UDC 37.013.31:811.111:[159.942
[37.013.321:004.8]:811.111:81’33
https://doi.org/10.18485/bells.2025.17.7

Igor Ivanović*
University of Montenegro
Faculty of Philology, Nikšić
https://orcid.org/0000-0002-7372-104X

OPINION MINING IN ESP CLASSROOMS: A COMPARATIVE ANALYSIS OF TRADITIONAL VS. AI-ASSISTED INSTRUCTION

Abstract
This study investigates sentiment analysis, specifically opinion mining, in English for Specific Purposes (ESP) classrooms by comparing traditional teaching methods with AI-assisted instruction. Using a within-subject repeated-measures design, the same cohort of 41 students (11 males, 30 females) first experienced traditional classroom methods during the winter semester (serving as the Control condition) and later participated in an AI-assisted classroom during the summer semester (serving as the Target condition). Employing advanced sentiment analysis techniques, we analysed 82 essays written by these students, focusing on their attitudes towards the provided learning materials. Our approach involved using a specialised sentiment lexicon for accurate scoring and classification. The findings reveal a clear shift towards more positive sentiment in the AI-assisted classroom (Target condition), underscoring AI’s potential to enhance the emotional dimensions of language learning. These insights are critical for understanding student perceptions across instructional modes and highlight the potential of AI in educational contexts.

Key words: opinion mining; sentiment analysis; sentiment lexicon; AI-assisted classroom; traditional paradigm; large language models in classroom

* iggybosnia@ucg.ac.me

1. Introduction

Artificial Intelligence (AI) has shown remarkable potential to revolutionise educational practices in recent years (Ayeni et al. 2024; Chiu et al. 2023). AI can analyse individual student's learning patterns, strengths, and weaknesses to provide personalised learning paths (Ayeni et al. 2024). By understanding the learner's proficiency level and preferred learning style, AI-powered systems can tailor content, exercises, and challenges to maximise the effectiveness of language learning. This ensures that learners are continually challenged and engaged at the appropriate level, leading to faster progress and improved retention. Apart from that, AI can automate language assessment, enabling faster and more accurate evaluation of learners' language proficiency. This includes automated grading of writing assignments, speaking assessments, and comprehension tests. Also, AI-powered chatbots can function as language practice partners, providing learners with conversational practice and instant feedback. These interactive language practice sessions can be available 24/7, allowing learners to practice at their convenience. However, while AI presents significant opportunities, its integration into educational settings also raises concerns, such as ethical considerations, data privacy, and potential biases embedded in AI algorithms (Chiu et al. 2023). Addressing these issues proactively is essential for successful and equitable implementation.

One more domain where AI can significantly impact language education is opinion mining, specifically sentiment analysis. Sentiment analysis is an interdisciplinary field that combines concepts and techniques from linguistics, psychology, computer science, and data science (Koufakou 2024). The integration of these theoretical foundations allows sentiment analysis systems to interpret and classify sentiments expressed in text data more effectively. However, given the complexity of human language and emotions, achieving highly accurate sentiment analysis remains an ongoing challenge (Shaik et al. 2022), and research in this area continues to evolve. For instance, sentiment analysis relies on linguistic theories to understand how words and phrases convey emotions and attitudes. Linguistic research helps identify sentiment-bearing words, phrases, and linguistic patterns that indicate positive, negative, or neutral sentiments. Additionally, psychological theories play a crucial role in sentiment analysis as they help to understand how humans express and interpret emotions and opinions. Emotion theories, like Plutchik's wheel of emotions or Ekman's six basic emotions, provide insights into the different dimensions of human emotions, which are essential for accurate sentiment classification. Psychological studies on sentiment expression and perception also influence the design of sentiment analysis algorithms. Furthermore, sentiment analysis often leverages semantic theories and techniques to handle word sense disambiguation. Different senses of a word can carry different sentiments, and understanding the correct sense in context is crucial for accurate sentiment analysis.

Sentiment analysis has numerous applications across various industries, including market research, social media monitoring, and student feedback analysis (Dalipi et al. 2021). It may help schools and educational organisations gain valuable insights from vast amounts of unstructured textual data, enabling them to make data-driven decisions and understand students' opinions (Weitl-Harms, Hastings & Lum 2024). Analysing students' sentiment in their written feedback or essays provides insight into their emotional engagement and attitudes towards the course content and teaching methods (Dake & Gyimah 2023). By identifying patterns of frustration, confusion, or enthusiasm, educators can adjust instructional strategies to address challenges and reinforce positive experiences, thereby improving the language learning process (Koufakou 2024).

There are several ways of how to perform sentiment analysis:

1. Rule-based approach relies on predefined rules and patterns to identify sentiment-bearing words and phrases. It can be useful for simple cases but may not capture more complex linguistic expressions.
2. Lexicon-based approach uses pre-built sentiment dictionaries that contain words or phrases annotated with their corresponding sentiment scores (positive, negative, or neutral). The overall sentiment of a piece of text is calculated based on the sentiment scores of its constituent words.
3. Machine learning-based approach involves training machine learning models on labelled datasets, where the sentiment of the text is already known. These models learn to recognise patterns and associations between words and sentiments, enabling them to predict the sentiment of unseen text.
4. Deep learning-based approaches, particularly recurrent neural networks (RNNs), and transformer-based models like BERT (Bidirectional Encoder Representations from Transformers),

> have shown remarkable performance in sentiment analysis tasks (Koufakou 2024). These models can capture complex contextual information and improve accuracy in sentiment classification.

For this paper, we opted for a document-level hybrid approach, combining machine learning models (especially deep learning, natural language processing, and AI) with rule-based and lexicon-based techniques. This hybrid approach to sentiment analysis combines the strengths of lexicon-based methods and machine learning techniques, leading to superior performance compared to using either method alone.

2. Literature Overview

In educational contexts, sentiment analysis is widely applied to student feedback and classroom interactions to gauge learners' attitudes. For example, researchers have applied sentiment analysis to discussion posts, social media comments, and course evaluations to assess student satisfaction and engagement (Ortigosa, Martín & Carro 2014; Colace et al. 2015). Recent studies confirm that the integration of sentiment analysis into educational analytics enables the identification of affective dimensions of learning, allowing educators to monitor engagement, frustration, or boredom across online and hybrid environments (Machová et al. 2023; Islam et al. 2024). Sentiment tracking has become particularly valuable in large-scale digital learning platforms such as MOOCs, where it supports early detection of disengagement and prediction of academic success (Gardner & Brooks 2018).

AI-assisted sentiment analysis has also gained attention for its potential to enhance language education. In AI-supported classrooms, sentiment analysis provides insights into how learners emotionally respond to AI feedback, adaptive learning systems, and automated language tutors (Khare et al. 2024). In English language learning contexts, particularly for ESP, it helps to detect learners' motivation and emotional states, offering data-driven support for curriculum adjustment and personalised feedback (Arnold 2021). These applications align with the affective turn in second language acquisition research, which emphasises the role of emotions in sustaining engagement and proficiency gains.

Common analysis methods include lexicon-based approaches, which use dictionaries of words with assigned sentiment scores, and supervised machine-learning models trained on labelled data (Feldman & Ungar 2012; Hutto & Gilbert 2014). While these classical methods remain foundational, current research increasingly leverages hybrid and transformer-based architectures such as BERT and RoBERTa for higher contextual accuracy in educational sentiment detection (Sivakumar & Rajalakshmi 2022; Koufakou 2024). Unlike purely technical studies, educational applications of sentiment analysis must balance computational accuracy with pedagogical interpretability, ensuring that emotion classification aligns with meaningful learning indicators (Chiu et al. 2023). Lexicon-based methods (e.g. VADER) are popular in education research because they are easily interpretable and do not require large annotated corpora (Liu 2010; Hutto & Gilbert 2014). However, these methods can misinterpret context or sarcasm, while machine-learning methods typically need extensive training data (Liu 2010; Feldman & Ungar 2012). Recent literature advocates a combined approach that integrates linguistic features with neural embedding models, thus capturing both sentiment polarity and contextual nuance. This methodological shift has been particularly influential in language-learning analytics, where emotional tone is closely tied to communicative competence and learner autonomy.

Understanding learner emotions is important, since affective states such as engagement, frustration or boredom have been linked to learning outcomes (Baker et al. 2010; D'Mello & Graesser 2012). Prior studies find that positive sentiments often correlate with motivation and success, whereas negative sentiments (e.g. confusion or anxiety) may signal learning difficulties. A growing body of work confirms these findings using deep sentiment models, which can detect subtle variations in learner mood and correlate them with vocabulary acquisition and participation rates (Li & Pan 2025). Although sentiment analysis has been used to examine online and classroom learning environments, few studies have directly compared different instructional modes. Only a limited number of comparative analyses exist exploring sentiment shifts between traditional, blended, and AI-assisted classrooms (Vistorte et al. 2024). These studies suggest that AI-mediated feedback tends to produce more positive emotional responses than traditional teacher-only instruction, yet evidence remains fragmented, especially in ESP domains. There is limited literature on how ESP learners respond emotionally to traditional versus AI-assisted teaching. This constitutes a crucial research gap, as ESP courses often involve goal-oriented professional communication tasks where emotional engagement and self-efficacy directly influence performance outcomes (Arnold 2021).

By applying sentiment analysis to student essays under both paradigms, the present study builds on this body of work and addresses this gap in the literature. Specifically, it contributes by integrating computational sentiment detection with pedagogical inquiry in ESP settings, examining whether AI assistance fosters measurable shifts in learner affect and engagement compared to traditional instruction. This dual focus – methodological and pedagogical – aims to situate sentiment analysis not merely as a diagnostic tool but as an instrument for understanding the emotional foundations of successful language learning in AI-supported environments.

3. Background

Several years ago, the author of this paper started implementing different computer-assisted technologies to enhance our teaching of both General English and English for Specific Purposes. We started with machine translation, then moved to gamification and game-based learning (Ivanović 2024), ending up with the use of AI in our classes. During this time, we noticed an interesting trend. Our research results and our students started reporting an increase in motivation, engagement, and student-initiated exchanges. This inspired us to design a new study which would probe into the reasons for such an increase. All participants were informed that their written coursework would be used anonymously for research aimed at improving teaching quality, and participation was strictly voluntary. No student faced any academic advantage or disadvantage based on their decision to participate. In line with ethical standards for educational research, informed consent was obtained at the beginning of the semester. Students were assured that their essays would be anonymised before analysis and that no identifying information would be stored or reported. Following completion of the study, a debriefing session was held to explain the research objectives and findings, allowing students to ask questions and withdraw their data if they wished. In the early planning stage, particular care was taken to avoid potential bias stemming from the dual role of teacher-researcher. The author's assistants, rather than the author, handled initial essay collection and anonymisation to ensure independence in data handling. This separation of roles helped mitigate possible perceptions of coercion or influence over students' decisions to participate.

We chose essays as assessment instruments since they are familiar to students and are usually perceived as less intrusive than direct question, thereby contributing to the genuineness of their expressed sentiments. Thus, those essays effectively served as indirect reflections of student attitudes composed of open-ended responses where our students were largely free from any undue influence. All students received clear submission guidelines, explicitly stating that essays should not contain any personal information that could identify them. Essays were submitted through an online platform, and after submitting the essays, our faculty assistants would assign a unique identifier to each essay. Our three assistants would sift through the received essays and remove any potential personal data unintentionally entered by the students. Those essays were then submitted to the author of this paper as plain text to remove or mask any metadata that might inadvertently contain identifying information from the submitted files. By combining strict anonymisation, institutional oversight, informed consent, and post-study debriefing, this research adheres to ethical principles of transparency, autonomy, and non-maleficence. These safeguards ensured that all participants' rights and dignity were fully protected while maintaining the methodological integrity of the study.

The essay tasks used for analysis were uniform across both semesters. Each student produced one 300–400-word essay responding to two open-ended prompts: (1) Describe how the course activities helped or hindered your learning this semester and (2) Discuss how technology influenced your motivation and participation in language classes. These prompts encouraged reflection on affective and cognitive aspects of learning without explicitly mentioning AI, thereby eliciting authentic expressions of sentiment. Essays were written in English and submitted electronically within one week of assignment.

4. Research Objective and Hypotheses

Our research objective is to investigate and compare the emotional experiences of ESP students in traditional classroom instruction and AI-assisted learning environments using sentiment analysis. The study aims to explore whether AI technologies in language classrooms have an impact on students' sentiments and to identify potential differences between the two instructional methods.

Our hypothesis is that there is a significant difference in the emotional experiences (sentiment scores) of ESP students between traditional classroom instruction (Control condition) and AI-assisted learning (Target condition). Clarifying the emotional impact of AI integration on students' experiences has significant practical implications. It may inform instructional design strategies, guide educators, and policymakers in decision-making about technology investments, and support efforts to foster emotionally supportive learning environments. This study also hypothesises that the AI-assisted group will display a higher average sentiment score due to increased engagement, personalised feedback, and adaptive pacing provided by AI-supported tools.

5. Study Design and Data Collection

Participants were recruited from a cohort of first-year faculty students at the University of Montenegro who had previously completed their first (winter) semester in a traditional classroom setting. At the end of the first semester, our students were asked to submit essays as part of their class assignments. The results from those essays served as a pre-intervention control. Moving on to their second (summer) semester, the same group of students were introduced to the concept of an AI-assisted classroom. To ensure methodological clarity and replicability, the AI-assisted classroom (Target condition) was based on a structured integration of several AI and digital learning tools. The core platform used was ChatGPT (OpenAI), employed under a supervised framework for generating model responses, grammar feedback, and vocabulary enhancement activities. Additionally, Quizizz AI and GrammarlyGO were used for adaptive quizzes and automated writing feedback, while IBM Watson Natural Language Understanding

(NLU)[1] was employed for sentiment tracking and analysis. Students interacted with these systems both synchronously (during instructor-led sessions) and asynchronously (for independent practice). Each 90-minute weekly class incorporated approximately 30–40 minutes of AI-supported activities, while the remainder was reserved for instructor-led discussion, peer correction, and reflection. AI tools were presented as instructional assistants rather than replacements for teacher interaction.

The teacher maintained a central facilitative role, guiding interpretation of AI-generated feedback and ensuring linguistic and ethical appropriateness in student use. Students were explicitly trained in responsible AI interaction to avoid overreliance and to promote critical engagement with machine-generated content (Chiu et al. 2023). This hybrid structure aligns with the definition of AI-supported instruction as a pedagogical model where intelligent systems provide adaptive scaffolding, continuous feedback, and affective support while human instructors retain decision-making and instructional authority (Luckin & Holmes 2016; Holmes et al. 2019). The traditional classroom (Control condition) followed a communicative, task-based approach typical of ESP courses at the Faculty of Material Science and the Faculty of Philology. Instruction relied on teacher-led explanations, printed ESP textbooks, pair work, and short oral presentations, with minimal use of digital tools beyond PowerPoint. Assessment consisted of weekly grammar and vocabulary exercises, mid-term tests, and a short essay at the end of the semester. Each 90-minute session combined a 30-minute lecture component, a 45-minute practice block, and a brief wrap-up discussion. This structure ensured consistent delivery and alignment with institutional standards, confirming that lower post-intervention ratings reflected pedagogical limitations of traditional formats rather than inconsistent teaching quality.

An AI-assisted classroom, also known as a smart classroom, is an educational environment where artificial intelligence (AI) technologies are integrated into the teaching and learning process to enhance the overall learning experience. In our case, AI helped us in four areas:

1. AI technologies were used to analyse students' strengths, weaknesses, and learning styles to tailor educational content and activities to everyone's needs. This personalisation allows students to progress at their own pace and receive targeted support where they need it most.

[1] https://cloud.ibm.com/apidocs/natural-language-understanding

2. AI-powered virtual tutors and chatbots were used to interact with students, answering questions, providing explanations, and helping in real time. These intelligent tutors can supplement the teacher's role, providing additional support and guidance.
3. AI was used to dynamically adjust the difficulty and complexity of learning materials based on students' performance and progress.
4. AI was used to automate the grading process for assignments, quizzes, and tests, providing instant feedback to students. This saves time for educators and enables students to receive prompt feedback on their performance.

Importantly, the AI integration focused on enhancing specific linguistic competencies – particularly writing and ESP vocabulary acquisition. For example, students were assigned weekly short writing tasks that were automatically analysed for lexical variety and syntactic complexity through GrammarlyGO, while IBM Watson provided sentiment feedback visualisations for reflective learning. This ensured that AI use was systematic, measurable, and pedagogically relevant.

Our theoretical population consisted of 47 enrolled students. Out of those, 41 participants were willing and/or able to submit their essays, thus only their results were included in this study. At the end of the second semester, the students were asked to submit another batch of essays with the same essay prompts. To maintain comparability, both Control and Target groups completed identical writing tasks under similar time constraints. Sentiments expressed in essays were considered the dependent variable, while the independent variable was the type of classroom (traditional vs. AI-assisted). The students were given a defined window of time (one week) to complete and submit their essays through an online platform. To maximise participation and reduce the likelihood of missing data, automated reminder emails were sent to the participants as the submission deadlines approached. The reminders were gentle yet persuasive, encouraging students to submit their essays promptly. No personal data were collected during this process, and the students were specifically instructed not to include any personally identifiable data. Pedagogical implementation and data collection were strictly separated. The classroom instruction and essay submission occurred as part of regular coursework, while data extraction and analysis took place only after anonymisation and consent confirmation. This procedural distinction ensured that research data were collected without interfering with pedagogical processes or student assessment. This comprehensive dataset we received from our students enabled a detailed analysis of the sentiments expressed by students in traditional and AI-assisted classrooms, shedding light on the emotional impact of AI integration in education.

6. Sentiment Analysis Tool and Sentiment Lexicon

To perform sentiment analysis, we needed a tool that would analyse our essay corpus for us. We opted for IBM Watson NLU, which is an integral component of IBM's suite of AI services designed to derive insights from text. The core features of this tool encompass a wide range of text analysis capabilities. It can determine the overall sentiment of a text, distinguishing between positive, negative, and neutral tones, and even go deeper to discern specific emotions such as joy, anger, sadness, fear, and disgust. In this study, IBM Watson NLU was selected because of its validated multilingual sentiment model and prior use in educational and psychological text analysis. The system's sentiment model relies on a proprietary lexicon and deep-learning classifier trained on a large corpus of human-annotated data. While the full lexicon is not publicly accessible, Table 1 presents a representative illustrative subset of sentiment-bearing words derived from Watson's English sentiment model to demonstrate the principle of valence scoring used in this study. These words were chosen to mirror common affective vocabulary observed in the student essays.

In the context of a sentiment lexicon, valence refers to the emotional orientation or polarity of a word or phrase. It represents the degree of positive or negative sentiment associated with a particular term. A word with a positive valence is associated with positive emotions, while a word with a negative valence is associated with negative emotions. Valence scores are often assigned to words in a sentiment lexicon based on human judgments or crowd-sourced annotations. These scores typically fall on a numerical scale, which assigns sentiment scores on a scale of -1 to +1, where -1 represents strong negativity, +1 indicates strong positivity, and 0 denotes neutrality. This scoring principle follows established approaches in computational linguistics and affective computing (Mohammad 2021).

Table 1. Lexicon sample words with their valences
(illustrative subset based on Watson's sentiment model)

1. „Joyful“ -> +0.9	11. „Bored“ -> -0.5
2. „Disappointed“ -> -0.7	12. „Grateful“ -> +0.8
3. „Confident“ -> +0.8	13. „Tense“ -> -0.6
4. „Frustrated“ -> -0.8	14. „Motivated“ -> +0.9
5. „Excited“ -> +0.7	15. „Uninterested“ -> -0.3
6. „Content“ -> +0.6	16. „Delighted“ -> +0.9
7. „Annoyed“ -> -0.6	17. „Disheartened“ -> -0.7
8. „Hopeful“ -> +0.8	18. „Curious“ -> +0.7
9. „Indifferent“ -> 0.0	19. „Apathetic“ -> -0.4
10. „Eager“ -> +0.7	20. „Proud“ -> +0.8

Before analysis, all essays were pre-processed using the following pipeline:

1. Conversion to lowercase, removal of punctuation and stop words (e.g., the, and, to), and normalisation of whitespace.
2. Segmentation into sentences and tokens using the Natural Language Toolkit (NLTK).
3. Reduction of inflected forms to base lemmas to ensure consistent sentiment scoring (e.g., studying → study).
4. Each essay was uploaded via the NLU Python SDK, and the output JSON included overall document sentiment (ranging from -1 to +1), sentence-level sentiment, and emotion intensity across five primary affective dimensions (joy, anger, sadness, fear, disgust).
5. Document-level sentiment scores were derived by averaging sentence-level scores weighted by sentence length. These composite sentiment values were then used for statistical comparison between the Control and Target groups.

All raw outputs were exported to CSV format, and data processing was conducted using Python (Pandas, NumPy) to compute descriptive and inferential statistics. Sentiment polarity (positive vs. negative) was determined based on thresholds of ≥ 0.05 and ≤ -0.05 following established practice (Hutto & Gilbert 2014). Scores between −0.05 and 0.05 were treated as neutral. The resulting metrics included mean sentiment, standard deviation, and distributional spread for each group. Inter-rater validation against manually annotated samples ensured reliability (Cohen's $\kappa > 0.80$). Although both groups responded to the same essay prompts across semesters, this repetition was controlled for in interpretation. The identical prompts were intended to standardise linguistic and thematic content for comparability; however, we acknowledge that familiarity with the task could independently influence sentiment expression. Therefore, causal inferences regarding the impact of AI tools are presented with caution. Sentiment scores reflect emotional tone in writing, not direct measures of satisfaction or attitude (Mohammad 2021). Consequently, findings are discussed as indicators of relative affective trends rather than definitive psychological states. The analytic framework applied in this study reflects best practices in text-based sentiment analysis for educational data, combining rigorous preprocessing, validated computational tools, and ethical interpretive restraint.

7. Data analysis

This section presents the findings of the sentiment analysis conducted on student essays from both the traditional and AI-assisted ESP classrooms. Sentiment scores were computed for each student, providing a quantitative measure of their attitudes towards the learning materials and instructional methods. Descriptive statistics reveal a marked difference in student attitudes between the two instructional paradigms. In the traditional classroom, the mean sentiment score was 0.28, with a median of 0.38 and a mode of 0.37. In contrast, the AI-assisted classroom achieved a higher mean sentiment score of 0.63, a median of 0.76, and a mode of 0.95. These statistics are summarised in Table 2 below.

Table 2. Summary of Sentiment Score Statistics
by Classroom Type

Statistic	Traditional Classroom	AI-Assisted Classroom
Mean	0.28	0.63
Median	0.38	0.76
Mode	0.37	0.95
Standard Deviation	0.38	0.33
Minimum	-0.69	-0.34
Maximum	0.67	0.96
Number Positive	31	35
Number Negative	10	6

Visual representations of the sentiment scores further illustrate the contrast between classroom types. Scatter plots were chosen specifically to visually capture the dispersion of sentiment scores and identify potential outliers or clustering patterns within each instructional setting. Figure 1 presents scatter plots of individual student sentiment scores for both instructional settings. In the traditional classroom (top panel), sentiment scores are widely dispersed, including a substantial number of negative outliers. Conversely, the AI-assisted classroom (bottom panel) demonstrates tighter clustering of positive sentiment scores, with fewer negative results.

To better capture the central tendency and spread, Figure 2 compares the mean, median, and mode sentiment scores across both groups. The consistently higher values for the AI-assisted classroom underscore a shift towards more positive and uniform student attitudes.

Figure 1. Scatter plots of sentiment scores
in traditional and AI-assisted classrooms.

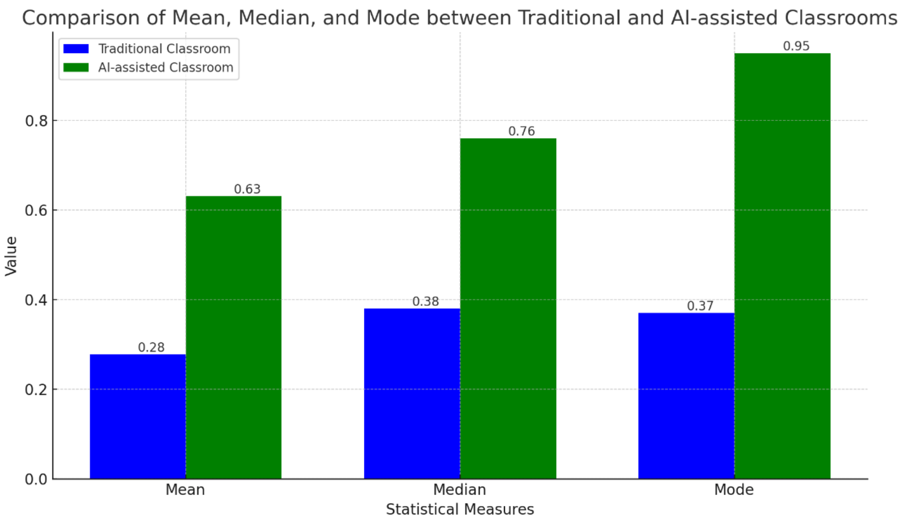

Figure 2. Bar chart comparing mean, median,
and mode between classrooms.

To assess the statistical significance of the observed differences in sentiment scores, an independent t-test was conducted. Prior to conducting the t-test, assumptions such as normality and homogeneity of variance were checked and satisfied, ensuring the validity of using this statistical method. The results indicated that the difference in mean sentiment scores between the AI-assisted classroom (mean = 0.63) and the traditional classroom (mean = 0.28) was statistically significant ($p < 0.01$). The magnitude of this effect, as measured by Cohen's d, was approximately 0.97, indicating a large effect size. This finding suggests that the introduction of AI tools into classroom instruction had a substantial and meaningful impact on student sentiment.

Table 3. Validation Metrics for Sentiment Classification

Metric	Control condition	Target condition
Classification Accuracy	77.5%	85.0%
Precision (Positive)	83.8%	87.2%
Recall (Positive)	91.2%	94.4%
F1-Score (Positive)	87.3%	90.7%
Cohen's Kappa	0.81	0.87

Beyond descriptive and inferential statistics, additional validation metrics were employed to evaluate the reliability and consistency of our sentiment analysis. Validation metrics presented in Table 3 were derived by comparing the automated sentiment classifications against manually annotated subsets of essays by trained raters. This manual coding served as the gold-standard benchmark for assessing the accuracy and reliability of our sentiment analysis tool. The overall classification accuracy, defined as the proportion of essays for which the sentiment label assigned by the analysis tool matched the human annotation, was 77.5% for the traditional classroom and 85.0% for the AI-assisted classroom. Further, the AI-assisted classroom outperformed the traditional setting in key measures of positive sentiment classification, achieving higher precision (87.2%), recall (94.4%), and F1-score (90.7%), compared to precision of 83.8%, recall of 91.2%, and F1-score of 87.3% in the traditional classroom. Inter-rater agreement, as measured by Cohen's Kappa, was high in both groups, indicating that the automated sentiment analysis produced labels that were consistent with manual human coding and thus confirmed the reliability of our analytical approach.

8. Supplemental Analysis

Having completed the sentiment analysis, we wanted to find the reasons behind such sentiment score distribution. When trying to understand the reasons behind the sentiment scores, both close and open-ended questions can provide deeper insights. Accordingly, two types of follow-up data were collected: (1) a post-intervention questionnaire containing close-ended Likert-scale items, and (2) a follow-up semi-structured interview comprising open-ended questions. These two instruments were administered sequentially to the same cohort of students.

Thus, we devised a post-intervention questionnaire that helped us probe into the reasons for the observed sentiment distribution. It consisted of nine Likert-scale items rated on a five-point scale (1 = strongly disagree to 5 = strongly agree) and was administered online to all 41 participants. Because the same group of students experienced both instructional modalities, this constituted a within-subject design, where each participant evaluated their experience across both the traditional (Control) and AI-assisted (Target) phases. To minimise recall bias, questionnaire items explicitly invited direct comparison between the two modes.

Table 4. Close-ended questions combined with the Likert scale (1-5)

Question Statement	**Traditional Classroom Average Rating**	**AI-assisted Classroom Average Rating**
The lessons were engaging.	3.2	4.5
I felt the teaching methods were effective.	3.0	4.6
Personalised learning opportunities were provided.	2.5	4.8

Question Statement	Traditional Classroom Average Rating	AI-assisted Classroom Average Rating
The classroom promoted interactive participation.	3.4	4.2
The pace of the lessons was appropriate for me.	3.1	4.0
The teaching tools used enhanced my understanding.	2.9	4.7
I faced technical difficulties during the lessons.	N/A	2.3
I felt comfortable asking questions and clarifying doubts.	3.5	4.3
I would prefer this method for my future courses.	2.7	4.9

Analysis of the questionnaire responses reveals differences in how students perceived the two instructional settings. Across nearly all statements, the AI-assisted (Target) condition received consistently higher average ratings than the traditional (Control) condition, often by a substantial margin. For instance, when asked whether the lessons were engaging, students in the traditional classroom reported a moderate average rating of 3.2, whereas those in the AI-assisted condition rated engagement much higher at 4.5. Similarly, perceptions of teaching effectiveness were significantly more positive in the AI-assisted classroom, with an average rating of 4.6 compared to 3.0 in the traditional condition. It is important to note that these comparisons do not reflect two separate student groups, but two instructional conditions experienced by the same participants. Therefore, statistical interpretation should be treated as indicative of relative preference and perceived improvement rather than independent group differences. The within-subject design allows for meaningful paired comparison of learners' perceptions but does not fully isolate the causal effect of AI tools from other factors, such as familiarity with essay tasks or semester progression.

One of the most striking contrasts emerged around personalised learning. While students in the traditional classroom gave this aspect a modest average score of 2.5, participants in the AI-assisted classroom rated it at 4.8, suggesting that AI integration played a critical role in addressing individual learning needs. The opportunity for interactive participation was also perceived more positively in the AI-assisted environment (4.2) than in the traditional one (3.4), indicating that the inclusion of AI tools may have fostered greater student involvement. Further, students in the AI-assisted classroom rated the appropriateness of lesson pacing (4.0 vs. 3.1) and the utility of teaching tools (4.7 vs. 2.9) considerably higher than in the traditional classroom. While technical difficulties were not relevant in the traditional setting, students in the AI-assisted group reported only modest issues (average 2.3 on a scale where lower is preferable), suggesting that technology integration was generally smooth and did not significantly hinder the learning experience. Comfort in asking questions and clarifying doubts was also reported as higher in the AI-assisted classroom (4.3) compared to the traditional classroom (3.5), reflecting an environment that was more supportive of active student engagement. Perhaps most compellingly, when asked about preferences for future courses, students in the AI-assisted classroom overwhelmingly favoured this method (average rating 4.9), while the traditional classroom received a much lower score (2.7).

However, these numerical trends must be interpreted with caution. The observed differences may reflect not only the influence of AI tools but also increased student familiarity with the course content or natural adaptation to university study between semesters. No inferential statistics were applied to the Likert data because the small sample size (n = 41) and ordinal nature of responses limited the validity of such analysis. Instead, descriptive comparisons were used to complement the sentiment data and provide qualitative context. Taken together, these results provide evidence that the integration of AI into classroom instruction not only improved student sentiment, as demonstrated by the earlier statistical analysis, but also transformed key dimensions of the learning environment. Students experienced greater engagement, more personalised support, increased interactivity, and higher satisfaction with instructional tools. Nonetheless, these results are best viewed as perceptual trends indicating relative affective improvement rather than causal proof of AI's pedagogical superiority.

The second part of the post-intervention study involved semi-structured interviews focusing on gathering in-depth, qualitative insights into students' experiences in both classroom settings. A subset of 15 volunteers (approximately 37% of participants) took part in these follow-up interviews, which were conducted online via Microsoft Teams. The interview protocol comprised seven open-ended questions inviting reflection on emotional engagement, interactivity, and perceived learning effectiveness. Participation was voluntary and anonymous (no cameras and no names). Instead of asking direct questions, the interviews encouraged students to reflect on their overall impressions, the most influential aspects of each classroom, and how specific instructional methods, tools, and opportunities shaped their learning journeys. Respondents also discussed their sense of engagement, opportunities for interaction and participation, and whether the classroom environment accommodated their individual learning preferences and pace. In addition, participants were invited to share any challenges or limitations they faced, and to compare the two instructional modes, expressing their preferences for future courses.

Qualitative data were analysed thematically following Braun and Clarke's (2006) six-phase model. Two independent coders reviewed all transcripts, achieving inter-coder reliability (Cohen's $\kappa = 0.86$). Themes were derived inductively from participant discourse, with discrepancies resolved through discussion. This coding protocol ensured transparency and strengthened interpretive validity. Our thematic analysis of these interview responses revealed several recurring themes that underpinned the more positive sentiment observed in the AI-assisted classroom. Foremost among these was the perception of personalised learning: students frequently cited the ability of AI-powered platforms to tailor content and activities according to their individual needs and proficiency levels, which led to greater engagement and a more positive learning experience overall. Many also noted the interactive nature of the AI-assisted classroom, highlighting gamified exercises, simulations, and other dynamic elements that made lessons more enjoyable and motivating. Another recurring theme was the benefit of immediate feedback. Students appreciated the real-time insights provided by AI systems, which helped them identify strengths and areas for improvement, thereby boosting confidence and sustaining motivation. The availability of enhanced language support, such as instant translation and pronunciation assistance, was also seen as instrumental in reducing language barriers and increasing comfort with participation. Furthermore, adaptive challenges, the ability of AI algorithms to adjust the difficulty level of tasks based on individual progress, helped maintain engagement and prevented both frustration and boredom. Continuous tracking and analysis of progress was also valued, as it enabled more targeted support and a clearer sense of achievement. These advantages of AI-assisted instruction translated into several positive outcomes. Students reported improvements in language proficiency, attributing this to increased motivation, engagement, and more focused practice. Many described a reduction in language learning anxiety and a greater willingness to communicate in the target language. Importantly, the positive emotional climate fostered by AI-supported instruction appeared to spark a long-term interest in language learning, with students expressing greater enthusiasm for continued study beyond the classroom. Enhanced engagement and enjoyment were also linked to better overall academic performance.

To illustrate these tendencies, we will show several representative student excerpts. One participant commented, "In the AI semester I felt more involving [sic] because I could see immediately if my writing was good or not. In the traditional class I have [sic] to wait for feedback." Another remarked, "I still prefer the teacher explaining grammar on the board, because sometimes the AI feedback was too mechanical." Such reflections reveal both the appeal of instant, adaptive feedback, and the continued value of direct human interaction. Nonetheless, it is important to recognise that while AI-assisted learning presents clear benefits, it may not suit all learners equally. Some students expressed a continued preference for traditional classroom environments, underscoring the importance of thoughtful implementation, ongoing educator training, and continual evaluation to ensure that AI integration effectively meets diverse student needs and delivers meaningful educational outcomes. These findings, though encouraging, should be interpreted within the limits of self-reported data, non-random sampling, and the absence of experimental control. Future research using mixed-method repeated-measures designs could more precisely quantify the emotional and linguistic effects of AI integration.

9. Conclusions

This study set out to explore the emotional experiences of ESP students in traditional and AI-assisted classroom settings using sentiment analysis and opinion mining. By analysing a dataset of student essays and conducting in-depth post-intervention questionnaires and interviews, we uncovered evidence suggesting that the integration of AI technologies into language education may enhance certain affective dimensions of the classroom experience. The results point to a general shift towards more positive sentiment in the AI-assisted (Target) condition compared to the traditional (Control) condition; however, these findings should be viewed as preliminary and correlational rather than conclusive.

Quantitative analysis revealed that students exposed to AI-supported instruction reported higher sentiment scores, greater engagement, and increased satisfaction with their learning experience. Nevertheless, sentiment values and self-reported perceptions represent indirect indicators of emotional states and cannot be interpreted as definitive measures of motivation, learning outcomes, or cognitive achievement. The statistical differences observed should therefore be understood as reflective of relative emotional tone rather than empirical proof of improved learning efficacy. The qualitative data complement these findings by providing contextual insight into students' perceptions of AI-supported learning. Participants described AI tools as facilitating more individualised feedback and flexible pacing, which they associated with higher comfort and engagement. Yet, such interpretations are subjective and may also be influenced by novelty effects, technological enthusiasm, or increased familiarity with university study by the second semester.

This research was exploratory in nature and conducted within a repeated-measures design, meaning that the same students experienced both instructional formats. While this design offers valuable within-subject comparability, it also introduces potential carryover effects from one semester to the next. The small sample size (n = 41), single-institution focus, and limited duration further restrict generalisability. Additionally, sentiment analysis, though validated for text-based affect detection, remains a probabilistic method sensitive to lexical variability, context, and cultural expression of emotion. Hence, while the hybrid analytic approach strengthens reliability, results should be interpreted cautiously.

Future studies should adopt more robust experimental or longitudinal designs, incorporating larger and more diverse participant samples across multiple institutions and disciplines. Mixed-method triangulation combining sentiment analysis with behavioural engagement metrics, performance data, and physiological indicators could also improve validity. Moreover, future work should include richer documentation of AI-assisted instructional procedures to allow for greater replicability and external evaluation. Despite these limitations, the current study contributes valuable preliminary insights into how AI-supported tools may influence the emotional climate of ESP learning environments. Its findings highlight the importance of understanding learner affect as part of responsible AI integration, offering a foundation for more rigorous inquiry rather than definitive conclusions.

References

Arnold, J. (2021). Affective factors in language learning: Making a difference. In M. Simons & T. F. H. Smits (Eds.), *Language education and emotions: Research into emotions and language learners, language teachers and educational processes* (pp. 3–17). Routledge/Taylor & Francis Group.

Ayeni, O. O., Al Hamad, N. M., Chisom, O. N., Osawaru, B., & Adewusi, O. E. (2024). *AI in education: A review of personalized learning and educational technology. GSC Advanced Research and Reviews, 18*(2), 261–271. https://doi.org/10.30574/gscarr.2024.18.2.0062

Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human–Computer Studies, 68*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 4,* Article 100118. https://doi.org/10.1016/j.caeai.2022.100118

Colace, F., Casaburi, L., De Santo, M., & Greco, L. (2015). Sentiment detection in social networks and in collaborative learning environments. *Computers in Human Behavior, 51*, 1061–1067. https://doi.org/10.1016/j.chb.2014.11.090

Dalipi, F., Zdravkova, K., & Ahlgren, F. (2021). Sentiment Analysis of Students' Feedback in MOOCs: A Systematic Literature Review. *Frontiers in Artificial Intelligence, 4*. https://doi.org/10.3389/frai.2021.728708

Dake, D. K., & Gyimah, E. (2023). Using sentiment analysis to evaluate qualitative students' responses. *Education and Information Technologies, 28*, 4629–4647. https://doi.org/10.1007/s10639-022-11349-1

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

Feldman, R., & Ungar, L. (2012). *Sentiment mining from user generated content.* (Tutorial presented at the Sixth International AAAI Conference on Weblogs and Social Media). (2 April 2025) <http://www.cis.upenn.edu/~ungar/AAAI/AAAI13.pdf>

Gardner, J., & Brooks, C. (2018). Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction, 28*(1), 127–203. https://doi.org/10.1007/s11257-018-9203-z

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Hutto, C., & 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

Islam, M. S., Kabir, M. N., Ghani, N. A., *et al.* (2024). Challenges and future in deep learning for sentiment analysis: A comprehensive review and a proposed novel hybrid approach. *Artificial Intelligence Review, 57*, Article 62. https://doi.org/10.1007/s10462-023-10651-9

Ivanović, I. (2024). Game-based learning and underachieving students: Replacing the traditional paradigm in Montenegro? *Education and Information Technologies*, 29. https://doi.org/10.1007/s10639-024-12740-w

Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, 102, Article 102019. https://doi.org/10.1016/j.inffus.2023.102019

Koufakou, A. (2024). Deep learning for opinion mining and topic classification of course reviews. *Education and Information Technologies, 29*, 2973–2997. https://doi.org/10.1007/s10639-023-11736-2

Li, G., & Pan, Y. (2025). Understanding trait emotional intelligence and L2 willingness to communicate: The serial mediation of perceived teacher support and L2 emotions. *Acta Psychologica, 260*, Article 105539. https://doi.org/10.1016/j.actpsy.2025.105539

Liu, B. (2010) Sentiment Analysis and Subjectivity. In: Indurkhya, N. and Damerau, F.J., Eds., *Handbook of Natural Language Processing*, 2nd Edition, Chapman and Hall/CRC, London, pp. 1–38.

Luckin, R., & Holmes, W. (2016). *Intelligence unleashed: An argument for AI in education.* UCL Knowledge Lab.

Machová, K., Szabóová, M., Paralič, J., & Mičko, J. (2023). Detection of emotion by text analysis using machine learning. *Frontiers in Psychology, 14*, 1190326. https://doi.org/10.3389/fpsyg.2023.1190326

Mohammad, S. M. (2021). Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion measurement* (2nd ed., pp. 323–379). Elsevier. https://doi.org/10.1016/B978-0-12-821124-3.00011-9

Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior, 31*, 527–541. https://doi.org/10.1016/j.chb.2013.05.024

Sivakumar, S., & Rajalakshmi, R. (2022). Context-aware sentiment analysis with attention-enhanced features from bidirectional transformers. *Social Network Analysis and Mining, 12*, Article 104. https://doi.org/10.1007/s13278-022-00910-y

Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. A. (2022). Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, *abs/2302.04359*, 100003. https://doi.org/10.1016/j.nlp.2022.100003

Vistorte, A. O. R., Deroncele-Acosta, A., Ayala, J. L. M., Barrasa, A., López-Granero, C., & Martí-González, M. (2024). Integrating artificial intelligence to assess emotions in learning environments: a systematic literature review. *Frontiers in Psychology*, *15*, 1387089. https://doi.org/10.3389/fpsyg.2024.1387089

Weitl-Harms, S., Hastings, J. D., & Lum, J. (2024). Using LLMs to establish implicit user sentiment of software desirability. In *Proceedings of the 2024 International Conference on Machine Learning and Applications (ICMLA)* (pp. 1645–1650). IEEE. https://doi.org/10.1109/ICMLA61862.2024.00254

Received: 19 June 2025
Accepted for publication: 20 October 2025