JOURNAL OF INTERDISCIPLINARY PERSPECTIVES

ISSN Print: 2984-8288, ISSN Online: 2984-8385
Vol. 3 No. 8, 503-513, August 2025

AI-Driven Insights from Student Feedback for Teacher Improvement

Marie Grace V. Ortiz*, Menchita F. Dumlao
Graduate School, Angeles University Foundation, Angeles City, Philippines

*Corresponding Author Email: ortiz.mariegrace@auf.edu.ph

Date received: May 17, 2025
Date revised: June 20, 2025
Date accepted: July 14, 2025

Originality: 88%
Grammarly Score: 99%
Similarity: 12%

Recommended citation:
Ortiz, M. G., & Dumlao, M. (2025). AI-driven insights from student feedback for teacher improvement. *Journal of Interdisciplinary Perspectives*, *3*(8), 503-513. https://doi.org/10.69569/jip.2025.418

Abstract. This study used Natural Language Processing (NLP) and Artificial Intelligence (AI) to analyze ten years of teacher evaluations. The research leveraged VADER and NRC for sentiment analysis and LDA for topic modeling to extract key themes. The Google Gemini AI model then generated actionable recommendations for pedagogical improvement. Analysis of 9,052 textual comments revealed a predominantly positive (71%) to neutral (27%) comments, and LDA identified eight distinct topics. The AI-driven analysis successfully provided targeted suggestions for pedagogical enhancement, offering a pathway toward data-informed professional growth for educators. However, multilingual feedback presented challenges for comprehensive analysis.

Keywords: Natural language processing; Artificial Intelligence, Google Gemini, Sentiment Analysis; Teacher evaluations.

1.0 Introduction

Teacher evaluations are a cornerstone of quality education, serving multiple critical functions within educational institutions, including informing personnel decisions, ensuring accountability, identifying professional development needs, and providing insights into teaching effectiveness from various perspectives. The Student Evaluation of Teaching (SET) is a widely adopted practice in higher education institutions due to its ease of implementation in collecting, presenting, and interpreting data (Spooren et al., 2013); however, a growing body of research highlights its significant limitations. Studies suggest that SETs are not consistently reliable or valid measures of teaching performance (Ed, 2000; Spooren et al., 2013; Zabaleta, 2007) and may have a limited impact on improving teaching practices or student satisfaction (Leguey et al., 2023). Critically, SETs are biased against female faculty, personality, physical appearance, age, experience, ethnicity, and race, as well as other fixed factors (Ching, 2018; Hutchinson et al., 2023; Kreitzer & Sweet-Cushman, 2022). Furthermore, concerns regarding bias against women and faculty of color raise serious ethical questions about their use in personnel decisions. Consequently, relying solely on potentially biased numerical ratings from SETs may lead to inequitable assessments of teacher effectiveness (Simmons, 1997), highlighting the critical need for more robust and nuanced evaluation methods. Unlike conventional numerical ratings, which often generalize complex feedback and can be inherently subjective or biased, advanced analytical approaches are required to extract deeper, more objective, and actionable insights from the rich qualitative data available.

Teacher evaluations serve as a vital tool for gathering feedback on instructional practices and identifying areas for growth. Among the various sources of evaluation data, student comments offer a rich and direct perspective on the classroom environment and the impact of teaching strategies. However, the sheer volume and unstructured nature of these textual responses often present a significant challenge in extracting meaningful and actionable insights for teacher improvement. Traditional qualitative analysis of such large datasets is labor-intensive, time-consuming, and prone to human subjectivity, often leading to valuable student perspectives being overlooked or misinterpreted. Current evaluation systems frequently struggle to effectively use and analyze these objective data to fairly and reliably measure teacher performance and provide teachers with valuable, constructive feedback (Nasim et al., 2017; Rajput et al., 2016). Recognizing the inherent value of student perspectives as direct recipients of instruction (Balam & Shannon, 2010; Bill & Melidan Gates Foundation, 2012; Wine & Wine, 2016), this research posits that the wealth of information contained within their open-ended comments offers a significant, yet often underutilized, resource for understanding and improving teaching. The challenge lies in effectively analyzing these large volumes of unstructured textual data to extract meaningful and actionable insights.

To directly address the limitations of traditional SETs—particularly their inherent biases, limited depth of feedback, and the difficulty in systematically extracting nuanced, actionable insights from rich textual feedback—this research explores the powerful capabilities of Natural Language Processing (NLP) and Artificial Intelligence (AI) methodologies. Specifically, it leverages sentiment analysis (SA) techniques to analyze comments in teacher evaluations that are full of sentiments, emotions, feelings, appraisals, or negative remarks expressed in student comments (Esparza et al., 2018; Newman & Joyner, 2018; Lalata et al., 2019) and topic modeling (TM) algorithms to recognize the key themes and subjects discussed. The integration of these NLP methods with an AI-powered recommender system, built upon models like Google Gemini, presents an unprecedented opportunity to synthesize these analytical outputs into concrete and actionable recommendations for enhancing teaching practices. This approach aims to move beyond potentially biased numerical ratings and unlock the valuable insights embedded within student narratives.

The lexicon-based sentiment analysis tool VADER, known for its high precision, recall, and F1 score in analyzing informal text and educational contexts (Yuliansyah et al., 2024), is well-suited for interpreting student evaluations due to its rule-based approach that accounts for conversational nuances like negation and punctuation (Hutto & Gilbert, 2014). The NRC lexicon not only provides polarity but also categorizes words into eightight primary emotions: anger, fear, anticipation, joy, sadness, surprise, disgust, and trust (Al-Garadi et al., 2021; Sullivan et al., 2020). This more nuanced approach, which categorizes sentiments, provides a a deeper understanding of the emotional drivers behind student comments, allowing educators to respond effectively to student needs (Ohtani, 2020).

Teachers are evaluated on various aspects of their teaching performance, as students discuss aspects such as teaching style, how teachers organize lectures, and subject mastery (Nasim et al., 2017). To gain a more detailed understanding of how to enhance the teaching and learning experience, the use of an evaluation tool that captures the various dimensions of teaching effectiveness is recommended (Abry et al., 2018; Philip, 2020; Tafazoli et al., 2022). To discover these aspects or topics discussed in student comments in teacher evaluations, the NLP machine learning technique, LDA topic modeling, shall be implemented. LDA is a topic modeling technique widely used in NLP for discovering hidden thematic structures in large, rich, and diverse collections of text, such as student comments in teacher evaluations. Word clouds are used to present visual data. Word clouds offer a valuable visual summary of student feedback, enabling teachers to quickly identify key themes, understand students' perceptions in their language, and pinpoint specific areas for revision and potential improvement in their teaching practices. They serve as a data-driven starting point for self-assessment and professional growth (Abry et al., 2018; Philip, 2020; Tafazoli et al., 2022).

Teacher evaluations are central to improving the teacher. However, improving teacher effectiveness does not end with knowing how positively or negatively students evaluate a teacher and on what aspect of teaching students need to improve their learning processes. Recommendations to improve the quality of teaching are essential for maximizing the benefits of teacher evaluations. The AI Google Gemini recommender system has the potential to develop teaching quality and student engagement within educational contexts. AI-driven insights can suggest areas for improvement, enabling educators to deliver targeted support and interventions where necessary (Seo et al., 2021); uncover student learning behavior patterns, helping teachers to redesign curriculum and instructional strategies (Hashim et al., 2022; Luan et al., 2020); and reveal trends in classroom dynamics and instructional effectiveness (Wang, 2025; Zhang & Zhang, 2024). The AI Google Gemini recommender system holds the potential to revolutionize education by improving teaching and learning.

This research aims to demonstrate a robust framework for leveraging NLP and AI to extract actionable insights from a substantial dataset of student comments collected over ten years of teacher evaluations. Specifically, this study involves: (1) cleaning and preprocessing the textual feedback; (2) applying NRC(National Research Counci) Word-Emoticon Association Lexicon and VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicons for sentiment analysis; (3) employing LDA (Latent Dirichlet Allocation) for topic modeling; (4) determining the sentiment associated with each identified topic; and (5) utilizing Google Gemini as an AI model to generate targeted recommendations, effectively functioning as a recommender system, for teacher improvement based on the synthesized sentiment and topic analysis. This approach seeks to transform raw student feedback into practical guidance for educators.

The results of this research are critically important for enhancing teacher development and improving the quality of instruction. By providing a systematic and AI-driven method for analyzing student feedback, this study offers a pathway towards data-informed professional growth for educators. Furthermore, the integration of advanced NLP techniques with AI-powered recommendation generation presents a novel approach to leveraging student voice for continuous improvement in teaching effectiveness. This work contributes to expanding academic understanding of educational data mining and the application of AI in enhancing learning and teaching practices.

2.0 Methodology

2.1 Research Design

The study employed a descriptive-quantitative approach utilizing NLP and AI to analyze 10 years of student comments from teacher evaluations, identifying sentiments, key topics, and generating AI-driven recommendations for teacher improvement. Quantitative NLP techniques quantify the sentiment expressed in student comments by assigning numerical scores by applying algorithms like VADER and determining topic distributions. By quantifying sentiment associated with specific topics, the research aimed to provide a data-driven foundation for the AI-powered recommender system.

2.2 Research Locale

The textual data analyzed in this research originates from student evaluations of teaching conducted at the University of the Cordilleras (UC), Philippines. The dataset encompassed student comments collected over a ten-year period (2009-2019), spanning 46 trimesters, across various undergraduate and graduate courses offered by different colleges within the institution. These evaluations are typically administered at the middle of each trimester to gather student feedback on instructor performance and the learning environment. All student comments were collected anonymously to encourage honest feedback, and the confidentiality of both student responses and instructor identities has been maintained throughout the data analysis process.

2.3 Research Participants

The participants of this study were undergraduate and graduate students currently enrolled in various courses offered during the conduct of teacher evaluation. Students provided the comments anonymously as part of the standard teacher evaluation process, and detailed written comments were voluntary. Sampling of comments was not employed, but rather, all available comments within the defined period were included. The analysis included a substantial volume of student feedback, totaling several thousands of comments.

2.4 Research Instrument and Data Pre-Processing

This research utilizes archival data in the form of textual comments from student evaluations of teaching, collected over a ten-year period (2009-2019). The analysis of this data was conducted using several computational tools and techniques. Sentiment was assessed using two lexicon-based approaches. NRC was employed to identify the presence of different emotions and overall sentiment in the text. Additionally, VADER was used, a sentiment analysis tool particularly effective for analyzing sentiment expressed in online text due to its sensitivity to valence and intensity. VADER provides positive, negative, neutral, and compound sentiment scores for each comment.

Prior to analysis, using the Natural Language Toolkit (NLTK) library in Python, the textual data was preprocessed. This involved steps such as removal of URLs to focus solely on the actual words and their meanings expressed by students; removal of hashtags to reduce redundancy; removal of numbers, which do not contribute to the linguistic meaning; removal of punctuation to standardize texts; removal of duplicate words leading to a more accurate representation of the actual word usage; and removal of stop words to highlight important words. By cleaning the text, the signal-to-noise ratio is improved, and NLP models extract more accurate and meaningful insights from the students' comments.

Despite these systematic steps, a primary challenge was handling non-English feedback. Student comments, while predominantly in English, occasionally included Filipino (Tagalog) and other local Cordilleran dialects. Since the primary computational tools, VADER for sentiment analysis and LDA for topic modeling, are largely optimized for English text, the presence of multilingual content presented a limitation. While these instances were observed, they were not systematically translated due to the volume and the focus on overall English-based patterns. Other common challenges involved inconsistencies in spelling and informal language use inherent in student feedback.

2.5 Data Gathering Procedure, Data Source, and Sampling Strategy
The researcher had no personal involvement or participation in the initial curation of this extensive 10-year dataset of raw student evaluations from the institutional database. The researcher's role was that of a secondary user of this existing institutional record. The dataset for this study comprises student textual comments during the conduct of the Student-Course-Teacher (SCT) evaluations collected over ten consecutive academic years, specifically from the 3rd trimester of the academic year 2008-2009 to the 2nd trimester of 2018-2019. From the perspective of the institutional records, this represents a census of available comments, not a sample derived by the researcher. The Human Resources Development Office (HRDO) forwarded via e-mail a set of 46 .csv files of students' comments as they existed in the institutional database. This comprehensive 10-year research approach was chosen specifically to capture the full spectrum of student feedback over a significant duration, allowing the determination of trends and patterns in student perceptions.

Upon receipt, the 46 .csv files of comments were then exported into a single .csv structured format, keeping individual copies for archival purposes. The process ensured that the textual data was isolated and prepared for subsequent analysis. All student comments were provided anonymously as part of the standard evaluation process, and the extracted data maintained this anonymity throughout. It is crucial to acknowledge the potential sources of bias inherent to the original primary data collection. The researcher had no control over student participation rates in the evaluations, and the administrative procedures of the SCT evaluation system, which may have evolved over the 10-year period, could influence the nature and scope of the comments provided. Further, the researcher has no control over the subjectivity inherent in qualitative feedback that may lead to skewed representation of highly positive or negative feedback.

While the researcher could not influence the primary data generation, managing these inherent biases in the analysis phase was clearly stated: this study is a secondary analysis of institutionally collected data. In the systematic processing of the vast volume of textual data, VADER (Valence Aware Dictionary and sEntiment Reasoner) was employed using Python for automated sentiment analysis, and Latent Dirichlet Allocation (LDA) was utilized in Python for automated topic modeling. The results from these computational methods, providing sentiment values and identified topics, were reported as generated by the algorithms.

2.6 Ethical Considerations
While research involving student feedback typically necessitates formal ethical clearance, the nature of data acquisition for this study followed an established institutional process that prioritized data privacy and anonymity. The dataset, comprising ten years of student textual comments from teacher evaluations, was obtained following an official request submitted to the Vice President of Administration and Student Services (VPASS). Crucially, HRDO was responsible for providing the dataset. Prior to release, the HRDO informed the researcher that the HRDO undertook comprehensive de-identification. This involved the complete removal of all personally identifiable information, including the names of faculty members and their corresponding quantitative ratings. Only the textual comments, completely stripped of any direct identifiers, were extracted from the secure institutional database for the purpose of this analysis.

Given that the data was already thoroughly anonymized by the HRDO and provided through an official institutional channel with built-in privacy safeguards, a separate ethical clearance protocol number was not formally pursued for this specific analysis of de-identified textual data. The study, therefore, operated under the premise that the data, as received, was in compliance with ethical guideline not requiring further institutional review board for human subjects. Despite the absence of a formal ethics approval number, this research strictly adhered to the principles of data privacy and confidentiality, ensuring that no personally identifiable information was accessed, processed, or reported throughout the study.

3.0 Results and Discussion

A comprehensive analysis of findings from applying Natural Language Processing (NLP) and Artificial Intelligence (AI) techniques to a decade of student comments from teacher evaluations is presented. Building on the established methodology, it systematically unveils key themes identified by Latent Dirichlet Allocation (LDA) topic modeling, their associated sentiment determined by VADER and NRC lexicons, and actionable recommendations for teacher improvement generated by the Google Gemini AI model. Sentiment analysis, employing tools like VADER and NRC, facilitated a nuanced interpretation of student feedback by categorizing comments into positive enthusiasm, negative, and neutral comments, allowing for analysis exceeding a superficial examination of textual data. The bar graph in Figure 1 shows unnoticeable differences in the counts for positive, negative, and neutral sentiment when analyzed by two different analysis tools, VADER (blue bars) and NRC (red bars).

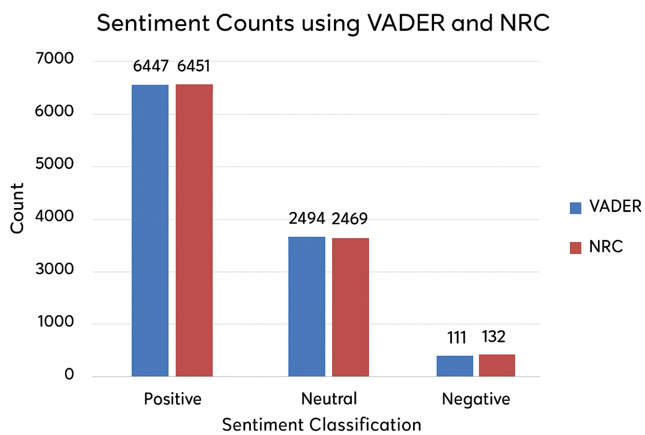

Figure 1. *Comparison of the Sentiment Classifications using VADER and NRC*

Both VADER and NRC identified a significantly large number of comments as positive, with VADER classifying slightly fewer (6447) than NRC (6451). The counts are very close for this category. Both tools also classified a substantial number of comments as neutral. Again, the counts are very similar, with VADER identifying 2494 and NRC identifying 2469. Both VADER and NRC classified a relatively small number of comments as negative compared to the positive and neutral categories. VADER identified 111 negative comments, while NRC identified slightly more, at 132. Both VADER and NRC show a similar distribution of sentiment across student comments. Many comments are classified as positive, followed by a significant number of neutral comments, and a much smaller number of negative comments. The counts for each sentiment classification are close between the two tools.

While both VADER and NRC serve as valuable resources in sentiment analysis, their design, underlying methodologies, and sensitivity to linguistic nuances can lead to similar outcomes in specific contexts (Heaton et al., 2023; Mujahid et al., 2021). Despite this, combining VADER and NRC is often lauded as a practically powerful approach, providing a multifaceted view of feedback that encompasses not just general sentiment but also emotional tone, thereby enhancing the interpretation of student comments (Mahmood et al., 2020). However, in contrast to this anticipated benefit, this powerful combination of VADER and NRC did not manifest as expected in the current study, as evidenced by Figure 2. Specifically, there was no notable increase in counts for neutral sentiments compared to using a single sentiment analysis tool.

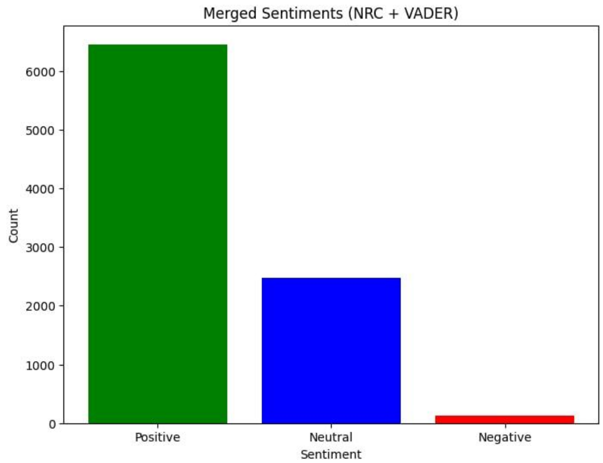

Figure 2. *Distribution of Sentiments using VADER and NRC*

The finding of a predominant neutral sentiment — over 27% of analyzed comments showed a predominant neutral sentiment, necessitating further examination. A key factor influencing this finding is the existence of non-English feedback stated in Filipino, Ilocano, or native dialects. Sample comments revealed frequent use of Filipino and local dialects. Sentiment analysis tools, such as VADER and NRC, which rely on English lexicons, often fail to recognize sentiment in languages other than English. This likely leads to non-English sentiment-bearing words being classified as neutral, thereby inflating the proportion of neutral sentiment. For example, phrases like "palagi kayo ingat" and "masaya," potentially positive for native speakers, would likely be deemed neutral. This linguistic characteristic reveals a limitation of using English-based sentiment analysis on multilingual feedback. The actual sentiment distribution may be more polarized than indicated. Future research could investigate multilingual sentiment analysis or translation to capture student sentiment more accurately.

Figure 3 presents a sample sentiment report generated by choosing a .csv textual comments of 183 rows of comments. This report highlights crucial aspects of the analysis, including the 'Original_Text' and 'Cleaned_Text' columns, sentiment VADER score, and the polarity of the comment (positive, negative, or neutral). This report provides transparency into the preprocessing steps, allowing users to cross-reference initial feedback with the processed version.

In the sample report (Figure 3), generally providing positive sentiment scores, with comment #175 having an almost perfect sentiment score of 0.9380, while a negative sentiment score for the comment in line #179, notably a comment combining Filipino and English. Another notable observation is the comment in line 5, 'calm, approachable', which was evaluated as 'neutral' by VADER. While 'calm' and 'approachable' are individually recognized as positive by standard VADER, the preprocessing step's removal of the punctuation, in this case, the comma (,), led to VADER's inability to parse this phrase for sentiment correctly. This specific instance highlights an area for future refinement in preprocessing to ensure better handling of punctuated positive phrases, thereby enhancing the detection of comment sentiment.

A critical consideration during the application of VADER was the multilingual nature of some student feedback. As depicted in Figure 3, the original comments included instances of Filipino phrases and code-switching alongside English text (e.g., sample comments in Rows 179 and 180 explicitly demonstrate this linguistic mixture). Since VADER is primarily calibrated for English, its direct application to these multilingual segments presented a limitation. While efforts were made during preceding data cleaning steps, the nuances of non-English expressions or code-switching could not be fully captured or accurately interpreted by VADER's English-centric lexicon. Therefore, the sentiment scores predominantly reflect the sentiment conveyed through the English-language components of the feedback, acknowledging that sentiment expressed solely through non-English phrases may not have been fully or accurately reflected in the quantitative analysis. Following sentiment analysis, LDA was employed to identify multiple topics, enabling a more comprehensive analysis. The performance and interpretability of Latent Dirichlet Allocation (LDA) are highly dependent on the selection of an appropriate number of topics, denoted as k. A higher coherence score generally indicates more interpretable and meaningful topics. Analysis of 9,052 textual comments spanning 10 academic years revealed a predominantly positive (71%)

to neutral (27%) sentiment and identified eight distinct topics that characterize student feedback on teacher evaluations.

First 5 rows:

	Original_Text	Feedback_Text	Cleaned_Text_Main	VADER_Score_Eng	VADER_Sentiment_Eng
0	Very professional and humble. He is very under...	Very professional and humble. He is very under...	professional humble motivating understanding	0.5367	Positive (VADER Eng)
1	He's considerate.	He's considerate.	considerate	0.4404	Positive (VADER Eng)
5	calm,approachable	calm,approachable	calmapproachable	0.0000	Neutral (VADER Eng)
6	sometimes very fast in speaking during lecture -	sometimes very fast in speaking during lecture -	lecture fast speaking	0.0000	Neutral (VADER Eng)
8	Thank you sir for all the teachings	Thank you sir for all the teachings	teaching	0.3612	Positive (VADER Eng)

Last 5 rows:

	Original_Text	Feedback_Text	Cleaned_Text_Main	VADER_Score_Eng	VADER_Sentiment_Eng	VAI
175	she is very helpful and very nice -she always ...	she is very helpful and very nice -she always ...	nice fun helpful energy lesson laugh morning	0.9380	Positive (VADER Eng)	
179	MAAYOS LAGI ANG PANANAMIT AT LAGING HANDA SA P...	MAAYOS LAGI ANG PANANAMIT AT LAGING HANDA SA P...	pagtuturo example lagi project assignment hand...	-0.2325	Negative (VADER Eng)	
180	WALA PO SIR <3 INGAT KAYO PALAGI <3	WALA PO SIR <3 INGAT KAYO PALAGI <3	palagi wala ingat	0.1197	Positive (VADER Eng)	
181	Mastery over his subjects	Mastery over his subjects	subject mastery	0.0000	Neutral (VADER Eng)	
182	excellent professor	excellent professor	excellent professor	0.5719	Positive (VADER Eng)	

Figure 3. *Sample Original Cleaned Comments*

To provide a clearer, qualitative illustration of these identified thematic structures and their manifestation within student feedback, Figure 4 presents a detailed summary of overall sentiment per topic, covering four topics, and offers a multifaceted view of student perceptions regarding teaching effectiveness. A comprehensive understanding of each topic and connecting the thematic insights with the sentiment expressed. This specific example highlights the key terms and their prominence within this particular topic, providing a detailed insight into the types of insights uncovered by the LDA model.

Overall Sentiment Per Topic:

	Topic ID	AI Label	Top Keywords	Num Comments	Avg VADER Eng Score	VADER Eng Dist (%)
0	0	Love the subject!	teaching, subject, student, love, experience	38	0.37	{'Positive (VADER Eng)': 60.5, 'Neutral (VADER Eng)': 34.2, 'Negative (VADER Eng)': 5.3}
1	1	Teacher & Subject Mastery \n Smart Class Scores High \n Ingat: Friendly, Smart Class\n Mas...	subject, teacher, matter, mastery, ingat	27	0.41	{'Positive (VADER Eng)': 70.4, 'Neutral (VADER Eng)': 29.6}
2	2	Awesome Teacher Lesson	lesson, magturo, considerate, magaling, teacher	33	0.27	{'Positive (VADER Eng)': 60.6, 'Neutral (VADER Eng)': 21.2, 'Negative (VADER Eng)': 18.2}
3	3	Calm English Activity Lesson	activity, improve, chill, lesson, calm	16	0.50	{'Positive (VADER Eng)': 75.0, 'Neutral (VADER Eng)': 25.0}

Figure 4. *Overall Sentiment Per topic*

The tabular presentation of overall sentiment analysis per topic in Figure 4 highlights each identified topic, its top keywords, the number of comments assigned to it, and its sentiment distribution as analyzed by VADER. For instance, Topic 3 labeled "Calm English Activity Lesson", qualitatively characterized by themes such as "activity", "improve", "chill", "lesson", and "calm", exhibited the highest sentiment score of 0.50. This strong positive sentiment is quantitatively supported by the fact that 16 clustering comments were dominated by 75% positive comments and contained no negative comments, indicating a highly favorable student perception of this specific aspect of teaching.

This positive finding stands in clear contrast to Topic 2, which focuses on themes such as "lesson", "magturo", "considerate", "magaling", and "teacher", labeled as "Awesome Teacher Lesson". Despite its larger cluster of comments, Topic 2 registered a lower average sentiment score of 0.27. This reduced score is quantitatively contributed to by a significant 18.2% of negative comments, suggesting specific areas within these themes where student experiences are less favorable or challenging.

Figure 4 thus provides granular insights into the sentiment expressed within each of the identified thematic areas, offering a crucial bridge between the qualitative outcomes of topic modeling (what students discuss) and the quantitative results of sentiment analysis (how students feel). This integrated view directly informs the actionable recommendations generated by the AI model, enabling targeted pedagogical improvements based on the specific strengths and weaknesses revealed by this dual analysis. The initial sentiment analysis using VADER consistently showed an overwhelmingly positive sentiment across all four identified topics, with positive comments ranging from 60.5% in Topic 0 to 75.0% in Topic 3. Delving deeper, however, it became clear that Topic 2 and Topic 0 also contained the highest proportion of negative comments. The prevalence of negative feedback in these specific areas highlights critical weaknesses, indicating a need for actionable insights to address these deficiencies effectively.

To gain a more granular qualitative understanding of the specific content and prominence of terms within these discussions—especially to contextualize the nuances in sentiment and pinpoint the nature of identified weaknesses—a word cloud analysis was subsequently conducted. As established, in a word cloud, the most frequent words in the corpus appear the largest and are considered the most central and important terms associated with the topic (Heimerl et al., 2014). Figure 5 visually presents four distinct word clouds, one for each identified topic, with concise labels generated by Google Gemini AI recommender, offering a direct visual representation of the key qualitative insights.

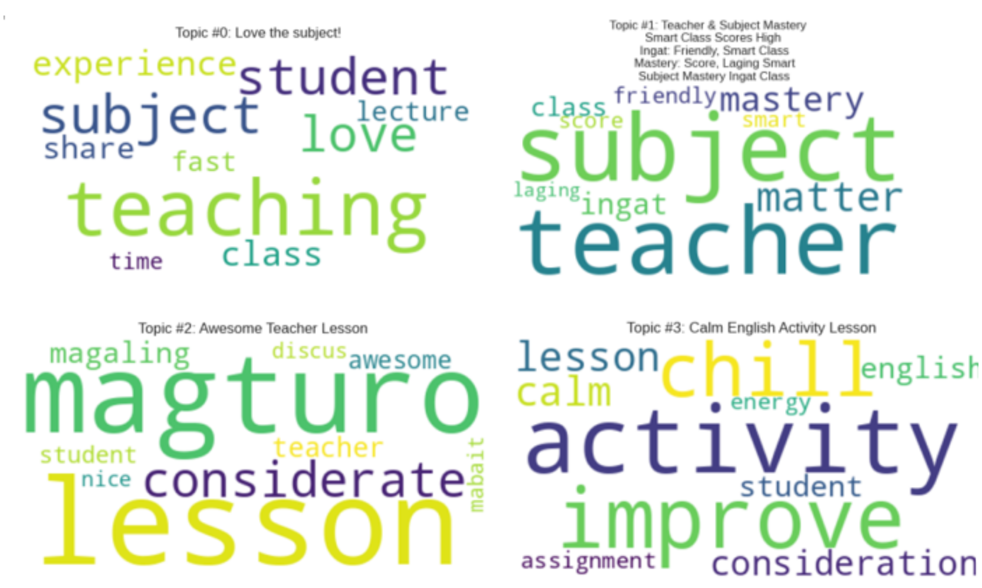

Figure 5. *Sample Word Clouds to Present the Topics*

The subsequent discussion will provide a detailed interpretation of Topic #0 as a representative example. Topic #0 ('Love the subject!') was selected for its positive sentiment and high number of comments, providing valuable insights into aspects of teaching that are well-received.

Topic #0: “Love the Subject” word cloud is dominated by terms such as "teaching," "subject", "love," “student,” and "experience." This theme indicates positive feedback and encapsulates students’ love or strong appreciation of the subject and how it is being taught. The theme grounds these sentiments in the actual learning environment, where student can share their experiences within the class. Topic 0 indicates students’ love for the subject, suggesting that “time” is running out “fast” to experience the class. This topic, with 38 comments and a sentiment score of 0.37, supported by 60.5% positive and a minor 5.3% negative comments, indicates practical pedagogical approaches that foster student enthusiasm and enjoyment in the learning process.

Words in a word cloud do not portray relationships, which may contribute to a shallow understanding of the content (Vrain & Lovett, 2019). Teachers may struggle to extract meaningful insights; therefore, adequate training is necessary in analyzing visual data, which can complicate the process of deriving meaningful information from word clouds (Boonen et al., 2016). Thus, an AI recommender system can be used to interpret word clouds to relay insights to teachers.

Utilizing the Google Gemini model, topics and their corresponding sentiment scores were identified. These data were then synthesized to generate actionable recommendations aimed at improving pedagogical practices, effectively transforming student feedback into pragmatic guidance for educators.

Figure 6 illustrates a sample of these actionable insights for Topic 0. Google Gemini AI generates a topic theme and provides actionable insights to enhance teaching effectiveness. The theme recommended by Google Gemini for Topic 0 is “Love the Subject,” which revolves around the largest word in Topic 0, “teaching”.

```
AI Recommendations for Selected Topics
Generating recommendations for the 3 topic(s) with the most negative/neutral sentiment.

------------------------------------------------------------
RECOMMENDATIONS FOR TOPIC 0: Love the subject!
------------------------------------------------------------
```

Here are 2-3 actionable teaching recommendations based on the topic "Love the subject!" (Avg Aug VADER Score: 0.22) and keywords like "teaching, subject, student, love, experience, class, share":

1. **Foster a Positive Learning Experience through Personal Connection:** Encourage students to connect the subject matter to their personal lives and interests. This can be achieved through:
   - **Action:** Assign projects or discussions that require students to relate the concepts to their hobbies, aspirations, or current events. For example, in a science class, have students design a related product that solves a problem in their daily lives. In a history class, have students research a historical figure that inspires them.
   - **Benefit:** This allows students to find personal meaning in the subject, increasing their engagement and love for it. Sharing these personal connections within the class (e.g., via presentations, small group discussions) builds a supportive learning community.
2. **Incorporate Active Learning and Collaborative Activities:** Move beyond lectures and rote memorization to incorporate active learning strategies that promote student interaction and enjoyment. This can be done by:
   - **Action:** Implement frequent group work, debates, games, simulations, or case studies that require students to actively use the concepts. Encourage peer teaching and student-led discussions. Use online platforms for interactive quizzes and collaborative projects.
   - **Benefit:** Active learning enhances comprehension, makes learning more fun, and provides opportunities for students to share their understanding and excitement with their peers, leading to an increased feeling of shared passion and love for the subject.
3. **Model Enthusiasm and Share Your Passion:** Demonstrate your own love and excitement for the subject matter. This will influence students' attitudes and inspire them to develop their own positive feelings. Try these approaches:
   - **Action:** Start class with a captivating anecdote, interesting fact, or relevant current event related to the subject. Share your own reasons for loving the subject, and explain why you find it fascinating. Use a dynamic and engaging teaching style (voice modulation, movement).
   - **Benefit:** Your enthusiasm is contagious. When students see your genuine passion, they are more likely to be inspired and develop a similar love for the subject, leading to a more positive class experience and increased engagement.

Figure 6. *Google Gemini AI Recommendations for Topic 0: “Love the Subject!”*

Based on the analysis of Topic #0, 'Love the subject!'—a theme consistently associated with positive student sentiment—Google Gemini generated key actionable recommendations aimed at fostering and enhancing this positive learning experience. Specifically, Google Gemini recommends teaching the subject by connecting it to students' personal lives, thereby enabling them to share relevant experiences in class. This includes encouraging projects that allow students to connect concepts to their hobbies, aspirations, or interests. Additionally, Google Gemini advises implementing frequent group work that encourages students to share their understanding with peers. Complementing these student-focused strategies, Google Gemini also emphasizes the teacher's role by recommending that they model enthusiasm and actively share their passion for the subject.

Beyond merely categorizing student feedback, this study's unique contribution lies in its innovative integration of advanced AI capabilities, specifically the Google Gemini recommender system, with a comprehensive analysis of teacher feedback. Traditional teacher evaluation methods often yield broad numerical scores or large volumes of unstructured textual comments that are challenging for educators to interpret and act upon effectively. This study’s approach moves beyond these limitations by leveraging AI to synthesize qualitative thematic insights with quantitative sentiment analysis, thereby providing highly contextualized and actionable pedagogical recommendations. This integration transforms raw feedback into precise, data-driven strategies for improvement, representing a significant advancement in how educators can utilize student evaluations for continuous professional development.

Building upon the insights provided by VADER’s sentiment scores and LDA’s topic models, which include sentiment scores per topic, Google Gemini leverages the insights from topic modeling and sentiment analysis of student feedback to generate actionable, pedagogical recommendations. It synthesizes the 'what' (topics students are discussing) with the 'how they feel' (sentiment) to provide targeted suggestions for improving teaching and learning. Therefore, with individual teacher evaluation results feeding the AI Google, Google Gemini recommender system, teachers gain targeted resources and insights to directly enhance their pedagogical approaches and student engagement directly, leading to significant improvements in teaching effectiveness and a more data-informed approach to professional development.

4.0 Conclusion

This research aimed to demonstrate a robust framework for leveraging NLP and AI to extract actionable insights from a substantial dataset of student comments collected over ten years of teacher evaluations. The sentiment analysis revealed a predominantly positive to neutral tone in student feedback across the decade, with similar distributions identified by both VADER and NRC. To visually represent the eight topics uncovered by LDA, word clouds were utilized. Notably, the presence of non-English feedback likely contributed to the neutral sentiment observed. The application of Google Gemini demonstrated its potential to generate relevant and actionable recommendations for teachers based on the identified topics and their associated sentiment, offering specific suggestions for pedagogical and engagement enhancements.

This study provides valuable insights into understanding and enhancing teacher effectiveness. The data-driven aspects of teaching with sentiments provide a better understanding of what students value, which shall guide teachers in their professional development. The potential of AI recommender systems to offer personalized insights and recommendations from a rich data source is an approach supporting continuous teacher growth. While providing initial insights, this study's methodology is limited by its reliance on English-language sentiment analysis tools, which potentially underrepresent sentiments in diverse, multilingual feedback. Subjectivity in topic interpretation by researchers also affects findings, suggesting a need for standardized, automated methods. The practical effectiveness of the AI system depends on the quality and representativeness of its training data, as well as the depth of pedagogical principles integrated, requiring ongoing assessment and refinement for optimal educational impact.

Future research directions aimed at expanding this investigation include integrating multilingual natural language processing tools. This research aimed to demonstrate a robust framework for leveraging NLP and AI to extract actionable insights from a substantial dataset of student comments collected over ten years of teacher evaluations. The sentiment analysis revealed a predominantly positive to neutral tone in student feedback across the decade, with similar distributions identified by both VADER and NRC. To visually represent the eight topics uncovered by LDA, word clouds were utilized. Notably, the presence of non-English feedback likely contributed to the neutral sentiment observed. The application of Google Gemini demonstrated its potential to generate relevant and actionable recommendations for teachers based on the identified topics and their associated sentiment, offering specific suggestions for pedagogical and engagement enhancements.

5.0 Contributions of Authors

Ortiz, Marie Grace: conceptualization, data gathering, writing, editing
Dumlao, Menchita: data analysis, data gathering

6.0 Funding

No funding agency was sought for this research.

7.0 Conflict of Interests

The authors declare no conflict of interests.

8.0 Acknowledgment

The researchers would like to acknowledge to the staff of UC-HRDO and Dr. Rhodora Ngolob.

9.0 References

Abry, T., Granger, K., Bryce, C., Taylor, M., Swanson, J., & Bradley, R. (2018). First grade classroom-level adversity: associations with teaching practices, academic skills, and executive functioning. School Psychology Quarterly, 33(4), 547–560. https://doi.org/10.1037/spq0000235

Al-Garadi, M. A. ., Yang, Y., Guo, Y., Kim, S., Love, J. S., Perrone, J., & Sarker, A. (2021). Large-scale social media language analysis reveals emotions and behaviours associated with nonmedical prescription drug use. Health Data Science. https://doi.org/10.34133/2022/9851989

Balam, E. M., & Shannon, D. M. (2010). Student ratings of college teaching: A comparison of faculty and their students. Assessment and Evaluation in Higher Education, 35(2), 209–221. https://doi.org/10.1080/02602930902795901

Bill & Melidan Gates Foundation. (2012). Asking students about teaching. Bill & Melinda Gates Foundation, 1–28. https://eric.ed.gov/?id=ED566384

Boonen, A. J. H., Reed, H. C., Schoonenboom, J., & Jolles, J. (2016). It's not a math lesson-we're learning to draw! teachers' use of visual representations in instructing word problem solving in sixth grade of elementary school. Frontline Learning Research, 4(5), 34–61. https://doi.org/10.14786/flr.v4i5.245

Ching, G. (2018). A literature review on the student evaluation of teaching. Higher Education Evaluation and Development, 12(2), 63–84. https://doi.org/10.1108/heed-04-2018-0009

Ed, E. I. (2000). ERIC t , Oil Digests. 1981, 1–6.

Esparza, G. G., de-Luna, A., Zezzatti, A. O., Hernandez, A., Ponce, J., Álvarez, M., Cossio, E., & Nava, J. de J. (2018). A sentiment analysis model to analyze students reviews of teacher performance using support vector machines. Advances in Intelligent Systems and Computing, 620, 157–164. https://doi.org/10.1007/978-3-319-62410-5_19

Hashim, S., Omar, M. K., Jalil, H. A., & Sharef, N. M. (2022). Trends on technologies and artificial intelligence in education for personalized learning: systematic literature review. International Journal of Academic Research in Progressive Education and Development, 11(1). https://doi.org/10.6007/IJARPED/v11-i1/12230

Heaton, D., Clos, J., Nichele, E., & Fischer, J. (2023). Critical reflections on three popular computational linguistic approaches to examine twitter discourses. Peerj Computer Science. https://doi.org/10.7717/peerj-cs.1211

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. Proceedings of the Annual Hawaii International Conference on System Sciences, 1833–1842. https://doi.org/10.1109/HICSS.2014.231

Hutchinson, M., Coutts, R., Massey, D., Nasrawi, D., Fielden, J., Lee, M., & Lakeman, R. (2023). Student evaluation of teaching: reactions of Australian academics to anonymous non-constructive student commentary. Assessment and Evaluation in Higher Education. https://doi.org/10.1080/02602938.2023.2195598

Hutto, C., & Gilbert, E. (2014). Vader: a parsimonious rule-based model for sentiment analysis of social media text. International Aaai Conference on Web and Social Media, 216–225.

Kreitzer, R. J., & Sweet-Cushman, J. (2022). Evaluating Student Evaluations of Teaching: a Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform. Journal of Academic Ethics, 20(1), 73–84. https://doi.org/10.1007/s10805-021-09400-w

Leguey, S., Cid-Cid, A. I., Guede-Cid, R., & Prieto, J. (2023). An Exploratory Analysis of Major Dropdowns in Student Evaluation of Teaching Ratings in Higher Education. Multidisciplinary Journal of Educational Research, 13(1), 91–113. https://doi.org/10.17583/remie.10419

Luan, H., Lai, H., Gobert, J. D., Yang, S. J., & Ogata, H. (2020). Challenges and future directions of big data and artificial intelligence in education. Frontiers in Psychology, 11. https://doi.org/10.3389/fpsyg.2020.580820

Mahmood, A., Kamaruddin, S., Naser, R., & Nadzir, M. (2020). A combination of lexicon and machine learning approaches for sentiment analysis on facebook. Journal of System and Management Sciences., 10(3), 140–150. https://doi.org/10.33168/JSMS.2020.0310

Mujahid, M., Lee, E., Rustam, F., Washington, P. B., Ullah, S., Reshi, A. A., & Ashraf, I. (2021). Sentiment analysis and topic modeling on tweets about online education during covid-19. Applied Sciences (Switzerland), 11(18). https://doi.org/10.3390/app11188438

Nasim, Z., Rajput, Q., & Haider, S. (2017). Sentiment analysis of student feedback using machine learning and lexicon-based approaches. International Conference on Research and Innovation in Information Systems, ICRIIS. https://doi.org/10.1109/ICRIIS.2017.8002475

Newman, H., & Joyner, D. (2018). Sentiment analysis of student evaluations of teaching. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10948 LNAI, 246–250. https://doi.org/10.1007/978-3-319-93846-2_45

Ohtani, S. (2020). Exploring public interest from twitter in 2021 using natural language processing for post-2020 biodiversity conservation strategies. https://doi.org/10.21203/rs.3.rs-1468450/v1

Philip, R. (2020). Word cloud analysis and single word summarisation as a new paediatric educational tool: results of a neonatal application. Journal of Paediatrics and Child Health, 56(6), 873–877. https://doi.org/10.1111/jpc.14760

Rajput, Q., Haider, S., & Ghani, S. (2016). Lexicon-Based Sentiment Analysis of Teachers' Evaluation. Applied Computational Intelligence and Soft Computing, 2016, 1–12. https://doi.org/10.1155/2016/2385429

Seo, K., Tang, J., Roll, I., Fels, S., & Yoon, D. (2021). The impact of artificial intelligence on learner–instructor interaction in online learning. International Journal of Educational Technology in Higher Education, 18(1), 54. https://doi.org/10.1186/s41239-021-00292-9

Simmons, T. L. (1997). Student Evaluation of Teachers: Professional Practice or Punitive Policy? JALT Testing & Evaluation SIG Newsletter, 1(1), 12–19. https://teval.jalt.org/test/sim_1.htm

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the Validity of Student Evaluation of Teaching: The State of the Art. In Review of Educational Research (Vol. 83, Issue 4). https://doi.org/10.3102/0034654313496870

Sullivan, K. J., Burden, M., Keniston, A., Banda, J. M., & Hunter, L. (2020). Characterization of anonymous physician perspectives on covid-19 using social media data. Biocomputing 2021, 26, 95–106. https://pmc.ncbi.nlm.nih.gov/articles/PMC7958992/

Tafazoli, D., Chirimbu, S., & Dejica-Carțiș, A. (2022). Web 2.0 in english language teaching: using word clouds. Professional Communication and Translation Studies, 167–172. https://doi.org/10.59168/XQDJ2288

Vrain, E., & Lovett, A. (2019). Using word clouds to present farmers' perceptions of advisory services on pollution mitigation measures. Journal of Environmental Planning and Management, 63(6), 1132–1149. https://doi.org/10.1080/09640568.2019.1638232

Wang, J. (2025). The impact of AI teaching on teaching quality. International Journal of Web-Based Learning and Teaching Technologies, 20(1), 1–22. https://doi.org/10.4018/IJWLTT.376489

Wine, D. (2016). Using student feedback to enhance teacher evaluation. Teacher Education, Educational Leadership & Policy ETDs. https://digitalrepository.unm.edu/educ_teelp_etds/49

Yuliansyah, H., Mulasari, S., Sulistyawati, S., & Ghozali, F. , S. B. (2024). Sentiment analysis of the waste problem based on youtube comments using vader and deep translator. Jurnal Media Informatika Budidarma, 8(1), 663. https://tinyurl.com/rkkwhefv

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. Teaching in Higher Education, 12(1), 55–76. https://doi.org/10.1080/13562510601102131

Zhang, J., & Zhang, Z. (2024). AI in teacher education: unlocking new dimensions in teaching support, inclusive learning, and digital literacy. Journal of Computer Assisted Learning, 40(4). https://doi.org/10.1111/jcal.12988