

You, W. & Horissian, K. (2026). AI-Assisted sentiment analysis of student voice: A human-validated comparison of lexicon- and transformer-based tools. *Intersection: A Journal at the Intersection of Assessment and Learning*, 7(5), 31–55. <https://doi.org/10.61669/001c.172638>

AI-Assisted Sentiment Analysis of Student Voice: A Human-Validated Comparison of Lexicon- and Transformer-based Tools

Wei You, Ph.D., Kevork Horissian, M.A.

Author Note

Wei You

Kevork Horissian

We have no conflicts of interest to disclose

Intersection: A Journal at the Intersection of Assessment and Learning

Volume 7, Issue 5 Special Issue: The Future of Assessment in an Era of Generative AI

Abstract: Open-ended survey responses offer rich insights into the student experience, yet analyzing this qualitative data at scale is often time-consuming and labor-intensive. This study evaluates whether off-the-shelf sentiment analysis tools can validly classify such data for institutional assessment. Analyzing 1,220 free-text responses from undergraduate students at a private liberal arts university, we compared the performance of four tools: two lexicon-based (TextBlob and VADER) and two transformer-based (BERT-nlptown and RoBERTa-cardiffnlp) against manual coding. Evaluation metrics included accuracy, macro F_1 score, and Cohen’s kappa (κ). While BERT-nlptown performed best, it achieved only moderate agreement with human coding ($\kappa = .42$). Furthermore, lexicon-based approaches misclassified roughly 20% of responses due to limited vocabulary coverage. Notably, all tools struggled to accurately classify negative, neutral and mixed sentiments. We recommend conducting lexicon-coverage diagnostics, few-shot prompting, domain-specific finetuning, and human-in-the-loop oversight to ensure valid, reliable, and trustworthy institutional assessment.

Keywords: *sentiment analysis, survey research, student voice, institutional assessment, natural language processing, validity, human-in-the-loop*

Introduction

In higher education, student perceptions of their educational experience shape their engagement, learning, and persistence. Extensive research indicates that students who feel a stronger sense of belonging are more likely to persist; student campus involvement promotes learning and development; and student engagement is associated with stronger academic performance and retention during the first year (Astin, 1984; Boyd et al., 2022; Hausmann et al., 2007; Kuh et al., 2008). Understanding how students perceive their educational experience is therefore crucial for informing institutional efforts to improve student success.

While Likert-scale survey items are widely used in institutional assessment to measure student perceptions and satisfaction, their closed-ended nature restricts respondents to pre-defined response categories established by the researcher. In contrast, open-ended survey questions allow students to describe their experience in their own words, freely expressing unanticipated concerns and identifying specific areas requiring institutional attention (Alhija & Fresko, 2009; Stupans et al., 2016). Despite their value and insight, qualitative feedback collected via open-ended survey questions is often left unanalyzed because it is labor-intensive and time-consuming (Geer, 1991; O’Cathain & Thomas, 2004;

Schuman & Presser, 1981). As higher education institutions increasingly rely on student voice data to inform strategic decision-making, there is a growing need for scalable analytic approaches that can efficiently process large volumes of qualitative feedback while maintaining methodological rigor and interpretative depth (Slotnick & Boeing, 2025). This need motivates the present study, which focuses on the systematic analysis of sentiments expressed in free-text student voice data.

This study draws on validity theory from educational and psychological measurement. From this perspective, sentiment labels generated by automated classification tools serve as measurement outcomes that must be validated rather than assumed (Kane, 2013; Messick, 1995). Using manual human coding as the criterion, this study evaluates the extent to which automated sentiment classifications support the intended inference that student responses reflect positive, negative, neutral or mixed sentiment. To assess the quality of these inferences, we report accuracy, macro F_1 score, and Cohen's kappa and interpret the results using established benchmarks for classification performance and agreement.

Accordingly, the purpose of this study is not simply to identify which tool performs best, but to assess the validity of each tool's classification in a specific institutional assessment context. Within this framework, the classification errors are conceptualized as potential threats to validity that assessment practitioners must consider when using automated sentiment analysis tools. By doing so, this framework explicitly connects the value of student voice as a source of assessment evidence with a rigorous evaluation of the automated methods used to interpret that evidence.

Defining Sentiment

Sentiment analysis refers to the use of computational methods to analyze opinions, attitudes, and emotional tone expressed in text (Medhat et al., 2014). Foundational work in this field commonly classifies text at the document or sentence level as positive, negative, or neutral, while also recognizing that a document may express both favorable and unfavorable evaluations of different aspects of the same target (Liu, 2012, p. 41). A response is coded positive when it expresses a favorable evaluation, emotion, or attitude toward the target, which in this study is students' favorable expression of their experience (e.g., "fun," "exciting," "supportive," "a welcoming community"). A response is coded negative when it expresses an unfavorable evaluation or attitude (e.g., "stressful," "isolating"). It is coded as neutral when a response conveys no evaluative content, such as factual statements or texts in which sentiment is simply absent. A fourth category, mixed, captures responses that express genuinely ambivalent sentiment, praising one aspect of the experience while criticizing another within a single response (Dave et al., 2003; Soboroff et al., 2008).

Sentiment Analysis in Educational and Assessment Contexts

While sentiment analysis originated in domains such as product reviews and social media, its application within educational research has grown rapidly over the past decade. Researchers have applied sentiment analysis to a wide range of contexts, from evaluating teaching effectiveness and enhancing learning systems to supporting institutional decision-making (Dalipi et al., 2021; Grimalt-Álvaro & Usart, 2024). For example, in a systematic review of 40 studies on massive open online

courses, Dalipi et al. (2021) identified key areas for sentiment analysis, including student evaluations of course content, detection of contradictory feedback, and understanding of course performance and student dropout. The authors also highlighted a shift from traditional lexicon-based approaches (e.g., TextBlob) toward deep learning methods to automate the analysis of student feedback.

More recently, researchers have extended these applications beyond course-level feedback to educational policy and broader measures of the student experience. Liu and Sun (2025) combined topic modeling, lexicon-based sentiment analysis, and manual coding to analyze stakeholder interview data regarding K-12 education policy. They concluded that human-GenAI collaboration enhanced efficiency while preserving human domain expertise. Grimalt-Álvaro and Usart (2024) examined the use of sentiment analysis as a formative assessment tool across 22 studies in online and hybrid higher education contexts. They noted that sentiment analysis can give instructors insight into their learning performance and emotional state, providing information that may not be captured by grades alone. Within this growing body of work, systematic reviews have documented TextBlob and VADER as widely used lexicon-based tools (Dalipi et al., 2021; Kastrati et al., 2021). These reviews further noted that more advanced transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT), have remained underexplored in educational research (Kastrati et al., 2021).

At the same time, comparative studies have shown that different sentiment analysis tools frequently produce inconsistent results and may diverge from human interpretation (Liu & Sun, 2025; Lu et al., 2023; Ma et al., 2025). Lu et al. (2023), for example, applied several lexicon-based sentiment analysis tools to more than 20,000 family medicine resident feedback comments and found that their performance was limited by ambiguous terminology. The authors concluded that domain-specific adaptations were necessary to improve accuracy. Similar findings were reported by Ma et al. (2025), who demonstrated that lexicon-based sentiment analysis could inform decision-making only when its validity limitations were carefully acknowledged. Liu and Sun (2025) likewise reported a discrepancy between sentiment classification generated from VADER and human expert judgment.

Notably, most of these studies relied on a single sentiment analysis tool or compared tools from the same lexicon-based family. As a result, relatively little is known about how alternative approaches, particularly more advanced transformer-based models, perform on institutional data in higher education. Even less empirical evidence exists regarding the extent to which sentiment classifications generated by lexicon- and transformer-based tools align with human judgment and support valid interpretation. This gap is especially critical for higher education assessment professionals, who increasingly need scalable and reliable methods to extract actionable insights from large volumes of qualitative student feedback.

The present study addresses this gap by conducting a human-validated comparison of multiple off-the-shelf sentiment analysis tools representing both lexicon-based and transformer-based approaches. We selected these two approaches to contrast traditional natural language processing methods that rely on dictionaries with more advanced deep learning methods architectures that are widely accessible to institutional researchers. Using open-ended student survey responses from a private liberal arts

university in the United States, the study investigates the following research question: “To what extent do two lexicon-based tools (TextBlob and VADER) and two transformer-based tools (i.e., BERT-nlptown and RoBERTa-cardiffnlp) align with the ground truth of manual coding?” In doing so, this study extends the application of AI-assisted methods in institutional assessment, providing practical recommendations for assessment professionals seeking to incorporate these tools into their workflows.

Method

Data

The data came from a homegrown survey designed to assess various aspects of students’ experiences, including their expectations and perceived overall fit, sense of community, well-being, housing, dining, and campus safety. In Fall 2025, the survey was sent to all undergraduate students ($N = 3,882$) at Bucknell University, a private liberal arts university in the Mid-Atlantic region in the United States. The survey remained open for five weeks, and a total of 1,486 students responded, with a 38% response rate. Among these respondents, 1,220 students responded to the open-ended question, “What words or phrases would you use to describe the Bucknell experience?” The demographic information of these respondents is presented in Table 1. The sample closely reflects the demographic characteristics of the overall student population.

Responses to this open-ended question ranged from single words (e.g., “great,” “stressful”) to more detailed, multi-sentence reflections (e.g., “It’s peaceful. Quiet. relaxation. away from noisy cities. supportive. small by space but huge by resources.”). While the longest response contained 63 words, the median was three words. For data integrity and participant anonymity, each respondent was assigned a unique random identifier (ID).

Comparative Evaluation of Sentiment Analysis Tools

To evaluate the performance of automated sentiment classification methods, this study used a multi-model comparative evaluation framework. We compared the outputs of four sentiment analysis tools, TextBlob, VADER, BERT, and RoBERTa, against manually coded labels. All personally identifiable information was removed prior to analysis to ensure participant confidentiality. The analysis consisted of two stages: First, the authors manually coded the anonymized responses through an iterative coding process; Then, each sentiment analysis tool was applied to the same set of responses in a Python environment to generate sentiment classification. Model outputs were compared with the manually coded labels using established measures of classification performance.

Manual Coding

Following the consensual qualitative research approach (Hill et al., 1997, 2005), we conducted an iterative coding process in which the lead author served as the primary coder and the second author served in the auditor role described in this approach, reviewing ambiguous cases, challenging interpretations, refining and finalizing the coding framework through discussion. We adopted this approach deliberately because many of the survey responses were short, often a single word or phrase, sometimes ambiguous, and heavily context-dependent; our goal was to develop a shared

interpretative standard, refined through audited discussion, which could serve as a defensible criterion for evaluating the automated tools. Accordingly, no inter-rater reliability was computed during the coding process. Instead, we documented the coding and audit process and the resulting decision rules presented in Table 2.

Table 1

Descriptive Characteristics of the Sample (N = 1,220)

Demographic characteristics	<i>n</i>	%
Gender		
Female	676	55
Male	544	45
Ethnicity		
White	812	67
Hispanic/Latino	140	12
U.S. Nonresident	85	7
Black/African American	62	5
Multiple Races	59	5
Asian	57	5
Student Classification		
First-year	327	27
Sophomore	294	24
Junior	276	23
Senior	311	26
Other classification	12	1
First-Generation Student		
No	1,002	82
Yes	218	18
Pell Recipient		
No	1,014	83
Yes	206	17

Note. Percentage may not sum to 100% due to rounding

Table 2

Manual Coding – Final Coding Standard and Decision Rules

Sentiment label	Definition/decision rule	Examples
Positive	The response conveys a favorable evaluation of the student experience or a clear positive valence.	“great”, “fun and engaging”, “amazing community”
Negative	The response conveys an unfavorable evaluation, or a clear negative valence.	“stressful”, “cliquey”, “isolating”
Mixed	The response contains both a clear positive and a negative element.	“Great academic, poor social life”
Neutral	The response is descriptive or factual.	“Small”, “rural”
Triangulation rule for ambiguous terms	For short, context-dependent responses (e.g., “challenging,” “busy”, “intense”), use the “Institution fit” and “Recommend Bucknell” questions for contextual information.	Positive if the student’s response is “challenging”, and their response to the fit question is 4/4 or 3/4, or their recommend score is 6/10, 7/10, 8/10, 9/10.
	For high fit (3-4) and/or highly recommend (6 -10), code the ambiguous term as “Positive”.	Negative if the student’s response is “challenging” and their response to the fit question is 1/4 or 2/4, or their recommend score is 1/10, 2/10, 3/10, or 4/10.
	For Low fit (1-2) and/or low recommend (1-4), code the ambiguous response as negative.	
One label only	Every response receives one label. If a response contains both positive and negative sentiment, code it as “Mixed”; if it is ambiguous, apply the triangulation rule.	

Note: One-hot encoding was applied to all the labels, with 0 indicating the absence of a label and 1 indicating its presence.

More specifically, during the first round of coding, the lead author, who served as the primary coder, assigned binary codes (0 = absent, 1 = present) to a sample of survey responses across four mutually exclusive categories: positive, negative, mixed, and neutral. Positive responses expressed a favorable evaluation, emotion, or attitude toward the student experience; negative responses expressed an

unfavorable evaluation or affect; neutral responses contained no clear evaluative content; and mixed responses contained both positive and negative evaluations. Each response received exactly one classification. Following the initial coding, both authors convened to review the coding process, with the second author serving as the auditor to resolve ambiguous comments, challenge interpretations, and refine the coding standard and decision rules. The audited review produced the finalized criteria describing the shared interpretative standard summarized in Table 2, which the lead author then applied in a second round of coding of all 1,220 responses.

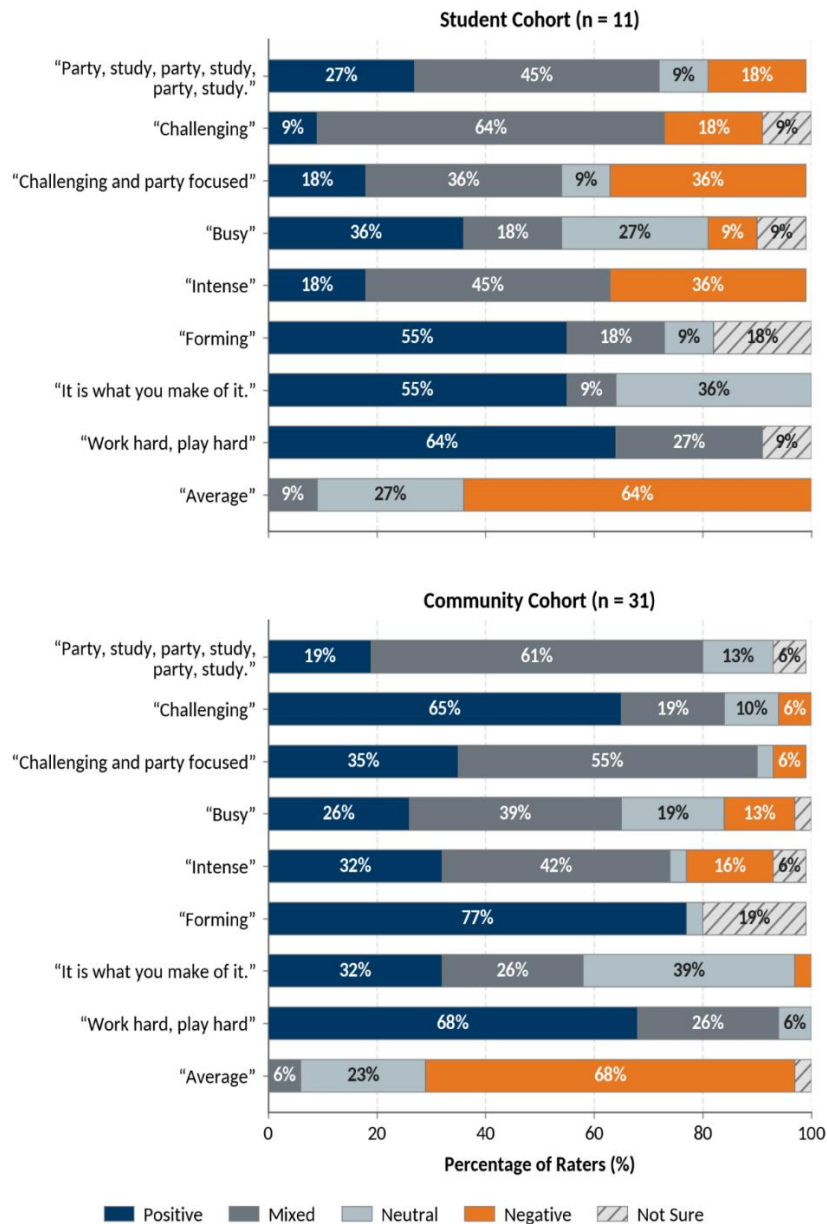
Several decision rules emerged from the consensus coding process. Sentiment was interpreted in the context of the survey question, “What words or phrases would you use to describe the Bucknell experience?” rather than solely from the conventional polarity of individual words. Clearly favorable or unfavorable evaluations were classified as positive or negative, respectively; responses containing both positive and negative evaluations were classified as mixed; and responses without discernible evaluative content were classified as neutral. When a word or phrase could reasonably carry different meanings in the institutional context, the authors jointly reviewed the response and considered additional contextual evidence when available. Decisions reached through these discussions were used to clarify the coding standard and guide the classification of substantively similar responses.

Two additional procedures provided contextual evidence for interpreting ambiguous language. First, we conducted two pilot studies in which a student cohort ($n = 11$) and a local community member cohort ($n = 31$) classified nine ambiguous comments using the same four sentiment categories, with an “Unsure” option when sentiment could not be determined. The pilot results are presented in Figure 1. Agreement among the cohort groups was low: Fleiss’s kappa $\kappa = .11$ for the student cohort and $\kappa = .19$ for the community cohort, confirming that some of these comments are genuinely ambiguous to independent readers (e.g., “busy,” “challenging”) and that sentiment for such responses cannot be determined from wording alone (Fleiss, 1971).

For such ambiguous and context-dependent responses, we used the triangulation method when respondent-level survey data could help clarify context-dependent responses, such as “challenging.” We cross-referenced these comments against two quantitative survey questions: 1. A 4-point scale assessing institutional “fit” (“Overall, how would you describe the University’s ‘fit’ for you?”); 2. A 10-point scale regarding the likelihood of recommending the institution (“How likely are you to recommend the University to a friend?”). These measures were treated as contextual evidence for interpretation of ambiguous sentiment communicated in student responses. For example, the comment “challenging” was interpreted as positive if the respondent reported high institutional fit (4/4) and/or a high recommendation score (9/10), whereas it was coded as negative when associated with a low institutional fit (1/4) and/or a low recommendation score (3/10). When ambiguity remained, both authors discussed the response and available evidence until agreement was reached. As shown in Table 2, once a decision rule was established, it was applied consistently to substantively similar cases.

Figure 1

Pilot Results



The objective of this consensual qualitative research approach adopted in this study was to develop a shared interpretation and construct a human reference standard through discussion and adjudication rather than estimate the reliability of two independent coders. Subsequent analyses of accuracy, macro F_1 , and Cohen's kappa quantify the extent to which each automated tool reproduced the adjudicated human reference standard.

Four sentiment analysis tools were applied to each response: two lexicon-based methods (TextBlob and VADER) and two transformer-based deep learning models (BERT-nlptown and RoBERTa-cardiffnlp). It is important to note that none of the four tools produces a mixed sentiment category by default, so we evaluated the outputs at the three-class granularity (positive, negative, neutral/mixed). To align with our manual coding, the neutral and mixed categories were combined into a single category for comparison. The implications of the absence of a distinct mixed category are explored in detail in the Limitations section.

Lexicon-Based Approaches

TextBlob (Loria, 2025) is a Python library for sentiment analysis that computes sentiment polarity (−1.0 to +1.0) and subjectivity (0.0 to 1.0), using a lexicon-based approach built on the Pattern Library (De Smedt & Daelemans, 2012). The Pattern Library’s sentiment lexicon contains approximately 2,900 entries, predominantly English adjectives, and assigns each entry a polarity score in the range between −1.0 and +1.0 and a subjectivity score in the range between 0.0 and 1.0. TextBlob estimates the overall polarity of a text by averaging the polarity values of all recognized tokens (De Smedt & Daelemans, 2012). It is important to note that the TextBlob library does not prescribe a standard classification threshold; its documentation defines only the polarity score range, leaving threshold selection to the researcher. A common convention in the literature classifies polarity values greater than 0 as positive, less than 0 as negative, and equal to 0 as neutral. However, a zero-point threshold may lead to false classifications of responses with very low polarity scores.

The Valence Awareness Dictionary and sEntiment Reasoner (VADER; Hutto & Gilbert, 2014) is a lexicon- and rule-based sentiment analysis tool specifically developed for social media. VADER generates a compound score between −1.0 and +1.0 along with separate positive, negative, and neutral sub-scores that sum to 1.0. The compound score is computed by summing the valence scores of each lexicon-hit word and applying heuristics adjustments for punctuation emphasis, capitalization, degree modifiers, conjunctions, and negation, after which the score is normalized to the −1.0 and +1.0 range (Hutto & Gilbert, 2014).

In contrast to TextBlob, VADER provides a recommended threshold for three-class sentiment classification. Therefore, the present study adopted a ± 0.05 threshold aligned with the VADER convention (Hutto & Gilbert, 2014) to enhance comparability across tools. Accordingly, responses with polarity $\geq .05$ were classified as positive; those with polarity ≤ -0.05 as negative; and those between $-.05$ and $.05$ were classified as neutral/mixed. Because the Pattern Library contains primarily adjectives, survey responses that fell outside of this lexicon were assigned to a default polarity of 0 regardless of their actual sentiment. The prevalence of these out-of-lexicon responses in the corpus, along with their impact on neutral/mixed classification, is reported in the Results section.

Transformer-Based Approaches

Transformer-based models address several limitations of the lexicon-based approaches by incorporating contextual information when interpreting text. Rather than assigning sentiment labels from predefined dictionaries, these models represent each token as a function of its surrounding

words and are better in capturing nuanced meaning. One of the most influential transformer architectures is Bidirectional Encoder Representations from Transformers (BERT), a language model designed for deep contextual language understanding that can be finetuned for downstream tasks such as sentiment classification (Devlin et al., 2019).

In this study, sentiment was analyzed using the `nlptown/bert-base-multilingual-uncased-sentiment` model (NLP Town, 2020), hereafter referred to as BERT-nlptown. This multilingual model was fine-tuned on a large corpus of product reviews across six languages (English, Dutch, German, French, Spanish, and Italian) to predict sentiment on a five-point scale. For each survey response, the model generated a probability distribution over the five rating categories. To generate a single continuous sentiment score, the predicted probabilities were converted into an expected-value estimate by multiplying each rating category (1-5) by its corresponding probability and summing the resulting values. The resulting scores ranged from 1 to 5, with higher values indicating more positive sentiment. Because the model generates a one-to-five-star rating, we mapped its output to three polarity classes using cut-points centered on the scale midpoint of 3. Ratings above 3.5 were coded positive, ratings below 2.5 were coded negative, and the symmetric band between 2.5 and 3.5 inclusive was coded neutral. Thresholding an ordinal rating scale into polarity classes is an established practice (Dave et al., 2003, p. 521; Hutto & Gilbert, 2014, p. 223); we set the boundaries symmetrically around the midpoint so that the neutral band reflects the scale's center rather than an arbitrary cut.

RoBERTa is developed as an optimization of the BERT architecture and relies on longer training and larger datasets to achieve better benchmark performance (Devlin et al., 2019; Liu et al., 2019). Like BERT, it relies on sub-word tokenization, a preprocessing technique that breaks words into smaller units rather than treating each word as a single token. This approach allows models to represent rare and previously unseen words by combining meaningful sub-word components learned during training. Sentiment classification was performed using the `cardiffnlp/twitter-roberta-base-sentiment-latest` model (Loureiro et al., 2022), an updated version of the original `cardiffnlp/twitter-roberta-base-sentiment` model introduced by Barbieri et al. (2020). Built on the RoBERTa architecture (Liu et al., 2019), it is finetuned on approximately 124 million tweets posted between January 2018 and December 2021 for three-class sentiment classification: positive, negative, or neutral (Loureiro et al., 2022), referred to in this study as RoBERTa-cardiffnlp. For each survey response, the model generated calibrated probabilities across the three sentiment classes, and the primary label was assigned to the class with the highest probabilities.

Evaluation Framework

The comparison between manual codes and the codes generated using the four sentiment analyses was conducted using a single-label evaluation framework. For each tool-category pair, the following metrics were computed: accuracy, precision, recall, F₁ score, and Cohen's kappa (Cohen, 1960). Accuracy is defined as the proportion of responses for which the tool's predicted label matches the manual code, computed across all categories. We report accuracy as the summary statistics for two reasons: first, it remains the standard metric in comparative sentiment analysis research, including the foundational benchmarks for the lexicon-based tools used here (Hutto & Gilbert, 2014), ensuring

comparability with prior studies. Second, among the metrics in this framework, accuracy is the most directly interpretable for non-specialist audiences. Accuracy, however, is sensitive to the marginal distribution of classes: when one category is substantially more prevalent than the others, high accuracy may reflect prevalence rather than meaningful agreement (Powers, 2020). Reporting accuracy alongside macro F_1 and Cohen's κ , therefore, provides complementary perspectives on tool-human agreement and aligns with current methodological recommendations for evaluating classification performance under class imbalance (Opitz & Burst, 2019).

Precision and recall are the two foundational classification metrics from the information retrieval tradition (Manning et al., 2008; van Rijsbergen, 1979). Precision measures the proportion of the tool's predicted labels that are correctly identified (i.e., how many of the tool's positive predictions matched the manual code), and recall measures the proportion of manually coded responses that the tool correctly detected. The F_1 score is the harmonic mean of precision and recall, penalizing tools that achieve high precision at the expense of low recall (Powers, 2020; van Rijsbergen, 1979). Agreement beyond chance was assessed using Cohen's κ , interpreted according to the benchmarks proposed by Landis and Koch (1977): poor (< 0), slight (.00 – .20), fair (.21– .40), moderate (.41– .60), substantial (.61– .80), and almost perfect agreement (.81–1.00). Cohen's κ is also sensitive to class prevalence and marginal distribution. As a result, substantial class imbalance can produce relatively low κ values when observed agreement is high, a phenomenon commonly referred to as the kappa paradox (Sim & Wright, 2005).

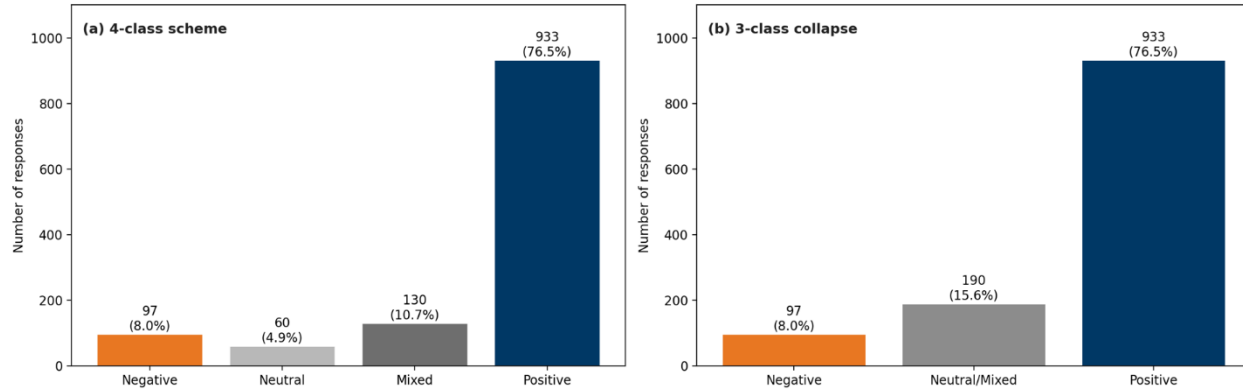
Because the manual labels in the present study are dominated by the positive class (76.5%), we reported the macro F_1 score and Cohen's kappa, since the two metrics show different aspects of the tool-human agreement under class imbalance. Macro F_1 score was computed as the unweighted arithmetic mean of pre-class F_1 scores using scikit-learn's classification report, consistent with Opitz and Burst's (2019) recommendation on using "Averaged F_1 " (arithmetic mean over harmonic means). This metric is less sensitive to skewed distribution and is particularly informative for evaluating performance in this imbalanced corpus, because each class contributes equally to the Macro F_1 score regardless of its prevalence.

Results

The distribution of sentiment analysis results, as coded by human raters, is presented in Figure 2. Because all four off-the-shelf tools generate only three-class output, the present study merges manual neutral and mixed codes into a single neutral/mixed class when reporting the results. In doing so, we keep the manual coding classification and the outputs generated by off-the-shelf tools on a common scale while preserving the theoretically meaningful distinction between polarized and non-polarized responses. As shown in Figure 2, the manual labels were strongly skewed toward positive: Under the 4-class scheme used during manual coding, 933 responses (76.5%) were positive, 130 (10.7%) were mixed, 60 (4.9%) were neutral, and 97 (8.0%) were negative. After collapsing neutral and mixed to align with the 3-class output of the four off-the-shelf sentiment analysis tools, 190 responses (15.6%) were neutral/mixed.

Figure 2

Manual Label Distribution Across the 4-Class and 3-Class Coding Schemes



Lexicon Coverage

A substantial coverage gap was identified across both lexical methods. A lexicon match was absent in 21.3% ($n = 260$) of responses when using TextBlob. In the left panel of Figure 3, the most prominent feature is the pronounced spike at score 0.00, where 329 of the 1,220 responses are concentrated. This spike represents “out-of-lexicon” survey responses containing no tokens recognized by TextBlob’s Pattern library; as a result, these responses received a default polarity score of zero and were classified as neutral/mixed regardless of their actual sentiment. Beyond the zero-score spike, the distribution is clearly right-skewed, with most non-zero scores concentrated in the positive range, reflecting the corpus’s overall positive composition (76.5% positive).

After the 259 responses with no TextBlob in-lexicon words were removed, the spike in the right panel declined sharply: only 70 responses remained in the zero bin, as shown in the right panel in Figure 3. These were cases in which TextBlob identified lexicon entries, but their polarity scores were set to zero. The distribution also became more strongly right-skewed once the out-of-lexicon responses were excluded. Together, the two panels show that roughly four out of five responses TextBlob classified as neutral/mixed in the full sample received that label because the lexicon did not recognize these words, not because of the substantive content conveyed in these responses.

Sentiment analysis results generated by VADER revealed a similar pattern even more distinctly. As shown in the left panel of Figure 4, 277 of the 1,220 responses were located on the zero bar. As a result, the out-of-lexicon rate of 21.6% was almost identical to TextBlob’s. The non-zero distribution was more strongly right-skewed than TextBlob’s, reflecting VADER’s heuristics that push responses farther from zero toward the positive direction. In the right panel, after dropping the 264 out-of-lexicon responses, the zero-bin collapsed to just 13 responses. This decline was substantially steeper than that observed for TextBlob (329 to 70 versus 277 to 13) and had a direct effect on classification performance: as shown in Tables 5 and 6, VADER’s neutral/mixed F_1 score decreased from .26 in the full sample to .14 in the hits-only subset. Once the out-of-lexicon responses were removed, VADER

essentially became a binary positive/negative classifier; it almost never assigned a non-zero compound score within the neutral/mixed band.

Figure 3

Lexicon Coverage in TextBlob

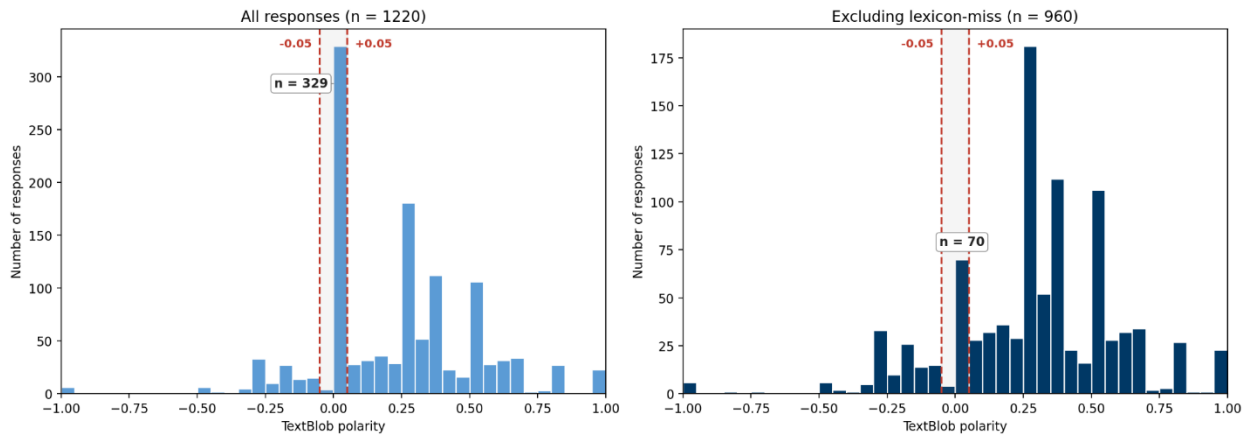
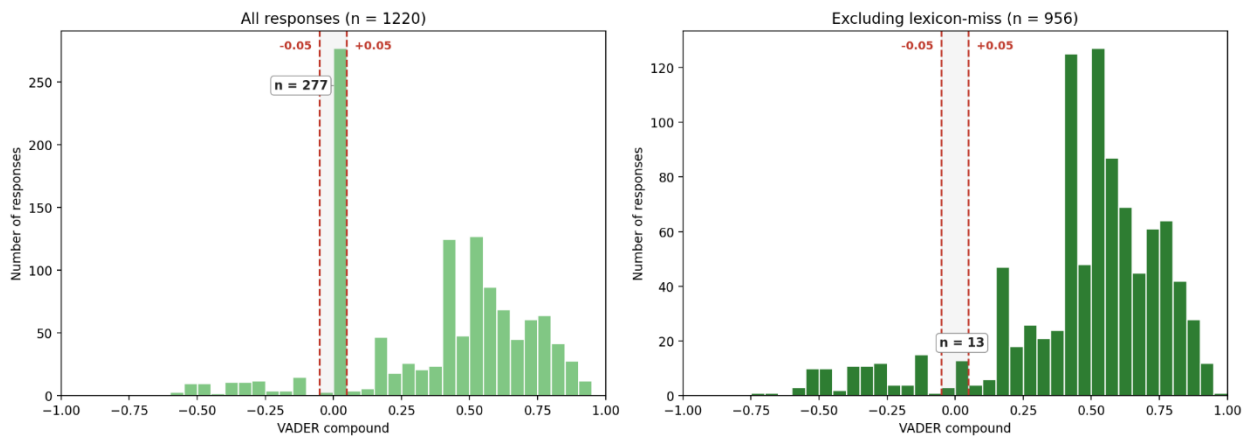


Figure 4

Lexicon Coverage in VADER



In aggregate, 29.5% ($n = 360$) of the total responses were missed by either TextBlob or VADER, and 13.4% ($n = 164$) remained undetected by both lexicon tools. To illustrate the substantive nature of the coverage gap, Figure 5 displays a word cloud of the most frequent responses not covered by both lexicon-based tools. These words and phrases demonstrate that out-of-lexicon responses were not inherently sentiment-neutral (e.g., “enriching”). Instead, they often contained institutionally and contextually meaningful language that conveyed clear evaluative sentiment, even though they were not covered by the tools’ predefined dictionaries.

lexicon token recognized by both TextBlob and VADER ($n = 860$). By excluding out-of-lexicon responses, this condition enabled a comparison of tool performance with lexicon coverage. To ensure comparability across methods, the two transformer-based models (BERT-nlptown and RoBERTa-cardiffnlp) were evaluated under the same full-sample and restricted-sample conditions. Evaluating all four tools under identical conditions enabled a more meaningful comparison of their performance and helped isolate the extent to which differences in classification accuracy were attributable to lexicon coverage rather than to sentiment classification itself.

Tool Comparison: Full Sample

Table 3 presents the performance metrics of the four sentiment analysis tools: accuracy, Macro F_1 , and Cohen's κ evaluated against the full labeled sample ($N = 1,220$). BERT-nlptown demonstrated the strongest performance across all metrics. Specifically, in terms of accuracy, BERT-nlptown ranked the highest (.79), followed by RoBERTa-cardiffnlp (.70), VADER (.69), and TextBlob (.63). Regarding Macro F_1 , BERT-nlptown also ranked the highest (.60), followed by RoBERTa-cardiffnlp (.55), VADER (.52), and TextBlob (.45). Regarding inter-rater reliability, κ values for TextBlob (.21) and VADER (.27) indicated "fair" agreement, while RoBERTa-cardiffnlp (.34) reached the upper bound of the "fair" range and BERT-nlptown (.42) approached "moderate" agreement. Notably, only BERT-nlptown achieved a "moderate" level of agreement ($\kappa > .40$) on the unweighted scale. These results suggest that even the best-performing off-the-shelf tools exhibit considerable disagreement with human coders when chance-corrected agreement is taken into account.

Table 3

Tool Performance on the Full Labeled Sample ($N = 1,220$)

Tool	Accuracy	Macro F_1	Cohen's κ
TextBlob	.63	.45	.21
VADER	.69	.52	.27
BERT-nlptown	.79	.60	.42
RoBERTa-cardiffnlp	.70	.55	.34

Note. Macro F_1 is the unweighted average of per-class F_1 scores. Cohen's κ is interpreted using the Landis and Koch (1977) bands: .21–.40 fair; .41–.60 moderate; .61–.80 substantial.

Tool Comparison: Lexicon-Hits-Only Subset

Table 4 repeats the comparison on the 860-response lexicon-hits-only subset, where the lexicon tools had at least some in-lexicon signal to analyze. As expected, the lexicon tools improved markedly: TextBlob accuracy rose from .63 to .78 and VADER from .69 to .83. The transformer-based models also improved, but more modestly. The ranking by Macro F_1 was the same as in the full sample: BERT-nlptown ranked the highest (.61), followed by RoBERTa-cardiffnlp (.58), VADER (.54), and TextBlob (.51). Kappa rankings also showed the same order across the two subsets, with BERT-nlptown ranking the highest (.42), followed by RoBERTa-cardiffnlp (.38), VADER (.37), and TextBlob (.31).

Table 4

Tool Performance on the Lexicon-Hits-Only Subset ($n = 860$)

Tool	Accuracy	Macro F_1	Cohen's κ
TextBlob	.78	.51	.31
VADER	.83	.54	.37
BERT-nlptown	.83	.61	.42
RoBERTa-cardiffnlp	.79	.58	.38

Note. Macro F_1 is the unweighted average of per-class F_1 scores. Cohen's κ is interpreted using the Landis and Koch (1977) bands: .21–.40 fair; .41–.60 moderate; .61–.80 substantial.

Per-Class Performance: Full Sample

Table 5 reports per-class F_1 on the full sample. Two patterns were evident. First, every tool achieved its highest F_1 score on the positive class due to both class imbalance and the survey context, in which positive sentiment is expressed in conventional, lexicon-friendly vocabulary (“exciting,” “fun”). Second, every tool struggled most in the neutral/mixed class, with F_1 ranging from .26 (VADER) to .37 (BERT-nlptown). The neutral/mixed class consisted of two types of responses: factual information and responses containing both positive and negative content. No off-the-shelf sentiment tools were designed to distinguish between them, a limitation that will be discussed in detail in the Limitations section.

Table 5

Per-Class F_1 on the Full Labeled Sample ($N = 1,220$)

Tool	Positive F_1	Neutral/Mixed F_1	Negative F_1
TextBlob	.78	.28	.30
VADER	.82	.26	.49
BERT-nlptown	.89	.37	.54
RoBERTa-cardiffnlp	.83	.36	.48

Note. F_1 is the harmonic mean of precision and recall, reported per class. Higher values indicate better balance between precision and recall in that class.

Per-Class Performance: Lexicon-Hits-Only Subset

Table 6 presents the per-class F_1 scores for the lexicon-hits-only subset ($n = 860$). A primary finding is that VADER's performance in the neutral/mixed class declined after excluding “out-of-lexicon” survey responses. Specifically, VADER's neutral/mixed F_1 score decreased from .26 in the full sample to .14 in this subset. VADER correctly identified only 13 of the 168 ground-truth neutral/mixed responses.

TextBlob’s neutral/mixed F_1 remained essentially around the same range (.28 to .29), supporting the conclusion that lexicon-based tools relied on “out-of-lexicon” survey responses to populate the neutral/mixed category in the full sample. Conversely, the transformer-based models demonstrated mixed results in this category; BERT-nlptown’s F_1 for neutral/mixed sentiment rose from .37 to .40, while RoBERTa-cardiffnlp showed a marginal decrease (.36 to .33), likely reflecting the loss of true positives on lexicon-miss rows it had previously classified correctly. Notably, three tools showed improved F_1 scores for the negative class: TextBlob rose from .30 to .36, VADER from .49 to .57, and RoBERTa-cardiffnlp from .48 to .52.

Table 6

Per-Class F_1 on the Lexicon-Hits-Only Subset ($n = 860$)

Tool	Positive F_1	Neutral/Mixed F_1	Negative F_1
TextBlob	.88	.29	.36
VADER	.92	.14	.57
BERT-nlptown	.91	.40	.52
RoBERTa-cardiffnlp	.89	.33	.52

Note: F_1 values are computed only on responses where both TextBlob and VADER had at least one in-lexicon token. Compare against Table 5 to see the effect of removing out-of-lexicon responses.

Summary of the Four-Tool Comparison

Across both evaluation conditions and all three metrics (accuracy, macro F_1 and Cohen's kappa), BERT-nlptown was consistently the top-performing off-the-shelf tool ahead of RoBERTa-cardiffnlp, VADER and TextBlob. However, even BERT-nlptown achieved only moderate agreement with human coders. This indicates that off-the-shelf sentiment analysis tools, regardless of their architecture, struggle to capture the nuances of this specific corpus. Furthermore, the deep-learning models failed to identify approximately half of the negative responses and frequently misclassified neutral or mixed sentiment as positive. A significant limitation was identified in the lexicon-based tools. Due to limited vocabulary coverage, VADER and TextBlob exhibited an “out-of-lexicon” misclassification rate that would cause a standard pipeline to silently misclassify roughly 20% of all data by default. Ultimately, these findings suggest that off-the-shelf sentiment analysis tools, whether advanced transformers or traditional lexicon-based approaches, introduce measurement error and systematic bias when applied directly to this dataset.

Discussion

This study evaluated four off-the-shelf sentiment analysis tools, two lexicon-based (TextBlob, VADER) and two transformer-based (BERT-nlptown, RoBERTa-cardiffnlp), applied to a corpus of 1,220 free-text undergraduate survey responses. The findings indicate that lexicon coverage represents the most critical limitation for the lexicon-based tools. In the current corpus, out-of-lexicon rates reached 21.3% for TextBlob and 21.6% for VADER. Notably, 13.4% of responses were completely not covered by either

tool. Under default classification rules, a polarity score of zero is automatically assigned to the neutral/mixed category, resulting in systematic “out-of-lexicon” misclassification regardless of the responses’ actual sentiment.

The performance disparity between the full-sample and “lexicon-hits-only” conditions directly quantified this limitation. When evaluating only responses containing lexicon matches, accuracy improved for both TextBlob and VADER. In contrast, the transformer models showed marginal gains. VADER’s performance on neutral/mixed cases in the full sample ($F_1 = .26$) was revealed as an artifact of the “out-of-lexicon” issue. In the hits-only subset, this metric dropped to .14, indicating that accuracy was driven primarily by “out-of-lexicon” survey responses rather than genuine classification accuracy. These findings align with prior research documenting lexicon-coverage constraints (Hamilton et al., 2016; Taboada et al., 2011; van Atteveldt et al., 2021) and recent work identifying similar issues in clinical and educational corpora (Cero et al., 2024; Lu et al., 2023). For example, van Atteveldt et al. (2021) found that dictionary-based methods often fail to achieve acceptable validity metrics compared to human coders due to limited vocabulary coverage.

While transformer-based models avoid coverage limitations through sub-word tokenization, they introduce training-domain bias. Although neither transformer tokenizer produced unknown tokens, their training domain differed substantially from the higher education survey data. RoBERTa-cardiffnlp was fine-tuned on approximately 124 million tweets (Loureiro et al., 2022), while the BERT-nlptown model was trained on multilingual product reviews. Observed error patterns reflected these biases: the BERT-nlptown model systematically over-predicted positive sentiment (labeling 66% of truly neutral/mixed and 31% of truly negative responses as positive), likely reflecting the positive skew in online review distributions. Conversely, RoBERTa-cardiffnlp over-predicted neutral/mixed sentiment: Assigning that label to 23% of positive and 59% of negative responses. This pattern is consistent with the conversational style and ambiguity prevalent in social media discourse. Because these errors were systematic rather than random, they may be mitigated through few-shot prompting and domain-specific fine-tuning.

Despite outperforming alternative tools, the strongest model, BERT-nlptown remained unsuitable for unsupervised deployment. Across both evaluation conditions and all three metrics, BERT-nlptown demonstrated the highest overall performance; however, it identified only approximately half of all negative responses (recall = .47 in the full sample and recall = .45 in the lexicon-hit only sample). Detailed information on precision and recall is included in the Appendix. For institutional assessment, where detecting student dissatisfaction is operationally important, such error rates are problematic. These findings indicate that off-the-shelf sentiment analysis tools exhibit limited validity for high-stakes institutional assessment applications without further adaptation.

Implications for Assessment Professionals

These findings offer three practical considerations for institutional researchers, assessment professionals, and faculty seeking to apply off-the-shelf sentiment analysis tools to analyze student feedback data. First, lexicon-based tools should not be used on survey data without first assessing

lexicon coverage, as approximately one in five responses in this dataset would have been silently misclassified as neutral/mixed. Second, transformer models trained on social media or consumer review data do not generalize seamlessly to higher education survey contexts; their systematic biases reflect the linguistic patterns of their original training domains. Third, no single evaluation metric adequately captures model performance on imbalanced educational datasets. Accuracy disproportionately rewards majority-class prediction, macro F_1 emphasizes balanced performance across classes, and Cohen's kappa adjusts for chance agreement. Collectively, these metrics provide a more comprehensive and transparent assessment of classifier performance on imbalanced educational data.

Limitations

Several limitations should be acknowledged. First, the decision to collapse neutral and mixed responses into a single category was largely driven by the comparative design of the sentiment tools used in this study. The off-the-shelf tools generate only three sentiment labels (positive, negative, neutral) by default and do not provide a "mixed" sentiment classification. Although the three-class framework remains the dominant convention in applied sentiment analysis (Liu, 2012), prior scholarship suggests that the boundary between neutral and mixed sentiment is often ambiguous. Pang and Lee (2008) observed that an overall neutral judgment may itself be "a mixture of positive and negative language" (p. 17). Similarly, Liu (2012) noted that a single text can simultaneously express "both positive and negative sentiments" toward different aspects of the same target (p. 41) and described prior three-class work that intentionally combined "no opinion or mixed opinion" into a single category (p. 42). Collapsing neutral and mixed responses into a single category in this study was consistent with established practice and the constraints of the tools used. However, this approach may have obscured meaningful distinctions between genuinely neutral responses and those that contain both positive and negative evaluations, potentially reducing the sensitivity of the sentiment classifier. Second, the present study was conducted at a single institution and analyzed a single survey question, limiting the generalizability of the findings to other institutional contexts or assessment domains. Third, the four off-the-shelf tools were evaluated using the literature-standard default thresholds. Threshold tuning was intentionally excluded because this study aimed to evaluate these tools under conditions that reflect typical practitioner use, rather than to optimize dataset-specific performance.

Future Directions

This study establishes an empirical performance baseline for off-the-shelf sentiment analysis tools, providing a benchmark for evaluating domain-specific optimization approaches. The most immediate extension involves few-shot prompting with frontier large language models (e.g., Claude Fable 5.1, ChatGPT 6) and/or fine-tuning transformer models on the labeled student responses dataset to assess whether these approaches improve classification performance, particularly for the underperforming negative and neutral/mixed classes. Furthermore, future work should operationalize "mixed" sentiment as a distinct category rather than collapsing it into a neutral classification. Approximately 10.7% of responses in the current corpus contained genuinely mixed sentiment, as evidenced by statements such as "it is stressful but rewarding." These responses are often the most diagnostically

informative for assessment purposes because they capture nuanced trade-offs rather than unidimensional positive or negative evaluations.

As mentioned earlier, future institutional deployment should adopt a human-in-the-loop framework. Automated classifiers are best used as decision-support systems rather than autonomous replacements for human expertise. Within this framework, high-confidence classifications are prioritized for rapid analyst review, while low-confidence or ambiguous classifications are routed for closer human review and inspection. Critically, responses that flag student distress, wellbeing, discrimination, or safety-related issues must automatically escalate to human evaluation regardless of the model's confidence. Future research should focus less on maximizing overall classification accuracy alone and more on developing empirically validated confidence thresholds, escalation protocols, and review workflows that align with professional standards for institutional assessment. This approach will support more efficient analysis of survey data by reducing the volume of responses that require intensive manual review, while still incorporating targeted human validation and oversight to ensure classification accuracy and model reliability.

An additional opportunity for institutional assessment is the development of a longitudinal, institution-specific language corpus. Rather than treating each year's open-ended responses as an entirely new classification problem, previously human-validated responses could serve as training or validation data for subsequent analyses. For transformer-based approaches, accumulated labeled responses could support domain-specific fine-tuning, enabling models to increasingly recognize how students at a particular institution describe their experiences. This approach should therefore be viewed as iterative rather than as a one-time training exercise. Language that appears stable across cohorts may provide useful institutional knowledge, while emerging expressions, slang, or context-dependent terminology would require periodic review. A human-in-the-loop process could identify newly occurring or low-confidence terms each year, adjudicate their meaning, and incorporate validated cases into an expanding institutional corpus. Over time, this process could test whether locally adapted models improve classification performance while remaining responsive to changes in student language.

Finally, multi-institutional replication across diverse survey types (e.g., curricular vs. co-curricular assessment) is necessary to isolate corpus-specific errors from model-specific biases. The systematic patterns documented in this study: BERT-nlptown's tendency to over-predict positive sentiment, RoBERTa-cardiffnlp's over-classification of neutral/mixed responses, and the approximately one-in-five misclassification rate for lexicon tools serve as testable hypotheses regarding the off-the-shelf models' behavior on comparable educational corpora. Cross-institutional replication and validation will ultimately clarify the extent to which these errors reflect generalizable algorithmic limitations versus corpus-specific characteristics.

Conclusion

Evaluation of four off-the-shelf sentiment analysis tools on undergraduate survey data indicates that none achieved the level of agreement with the human reference standard needed to support unsupervised deployment. Lexicon-based tools are prone to misclassification due to coverage gaps in the lexicon, while transformer-based models exhibit domain biases. These findings do not imply that

automated sentiment analysis is inherently unsuited for educational research or institutional assessment; rather, they demonstrate that off-the-shelf tools are insufficient for high-stakes assessment contexts. Bridging this gap will require combining human-in-the-loop oversight with domain-specific optimization, including fine-tuning and few-shot prompting with frontier large language models such as Claude Fable 5.1 and ChatGPT 6.

References

- Alhija, F. N. A., & Fresko, B. (2009). Student evaluation of instruction: What can be learned from students' written comments? *Studies in Educational Evaluation*, 35(1), 37–44. <https://doi.org/10.1016/j.stueduc.2009.01.002>
- Astin, A. W. (1984). Student involvement: A developmental theory for higher education. *Journal of College Student Personnel*, 25(4), 297–308.
- Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1644–1650). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.findings-emnlp.148>
- Boyd, N. M., Liu, X., & Horissian, K. (2022). Impact of community experiences on student retention perceptions and satisfaction in higher education. *Journal of College Student Retention: Research, Theory & Practice*, 24(2), 337–365. <https://doi.org/10.1177/1521025120916433>
- Cero, I., Luo, J., & Falligant, J. M. (2024). Lexicon-based sentiment analysis in behavioral research. *Perspectives on Behavior Science*, 47(1), 283–310. <https://doi.org/10.1007/s40614-023-00394-x>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Dalipi, F., Zdravkova, K., & Ahlgren, F. (2021). Sentiment analysis of students' feedback in MOOCs: A systematic literature review. *Frontiers in Artificial Intelligence*, 4, Article 728708. <https://doi.org/10.3389/frai.2021.728708>
- Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web* (pp. 519–528). <https://doi.org/10.1145/775152.775226>
- De Smedt, T., & Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research*, 13, 2063–2067.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. <https://doi.org/10.1037/h0031619>
- Geer, J. G. (1991). Do open-ended questions measure “salient” issues? *Public Opinion Quarterly*, 55(3), 360–370. <https://doi.org/10.1086/269268>

- Grimalt-Álvaro, C., & Usart, M. (2024). Sentiment analysis for formative assessment in higher education: A systematic literature review. *Journal of Computing in Higher Education*, 36, 647–682. <https://doi.org/10.1007/s12528-023-09370-5>
- Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 595–605). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D16-1057>
- Hausmann, L. R. M., Schofield, J. W., & Woods, R. L. (2007). Sense of belonging as a predictor of intentions to persist among African American and White first-year college students. *Research in Higher Education*, 48(7), 803–839. <https://doi.org/10.1007/s11162-007-9052-9>
- Hill, C. E., Thompson, B. J., & Williams, E. N. (1997). A guide to conducting consensual qualitative research. *The Counseling Psychologist*, 25(4), 517–572. <https://doi.org/10.1177/0011000097254001>
- Hill, C. E., Knox, S., Thompson, B. J., Williams, E. N., Hess, S. A., & Ladany, N. (2005). Consensual qualitative research: An update. *Journal of Counseling Psychology*, 52(2), 196–205. <https://doi.org/10.1037/0022-0167.52.2.196>
- Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216–225. <https://doi.org/10.1609/icwsm.v8i1.14550>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K., & Wani, M. A. (2021). Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study. *Applied Sciences*, 11(9), Article 3986. <https://doi.org/10.3390/app11093986>
- Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *The Journal of Higher Education*, 79(5), 540–563. <https://doi.org/10.1080/00221546.2008.11772116>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Liu, A., & Sun, M. (2025). From voices to validity: Leveraging large language models (LLMs) for textual analysis of policy stakeholder interviews. *AERA Open*, 11, Article 23328584251374595. <https://doi.org/10.1177/23328584251374595>
- Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool. <https://doi.org/10.2200/S00416ED1V01Y201204HLT016>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. <https://arxiv.org/abs/1907.11692>
- Loria, S. (2025). TextBlob: Simplified text processing (Version 0.19.0) [Computer software]. <https://textblob.readthedocs.io/>
- Loureiro, D., Barbieri, F., Neves, L., Espinosa Anke, L., & Camacho-Collados, J. (2022). TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the*

- Association for Computational Linguistics: System Demonstrations* (pp. 251–260). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-demo.25>
- Lu, K. J. Q., Meaney, C., Guo, E., & Leung, F.-H. (2023). Evaluating the applicability of existing lexicon-based sentiment analysis techniques on family medicine resident feedback field notes: Retrospective cohort study. *JMIR Medical Education*, 9, Article e41953. <https://doi.org/10.2196/41953>
- Ma, I., Paget, M., Desy, J., Harvey, A., Bendiak, G., Naugler, C., & McLaughlin, K. (2025). A validity evaluation of lexicon-based sentiment analysis of medical students' clinical performance from in-training evaluation reports. *Medical Education*, 60(6), 685–692. <https://doi.org/10.1111/medu.70101>
- Soboroff, I., Macdonald, C., & Ounis, I. (2008). *Overview of the TREC 2007 Blog Track, Special Publication (NIST SP)*, National Institute of Standards and Technology. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=152150
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113. <https://doi.org/10.1016/j.asej.2014.04.011>
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- NLP Town. (2020). bert-base-multilingual-uncased-sentiment [Machine learning model]. *Hugging Face*. <https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment>
- O'Cathain, A., & Thomas, K. J. (2004). "Any other comments?" Open questions on questionnaires – a bane or a bonus to research? *BMC Medical Research Methodology*, 4, 25. <https://doi.org/10.1186/1471-2288-4-25>
- Opitz, J., & Burst, S. (2019). Macro F1 and macro F1. *arXiv*. <https://doi.org/10.48550/arXiv.1911.03347>
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135. <https://doi.org/10.1561/1500000011>
- Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv*. <https://doi.org/10.48550/arXiv.2010.16061>
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Academic Press.
- Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3), 257–268. <https://doi.org/10.1093/ptj/85.3.257>
- Slotnick, R. C., & Boeing, J. Z. (2025). Enhancing qualitative research in higher education assessment through generative AI integration: A path toward meaningful insights and a cautionary tale. *New Directions for Teaching and Learning*, 2025(182), 97–112. <https://doi.org/10.1002/tl.20631>
- Stupans, I., McGuren, T., & Babey, A. M. (2016). Student evaluation of teaching: A study exploring student rating instrument free-form text comments. *Innovative Higher Education*, 41(1), 33–42. <https://doi.org/10.1007/s10755-015-9328-5>
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267–307. https://doi.org/10.1162/COLI_a_00049

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

<https://doi.org/10.1080/19312458.2020.1869198>

van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). Butterworths.

About the Authors

Wei You, Ph.D., Assessment Coordinator, Office of Institutional Research & Analytics, Bucknell University, wy005@bucknell.edu

Kevork Horissian, MBA, Associate Vice President & Chief Analytics Officer, Bucknell University, kth005@bucknell.edu

Appendix

Table A1

Per-Class Precision and Recall of the Four Sentiment Analysis Tools Against Manual Coding

Tool and sentiment class	Full sample (<i>N</i> =1,220)		Lexicon-hits-only (<i>n</i> = 860)	
	Precision	Recall	Precision	Recall
TextBlob				
Negative	.27	.34	.26	.62
Neutral/Mixed	.22	.39	.55	.20
Positive	.86	.71	.86	.90
VADER				
Negative	.53	.46	.46	.75
Neutral/Mixed	.22	.32	.76	.07
Positive	.86	.79	.86	.98
BERT-nlptown				
Negative	.63	.47	.62	.45
Neutral/Mixed	.42	.33	.50	.33
Positive	.86	.92	.87	.94
RoBERTa-cardiffnlp				
Negative	.62	.39	.52	.52
Neutral/Mixed	.27	.53	.33	.34
Positive	.91	.76	.90	.89

Note. Precision = proportion of the tool's predictions for a class that matched the manual code; Recall = proportion of manually coded responses in a class that the tool correctly identified. Manual neutral and mixed codes were combined into a single neutral/mixed class to match the three-class output of the tools.