause the LLM to generate inaccurate content. Also, we removed a total of 2,471 professors with a large number of reviews accounting for more than 12,000 characters in total. As these professors would take too long to process, we prioritized shorter reviews to test the feasibility of our system. Therefore, we restricted our initial analysis to a total of 34,577 professors (i.e., 70.35% of the dataset). We did not process individual course reviews because it would result in a higher data sparsity in terms of the number of reviews and content and, consequently, limit the generalizability of our findings. In fact, reviews of 155,796 courses (i.e., 68.42% of the dataset) had less than 500 characters and, thus, would not be suitable for a comprehensive analysis.

LLM selection

For our analysis, we utilized Llama 3, an open-source LLM developed by Meta. Compared to its predecessors, Llama 3 exhibits better alignment with user instructions, leading to more accurate and relevant responses, and offers a more diverse range of answers. Before choosing Llama 3, and specifically, the model trained with 8 billion parameters, we tested several other open-source LLMs, including Gemma, Mistral, and Phi3, on a subset of the dataset consisting of 100 reviews.



Number of reviews by course

Total characters per course reviews

	0	3	6	9	12	15	18	21	24	27	30	33
0	85414	1788	74	4	0	0	0	0	0	0	0	0
250	63398	4756	297	57	6	2	0	0	0	0	0	0
500	14248	8562	498	75	36	10	4	0	0	0	0	0
750	19	10100	908	128	36	12	5	3	0	0	0	0
1000	1	6394	1328	179	45	20	10	1	0	0	0	0
1250	0	3339	2092	247	51	26	9	3	2	0	0	0
1500	0	1196	2393	445	81	23	10	5	1	2	0	0
1750	0	6	2145	585	136	31	11	6	3	2	0	0
2000	0	0	1350	813	170	49	20	4	5	1	5	0
2250	0	0	741	858	253	65	19	11	2	2	1	1
2500	0	0	232	968	267	80	25	12	10	1	0	0
2750	0	0	8	687	368	106	35	22	6	6	3	1
3000	0	1	0	444	404	128	47	12	7	3	1	2
3250	0	0	0	199	423	188	62	26	10	4	4	1
3500	0	0	0	57	404	205	67	22	10	8	3	1
3750	0	0	0	4	298	244	83	32	17	9	2	2
4000	0	0	0	0	165	226	85	43	22	7	5	2
4250	0	0	0	0	71	183	129	61	23	7	1	0
4500	0	0	0	0	23	183	133	61	20	10	8	2
4750	0	0	0	0	2	114	124	68	23	13	4	6
5000	0	0	0	0	1	64	128	56	31	13	5	3
5250	0	0	0	0	0	26	109	69	35	21	4	4
5500	0	0	0	0	0	10	79	82	46	16	9	5
5750	0	0	0	0	0	0	53	79	46	24	19	3
6000	0	0	0	0	0	0	38	78	48	23	16	7
6250	0	0	0	0	0	0	15	68	54	31	17	8
6500	0	0	0	0	0	0	8	44	58	26	14	11
6750	0	0	0	0	0	0	1	22	58	26	21	7
7000	0	0	0	0	0	0	0	20	37	26	16	8

Figure 1 Distribution of CS reviews by number of reviews per professor and total characters (excerpt).

LLM priming

To ensure the relevance and accuracy of the extracted themes, we initially extracted a set of pedagogical keywords and themes that guided the design of our system prompt to the LLM. To this end, we asked GPT-4 to analyze reviews for over 10,000 professors and extract key themes representing various aspects of teaching and learning. The LLM priming process involved an initial extraction of pedagogical keywords and themes from 10,000 rows of review data using GPT-4. This approach was validated through manual cross-verification to ensure that the themes accurately represented key dimensions of teaching quality, such as teaching style, student interaction, and assessment fairness. Figure 2 represents a word cloud of the most common elements found in reviews. This step was key to informing our coding process. In fact, based on these pedagogical themes, we identified the following five dimensions that were most pertinent to a student's experience:

- *Teaching style and classroom environment*: teaching methods, ability to engage students, and positive learning atmosphere.
- *Learning approach and course content*: organization and presentation of relevant course content, use of assignments and projects.
- *Participation and interaction*: encouraging student participation, being responsive to feedback, and availability outside of class.
- *Workload and expectations*: clear communication of course requirements, reasonable workload distribution, appropriate academic challenge, fair grading practices, and clear workload and expectations
- *Overall experience*: the overall classroom experience is determined by the professor's teaching effectiveness, and ability to enhance student interest and engagement.

The following system prompt was utilized to prime the LLM.
You will be given a professor's review and you will produce a description of the professor based on all the following aspects: - teaching style and classroom environment; - learning approach and course content; - participation and interaction; - workload and expectations; - overall experience. For each dimension, calculate a score from 1 to 5 based on the sentiment of the review. Absolutely describe all the 5 aspects. Finally, produce a list of suggestions for prospective students taking the professor, especially in computer science disciplines. Avoid mentioning the name of the professor and the reviews.

Data processing

We utilized the model on a client using Ollama, an open-source project designed to simplify the process of running LLMs on local machines. Ollama acts as a standard interface for interacting with an LLM, and it supports a growing number of models, many of which Open Source. To process the dataset, we developed a custom JavaScript program that utilized Ollama's node package as an interface to query the LLM. The script was executed in a NodeJS environment on a computer equipped with a multi-core 12th gen Intel(R) i7-12800H processor with an NVidia RTX A2000 graphic card equipped with 8GB RAM and Cuda-enabled GPU.

Post-processing

After processing the data, we assessed the LLM's output based on the following dimensions.

1. *Completeness*, that is, the presence of all the required elements, that is, (1) an analysis of each of the five key dimensions of teaching, (2) a numeric score for each dimension, and (3) the list of suggestions on how to succeed.

2. *Correctness*: whether the summary generated by the LLM reflected the content of students' original review.
3. *Consistency*: the LLM's ability to generate consistent output, including formatting of text, scores, and lists.
4. *Appropriateness*, including relevance of the information, use of an appropriate tone, and absence of inappropriate comments.
5. *Efficiency*, that is, the ability of the LLM to produce an effective summary without being too dry or verbose.

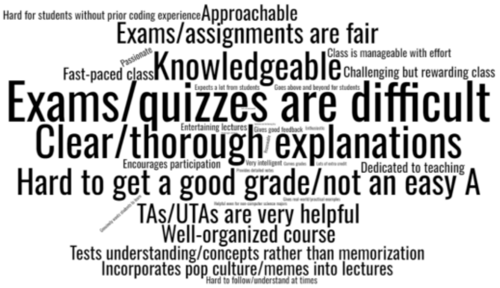


Figure 2 Word cloud of salient themes

4. RESULTS AND ANALYSIS

Processing all the reviews required approximately 84 hours. The LLM generated responses consisting of an average of 182.97 tokens, which took an average of 8805.30ms to generate.

The list below presents some examples of the summaries generated by the LLM based on the reviews. Specifically, the examples present excerpts of each of the five dimensions from different professors' reviews.

- *Teaching style and classroom environment: 5/5. The professor gives great lectures, uses YouTube videos as a teaching tool, provides easy-to-understand overviews of the weeks, and is a great teacher, which really helps you learn the material rather than memorize it.*
- *Learning approach and course content: 3/5. The professor is knowledgeable, but the way he teaches may require students to rely on outside sources or teach themselves the material. The course content is challenging in the second half.*
- *Participation and interaction: 3.5/5. While the professor is not overly engaging or interactive in his lectures, he does seem to be available for one-on-one office hours, which can be helpful for students who need additional support.*
- *Workload and expectations: 4/5. The professor sets realistic expectations for assignments and projects, with clear guidelines and deadlines. While the workload can be challenging at times, it is manageable, and students are encouraged to ask for help if needed.*
- *Overall experience: 2.5/5. Students have mixed opinions about the professor's teaching style, with some finding him boring or dry and others appreciating his willingness to help.*

The examples show the LLM's ability to summarize effectively the key dimensions of professors' classroom experience, providing insight into aspects of teaching that can be useful for instructors and students. Simultaneously, the excerpts demonstrate the LLM's ability to abstract from reviewers' sentiments and produce a balanced reflection of teachers' qualities. For instance, in the fifth item of the list above, some students' reviews incorporated angry comments and inappropriate remarks, which the LLM successfully filtered out and reframed using a constructive tone. The latter aspect highlights the importance of using LLMs with high alignment and proper safeguards.

Performance evaluation

The post-processing step described in the previous section enabled us to evaluate the LLM's performance and, consequently, the feasibility and efficacy of our approach. To this end, using data produced from the post-processing parser described in the previous section, we analyzed quantitative dimensions (i.e., completeness, consistency, and efficiency) in all the 34,577 summaries generated by the LLM.

Figure 3 represents the completeness of the output of the LLM. Most summaries (i.e., 73%) included all five elements, whereas the remaining 27% lacked comments on one or more of the dimensions of teaching qualities. This is because some students' reviews did not include comments that enabled the LLM to generate an appropriate summary. Also, 68% of LLM-generated reviews included a score for each dimension. A closer look at the content of some reviews revealed that although the information generated by the LLM is incomplete, in these circumstances, the system behaved correctly: instead of making up content, it simply avoided producing any. The score was completely missing in 19% of the reviews. This is because of the missing information described previously. However, in this case, the issue is also caused by an inconsistency in the results produced by the LLM. A mitigation strategy, in this case, would consist of either requiring the LLM to regenerate the review entirely or prompting the LLM to produce a score for each dimension present in the generated output. As far as the completeness of suggestions is concerned, the system provided two or more suggestions in 81% of the cases, whereas 17% of the reviews did not incorporate any recommendations. As in the previous case, this issue can be mitigated by requiring the LLM to process the original review and by deliberately asking it to only produce suggestions by conditioning the system prompt accordingly.

As far as the consistency of the output is concerned, our analysis primarily focused on syntactical aspects such as the formatting of lists and scores. LLMs produce Markdown-formatted output. Specifically, lists, including the dimensions of teaching quality and suggestions for academic success, were represented using the "-" symbol (i.e., unordered) and numbers (i.e., ordered) in 44% and 47% of the cases, respectively. In the remaining 9% of the cases, the output was unstructured. In the former situation, the parser was able to reconcile the items in the lists, in the latter scenario, the solution is to prompt the LLM to regenerate the output. Furthermore, when present (i.e., in 81% of the cases, as discussed above), scores were represented as a number (i.e., 3, or 5) in 42% of the cases and as a number with respect to its maximum value (i.e., 3/5, or 2.5/5) in 58% of the cases. The parser could handle such cases without requiring further processing.

For cases where the LLM-generated summaries were incomplete or inconsistent, a more detailed review revealed that this typically occurred in reviews with sparse content or ambiguous language. When a review lacked sufficient detail, the LLM occasionally omitted one or more dimensions of teaching quality, leading to incomplete summaries. Similarly, inconsistencies in formatting were more common in reviews with non-standard phrasing or excessive repetition of themes. A potential strategy for improving incomplete outputs would involve prompting the LLM to regenerate the summary when key dimensions are missing. This could be achieved by setting minimum thresholds for data content, requiring the model to extract themes from multiple reviews rather than relying on sparse or brief input. Additionally, a fallback mechanism could request the LLM to provide suggestions for improving the reviews when a lack of data prevents a complete analysis, though this could result in content that is not present in the original review. Inconsistent formatting could be addressed through better prompt engineering. For example, by enforcing specific formatting rules within the system prompt (e.g., always use numbered lists for suggestions), we can ensure a more consistent structure across all outputs. Also, in our future work, we plan to integrate post-processing tools to standardize the final output format, resolving inconsistencies without requiring reprocessing of the original data. r professors with limited reviews, the LLM struggled to provide complete summaries due to a lack of data. One strategy to improve accuracy in these cases would be to aggregate reviews over multiple courses or time periods, allowing the LLM to analyze a broader dataset and generate more complete summaries. However, our strategy of choice is to include a fallback option to indicate that insufficient data is available to generate a fully detailed summary, ensuring that the output remains informative without misrepresenting the review data. We will implement this in our future work.

The last quantitative dimension considered in our analysis is the efficiency of the system, measured as the ability of the LLM to produce comprehensive reviews in a concise format. The average review length was 2054±923 characters with a mode of 1926 characters. In 27,065 cases (i.e., 78%), the LLM generated reviews ranging between 1,000 and 3,000 characters, which is an appropriate length. In 3,394 cases (i.e., 9%), reviews were considered too short, whereas in 4118 instances (i.e., ~12%), they were too long.

Moreover, we evaluated correctness and appropriateness by sampling 500 LLM-generated reviews at random from six categories, that is, reviews with high and low completeness scores, consistency, and efficiency. As far as the correctness of the reviews is concerned, we did not find any LLM-generated summary that did not match the content of the original review. This is an indication of the performance of the LLM, its ability to limit hallucinations, and its high alignment. Some items included in the suggestions consisted of general advice that was not necessarily part of the original review, which is not necessarily a concern, given the purpose of our approach. We found a strong correlation between appropriateness and the other dimensions of our analysis, with specific regard to completeness and consistency: out of the 500 summaries produced by the LLM and analyzed manually, all the outputs that scored 70% and above in the quantitative dimensions had appropriate content and did not raise any specific concern in terms of appropriateness. On the contrary, we found that in three cases, our LLM-generated summaries contained a somewhat negative tone resulting from the original student's comment, which was left unfiltered (e.g., "*If you really wanna learn from the class, it's all up to you*"). Based on our evaluation, these circumstances can be addressed by filtering out any output ranking low in completeness, correctness, and consistency.

	0	1	2	3	4	5
Summary	0.06	0.01	0.02	0.06	0.12	0.73
Score	0.19	0.01	0.01	0.05	0.06	0.68
Suggestions	0.17	0.02	0.09	0.28	0.25	0.19

Figure 3 Performance evaluation statistics

5. DISCUSSION

Our study focused on identifying key themes and aspects relevant to pedagogy in CS education, regardless of whether the reviews were positive or negative, by abstracting from arbitrary quantitative measures of teaching quality or bias caused by reviewers' sentiment. This approach has several advantages. By ignoring quantitative scores, the study provides a more comprehensive understanding of the key factors that influence student learning experiences. This holistic approach ensures that the identified themes are not biased towards only favorable aspects of teaching. Furthermore, considering both positive and negative reviews offers a balanced perspective on educators and their teaching practices. This approach acknowledges that even highly regarded professors may have areas where they can enhance their teaching, while professors with mixed reviews may still exhibit strengths in certain aspects of pedagogy. Finally, analyzing reviews across the spectrum of sentiment helps extract suggestions relevant to students' academic success.

Indeed, our study suffers from the same limitations as other works based on RMP. As discussed in previous literature, publicly available reviews left spontaneously by a relatively limited number of individuals may not be representative of all experiences. For instance, students who are highly satisfied or dissatisfied may be more likely to leave reviews, leading to a potential bias in the data. Although this could influence the identified categories and themes captured in the paper and their relative importance, we addressed this concern by expanding our sample to many reviews across professors teaching different courses at numerous institutions. Furthermore, by abstracting from sentiment, our approach enables leveraging negative reviews as items students can consider. Another limitation lies in the LLM's ability to interpret subjective student feedback. While the model filters out inappropriate or biased language, there is still the potential for subtle biases in the data to influence the output. The LLM's reliance on sentiment analysis to score teaching dimensions may inadvertently overemphasize negative reviews, as students who are dissatisfied are more likely to leave detailed feedback.

It is important to clarify that the final dataset was indeed aggregated based on individual professors, but our objective was to distill general pedagogical themes rather than provide course-specific guidance. While this aggregation could limit granularity at the course level, we believe that patterns in teaching style, classroom engagement, and assessment methods often transcend specific courses. Thus, while the system produces summaries for professors across all courses they teach, these summaries reflect common pedagogical elements relevant to students' overall success. Nevertheless, we acknowledge this limitation and suggest future work could focus on extracting course-specific insights by refining the granularity of the data to individual course reviews, particularly for professors with a larger dataset of comments across various courses.

Another limitation in our study is related to the limited contextual information about the specific course, student background, or learning conditions. As the context is rarely captured in reviews, the lack of information could lead to an oversimplification of the complex dynamics of teaching and learning. Therefore, our analysis could fail to fully understand the factors contributing to a student's positive or negative experience. However, this problem is inherent in other forms of evaluations of teaching, including SETs, which rarely capture contextual information. Nevertheless, the categories and themes identified in our study provide further studies with a taxonomy for qualitative and quantitative research studies on contextual factors, including courses, student demographics, and learning conditions.

Despite these limitations, the study's approach of focusing on key themes and aspects relevant to pedagogy, regardless of the sentiment of the reviews, provides valuable insights into the factors that shape student learning experiences in CS education. Educators can use these findings to reflect on their own pedagogical approaches and develop strategies to enhance student learning outcomes. Simultaneously, our approach provides prospective students with a more in-depth analysis of reviews left by past students, offering insight into the classroom experience and suggesting ways to prepare for the course. While previous studies analyzed RMP's reviews to investigate the dimensions of teaching, offering actionable items based on students' reviews is an original contribution to our approach.

Several aspects of our paper are innovative with respect to the state of the art. The previous use of RMP data has been limited to individual instructor evaluations without systematically identifying generalizable teaching themes across disciplines. Our approach differentiates itself by focusing on extracting broader pedagogical insights that are applicable across courses and instructors, aiming to provide actionable feedback to students on how to succeed in specific courses. This is in contrast to previous studies, which primarily assessed individual instructor performance based on RMP scores (Timmerman, 2008). By utilizing Large Language Models (LLMs), our methodology abstracts from the individual biases present in RMP reviews and identifies recurring pedagogical themes, such as teaching style and classroom management, which can inform both students and instructors.

Additionally, the literature demonstrates that RMP data can be biased by factors unrelated to teaching quality, such as professor attractiveness, gender, or discipline (Legg & Wilson, 2012). Our proposed method addresses these biases through a multi-step filtering process that removes irrelevant content, such as personal remarks or emotionally charged comments, ensuring that the focus remains on pedagogical aspects that contribute directly to educational outcomes. The LLM also abstracts sentiment and evaluates reviews based on themes of teaching effectiveness, rather than subjective judgments that often dominate online evaluations.

While previous works, such as Sutoyo et al. (2020), have applied sentiment analysis and NLP frameworks like BERT to educational reviews, their focus was primarily on identifying sentiments and themes related to student satisfaction. Our study improves upon this by shifting the focus from student satisfaction to actionable pedagogical insights aimed at enhancing both teaching effectiveness and student success. Unlike sentiment analysis, which often overemphasizes emotional responses, our LLM-based approach seeks to provide a balanced and constructive analysis of teaching practices, offering not only a thematic breakdown but also concrete recommendations for both instructors and students. This methodological shift addresses the gaps left by prior studies, which often overlook the deeper pedagogical implications of student feedback.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented a study aimed at providing teachers and students with actionable insights into classroom experiences, to offer suggestions for improving the quality of teaching and, simultaneously, helping students succeed in their courses. To this end, we leveraged the vast amount of information available on RMP, a popular platform where students rate their professors on various criteria such as helpfulness, easiness, and quality of lectures. Several previous studies focused on the analysis of aspects such as the validity of the data collected by the platform, the assessment of professors' quality, and the sentiment of the reviews. On the contrary, our methodology introduces a novel approach to processing students' comments and extracting meaningful content that contributes to teaching effectiveness and student success rather than focusing on elements that do not directly impact educational outcomes.

To this end, after gathering the entire dataset of professor reviews, we filtered them to include only instructors teaching CS courses. Then, our analysis employed a mixed-methods approach based on the use of LLMs to analyze the qualitative reviews and the quantitative evaluation of the performance of the LLM. The primary objective of our study was to extract insights into teaching quality, professor-student interactions, and course content from user-generated reviews. We utilized large language models, particularly Llama3, for natural language processing tasks to handle the vast amount of unstructured text data. Specifically, we asked the LLM to create a summary that represented the classroom through five key dimensions, that is, (1) teaching style and classroom environment, (2) learning approach and course content, (3) participation and interaction, (4) workload and expectations, and (5) overall experience. For each dimension, the LLM also assigned a quality score on a scale from 1 to 5 to provide students with a numeric indicator. Finally, based on the instructor's classroom experience, the LLM identified suggestions to help the students succeed.

Our findings demonstrate the potential of LLMs and data-driven approaches to analyze a vast number of reviews, identify best practices, and offer practical guidance for improving CS education and student outcomes. For educators, our analysis highlights effective teaching strategies and areas for improvement. For students, we offer suggestions and tips to excel in their chosen CS courses based on the collective experiences shared by their peers.

Based on the findings of this study, we propose several practical recommendations for implementing LLM-generated insights in educational practice. Educators could use LLM-generated insights as a complementary tool to improve their teaching practices. The summaries can provide a high-level view of student feedback, offering a more comprehensive understanding of their teaching effectiveness. The ability of LLM-based reviews to focus on recurring themes, such as classroom interaction and workload expectations, can help them make targeted adjustments that enhance student engagement and learning outcomes. As it relates to students, LLM-generated summaries can help students make more informed decisions when selecting courses or preparing for classes. By reviewing the pedagogical themes and recommendations, students can better understand what to expect in a course and how to succeed, rather than being influenced by the sentiment of the review, as reported by Boswell & Sohr-Preston (2020). For example, insights about workload expectations or participation requirements can help students plan their time more effectively. Finally, institutions could leverage LLM-generated insights to inform curriculum development and faculty evaluations. Thematic analysis of student feedback can identify broader trends in teaching quality, allowing departments to address systemic issues that may be hindering student success. Additionally, institutions could use these insights to develop professional development programs tailored to the specific needs of educators, enhancing teaching practices across departments.

9. REFERENCES

Boswell, S. S., & Sohr-Preston, S. L. (2020). I checked the prof on ratemyprofessors: effect of anonymous, online student evaluations of professors on students' self-efficacy and expectations. Social Psychology of Education, 23(4), 943–961. https://doi.org/10.1007/s11218-020-09566-y

Coladarci, T., & Kornfield, I. (2007). Ratemyprofessors.com versus formal in-class student evaluations of teaching. Practical Assessment, Research & Evaluation, 12(6), 1–15.

Felton, J., Koper, P. T., Mitchell, J., & Stinson, M. (2008). Attractiveness, easiness and other issues: Student evaluations of professors on ratemyprofessors. com. Assessment & Evaluation in Higher Education, 33(1), 45–61. https://doi.org/10.1080/02602930601122803

Gordon, N., & Alam, O. (2021). The role of race and gender in teaching evaluation of computer science professors: A large scale analysis on ratemyprofessor data. Proceedings of the 52nd ACM Technical Symposium on Computer Science Education, 980–986. https://doi.org/10.1145/3408877.3432369

Hartman, K. B., & Hunt, J. B. (2013a). What RateMyProfessors. com reveals about how and why students evaluate their professors: A glimpse into the student mind-set. Marketing Education Review, 23(2), 151–162. https://doi.org/10.2753/MER1052-8008230204

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. Cogent Education, 4(1), 1304016. https://doi.org/10.1080/2331186X.2017.1304016

Kindred, J., & Mohammed, S. N. (2005). "He will crush you like an academic ninja!": Exploring teacher ratings on ratemyprofessors. com. Journal of Computer-Mediated Communication, 10(3), JCMC10314. https://doi.org/10.1111/j.1083-6101.2005.tb00257.x

Legg, A. M., & Wilson, J. H. (2012a). RateMyProfessors. com offers biased evaluations. Assessment & Evaluation in Higher Education, 37(1), 89–97. https://doi.org/10.1080/02602938.2010.507299

Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. Studies in Educational Evaluation, 54, 94–106. https://doi.org/10.1016/j.stueduc.2016.12.004

Luxton-Reilly, A., Albluwi, I., Becker, B. A., Giannakos, M., Kumar, A. N., Ott, L., … Szabo, C. (2018). Introductory programming: A systematic literature review. Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, 55–106. https://doi.org/10.1145/3293881.3295779

Robins, A., Rountree, J., & Rountree, N. (2003). Learning and teaching programming: A review and discussion. Computer Science Education, 13(2), 137–172. https://doi.org/10.1076/csed.13.2.137.14200

Rosen, A. S. (2018). Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors. com data. Assessment & Evaluation in Higher Education, 43(1), 31–44. https://doi.org/10.1080/02602938.2016.1276155

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. Review of Educational Research, 83(4), 598–642. https://doi.org/10.3102/0034654313496870

Stephenson, C., Miller, A. D., Alvarado, C., Barker, L., Barr, V., Camp, T., … Others. (2018). Retention in Computer Science Undergraduate Programs in the U.S.: Data Challenges and Promising Interventions. ACM New York, NY, USA. https://doi.org/10.1145/3406772

Sutoyo, E., Almaarif, A., & Yanto, I. T. R. (2020). Sentiment analysis of student evaluations of teaching using deep learning approach. The International Conference on Emerging Applications and Technologies for Industry 4.0, 272–281. Springer. https://doi.org/10.1007/978-3-030-80216-5_20

Timmerman, T. (2008). On the validity of Ratemyprofessors.com. Journal of Education for Business, 84(1), 55–61. https://doi.org/10.3200/JOEB.84.1.55-61

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty’s teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. Studies in Educational Evaluation, 54, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Wang, W., Zhuang, H., Zhou, M., Liu, H., & Li, B. (2020). What makes a star teacher? A hierarchical BERT model for evaluating teacher’s performance in online education. arXiv Preprint arXiv:2012. 01633.