JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH AND ANALYTICS

Volume 18, No.4
December 2025
ISSN: 1946-1836

In this issue:

iscap
Information Systems & Computing Academic Professionals

The **Journal of Information Systems Applied Research and Analytics** (JISARA) is a double-blind peer reviewed academic journal published by ISCAP, Information Systems and Computing Academic Professionals. Publishing frequency is three issues a year. The first date of publication was December 1, 2008. The original name of the journal was Journal of Information Systems Applied Research (JISAR).

JISARA is published online (https://jisara.org) in connection with the ISCAP (Information Systems and Computing Academic Professionals) Conference, where submissions are also double-blind peer reviewed. Our sister publication, the Proceedings of the ISCAP Conference, features all papers, teaching cases and abstracts from the conference. (https://iscap.us/proceedings)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%) and other submitted works. The non-award winning papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in JISAR. Currently the acceptance rate for the journal is approximately 35%.

Questions should be addressed to the editor at editor@jisara.org or the publisher at publisher@jisara.org. Special thanks to members of ISCAP who perform the editorial and review processes for JISARA.

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH AND ANALYTICS

Editors

2025 JISARA Editorial Board

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

Rajani Khatri
khatrir2@mymail.nku.edu

Will Pond
pondw1@mymail.nku.edu

Na Le
len4@mymail.nku.edu

Northern Kentucky University
Highland Heights, KY 41099

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. Although we realized our analysis or publicly available professor reviews the proposed methodology can be utilized in the context of official Student Evaluation of Teaching. We discuss how the proposed method can be utilized to process instructors' reviews, highlight teaching strategies, and elicit actionable information for both students and educators. Also, we describe how the same approach could also be utilized to identify areas for potential improvement.

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

Recommended Citation: Caporusso, N., Doan, M., Acharya, B., Pandit, P., Shrestha, S., Khatri, R., Pond, W., Le, N., (2025). From Angry Reviews to Classroom Success: Using LLMs to Synthesize RateMyProfessors.com Data. *Journal of Information Systems Applied Research and Analytics*. v18, n4, pp 20-35. DOI# https://doi.org/10.62273/EBAP6519

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

Nicholas Caporusso, My Ham Doan, Bikash Acharya, Priyanka Pandit, Sushant Shrestha, Rajani Khatri, Will Pond and Na Le

1. INTRODUCTION

Education is continuously 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 institutions 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 students (Stephenson et al., 2018). Although there is a growing body of academic literature on pedagogical best practices, student voices and perspectives are often missing from this discourse (Robins et al., 2003). Indeed, student feedback is essential for professors to improve their teaching effectiveness and enhance learners' experiences. It helps professors identify their strengths and weaknesses, refine course design and content, adapt teaching methods and styles, address student concerns and challenges and promote student engagement and motivation, encourage self-reflection and professional growth, and align teaching with student needs and expectations.

To effectively solicit student feedback, institutions in higher education usually collect Student Evaluation of Teaching (SET) at the end of each course. SET are usually administered in the form of surveys with questions aimed at capturing students' view on aspects of teaching that are deemed as important such as clarity, competence, and classroom environment. Among various applications, SET provides instructors with insights that can be used to improve teaching quality and identify areas for professional development. In addition to feedback to professors, SET also informs personnel decisions at the administrative level (Coladarci & Kornfield, 2007). However, despite the widespread use of SET, the design of SET questionnaires and the analysis of the collected data often lacks a systematic approach, leading to fragmented and inconsistent utilization of the information across departments and institutions due to several factors. First, SET questionnaires generate a large amount of qualitative and quantitative data, making it challenging to process and interpret the information effectively (Spooren et al., 2013). Specifically, qualitative data requires careful coding and analysis to identify common themes and patterns in student feedback. As a result, without a standardized approach to data analysis, different departments and institutions may employ varying methods, leading to a lack of comparability and consistency in how SET data is used (Uttl et al., 2017). Also, the results of SET evaluations are often not publicly shared, leading to a lack of transparency and consistency, and poor student involvement in the debriefing process. As a result, they are perceived as being primarily used for evaluating individual instructors' performance rather than identifying broader trends and best practices in teaching (Hornstein, 2017). This, in turn, limits the potential for SET data to inform institutional policies, professional development initiatives, and the sharing of effective teaching strategies across departments and institutions. Furthermore, 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).

In the past decade, professor reviews platforms such as RateMyProfessors.com (RMP) have gained popularity because they address the unmet need of students to be able to access professor reviews before making enrollment decisions. Websites like RMP enable students to anonymously and publicly share their ratings, comments, and opinions on their teachers. To this end, and similarly to SET, RMP utilizes various quantitative criteria, including clarity, helpfulness, and easiness (Timmerman, 2008). Although its validity and usefulness have been questioned by scholars and educators, RMP offers a wealth of student reviews and opinions about courses and instructors. Indeed, RMP and similar platforms are not an official instrument, and SET surveys remain the most comprehensive and reliable source of student feedback for educators. Also, the reviews published on unofficial professor reviews websites are not moderated, and many reviews contain elements unrelated to pedagogy, including personal retaliation, inappropriate comments, and swear words. However, thanks to

their extensive publicly available longitudinal datasets, unofficial platforms like RMP could be utilized as a resource for experimenting novel solutions, particularly when SET data are not readily available, which is mostly the case. By examining RMP reviews, researchers focusing on SET can design, develop, and test novel systems for supporting students' and professors' experiences, ultimately potentially enhancing the quality of their instruction. Specifically, the similarities between SET and the reviews published on RMP make it possible to use RMP as a testbed to evaluate, for instance, whether solutions based on Natural Language Processing (NLP) can process unstructured information from textual and distill overarching themes and evidence-based insights.

This paper proposes a novel approach to analyzing data collected using SET surveys and extracting relevant information that can make it easier for professors, students, and administrators to draw insight from reviews. Specifically, in our work, we use Large Language Models (LLMs) and their capabilities in NLP tasks, including text classification and summarization. Our methodology utilizes an LLM-based pipeline that, starting from a large body of instructor reviews, (1) extracts a summary of the key dimensions and aspects of the learning experience (e.g., teaching style and classroom environment, learning approach and course content, participation and interaction, workload and expectations, and overall experience), (2) utilizes the key dimensions of teaching to process an instructor's reviews and generate a relevant summary of the learning experience provided to the students, (3) evaluates the students' sentiment on the dimensions of teaching effectiveness, and (4) suggests feedback for improvement. By leveraging the power of LLMs, our proposed method focuses on key pedagogical themes rather than on aspects that are not related to academic success, and it filters out irrelevant or biased information, including angry comments.

In our study, we focused on demonstrating the potential of LLMs and data-driven approaches to analyze a vast number of reviews, identify best practices, and offer practical guidance to students and professors. To this end, we validated our method using publicly available reviews posted on RMP. Nevertheless, the same approach can be utilized on official SET data.

2. RELATED WORK

SET surveys have emerged as the primary tool for assessing teaching effectiveness in higher education. However, the rise of online platforms like RateMyProfessors.com has provided students with an alternative avenue to share their opinions and experiences with professors and courses. Although SET remains the most comprehensive and institutionally recognized source of student feedback, the lack of availability of data hinders researchers' ability to investigate students' ratings, comments, and reviews. 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).
As a result, in the past decades, unofficial professor review platforms like RMP achieved increasing popularity thanks to their accessibility to end-users (i.e., students and professors) and researchers. To this date, RMP remains the largest dataset of professors' reviews, and it has attracted the attention of researchers interested in understanding its validity and potential utility for a variety of purposes.

In particular, 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, different 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.

Indeed, RMP has several limitations, and it should not be utilized as an official source of information for research regarding teaching experiences. However, its vast dataset of reviews offers an excellent resource for developing and testing systems before they are deployed on official SET. 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). Also, it offers valuable insights into student perceptions and preferences. Also, in addition to students using RMP for enrollment decisions, instructors and institutions might already be leveraging the data available on RMP for hiring decisions or to gain insight on various aspects of teaching, including rapport with students, communication skills, and classroom management.

Another aspect that makes RMP's dataset an interesting source of information for research studies is the nature of the data collection process, which is not mediated by questions designed by a specific institution. Therefore, by analyzing RMP data, researchers can obtain a deeper and broader understanding of the factors that students consider important in their learning experience. This information can be used to design solutions that improve teaching practices and enhance student satisfaction. To this end, although studies have suggested that RMP comments and qualitative feedback can provide insights into effective teaching practices (Hartman & Hunt, 2013), limited research has explored its use as a tool for identifying best practices in teaching. While the majority of research has centered on the validity and impact of RMP ratings, few studies utilized the content of RMP reviews as a source of insights for professors. One study utilized text analysis techniques to predict professor classifications based on student comments, revealing differences in the language used to describe "good" professors across various student groups. This study suggests that RMP reviews contain valuable information about student perceptions and priorities, which could be leveraged by professors to understand and adapt to their students' needs (Azab et al., 2016). The lack of studies analyzing the content of RMP reviews presents an opportunity for further research. By examining the themes, sentiments, and specific feedback contained within RMP comments, researchers could uncover actionable insights for professors looking to improve their teaching practices and better connect with their students. Such analyses could also shed light on the factors that contribute to student satisfaction and perceptions of teaching effectiveness, complementing the findings of traditional SET surveys.

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. Thus, there remains a gap in leveraging the rich qualitative data available in

RMP reviews to identify the best practices.

In this paper, we use the data from RMP as a testbed for an LLM-based solution ultimately aimed at processing reviews collected through SET surveys. Consequently, in our results we derive insights based on the content of the reviews from RMP to demonstrate the viability of our approach and validate our methodology rather than extracting information from the reviews. Nonetheless, the interaction dynamics of RMP, with specific regard to the ability of students to publish their comments anonymously, might also result in useful insights into learning experiences.

3. MATERIALS AND METHODS

The objective of our work is to automatically extract information from SET to enhance the assessment of professors' teaching quality to benefit instructors and students. Specifically, our goal is to leverage LLMs' capability of understanding and generating human-like text very accurately to analyze 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 intuitive and actionable information.

In this paper, we present the results of a study in which we investigated the use of LLMs to analyze professors' reviews and extract key features that can inform and improve pedagogical practices as well as guide students in succeeding in academic courses. 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 professor reviews, whether from RMP or official SETs. We designed a multi-step process for extracting different types of information from professor reviews, and we utilized publicly available data from an online website to validate our approach. To this end, the massive dataset offered by RMP is an exceptional testbed to evaluate different approaches based on LLMs, their feasibility, and their performances. In this phase, we are focusing on RMP because the nature of its data (i.e., the amount and it being publicly available) enables testing our method on a large number of reviews, validate our approach, and evaluate necessary improvements. In the next phase of our work, we aim to support data from official SET surveys.
After defining key dimensions of teaching effectiveness and student success, our proposed methodology consists in using LLMs to process individual professors' reviews, filter out irrelevant or inappropriate content, and extract the following outputs for each instructor or courses. The outputs are described in Figure 1.

1. A summary of the learning experience that students are expected to have with the professor or on the course and tips to perform well in the class. This primarily benefits students in their enrollment decisions, when they seek to know what kind of learning environment they will be in. In addition to providing prospective students with insights into course selections, this information can be utilized by the instructor to improve their teaching.
2. An analysis of the sentiment of the students, which can be utilized by professors to evaluate students' general perceptions and responses to their teaching style.
3. A list of actionable improvement items based on relevant students' suggestions. The instructor can use this information to quickly identify adjustments needed to accommodate an evolving audience.

In our study, we evaluated whether LLMs could assist in every step of this process, including summarizing a large number of reviews into an essential list of relevant feedback, capturing the expected classroom experience, and achieving insights that can be converted into suggestions for student success. By using LLMs, we aim to abstract aspects of the original reviews that can influence students and instructors negatively, such as the sentiment of the reviewer and their ability to articulate their opinions. Furthermore, this approach could also be utilized to filter out inappropriate information, including sexist comments (Boswell & Sohr-Preston, 2020), and provide the audience with a more polished digest.

In the context of official SET, each institution creates a survey with questions designed based on a predefined set of dimensions of teaching excellence and student success identified by a specific committee or unit. As a result, students' answers and reviews contain information collected from several questions each investigating one or more aspects. Therefore, using this top-down approach, the data collected from students' comments in the context of official SET reflect the aspects that are relevant for the institution. Conversely, in our case we utilized publicly available reviews collected in a bottom-up fashion from students. Therefore, the content was not guided or directed by any specific

dimensions, because RMP provides users with one text field only where they can enter their review. As we had no control over the data collection process, we could not make assumptions on the dimensions considered relevant by the students. Consequently, we used LLMs to also extract the most recurring topics and code and infer the relevant dimensions based on the content of students' reviews.

As a result, the steps in our process (also described in Figure 1) can be summarized as follows:

1. Collect the dataset. In the study presented in this paper, we utilized RMP's data. However, in regular application scenarios, the dataset is already collected by the institution and consists of course evaluations from SET.
2. Selection of professors. For the purpose of this study, we utilized a representative sample of RMP's dataset.
3. Pre-process the data to eliminate reviews that do not contain relevant information.
4. Extract the main themes from the content of the reviews.

Conversely, when applied to data from SET surveys, the process would be as follows.

1. Definition of quality metrics, which informs the creation of survey questions. Quality metrics would be defined top-down by the institution, whereas in our study they were extracted bottom-up from the content of the reviews.
2. Collection of the dataset, that is, administer course evaluations questionnaires to students and ensure a representative sample fills them out.
3. Process the data in a way similar to step 3 described above.

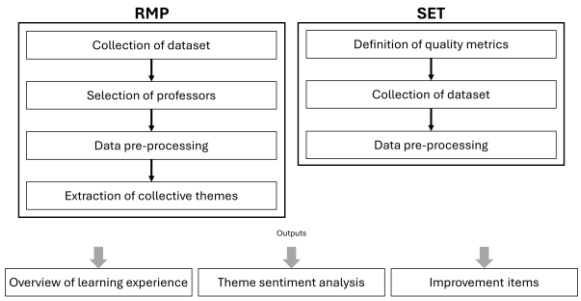

Figure 1 – An overview of our methodology and its different application with RMP's dataset and data from SET

3.1 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 in one discipline only. We focused on a single academic field, that is, Computer Science (CS), 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. Nevertheless, the methodology could be utilized for other disciplines or generalized and applied in transdisciplinary fashion, regardless of a particular academic area.

Then, we aggregated and processed all the reviews on an individual professor basis. Although our initial goal is to process single courses, the data collected by RMP consists of very few reviews for most courses and in a large number of reviews in a limited subset of courses. As the high variance and sparse number of reviews per course would result in many courses having insufficient information, which would result in a poor outcome. However, this limitation would not affect data collected via SET, which has significantly higher response rates.

3.2 Pre-processing

Subsequently, we pre-processed our data to filter out irrelevant reviews. To this end, we analyzed the distribution of reviews per course and number of characters per reviews, which is shown in Figure 2. As shown in the image, a large number of courses have less than 3 reviews and less than 250 characters, resulting in very limited information. 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 2. 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). As discussed earlier, we did not process individual course reviews because it would result in 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.

This step was realized manually, by filtering reviews based on their length and content. Working with SET datasets would require the same type of preprocessing, which could be realized by analyzing the text with quantitative techniques or using traditional NLP approaches.

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 2 Distribution of reviews by number of reviews per professor and total characters (excerpt).

3.3 LLM selection

The third step in our process was to select an LLM suitable for text summarization, sentiment analysis, and text generation tasks. Many recent models, including free and open-source models, are equipped to perform well in these tasks. The goal of our work was to study the feasibility of our approach and validate our methodology rather than evaluating and comparing the LLMs performances of a series of models. As a result, our criteria in choosing the model were primarily guided by the feasibility integrating the LLM into the process. We decided to utilize 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. Although their performances were similar, we chose Llama 3 because of its interoperability and openness to fine-tuning, which could be useful in our future work.

In our approach, we considered the LLM as a processing tool. Therefore, the model utilized in our study can be replaced by a different LLM that more appropriately or conveniently supports the specific use case or application scenario of the proposed approach.

In our study, 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 are 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.

3.4 Extraction of collective themes

In official SET surveys, students answer questions that investigate specific dimensions, which, in turn, can be utilized to guide the analysis of the content of the reviews. Conversely, as mentioned previously, one of the main limitations of using data from RMP as a test dataset is the unstructured way in which feedback is collected from users, with reviews being the result of one general text entry. In this context, prompting the LLM to analyze a review without any specific pointers results in a very general and inconsistent summary. Also, arbitrarily choosing dimensions of teaching excellence and student success would result in incorrect assumptions or in the LLM potentially generating text to fill out elements requested in the prompt that are missing in the data.

Therefore, we utilized the LLM to extract overarching themes that emerge across multiple reviews, in a process similar to manual coding in qualitative research. These themes could include common praise points, recurring concerns, or specific aspects of teaching that students frequently mention when providing feedback

about their professors. Identifying these collective themes helps understanding the broader patterns and trends. For example, themes could include the clarity of explanations, the helpfulness of feedback, the engaging nature of lectures, or the availability of resources.

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 3 represents a word cloud of the most common elements found in reviews. This step was key to informing our coding process.

A number of themes emerged from the analysis of all the professors' reviews. We initially grouped them into 12 overarching areas related to teaching effectiveness and student success. These represent the key aspects that students frequently mention when providing feedback about their professors. While most themes are applicable across disciplines, industry, and real-world connections emerged as particularly relevant to CS education, especially in contexts such as software engineering.

1. Teaching methods and styles, representing whether the professor uses clear communication, structured learning, technology integration, interactive and hands-on activities, flipped classroom models, digital tools, multimedia resources, visual aids, animations, interactive lectures, and dynamic teaching techniques.
2. Course content and design, which incorporates real-world examples, updates content regularly, uses interdisciplinary perspectives, practical applications, case studies, varied assessments, project-based learning, reflective assignments, cross-departmental projects, current research, and podcasts.
3. Student engagement and participation, describing whether the instructor utilizes gamification, provides incentives, encourages active participation through discussions and coding sprints, and uses interactive simulations, real-time polls, collaborative learning, peer-to-peer teaching, peer review, student-led discussions, study groups, student showcase events, and infographics.
4. Feedback and assessment, which represents whether the professor provides timely and constructive feedback, uses clear grading rubrics, conducts formative and frequent assessments, offers self-paced learning options, sets transparent expectations, and monitors student progress.
5. Classroom environment and management, that is, whether the professor maintains a structured and respectful environment, uses inclusive teaching practices, creates an engaging atmosphere, maintains open communication, focuses on student-centered learning, uses active learning techniques, and adapts to different learning styles and paces.
6. Student support and development, including whether the instructor establishes mentorship programs, provides resources and support, promotes well-being, offers professional development, encourages growth mindset, continuously improves, fosters partnerships, provides growth opportunities, uses early alert systems, encourages learning from mistakes, and helps balance academic and personal life.
7. Collaboration and interaction, evaluating whether the professor assigns group projects, encourages collaboration, solicits student input, uses peer review, promotes collaborative projects, and uses communication platforms, online collaboration tools, and interactive workshops.
8. Use of technology in teaching, measuring how the instructor incorporates relevant technology tools and platforms, uses digital tools, integrates technology seamlessly, uses online learning platforms, virtual and augmented reality, learning management systems, and adaptive learning technology.
9. Content delivery and resources, describing teaching methods, the instructor’s level of presence in the classroom, the use of modular assignments, online platforms, digital resource libraries, supplementary materials, recorded lectures, and optional workshops.
10. Industry and real-world connections, which are particularly relevant in technical disciplines, describing whether the instructor incorporates elements such as guest lectures,

builds industry connections, emphasizes real-world applications, aligns with professional standards, organizes guest speaker series, and collaborates with industry.

11. Continuous learning and improvement, representing whether students think that the instructor regularly updates content and methods, encourages professional development, promotes a growth mindset, implements feedback mechanisms, uses reflective assignments, and provides ongoing learning opportunities.
12. Flexibility and adaptability, representing whether the professor offers flexible deadlines, adapts teaching methods, uses adaptive learning technology, communicates expectations clearly, provides self-paced learning options, and implements early alert systems.

We did not quantify the occurrence of each theme in the reviews and weigh them based on the number of occurrences. This is because our goal was to identify all the key themes without necessarily setting a relevance threshold to scope the landscape of students' comments. Furthermore, associating any quantifiers to themes would introduce validity problems in our study, considering the concerns expressed by previous studies about the lack of completeness of RMP's data. Ultimately, this step was necessary only because of the characteristics of the RMP dataset.

Then, based on these pedagogical themes, we identified the following five dimensions that were most pertinent to a student's experience. This is to provide students with a more succinct summary highlighting the main aspects only.

1. *Teaching style and classroom environment*: the professor's teaching methods, ability to engage students, and create a conducive learning atmosphere define the classroom environment.
2. *Learning approach and course content*: the professor's organization and presentation of relevant, applicable course content, along with the use of assignments and projects, shape the learning approach.
3. *Participation and interaction*: whether the professor encourages student participation, being responsive to feedback, and maintaining availability outside of class characterize effective participation and interaction.
4. *Workload and expectations*: whether the professor establishes clear communication of course requirements, reasonable workload distribution, appropriate academic challenge, and fair grading practices define the workload and expectations.
5. *Overall experience*: the overall classroom experience is determined by the professor's teaching effectiveness, ability to enhance student interest and engagement, supportiveness, and the sense of accomplishment students gain from the course.

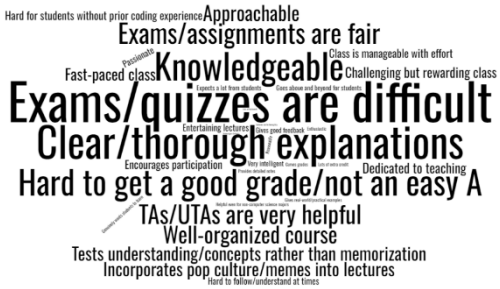

Figure 3 Word cloud of salient themes

3.5 Summary generation

After defining the five dimensions, we started feeding each professor's reviews into the LLM to generate a summary of their teaching experience. To this end, we used the five dimensions to generate the following system prompt, which 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." This prompt was designed to elicit a comprehensive analysis of the professor's performance across five key dimensions, along with a numerical score for each aspect and a list of suggestions for improvement. The model was reset before processing each review to prevent any influence from previous inputs on the LLM's output.

3.5 Sentiment analysis

Subsequently, we analyzed the sentiment associated with each of the 13 initial themes, with the aim to determine whether the themes are generally addressed by students as positive, negative, or neutral.

Our goal was to evaluate whether the LLM is able to capture the overall sentiment of students with respect to each dimension to provide the instructor with a prioritized list of improvement items, if any. To obtain an analysis of the students' sentiment in each professor's class, we processed individual professors' reviews using the following system prompt.

"*Based on the professor's review, rate each of the following aspects (if present) on a scale from -2 to +2 based on the sentiment of the review. 1. Teaching Methods and Styles; 2. Course Content and Design; 3. Student Engagement and Participation; 4. Feedback and Assessment; 5. Classroom Environment and Management; 6. Student Support and Development; 7. Collaboration and Interaction; 8. Use of Technology in Teaching; 9. Content Delivery and Resources; 10. Industry and Real-World Connections; 11. Ethics and Professionalism; 12. Continuous Learning and Improvement; 13. Flexibility and Adaptability. Only list the criteria and the numeric value*".

4. RESULTS AND ANALYSIS

4.1 Summary generation

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.

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.

This post-processing step 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 4 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 output 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 4 Summary generation - Performance evaluation statistics

4.2 Sentiment analysis

Analyzing the sentiment of each professor's reviews took a total of 4 hours. The performance of the model was evaluated by randomly sampling approximately 10% of the output, that is, 3,000 professors, and manually comparing the content of the reviews and the extracted sentiment, assigning a score from 1 to 5 based on the accuracy of the LLM in classifying the sentiment. On average, the resulting score was 2.7, which was unexpectedly low, considering LLMs' ability in sentiment analysis tasks. However, the main issue was that in many cases, the model inferred a sentiment score for all the 12 dimensions even if the review did not have any content related to some of the teaching evaluation and student success metrics. This was due to the following factors:

1. The nature of the dataset and, specifically, the data collection process, which did not capture content for each of the dimensions. We extracted the 12 themes by aggregating the content of all the reviews. However, some themes were not mentioned in many professors' reviews. This issue is inherently solved using data from official SET surveys, where each dimension has a corresponding answer.
2. The number and specificity of the dimensions was too high for the LLM to find enough content in each review. This issue would be solved as in 1.
3. The inherent nature of LLMs, which makes them "fill in the blank" in case of missing or incomplete input. In addition to the solution mentioned in the previous two points, this issue might be solved by using prompt-engineering and fine-tuning techniques.

Overall, this aspect requires further investigation and will be analyzed in a follow-up study.

4.3 Improvement items

As the sentiment analysis could not provide an accurate representation of each professor's performance over the 12 dimensions, instead of analyzing improvement items on a per-professor basis, we aggregated the results and asked the LLM to analyze the sentiment on the entire set of aggregated reviews and identify suggestions for improvement. The results of our sentiment analysis (see Figure 5) show that, on average, students have an overall slightly positive attitude toward their instructors. Specifically, teaching methods and styles (+1.52) and ethics and professionalism (+1.50) received the highest positive sentiment scores. By manually comparing the reviews and the generated sentiment analysis score, we found that professors who employ engaging, interactive, and well-structured teaching methods while also emphasizing the importance of professional ethics are likely to be viewed favorably by students. Course content and design (+0.88), use of technology in teaching (+1.13), content delivery and resources (+1.13), and continuous learning and improvement (+1.06) also received positive sentiment scores, suggesting that students appreciate well-organized and relevant course content the effective integration of technology, accessible learning resources, and a commitment to ongoing improvement. Professors who keep their course content up-to-date, leverage technology to enhance learning experiences, provide comprehensive resources, and demonstrate a willingness to adapt and improve their teaching are likely to be positively perceived by students. Classroom environment and management (+0.78), student support and development (+0.63), collaboration and interaction (+0.69), industry and real-world connections (+0.69), and flexibility and adaptability (+0.50) received moderate positive sentiment scores. These results suggest that students value a positive and inclusive classroom atmosphere, supportive and development-oriented learning environments, opportunities for

collaboration, connections to real-world applications, and a degree of flexibility in the learning process. While professors are generally doing well in these areas, there may be room for further improvement to enhance student experiences and outcomes. On the other hand, student engagement and participation (-1.00) and feedback and assessment (-0.83) received negative sentiment scores, indicating potential areas of concern for students. These results suggest that students may feel less satisfied with the level of engagement and interaction in their courses and may desire more effective feedback and assessment practices. Professors should focus on strategies to promote active learning, encourage student participation, and provide timely, constructive, and actionable feedback to address these concerns and improve student sentiment in these areas.

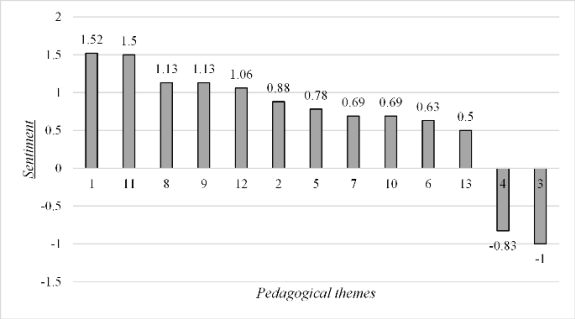

Figure 5 Sentiment analysis based on the aggregated sample of reviews

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.

After validating the feasibility of our methodology, in our future work, we will apply our proposed method to data from official SET surveys. This would enable us to address some of the limitations we encountered using RMP's dataset, with specific regard to the lack of specificity with respect to key dimensions of teaching excellence and student success.

9. REFERENCES

Azab, M., Mihalcea, R., & Abernethy, J. (2016). Analysing RateMyProfessors Evaluations Across Institutions, Disciplines, and Cultures: The Tell-Tale Signs of a Good Professor. Social Informatics, 438–453. https://doi.org/10.1007/978-3-319-47880-7_27

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.