Engineering Educators Bringing the World Together

2025 ASEE Annual Conference & Exposition

Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025 ASEE

Paper ID #47836

BOARD # 73: Leveraging Transformer-Based Models for Sentiment Analysis in Educational Reviews: A Comparative Study with Lexicon Models Using Coursera Data

Priyanshu Ghosh, Mission San Jose High School

Priyanshu Ghosh is a high school student at Mission San Jose High School in Fremont, California. He has a strong interest in data science and artificial intelligence, with a focus on sentiment analysis and machine learning applications. As the primary author of this research paper, Priyanshu is actively engaged in exploring advanced computational techniques to derive actionable insights from large datasets, particularly in the education sector.

Dr. Mihai Boicu, George Mason University

Mihai Boicu, Ph.D., is Associate Professor in the Information Sciences and Technology Department at George Mason University. He is an expert in artificial intelligence, structured analytical methods, probabilistic reasoning, evidence-based reasoning, personalized education, active learning with technology, crowd-sourcing, and collective intelligence. He is the main software architect of the Disciple agent development platform and coordinates the software development of various analytical tools used in IC and education. He has over 150 publications, including 2 books and 3 textbooks. He has received the Innovative Application Award from the American Association for Artificial Intelligence, and several certificates of appreciation from the U.S. Army War College and the Air War College. He is a GMU Teacher of Distinction.

Title: Quantifying Sentiment in Educational Reviews: A Comparative and Aspect-Based Analysis Using Lexicon and Transformer Models on Coursera Data

Abstract

This study examines the effectiveness of sentiment analysis tools, VADER, AFINN, and BERT, on a dataset comprising over 1.45 million Coursera course reviews, representing diverse academic disciplines and institutions. The research addresses three crucial questions: How does BERT's ability to account for sentiment nuances compare to traditional lexicon-based tools? What temporal factors influence sentiment patterns, such as differences in feedback provided on weekdays versus weekends? How do sentiment trends vary across dimensions like course content quality and instructor performance? By integrating BERT's advanced contextual capabilities with conventional methods, the study offers a deeper understanding of student feedback. The findings underscore BERT's superior capacity to analyze complex, aspect-specific sentiments. Additionally, temporal trends reveal distinct variations in sentiment based on feedback timing, while weak correlations between specific course elements and overall ratings highlight the importance of holistic course development. This research offers practical recommendations for educational platforms, underscoring strategies to refine feedback mechanisms, elevate course design, and enhance the overall learning experience.

1. Introduction

Online education platforms like Coursera rely heavily on user-generated reviews to attract prospective students and provide valuable insights to course developers (Dalipi et al., 2021). While numerical ratings offer a quick overview of student satisfaction, written reviews expand opinions, capturing nuanced feedback that can be analyzed for richer insights. Traditional sentiment analysis tools, such as VADER (Hutto & Gilbert, 2014) and AFINN (Nielsen, 2011), are popular for their simplicity and efficiency, but they often fall short in interpreting the subtle context and intricacies of detailed reviews. This limitation adds difficulty to addressing sentiment at a detailed level, particularly when analyzing specific course elements such as instructor performance, content quality, or structure.

In this study, "sentiment" refers to the overall emotional tone expressed in student reviews, classified as positive, neutral, or negative. Sentiment analysis tools process textual feedback to extract meaning and assign sentiment scores, allowing researchers to assess trends in student satisfaction. However, sentiment in educational reviews can be skewed by outside factors, including instructor demographics (e.g., race or gender), course pricing models (paid vs. free enrollments), and completion status. Earlier work has considered the impact of instructor demographics on course evaluations; for instance, an online‑teaching experiment showed that the same instructors received significantly lower ratings when students perceived them as female rather than male (MacNell, Driscoll, & Hunt, 2015). However, the dataset used in this research does not contain demographic information, so the present study focuses exclusively on textual sentiment trends in the available review data.

Developments in natural language processing, particularly transformer-based models like BERT (Devlin et al., 2019), have improved the field by enabling a more sophisticated understanding of language. Unlike lexicon-based tools, BERT's ability to process text bidirectionally allows it to handle long-form, context-rich reviews with greater precision. Although BERT has been applied to various domains, including movie reviews and ideological texts, its utility in educational platforms like Coursera remains underexplored. This paper sets out to fill that gap by assessing and comparing the performance of traditional sentiment analysis tools (VADER, AFINN) and BERT on a dataset of over 1.45 million Coursera reviews.

Extending beyond basic sentiment classification, this research proposes a hybrid framework that combines BERT's deep contextual understanding with the efficiency of lexicon‑based tools to probe sentiment patterns across important course aspects—such as content quality and instructor performance—while also mapping temporal shifts, including weekday‑versus‑weekend differences. This integrated approach yields a more detailed view of student feedback than either method can achieve alone, leveraging a uniquely large and diverse Coursera dataset to highlight how aspect‑level and time‑based trends can inform concrete improvements in feedback loops, course design, and overall learner experience, thus advancing the field beyond earlier lexicon‑dependent studies (Zhang et al., 2018).

2. Literature Review

2.1 Sentiment Analysis in Educational Reviews

Sentiment analysis is a capable tool for interpreting user-generated feedback across various fields, including education. Within Massive Open Online Courses (MOOCs), it has been widely applied to assess student engagement, course performance, and overall satisfaction levels. Research by Dalipi et al. (2021) underscores the role of sentiment analysis in enhancing review systems and improving user experiences on platforms like Coursera. While traditional sentiment analysis methods, such as lexicon-based approaches, effectively categorize feedback into broad sentiment classes—positive, negative, and neutral—they often struggle to capture nuanced opinions related to specific course elements, such as instructor effectiveness, course content, or technical support (Zhang et al., 2018).

One important consideration in sentiment analysis of student feedback is the potential influence of external factors on review patterns. Instructor demographics (e.g., race or gender), course pricing models (paid vs. free enrollments), and completion status (whether the reviewer finished the course) could all play a role in shaping sentiment scores (MacNell, Driscoll, & Hunt, 2015). However, such data is often unavailable due to privacy policies and platform limitations, making it difficult to directly assess these influences (Dalipi, Ahlgren, & Zdravkova, 2021). Prior studies like MacNell et al., 2015 have acknowledged potential biases in online course evaluations, but this paper targets analyzing sentiment trends using available data rather than examining demographic-driven sentiment variations.

Previous studies, such as Xie et al. (2024), have demonstrated the potential of sentiment analysis for educational feedback but have largely concentrated on classification tasks, offering limited insights into temporal sentiment trends, aspect-specific feedback, or contextual factors affecting sentiment expression. This study addresses these gaps by leveraging a large-scale dataset of Coursera reviews to conduct a multidimensional sentiment analysis. In particular, it explores how sentiment patterns shift over time (e.g., weekday vs. weekend variations) and how sentiment correlates with course aspects like content quality and instructor performance (Li et al., 2019).

2.2 Methodological Advancements: From Lexicons to Transformer Models

The development of sentiment analysis tools showcases a significant shift from basic lexicon-based methods to advanced machine learning models, each tailored to different needs and challenges:

- **VADER (Valence Aware Dictionary for Sentiment Reasoning):** Popular for analyzing short, casual text like tweets, VADER is often the go-to for quick sentiment evaluations. However, its dependence on predefined lexicons means it can miss subtleties such as sarcasm or layered emotional tones—common in educational reviews where feedback is often detailed and nuanced (Hutto & Gilbert, 2014).

- **AFINN (Affective Norms for English Words):** Known for assigning sentiment scores to individual words, AFINN is fast and scalable for large datasets. But, without considering the context around each word, it struggles with feedback that relies on specific jargon or implied meaning, such as a student's critique of course content. (Nielsen, 2011).

- **BERT (Bidirectional Encoder Representations from Transformers):** BERT revolutionized sentiment analysis by understanding both the words that come before and after a phrase. This context-aware processing makes it highly effective for analyzing long, detailed reviews, where feedback often spans multiple ideas or sentiments (Li et al, 2019).

Earlier works, such as those by Xie et al. (2024) and Li et al. (2019), explored BERT in educational settings but focused on short-form text like course summaries. Building on their insights, this study evaluates how BERT performs with long-form reviews on Coursera. By pairing BERT with traditional tools, we aim to create a framework that combines efficiency with depth, meeting the unique challenges of sentiment analysis in educational platforms.

To illustrate how these approaches diverge in practice, consider the following student review:

"The instructor was knowledgeable, but the course content felt outdated, and I struggled with the assignments."

Below is a comparison of how VADER, AFINN, and BERT classify sentiment for this review:

Sentiment Tool	Output
VADER	Compound Score: 0.12 (Neutral)
AFINN	Sentiment Score: -1 (Slightly Negative)
BERT	Instructor Sentiment: Positive, Content Sentiment: Negative, Overall Sentiment: Mixed

- VADER assigns a near-neutral sentiment score because it balances positive and negative words without fully capturing the contrast in sentiment between different parts of the review.
- AFINN assigns a slightly negative score, as it treats words individually rather than analyzing the full sentence structure.
- BERT correctly distinguishes between different aspects of the review, assigning positive sentiment to the instructor while identifying negative sentiment towards the course content and assignments.

This example illustrates BERT's superiority in aspect-based sentiment analysis, making it particularly useful for analyzing detailed educational reviews where multiple sentiments may coexist.

This study introduces a hybrid framework that combines BERT's deep contextual understanding with the computational efficiency of traditional sentiment tools to enhance sentiment accuracy, interpretability, and scalability. By leveraging aspect-based sentiment analysis, this approach enables granular insights into student feedback, helping course designers and platform managers make data-driven improvements. This methodology offers a more precise, structured, and holistic approach to sentiment analysis in educational settings than existing models allow.

2.3 Temporal Factors in Sentiment Expression

The timing of feedback can reveal fascinating trends in sentiment expression that are often overlooked. For instance, research by Lundqvist et al. (2020) suggests that reviews written on weekends tend to be more positive than those written during weekdays. This pattern could reflect students' emotional states, with weekends potentially offering a more relaxed environment for reflection. Similarly, external events like holidays or exam periods can shape the tone of feedback, often leading to spikes in either praise or frustration.

For platforms like Coursera, understanding these temporal patterns is crucial. By analyzing how sentiment changes between weekdays, weekends, and across different months, this study attempts to identify optimal times for collecting feedback. These insights can help course providers strategically schedule surveys or launch updates during periods when students are likely to be more receptive and positive.

2.4 Aspect-Based Sentiment in Educational Contexts

Aspect-based sentiment analysis has emerged as a powerful tool for dissecting feedback into its core components, such as course content, instructor effectiveness, and structure. While traditional tools often focus on providing an overall sentiment score, they lack the precision needed to identify specific areas of praise or concern (Hutto and Gilbert, 2014).

BERT, with its ability to assign sentiment scores to individual aspects of a review, bridges this gap. For example, consistently low scores for instructor performance might highlight the need for additional training, while high scores for course structure could point to best practices worth replicating. Studies like those by Alaparthi and Mishra (2021) have applied BERT to general sentiment analysis but have not fully explored its potential in educational settings.

This research takes the next step by applying aspect-based analysis to Coursera reviews. By doing so, it attempts to provide course providers and platform managers with targeted insights that go beyond broad sentiment classifications, offering suggestions for meaningful improvements in course design and delivery.

3. Data Description

The dataset used in this research is sourced from Kaggle as the “Course Reviews on Coursera” dataset, compiled by Muhammad (2023). It comprises 1.45 million Coursera reviews, spanning a wide range of 482 courses offered by 141 institutions worldwide. This rich dataset serves as a strong foundation for analyzing sentiment across various courses, institutions, and time periods. The key elements include *Review Text:* User-generated feedback that captures detailed reflections on course experiences, providing the basis for qualitative sentiment analysis. *Reviewer ID:* An anonymized identifier for each reviewer, ensuring privacy while enabling longitudinal studies of sentiment trends. *Date of Review*: Timestamps for each review, allowing the exploration of temporal patterns in sentiment across weekdays, weekends, and months. *Rating:* A numerical score (ranging from 1 to 5) assigned by reviewers, serving as a quantitative measure of overall course satisfaction. *Course ID and Name:* Identifiers for specific courses, enabling comparisons across different disciplines and topics. *Institution:* The offering institution, facilitating institutional-level sentiment analysis to uncover variations in perceived course quality. *Course URL:* Links to course webpages, offering additional context for validation and future research.

This comprehensive dataset allows for the detailed analysis of sentiment trends, ratings, and key course aspects, such as content quality and instructor performance. Its size and diversity set it apart from prior studies that often relied on smaller, domain-specific datasets (Devlin et al., 2019).

4. Key Concepts and Importance

4.1 Quantifiable Sentiment Metrics

Quantifiable sentiment metrics were computed with the use of advanced natural language processing tools, VADER, AFINN and BERT to measure sentiment polarity. The key metrics analyzed include:

- **Sentiment Scores**: Derived from VADER, AFINN, and BERT, these scores measure sentiment polarity (positive, neutral, or negative) and are used to analyze specific aspects of a course, such as: *Course Content Quality, Instructor Effectiveness, Course Structure,* and *Overall User Satisfaction.* By comparing sentiment scores across tools, the study assesses their capacity to digest subtle feedback and contextual sentiment effectively.
- **Ratings**: Numerical ratings (1–5) assigned by students act as a benchmark for evaluating the alignment between machine-generated sentiment scores and human judgments of course quality. This comparison points out the strengths and limitations of each sentiment analysis tool.
- **Temporal Factors**: The timing of reviews (e.g., weekdays versus weekends) is examined to identify temporal patterns in sentiment expression. These insights explain how external factors, such as workload or leisure time, influence the tone of student feedback.

Together, these metrics give a solid framework for investigating aspect-specific and temporal sentiment trends, offering valuable input for enhancing feedback systems and refining course design.

4.2 Managerial Implications

The sentiment metrics derived from this study may have meaningful implications for educational platforms like Coursera, offering actionable strategies to improve user experience and course quality:

- **Personalized Course Recommendations**: By identifying the factors that drive satisfaction, such as engaging content or effective pacing, platforms can tailor course recommendations to individual users, boosting engagement and retention rates.
- **Targeted Feedback Systems**: Aspect-based sentiment analysis enables course providers to determine and address specific areas needing improvement, such as instructor performance or outdated materials. This focused approach promotes more efficient resource allocation and impactful interventions.

- **Real-Time Feedback for Instructors**: Continuous monitoring of sentiment trends allows instructors to track the effects of course updates or new teaching methods. This immediate feedback equips them to address student concerns proactively and fine-tune their approach.
- **Benchmarking and Performance Analysis**: By analyzing sentiment trends across courses and institutions, platforms can identify standout performers and areas requiring improvement. These benchmarks could guide decision-making, ensuring competitive course offerings and encouraging steady progress.

In summary, our results offer educational platforms to optimize their feedback mechanisms, refine course design, and generally improve learning experience. By acting on these insights, platforms can better meet the changing needs of students and educators alike.

5. Research Hypotheses, Findings and Analysis

The hypotheses in this study aim to assess the predictive power of sentiment analysis tools, explore temporal and contextual sentiment variations, and examine how factors such as review length, course content, and institutional characteristics influence sentiment. The results give actionable insights for enhancing feedback systems and help improve course offerings and student satisfaction.

Hypothesis 1: BERT + Covariables explain variance more accurately than traditional tools

Rationale: Traditional sentiment analysis tools such as VADER and AFINN rely on fixed lexicons and heuristic rules, which often fail to capture the nuanced context in detailed student reviews. In contrast, BERT's bidirectional architecture can more effectively interpret complex, context-dependent language, yielding a normalized sentiment score (bert_norm) that captures a holistic view of student feedback. Moreover, adding contextual covariates such as weekday, month, year, institution, and course name (via one-hot encoding) and review length enables the model to account for temporal and course-specific factors. This extended approach, labeled as *BERT + Covariables* is designed to boost the model's predictive power and provide a more accurate reflection of overall course reception.

Findings: In explaining sentiment variations, the differences by model are as follows: *Traditional Sentiment Tools (VADER + AFINN)*: $R^2 = 0.121$; *BERT + Covariables*: R^2: 0.394 and MSE: 0.393. Incorporating additional contextual variables has significantly improved the model's ability to explain the variation in course ratings. An R^2 of 0.394 indicates that nearly 39.4% of the variation in ratings is accounted for by the extended feature set.

Analysis: The improvement in predictive performance underscores the importance of integrating holistic sentiment analysis (via BERT) with contextual factors. Variables such as weekday, month, year, institution, and course name add explanatory value and help refine our predictions by capturing external influences and course-specific characteristics. This guided our next hypotheses.

Hypothesis 2: Temporal Variations Exist in Sentiment Expression

Rationale: Temporal factors such as academic workload, exam schedules, and holidays can influence student sentiment. For example, reviews written on weekends might reflect a different mood compared to those written on weekdays due to changes in stress levels or available time. Similarly, seasonal effects (e.g., summer versus non-summer periods) may capture variations in student experiences driven by the academic calendar.

Findings: *Overall Difference Between Weekend and Weekday Reviews:* Weekday Average Rating = 4.67 Weekend Average Rating: 4.68 (t-statistic = 4.652, p-value = 0.00000). Statistically significant difference between weekend and weekday reviews. Although the difference in mean rating is small (0.01 point), the statistical test confirms it is reliably detectable given the large sample size. *Seasonal Comparison (Summer vs. Non-Summer):* Summer Average Rating: 4.69 Non-Summer Average Rating: 4.66 (t-statistic = 12.892, p-value = 0.00000) - Statistically significant difference between summer and non-summer reviews. This difference (approximately 0.03 points) suggests that external factors tied to the season—such as varying academic pressures, holidays, or workload changes—may influence student sentiment more noticeable than the day of the week.

Analysis: The data spans from 2015 through 2020, providing robust evidence that these temporal variations are not due to random chance. Future studies should further explore these temporal influences by investigating additional contextual factors, such as holiday breaks or major course announcements and deadlines (Dalipi et al., 2021).

Hypothesis 3: Course ratings are better predicted by holistic than isolated measures

Rationale: Previous research indicates that students often struggle to pinpoint specific negative aspects of a course, tending instead to provide a generalized, overall impression (Marsh & Roche, 1997). In our dataset, aspect-specific sentiment measures (e.g., instructor, content, and structure) showed only weak correlations with overall ratings, suggesting that students' feedback on these individual components does not strongly reflect their overall course evaluation. In contrast, holistic sentiment measures—such as the normalized BERT score—capture a more unified impression of the course. Furthermore, temporal analyses reveal that external factors (e.g., workload, deadlines, and seasonal effects) also influence how students rate courses. Together, this suggests that overall course reception is more dependent on general, holistic sentiment and contextual, temporal factors than on detailed, aspect-specific evaluations.

Findings: Pearson correlation coefficients were computed between the overall course rating and different sentiment metrics. The findings are summarized as follows: The holistic sentiment measure, represented by the normalized BERT score (bert_norm), exhibited a moderate positive correlation with the overall course rating (r = 0.508, p < 0.00001). In comparison, the aspect-specific sentiment measures demonstrated almost no correlations: *Instructor sentiment*: r = 0.119, p < 0.00001; *Content sentiment*: r = 0.047, p < 0.00001; *Structure sentiment*: r = 0.033, p < 0.00001.

Analysis: These results indicate that the holistic sentiment measure encapsulates a more robust signal related to overall course evaluation relative to the finer-grained, aspect-specific indices. Follow-up studies should separate positive and negative clauses or apply aspect-level models before retesting the link between review length and sentiment.

Hypothesis 4: Sentiment varies between different courses of the same institution

Rationale: Variations in course content, delivery style, and how well a course aligns with student expectations can lead to differences in overall sentiment. If students provide holistic evaluations that reflect these factors, then normalized sentiment measures (e.g., bert_norm) should vary significantly across courses within the same institution.

Findings:

ANOVA Results for Normalized BERT Sentiment Across Courses by Institution

Institution	F-statistic	p-value	No. of Courses
Yale University	10.59	1.95e-27	17
University of Michigan	101.12	0.00e+00	30
DeepLearning.AI	105.57	7.91e-316	16
Johns Hopkins University	75.84	1.19e-176	13
IBM	61.49	1.51e-199	18
University of Virginia	4.70	1.46e-09	17
University of California, Irvine	7.21	5.27e-15	15
Stanford University	27.57	6.28e-53	11

Google Cloud	7.80	1.20e-23	22
University of Pennsylvania	24.53	4.79e-180	43
University of London	5.28	2.83e-08	12
University of California, Davis	14.83	7.21e-31	13
Duke University	31.48	1.10e-125	23
University of Illinois at Urbana-Champaign	7.99	6.94e-16	14

The low p-values (< 0.05) across institutions confirm that the differences in normalized BERT sentiment scores across courses are statistically significant. This indicates that course-specific factors (e.g., course design, content, delivery) meaningfully affect overall sentiment. The analysis spans 6 years (from 2015 to 2020), ensuring robust temporal coverage. Additionally, only institutions with more than 10 courses (each with at least 2 reviews) were included, strengthening the reliability of these findings.

Analysis: These results imply that overall course reception is more strongly influenced by holistic factors rather than the institution offering the course alone, and that course decisions should be made based on a course-by-course basis rather than by its offering institution. Educational platforms can use this insight to focus on identifying underperforming courses to improve student satisfaction and make more informed course marketing decisions.

Key Takeaways:

The analysis demonstrates that incorporating BERT into sentiment analysis could potentially enhance predictive accuracy compared to traditional tools like VADER and AFINN. While BERT excels at capturing nuanced and contextual sentiment, traditional tools remain valuable for tasks where computational efficiency is essential as suggested by Xie et al. (2024). The weak correlations observed between aspect-based sentiment metrics (e.g., instructor quality, content, and structure) and overall course ratings suggest that student satisfaction is influenced by a combination of holistic factors rather than isolated course elements. Temporal variations in sentiment further reveal how feedback may fluctuate over time, highlighting the strategic importance of timing in feedback collection to capture more representative and actionable insights. The three key takeaways for course designers are as follows:

1. **Adoption of Hybrid Sentiment Models**: A hybrid approach that combines BERT with traditional tools offers a balanced solution for sentiment analysis, leveraging BERT's contextual depth alongside the efficiency of traditional methods.
2. **Timing Feedback Cycles**: Analyzing temporal sentiment trends could potentially help platforms strategically schedule feedback collection and course updates, increasing engagement during high-sentiment periods.
3. **Holistic Course Development**: The weak correlations between individual course aspects and overall satisfaction underscore the importance of tackling all course components comprehensively—content, structure, and instructor quality—for meaningful improvements.

By integrating these findings into their feedback mechanisms, educational platforms could potentially enhance student satisfaction, improve course design, and increase sustained engagement over time.

5. Limitations

While this study may give information in sentiment analysis in educational reviews, several limitations should be acknowledged:

1. Lack of Instructor Demographic Data: One key limitation is the absence of instructor demographic data (e.g., race, gender), which may influence student sentiment. Prior research by Dalipi et al. (2021) suggests that implicit biases can affect course evaluations, with students potentially rating instructors differently based on identity factors rather than objective course quality. However, due to privacy restrictions, our dataset does not include demographic attributes, making it impossible to measure these effects directly. Future studies could address this gap by incorporating demographic data where available, enabling a more thorough understanding of sentiment patterns in educational reviews.

2. Changes in Coursera's Review Policies: Coursera's review policies have changed over time, which may introduce selection bias in sentiment trends. In earlier versions of the platform, students could leave reviews without completing a course, while more recent policies may restrict reviews to students who have completed or paid for a course. These changes may influence sentiment distributions, as students who complete a course are generally more likely to leave positive feedback compared to those who drop out early (Lundqvist et al., 2020). Since our dataset does not differentiate between paid vs. free users or completers vs. non-completers, we cannot assess how these factors impact sentiment. Future work should investigate how sentiment trends vary based on student enrollment type (paid vs. free) and course completion status to better understand potential biases.

3. Dependence on Textual Data: This study relies solely on text-based sentiment analysis, without incorporating non-textual feedback such as course engagement metrics, quiz

performance, or discussion forum interactions. These additional data sources could give a fuller picture of student experiences. A multimodal approach—combining textual analysis with behavioral data—could improve sentiment prediction accuracy and offer clearer takeaways in regards to student satisfaction (Baltrušaitis et al., 2019). Future research could integrate quantitative course performance indicators alongside sentiment analysis to enhance predictive models.

Despite these limitations, this study attempts to make meaningful contributions to the field of educational sentiment analysis by evaluating sentiment trends at scale, comparing traditional lexicon-based tools with transformer models, and providing actionable insights for course designers and platform managers. Addressing these limitations in future work will further refine sentiment analysis methodologies and improve the understanding of student feedback in online education platforms.

6. Future Research Directions

Future research can build upon this study by expanding aspect‑based sentiment analysis, examining longitudinal trends, conducting cross‑platform comparisons, and integrating multimodal data for a better understanding of student feedback in online education (Baltrušaitis et al., 2019).

1. Enhanced Aspect‑Based Sentiment Analysis: Future studies should extend aspect‑based sentiment analysis beyond core elements like instructor performance and course content to include course relevance, peer interactions, instructional style, and engagement metrics. Analyzing these additional aspects may uncover new drivers of student satisfaction and provide targeted recommendations for course improvement. Moreover, advanced transformer models like BERT can be further refined to capture subtle contextual variations in student sentiment, particularly in areas not well‑detected by traditional sentiment tools (Devlin et al., 2019).

2. Longitudinal Sentiment Trends: A long-term temporal analysis of sentiment across multiple semesters or years could help platforms identify seasonal trends, track shifts in sentiment following course modifications, and assess the impact of evolving platform policies. Such investigations can enhance curriculum design, improve instructor responsiveness, and optimize the timing of course updates and promotional efforts. In addition, (Lundqvist et al., 2020) provide insights into dynamic sentiment patterns in evolving online communities that can inform future longitudinal studies in MOOCs.

3. Cross‑Platform Sentiment Comparison: Comparative sentiment analysis across multiple online learning platforms (e.g., Coursera, edX, Udemy, Khan Academy) could highlight platform-specific trends, engagement drivers, and sentiment variations (Xie et al., 2024). By identifying differences in course delivery, user interaction, and rating behaviors, future research can help platforms refine their recommendation algorithms, tailor course structures, and improve student retention strategies. Recent empirical study by (Lundqvist et al., 2020) has demonstrated

the value of cross-platform analysis in identifying unique user sentiment trends that differ from platform to platform.

4. Generalizability Across Languages and Disciplines: To assess the robustness of BERT-based sentiment models, future research should test their effectiveness across diverse languages, subject areas, and cultural contexts. Sentiment models trained primarily on English-language reviews may not generalize well to non-English courses or underrepresented disciplines. Evaluating these models in multilingual educational settings could potentially improve cross-cultural sentiment interpretation and promote inclusivity in online learning analytics.

5. Multimodal Sentiment Analysis: Future research could integrate multimodal data sources, such as audio feedback, video responses, and facial expression analysis, to enhance sentiment classification beyond textual reviews. Machine learning models that incorporate voice tone, visual engagement, and emotional cues may offer a richer, more holistic view of student experiences, allowing platforms to develop personalized learning pathways and adaptive course recommendations. Complementary surveys on multimodal machine learning (Baltrušaitis et al., 2019) underscore the potential benefits and challenges associated with integrating diverse data modalities for sentiment analysis.

7. Conclusion

This study underscores the benefits of integrating BERT-based models with traditional lexicon-based sentiment tools like VADER to analyze student feedback on online education platforms. While traditional tools effectively capture overall sentiment polarity, BERT's advanced transformer-based architecture excels in recognizing contextual nuances, enabling a more detailed and precise analysis of feedback.

The findings demonstrate the strength of a hybrid approach, where combining BERT with lexicon-based methods significantly improves both predictive accuracy and interpretive depth. This dual strategy provides course designers and platform managers with actionable insights, helping them better understand and respond to the needs of their students. However, the weak correlations observed between aspect-based sentiment (e.g., instructor quality, course content, structure) and overall ratings suggest that student satisfaction is influenced by a combination of interrelated factors. Focusing on isolated aspects is unlikely to drive meaningful improvements without a more holistic strategy for enhancing the overall learning experience.

Temporal patterns in sentiment further highlight the importance of timing in feedback collection. Periods with heightened positive sentiment, such as specific months or weekends, present valuable opportunities for platforms to optimize their feedback cycles, solicit more favorable reviews, and schedule key course updates or launches to align with these high-sentiment intervals.

In summary, this research highlights the potential of hybrid sentiment analysis models that integrate BERT with traditional tools. Together, these methods create a comprehensive framework for extracting insights from feedback, improving course quality, and strategically planning initiatives to boost student engagement. By leveraging aspect-specific feedback and temporal sentiment trends, online education platforms can make informed, data-driven decisions to better address the evolving needs of their learners.

8. References

1. F. Dalipi, F. Ahlgren, and K. Zdravkova, "Sentiment analysis of students' feedback in MOOCs: A systematic literature review," *Frontiers in Artificial Intelligence*, vol. 4, 2021.

2. C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, Ann Arbor, MI, Jun. 2014, pp. 216-225.

3. F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," *Proceedings of the ESWC Workshop on Making Sense of Microposts*, Heraklion, Greece, May 2011, pp. 93-98.

4. L. MacNell, A. Driscoll, and A. N. Hunt, "What's in a name? Exposing gender bias in student ratings of teaching," *Innovative Higher Education*, vol. 40, no. 4, pp. 291-303, 2015.

5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the NAACL-HLT Conference*, Minneapolis, MN, Jun. 2019, pp. 4171-4186.

6. L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, e1253, 2018.

7. P. Xie, H. Gu, and D. Zhou, "Modeling sentiment analysis for educational texts by combining BERT and FastText," *Proceedings of the International Conference on Computer Science and Technologies in Education (CSTE)*, Xi'an, China, Jul. 2024, pp. 1-6.

8. X. Li, H. Zhang, Y. Ouyang, X. Zhang, and W. Rong, "A shallow BERT-CNN model for sentiment analysis on MOOCs comments," *Proceedings of the IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, Yogyakarta, Indonesia, Dec. 2019, pp. 1-6.

9. K. Ø. Lundqvist, T. Liyanagunawardena, and L. Starkey, "Evaluation of student feedback within a MOOC using sentiment analysis and target groups," *International Review of Research in Open and Distributed Learning*, vol. 21, no. 3, pp. 140-156, 2020.

10. S. Alaparthi and M. Mishra, "BERT: A sentiment analysis odyssey," *Journal of Marketing Analytics*, vol. 9, no. 2, pp. 118-126, 2021.

11. M. Muhammad, “Course Reviews on Coursera,” Kaggle dataset, 2023. Available: https://www.kaggle.com/datasets/iamspruce/course-reviews-on-coursera (accessed June 20 2024).

12. H. W. Marsh and L. A. Roche, “Making students’ evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility,” *American Psychologist*, vol. 52, no. 11, pp. 1187‑1197, 1997.

13. T. Baltrušaitis, C. Ahuja, and L.‑P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423‑443, 2019.