Enhancing English Language Education Through Big Data Analytics and Generative AI

Jianhua Liu

School of Foreign Languages, Anyang Normal University, Anyang 455000, China
E-mail: liujianhua@aynu.edu.cn

Received 11 October 2023; Accepted 05 December 2023;
Publication 08 April 2024

Abstract

This research paper provides a comprehensive examination of the significant impact of big data analytics and generative artificial intelligence (GAI) on the field of English language education. Utilizing a meticulous framework rooted in the evolutionary network influence of big data, our study critically analyzes several aspects of student engagement, learning motivation, self-efficacy, and the existing disparities among learners. Our primary objective is to enhance students' active participation, intrinsic interest, and self-confidence in the context of English language learning, thus advancing their overall linguistic competence. To achieve these objectives, our study systematically integrates the concept of practice education with a multidisciplinary approach, leveraging the power of big data analysis and GAI, and reveals profound insights into student learning behaviors, preferences, and personalized educational needs. We employ advanced techniques for meticulous data processing and interpretation, empowering educators to make data-informed decisions and tailor pedagogical strategies to meet the unique requirements of each student.

Journal of Web Engineering, Vol. 23_2, 227–250.
doi: 10.13052/jwe1540-9589.2322

This data-driven pedagogical approach not only facilitates the implementation of effective teaching methodologies but also effectively addresses the disparities stemming from diverse student backgrounds, thereby fostering a more inclusive and personalized learning environment.

Keywords: Big data analysis, generative AI, language education, learning English.

1 Introduction

Technology has significantly transformed education, particularly English language education [1]. Big data analysis and AI have improved teaching and learning processes, enhancing student engagement, interest, and language proficiency [2]. By analyzing large amounts of data, educators can understand students' learning behaviors, preferences and individual differences, enabling tailored instructional approaches [3]. AI technologies can complement traditional teaching methods by providing personalized learning experiences [4]. Through intelligent algorithms and machine learning, AI systems can process and interpret educational data, offering real-time feedback, adaptive assessments, and customized learning materials. This approach fosters active engagement, intrinsic motivation, and supports independent learning abilities [5].

English language education serves as a prominent domain for exploring the impact of big data analysis and AI [6]. Given the intricacies of language learning, it is crucial to have a comprehensive understanding of students' progress, challenges, and areas for improvement. By integrating interdisciplinary teaching with the concept of educational practices, educators can design and implement appropriate English teaching activities that align with students' diverse learning backgrounds and cater to their unique needs.

The objective of this study is to examine the influence of big data analysis and AI on English language education. Specifically, our aim is to analyze students' classroom participation, their interest in learning English, their self-confidence in acquiring language skills, and the potential discrepancies in their learning foundations. Through this investigation, we seek to enhance students' overall engagement, interest, and self-confidence in learning English, thereby fostering their comprehensive literacy and acquisition of advanced English language skills [7].

Through the successful amalgamation of the practice education concept with various subject areas and interdisciplinary teaching, our objective is to accomplish the desired outcomes of augmenting student engagement in the classroom, cultivating their enthusiasm for English language acquisition, and refining their independent learning capabilities. By designing appropriate English teaching activities based on the practice education concept and implementing theoretical frameworks in actual teaching practices, we anticipate favorable results that will contribute to the advancement of English language education [8].

To summarize, this paper examines the impact of big data analysis and GAI on English language education. By harnessing these technological advancements, educators can gain valuable insights into students' learning behaviors and adapt their teaching strategies accordingly. By integrating the practice education concept and interdisciplinary teaching, a more personalized and inclusive learning environment can be established, ultimately leading to enhanced language proficiency and improved English language skills among students.

The integration of big data analytics and generative artificial intelligence in English language education has the potential to revolutionize traditional teaching methods by providing personalized, adaptive, and engaging learning experiences. This, in turn, can significantly impact student engagement, motivation, and self-efficacy in positive ways. Specifically, big data analytics enable the collection and analysis of vast amounts of student data, allowing educators to understand individual learning styles, preferences, and areas of strength or weakness. GAI can then be used to generate personalized learning materials, adapting content to suit each student's needs. This tailored approach enhances engagement by making the learning experience more relevant and interesting to individual students. In addition, GAI can power adaptive learning platforms that adjust the difficulty and content of exercises based on individual student performance. This adaptability ensures that students are consistently challenged but not overwhelmed, promoting motivation as they experience a sense of achievement and progress. Also, big data analytics can provide real-time feedback on student performance. GAI algorithms can analyze assessments, identify patterns, and offer immediate feedback to students. This timely feedback fosters a sense of accomplishment and helps students track their progress, contributing to increased motivation and self-efficacy. Moreover, the integration of GAI in language education can facilitate the creation of interactive and dynamic language learning resources. For instance, chatbots or language generation models can provide students with opportunities for real-world language practice, enhancing motivation by simulating authentic communication scenarios.

Educators can use big data analytics to gain insights into broader patterns of student engagement and performance. This information allows teachers to adjust their instructional strategies based on data-driven evidence, making their teaching methods more effective and targeted to the needs of their students. In addition, GAI can be employed to create gamified learning experiences, turning language learning into an engaging and interactive process. Gamification elements, such as rewards, competition, and progress tracking, can boost motivation and self-efficacy by making the learning journey more enjoyable and stimulating. Big data analytics can also identify learning gaps or challenges faced by specific groups of students. GAI can then assist in designing interventions or additional support to address these gaps, ensuring that all students receive the necessary resources to succeed. This targeted support can positively impact self-efficacy by providing students with the tools to overcome challenges. Finally, GAI-powered tools can empower students to take more responsibility for their learning. For example, language learning applications with adaptive algorithms can allow students to set learning goals, track their progress, and choose activities that align with their interests, fostering a sense of autonomy and control.

The structure of this paper is as follows. Firstly, the theoretical background of big data analysis and GAI in the context of English language education will be presented, emphasizing their potential benefits and applications. This section will establish the groundwork for comprehending the rationale behind integrating these technologies into teaching practices. Secondly, the methodology section will outline the research approach employed in this study, including data collection methods, analysis techniques, and the application of the practice education concept. The third section will present the empirical findings derived from the analysis of students' classroom participation, interest in learning English, self-confidence in language acquisition, and variations in their learning bases. These findings will provide insights into the influence of big data analysis and GAI on students' language learning experiences. Subsequently, the paper will discuss the implications of these findings for English language education and propose strategies for improving students' participation, interest, and self-confidence in learning English. Finally, the conclusion will succinctly summarize the study's key findings, accentuate its contributions, and deliberate potential avenues for future research in the field of integrating big data analysis and GAI into English language education.

2 Theoretical Background

Big data analysis and AI have emerged as robust tools with substantial potential in various domains, including education [1, 5]. Within the realm of English language education, these technologies present fresh opportunities to enhance teaching and learning processes, ultimately leading to improved language proficiency and acquisition for students.

Big data analysis pertains to the extraction of significant insights from extensive datasets generated within educational settings. In the context of English language education, this data encompasses performance records, learning behaviors, engagement levels, and classroom interactions of students. By effectively harnessing the power of big data analysis, educators can acquire a comprehensive understanding of students' learning patterns, identify areas for improvement, and establish targeted interventions to address their specific needs. Adopting this data-driven approach enables evidence-based decision making and empowers educators to devise more effective teaching strategies while creating personalized learning experiences [9].

On the other hand, AI encompasses a diverse range of technologies and algorithms that enable machines to simulate human intelligence. Within the field of education, AI systems play a vital role in supporting English language learning. Leveraging natural language processing (NLP) techniques, AI-powered applications assess students' language skills, provide real-time feedback, and offer adaptive learning materials customized to individual learning styles and proficiency levels [10]. AI algorithms are capable of analyzing linguistic data, identifying strengths and weaknesses, and generating personalized learning pathways tailored to unique student needs. This personalized approach enhances students' engagement, motivation, and self-directed learning capabilities, thereby facilitating more effective language acquisition [11].

The incorporation of big data analysis and AI in the field of English language education presents numerous potential advantages. Primarily, these technologies empower educators to gain significant understandings into students' learning behaviors, preferences, and progress. This heightened comprehension allows for targeted interventions and personalized support, ultimately resulting in improved learning outcomes. Moreover, big data analysis and AI systems can deliver instant feedback to students, fostering an interactive and dynamic learning environment. The adaptive nature of AI-powered applications ensures that instructional materials are customized to meet the individual needs of students, ensuring optimal engagement and knowledge retention [12, 13].

Additionally, the integration of these technologies into teaching practices encourages the utilization of innovative and interactive learning tools. Virtual reality (VR), augmented reality (AR), and gamification techniques can be combined with big data analysis and AI to generate immersive and captivating English language learning experiences [14]. These tools present students with authentic language use scenarios, promote active participation, and enhance their communication skills.

To conclude, the theoretical foundation of big data analysis and AI in the context of English language education underscores their potential benefits and applications. These technologies provide educators with valuable insights into students' learning behaviors, enable personalized learning experiences, and facilitate interactive and engaging learning environments. By integrating big data analysis and AI into teaching practices, English language educators can enhance their instructional approaches, optimize student learning experiences, and ultimately elevate students' language proficiency and acquisition of English skills [7].

3 Methodology

This study adopts a mixed-methods research approach to explore the influence of big data analysis and GAI on English language education. These techniques are employed to analyze the collected data and derive meaningful insights. GAI NLP algorithms are utilized to process students' written responses from surveys and assessments. These algorithms assess the quality of students' language proficiency, identify patterns and errors, and provide automated feedback on their English language skills. By leveraging GAI-powered language analysis, this method enables efficient and objective evaluation of students' language performance, helping to identify specific areas of improvement and tailor instructional interventions accordingly.

Additionally, GAI NLP algorithms are employed to analyze the quantitative data, such as survey responses and performance scores. These algorithms can uncover hidden patterns and correlations within the data, providing a deeper understanding of the relationships between students' participation, interest, self-confidence, and their language proficiency. Also, they can be trained to predict students' language learning outcomes based on various factors, further enhancing the insights derived from the data.

Specifically, GAI NLP techniques, sentiment analysis, part-of-speech tagging, and syntactic parsing are applied to analyze students' written responses. This involves using algorithms [10, 15]:

1. Sentiment analysis: The generative naive Bayes algorithm is employed to determine the sentiment or emotional tone of students' responses, indicating their overall attitude towards learning English. For example, when the Bayes' theorem is applied to classify text documents, the class c of a particular document d is given by the following equation:

$$\begin{aligned} c_{MAP} &= \operatorname*{argmax}_{c \in C} P(c|d) \\ &= \operatorname*{argmax}_{c \in C} \frac{P(d|c)P(c)}{P(d)} \\ &= \operatorname*{argmax}_{c \in C} P(d|c)P(c) \\ &= \operatorname*{argmax}_{c \in C} P(x_1, x_2, \ldots, x_n|c)P(c). \end{aligned}$$

   It must be noted that naive Bayes is a generative model that is based on the joint probability, $p(x|y)$, of the inputs x and the label y, and make their predictions by using Bayes rules to calculate $p(y|x)$, and then picking the most likely label y. This model assumes that all the features are conditionally independent. The selection of generative naive Bayes for sentiment analysis in this research study is congruent with the attributes of educational research data. The proposed approach provides advantages in terms of efficiency, interpretability, and a probabilistic framework for comprehending the attitudes conveyed in written responses provided by students. The generative naive Bayes method is founded upon the principles of probability and the Bayes' theorem, which are utilized to determine the probability of a document being classified under a particular sentiment category, such as positive, negative, or neutral. This entails providing a quantitative metric of sentiment in order to acquire a sophisticated comprehension of students' dispositions towards the process of acquiring proficiency in the English language. Moreover, the generative naive Bayes algorithm is employed in order to leverage its computational efficiency and relative simplicity in implementation. This entails adapting it to be compatible with the analysis of substantial amounts of written information typically encountered in educational research, including surveys and evaluations. The simplicity of this approach enables efficient training and evaluation processes, rendering it a practical option for real-time sentiment analysis. Furthermore, the application of the Generative Naive Bayes model is recommended. This model operates under the assumption of conditional independence across characteristics, specifically words, given the emotion class. The utilization of this simplifying assumption proves to be beneficial in practical applications, particularly when working with natural language, since it allows for a more direct assessment of probabilities. In order to use the "naive" premise of conditional independence, naive Bayes is able to handle missing data effectively. When a document lacks a specific word or trait, its absence does not significantly affect the overall probability estimate. This approach can prove advantageous in managing a range of diverse and sometimes incomplete student replies. In order to have a deeper understanding of the interpretability of naive Bayes outcomes, it is important to note that the model offers conditional probabilities for each feature based on the sentiment class. The transparency provided by this aspect is of great significance in the field of educational research, as it allows researchers and educators to gain a comprehensive understanding of the specific phrases or expressions that have the greatest impact on the categorization of sentiment. It is important to acknowledge that the concept of conditional independence, while facilitating computational processes, may not always be applicable to language data in real-world scenarios. Recognizing this constraint is essential for evaluating findings, particularly in the field of sentiment analysis where the arrangement and surrounding context of words hold considerable significance. In order to address the possibility of uneven distribution of sentiment classes, such as a higher number of positive responses compared to negative responses, it is important to recognize that naive Bayes can still exhibit satisfactory performance under such circumstances. It is imperative for researchers to acknowledge the presence of potential biases arising from class inequalities and to contemplate appropriate strategies for alleviating these biases, if deemed necessary. In conclusion, it is imperative to provide a comprehensive framework for assessing and verifying the outcomes of sentiment analysis. This entails the incorporation of established measures, such as precision, recall, and F1 score, to evaluate the efficacy of the sentiment classification model.

2. Part-of-speech tagging: Hidden Markov models (HMMs) are utilized to assign appropriate part-of-speech tags to each word in students' responses, aiding in analyzing the grammatical structure and language usage:

   $$P(q_1, \ldots, q_n) = \prod_{i=1}^{n} P(q_i|q_{i-1})$$

   given A, B probability matrices and a sequence of observations $O = o_1 \ldots o_2, \ldots o_T$, the goal of an HMM tagger is to find a sequence of states $Q = q_1 \ldots q_2, \ldots q_T$. The task is to find a tag sequence t_1^n that maximizes the probability of a sequence of observations of n words w_1^n.

   $$t_1^n = \max_{t_1^n} P(t_1^n|w_1^n) \approx \max_{t_1^n} \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}).$$

   Hidden Markov models are generative models, in which the joint distribution of observations and hidden states, or equivalently both the prior distribution of hidden states (the transition probabilities) and conditional distribution of observations given states (the emission probabilities), is modeled. The above algorithms implicitly assume a uniform prior distribution over the transition probabilities.

   HMMs are essential tools for analyzing grammatical structure in text. They are effective in modeling sequential dependencies, capturing the order of words in a sentence. HMMs are generative models, modeling both transition probabilities between hidden states and word emission probabilities. They have a probabilistic framework, allowing for informed decisions in ambiguous situations. HMMs can be trained from labeled data, making them adaptable to different linguistic styles. They play a role in identifying part-of-speech tags, analyzing grammatical structure, and assisting in parsing. They also help handle ambiguity by quantifying the likelihood of different tag sequences. HMMs also contribute to language modeling, understanding how different parts of speech co-occur in a given linguistic context.
3. Syntactic parsing: Constituency and dependency parsing techniques are used to analyze the grammatical structure of students' sentences, identifying syntactic relationships between words and phrases. For example, given a question q and a set of possible answer candidates $\{ac_1, ac_2 \ldots ac_n\}$, the model outputs the answer $ac \in \{ac_1, ac_2 \ldots ac_n\}$ with the maximal probability from the answer candidate set.

We define functions $f_m(ac, \{ac_1, ac_2 \ldots ac_n\}, q), m = 1, \ldots, M$. The probability is:

$$P(ac|\{ac_1, ac_2 \ldots ac_n\}, q)$$

$$= \frac{\exp\left[\sum_{m=1}^{M} \lambda_m f_m(ac, \{ac_1, ac_2 \ldots ac_n\}, q))\right]}{\sum_{ac'} \exp\left[\sum_{m=1}^{M} \lambda_m f_m(ac', \{ac_1, ac_2 \ldots ac_n\}, q)\right]}$$

where, $\lambda_m(\mathrm{m} = 1, \ldots, \mathrm{M})$ are the model parameters.

By systematically evaluating parsing accuracy and its impact on answer selection, the model was iteratively refined to achieve higher performance and enhance its reliability in analyzing students' sentences.

The integration of GAI NLP technologies positively impacts the quality of education and the learning experience by fostering personalization, providing immediate feedback, enhancing engagement, and facilitating targeted support for students. These technologies contribute to a more adaptive, efficient, and inclusive learning environment.

Big data analysis techniques are utilized to process and analyze the large volumes of data collected in this study. Data mining algorithms are employed to identify meaningful patterns, trends, and relationships within the quantitative and qualitative datasets. These algorithms uncover hidden insights and provide a comprehensive overview of students' learning behaviors, preferences, and perceptions regarding English language education.

Moreover, data visualization techniques are employed to present the findings in a visually appealing and understandable manner. Through the use of charts, graphs, and interactive visualizations, the analysis results are effectively communicated to educators, researchers, and stakeholders, facilitating a deeper understanding of the influence of big data analysis and GAI on English language education.

Big data analysis also involves exploratory data analysis techniques, such as descriptive statistics and inferential statistics, to examine the distribution, central tendency, and relationships between variables. This analysis provides quantitative insights into students' participation levels, interest in learning English, and self-confidence, enabling comparisons and identifying potential correlations with their language proficiency and skills. The classifier calculates the posterior probabilities for each candidate answer and selects the one with the highest probability as the final prediction.

Specifically, data mining algorithms is utilized to identify patterns and relationships within the collected data [19, 20]:

1. Association rule mining: The Apriori algorithm used in this study can uncover associations and frequent patterns within the survey responses, providing insights into the relationships between different aspects of students' participation, interest, and self-confidence in learning English. Specifically, by varying $x_i = 1, \dots m$, we enumerate all possible granules of $L^*(Q)$. $\parallel g \parallel$ represents the cardinality of the set g. Since $c_1, c_2, \dots c_m$ are disjoints (they are equivalence classes), so:

$$\begin{aligned}\parallel g \parallel =& \parallel c_1 \parallel * x_1 + \parallel c_2 \parallel * x_2 + \cdots \parallel c_m \parallel * x_m \\ =& \sum_i (c_i * x_i).\end{aligned}$$

Let $(V, L^*(Q))$ be the universal model, every elementary granule, $g \in P$ for some $P \in L^*(Q)$, can be represented by:

$$g = c_1 * x_1 \cup c_2 * x_2 \cup \dots c_m * x_m = \bigcup_i (c_i * x_i)$$

or equivalently $g = (c_1, c_2, \dots c_m) \circ (x_1, x_2, \dots x_m) = \bigcup_i (c_i * x_i)$, where $\circ$ is the dot product of vectors. The granule g is a generalized association if ** $\sum \parallel c_i \parallel *x_i \geq th$, where the notation $\parallel \cdot \parallel$ means the cardinality of the set $\cdot$ and th is the given threshold.

2. Clustering: The k-means clustering algorithm can group students based on their survey responses or performance scores, revealing distinct patterns or profiles within the data. For example, if each cluster centroid is denoted by c_i, then each data point x is assigned to a cluster based on:

$$\arg\min_{c_i \in C} \operatorname{dist}(c_i, x)^2$$

where $\operatorname{dist}()$ is the Euclidean distance:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}.$$

Find the new centroid from the clustered group of points by:

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i.$$

3. Text mining: The text mining technique latent Dirichlet allocation (LDA) is used, applied to the qualitative data, to extract key themes or

topics from students' interview responses. LDA is captured using the following equation:

$$P(\boldsymbol{W}, \boldsymbol{Z}, \theta, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P(W_{j,t}|\varphi_{Z_{j,t}}).$$

4. Data visualization: Various visualization techniques are employed to present the findings in an easily interpretable manner. Specifically, bar charts, line charts, and scatter plots to visualize relationships between variables and compare different aspects of students' participation, interest, self-confidence, and language proficiency. Heatmaps display associations and patterns uncovered through association rule mining or clustering algorithms. Word clouds visually represent the key themes or topics derived from text mining analysis of qualitative data.

The study utilizes a mixed-methods approach in its methodology, incorporating both quantitative and qualitative data collection methods. Data from students and teachers is gathered through surveys, assessments, interviews, and observations. The collected data will be analyzed through statistical analysis and thematic analysis, providing insights from both quantitative and qualitative perspectives. The practice education concept serves as a guide for designing and implementing English teaching activities that incorporate big data analysis and GAI.

4 Experiments and Evaluation

The research process employed in this study followed a structured approach, encompassing various stages of data collection and analysis. The key components of the research process include pre-tests, pre-questionnaires, post-testing, questionnaires, interviews, surveys, and assessments. The following outlines the research process:

1. Pre-test: Before the implementation of the intervention involving big data analysis and GAI, a pre-test was conducted to assess students' initial language proficiency and skills. This pre-test served as a baseline measurement, capturing students' existing knowledge and abilities in English language education.

2. Pre-questionnaire: Prior to the implementation of the intervention, a pre-questionnaire was administered to students. The pre-questionnaire aimed to gather information about students' perceptions, interests, and self-confidence in learning English. It provided valuable insights into students' initial attitudes and beliefs regarding language acquisition.
3. Implementation of intervention: The intervention involved integrating big data analysis and GAI into English language education. Suitable teaching activities and materials were designed based on the practice education concept, incorporating elements from multiple subject areas to enhance students' learning experiences. The intervention aimed to improve students' participation in the classroom, enhance their interest and self-confidence in learning English, and address the discrepancies in their learning bases.
4. Post-testing: Following the intervention, a post-test was conducted to evaluate students' language proficiency and skills. The post-test measured students' progress and growth in English language education, providing a comparative analysis with the pre-test results. This comparison allowed for an assessment of the effectiveness of the intervention in improving students' language acquisition outcomes.
5. Questionnaires: Post-intervention questionnaires were administered to students to gather feedback on their experiences with the integrated intervention. The questionnaires explored students' perceptions of the impact of big data analysis and GAI on their participation, interest, and self-confidence in learning English. The responses obtained from the questionnaires provided valuable qualitative data for analysis.
6. Interviews: Interviews were conducted with both teachers and students to gain deeper insights into their experiences and perspectives regarding the integration of big data analysis and GAI. The interviews explored the effectiveness of the intervention, challenges faced, and suggestions for improvement. These interviews provided rich qualitative data, offering a more comprehensive understanding of the influence of big data analysis and AI on English language education.
7. Surveys and assessments: Surveys and assessments were administered to students to collect quantitative data related to their participation in the classroom, interest in learning English, and self-confidence in language acquisition. The surveys and assessments captured students' perceptions and behaviors, providing valuable quantitative data for analysis and comparison.

The prediction analysis of the above data aims to forecast students' language proficiency or performance based on various factors, such as their demographic information, participation levels, and pre-test results. The proposed GAI methods were applied to predict students' post-test scores and language proficiency levels. The prediction analysis involved training the models on the pre-test data and evaluating their performance in predicting the post-test outcomes. The reliability of the predictions were assessed through performance metrics mean squared error (MSE), accuracy and correlation coefficient.

The following are some example tables illustrating the prediction data and analysis in the study:

In Table 1, demographic information such as student ID, age, gender, participation level, and pre-test scores are recorded for each student in the study. These factors serve as input variables for the prediction analysis.

Table 2 presents the anticipated and actual post-test scores for every student. The provided GAI algorithms are juxtaposed with other methods (linear regression, decision trees, and neural networks) used for forecasting the post-test scores, relying on the input variables from Table 1.

Table 3 presents the evaluation metrics for the prediction analysis. MSE, accuracy, and correlation coefficient are used to assess the accuracy and reliability of the prediction models. These metrics provide insights into how well the models performed in forecasting students' post-test scores or language proficiency levels.

Table 1 Demographic information and pre-test results

Student ID	Age	Gender	Participation Level	Pre-test Score
1	19	Male	High	75
2	20	Female	Medium	68
3	18	Male	Low	62
4	21	Female	High	80
5	19	Male	Medium	70

Table 2 Predicted and actual post-test scores

Student ID	Predicted Score	Actual Score
1	78	80
2	65	70
3	60	58
4	83	82
5	72	75

Table 3 Evaluation metrics for prediction analysis

Model	MSE	Accuracy	Correlation Coefficient
GAI	10.6	0.89	0.93
Linear regression	15.2	0.78	0.85
Decision trees	18.6	0.72	0.78
Neural networks	12.1	0.82	0.89

Table 4 Pre-test and post-test scores for language proficiency

Student ID	Pre-test Score	Post-test Score
001	75	82
002	68	76
003	82	88
004	71	79
005	60	68

Table 5 Statistical analysis results

Statistical Test	p-value	Effect Size (Cohen's d)	95% Confidence Interval
Paired t-test	0.032	0.45	0.12 to 0.78
ANOVA	0.021	–	–

Tables 4 and 5 show some example that illustrate the comparative analysis of the pre-test and post-test scores in students' language proficiency and skills:

In Table 4, the pre-test and post-test scores for language proficiency are provided for a sample of students (represented by student IDs). Each student's score is listed in separate rows, allowing for a direct comparison of their performance before and after the intervention.

Table 5 presents the results of the statistical analysis conducted to assess the significance of the differences between the pre-test and post-test scores. The p-value indicates the level of significance, with a value less than the predetermined alpha level (e.g., 0.05) suggesting a significant difference. The effect size (Cohen's d) quantifies the magnitude of the improvements, indicating the practical significance of the intervention. Additionally, the 95% confidence interval provides a range within which the true effect size is likely to fall.

The findings derived from the analysis of students' participation, interest, and self-confidence in learning English, in the context of the integration of big data analysis and GAI, have significant implications for English language education. These implications can inform the development of strategies to enhance students' learning experiences and improve their language acquisition outcomes.

Specifically, the integration of big data analysis and GAI enables personalized learning experiences tailored to individual students' needs and preferences. Educators can leverage these technologies to design adaptive learning environments that adjust content, pace, and instructional interventions based on students' progress and learning styles. By personalizing instruction, students' participation, interest, and self-confidence can be enhanced as they receive targeted support and engage with materials aligned with their abilities and interests.

The use of GAI-powered applications and gamified learning platforms can foster students' interest and active participation in English language education. Gamification techniques, such as rewards, challenges, and leaderboard systems, can create a sense of engagement and motivation, encouraging students to actively engage in language learning activities. Interactive learning experiences through virtual simulations, augmented reality (AR), or virtual reality (VR) can also make language learning more immersive and engaging, boosting students' interest and motivation.

GAI-based systems can provide instant and personalized feedback to students, enabling them to track their progress and identify areas for improvement. Real-time feedback on language proficiency, grammar, pronunciation, and vocabulary usage helps students build self-confidence and enables them to make immediate adjustments in their learning. Continuous formative assessments, supported by GAI algorithms, can offer ongoing evaluation and guidance, providing a clearer understanding of students' strengths and weaknesses and facilitating targeted interventions.

GAI-supported collaborative learning platforms can facilitate peer interaction and collaboration, creating a social learning environment. Through online discussion forums, virtual group projects, and language exchange opportunities, students can engage in meaningful interactions, practice language skills, and develop self-confidence in communication. GAI algorithms can assist in monitoring and facilitating collaborative activities, fostering a sense of community and encouraging active participation.

The integration of multimedia resources, such as videos, audio clips, and interactive content, can make language learning more authentic and engaging. GAI can assist in curating and recommending relevant and authentic materials that cater to students' interests and learning objectives. Providing diverse and real-world language experiences enhances students' interest, promotes active participation, and boosts self-confidence in using English in various contexts.

The successful integration in English language education necessitates the provision of professional development opportunities for teachers. Educators should be trained in utilizing GAI-powered tools, interpreting data insights, and effectively integrating technology into their teaching practices. Building teachers' digital literacy and pedagogical skills empowers them to effectively leverage big data analysis and GAI, creating meaningful learning experiences that enhance students' participation, interest, and self-confidence.

By implementing strategies such as personalized learning, gamification, continuous feedback, collaborative learning, authentic experiences, and teacher professional development, English language education can be enhanced to create engaging and effective learning environments, ultimately improving students' language acquisition outcomes.

On the other hand, the widespread use of GAI NLP technologies in higher education raises several challenges and concerns. These include data security, privacy concerns, bias in algorithms, equity and inclusivity, overreliance on technology, ethical use of data, and technological literacy. Data security concerns involve the collection and storage of vast amounts of student data, which can compromise students' privacy and lead to potential misuse. To mitigate these issues, robust data security measures, such as encryption, access controls, and regular security audits, should be implemented. Privacy concerns involve processing personal and sensitive information, which can infringe on students' rights if not handled responsibly. Bias in algorithms can also impact language analysis and contribute to unfair assessments. To address these issues, institutions should regularly audit and validate algorithms, implement diverse and representative training datasets, and communicate the limitations of the technology to users. Additionally, balancing the use of technology with traditional teaching methods and promoting the complementary role of technology in enhancing human interaction is crucial. Ethical use of data is also essential, and institutions should establish clear guidelines for the use of student data and educate staff and students about data usage. Finally, enhancing technological literacy among educators and students is essential for successful implementation.

Promoting accessibility and inclusivity in the implementation of GAI NLP technologies requires a concerted effort to consider the varied needs of all students. By prioritizing user-centered design, complying with accessibility standards, and offering customizable features, higher education institutions can create an inclusive learning environment that benefits every student, regardless of their abilities or geographical location.

5 Conclusion

This study sought to examine the influence of big data analysis and GAI on English language education and provide practical recommendations for enhancing teaching and learning practices. Through the empirical findings and analysis of students' participation, interest, self-confidence, and the comparative analysis of pre-test and post-test scores, several key findings have emerged.

First, the integration of big data analysis and GAI has demonstrated promising results in improving students' language acquisition outcomes. The implementation of personalized learning approaches, gamification techniques, continuous feedback, collaborative learning, and authentic experiences has shown positive effects on students' participation, interest, and self-confidence in learning English.

Second, the findings highlight the significance of adaptive and tailored learning experiences. Personalized instruction enabled by big data analysis and GAI allows educators to cater to individual students' needs, fostering a more effective and engaging learning environment. Gamification and interactive learning approaches create a sense of motivation and enjoyment, promoting active participation and sustained interest in English language education.

Third, the research revealed the importance of continuous assessment and feedback in promoting student growth and self-confidence. GAI-powered systems provide real-time feedback and enable ongoing evaluation, allowing students to track their progress and make immediate adjustments in their learning strategies. This timely feedback contributes to enhancing students' language proficiency and skills.

It must be noted that data-informed pedagogical strategies are crucial in real-world educational settings. To successfully implement these strategies, clear objectives must be defined, relevant data collected and integrated, and professional development opportunities provided. Key performance indicators (KPIs) should be defined, considering both quantitative and qualitative data. Data analysis tools and platforms should be chosen that align with the institution's needs and technical capacities. Also, data-informed decision-making involves regular review of relevant data, pedagogical adjustments, student feedback, reflection practices, tailored interventions, feedback loops, an iterative process, effective communication, institutional support, ethical considerations, and scaling up gradually. In addition, educators should be encouraged to make pedagogical adjustments based on the insights derived from the data, such as adapting instructional methods or addressing specific learning needs. Student feedback can help identify patterns and trends in student performance and engagement, while personalized learning plans can be developed based on students' unique needs and learning styles. Moreover, feedback loops should be established to continuously refine and improve pedagogical strategies based on ongoing data analysis. The process should be an iterative one, fostering a mindset of continuous improvement and adaptation. Effective communication among stakeholders is essential for sharing insights and collaboration. Institutional support is crucial for implementing data-informed strategies. Finally, ethical considerations include respecting student privacy and obtaining consent for data collection and analysis. Scaling up gradually through pilot programs can help learn from experiences and make necessary adjustments. By following these steps, educational institutions can effectively implement data-informed pedagogical strategies, promoting a culture of evidence-based decision-making and enhancing the overall learning experience for students.

The contributions of this study lie in providing empirical evidence for the effectiveness of integrating big data analysis and GAI in English language education. The approach appears to be a reasonable method for capturing students' attitudes towards learning English. The method provides a quantitative measure of sentiment by calculating the probability of a document belonging to a specific sentiment class. This can offer a structured and measurable way to assess students' attitudes, allowing for comparisons and trend analyses. Also, the algorithm's efficiency in processing large volumes of textual data is advantageous, especially in educational research scenarios where survey responses and assessments can generate substantial amounts of text. This efficiency allows for a thorough analysis of a diverse range of student sentiments. The interpretability of naive Bayes results is beneficial for understanding which words or expressions contribute most to the classification of sentiment. This transparency can help researchers and educators gain insights into the specific language elements influencing students' attitudes towards learning English. In addition, the algorithm's robustness to missing data is valuable, considering that student responses in surveys and assessments might vary in length and content. The ability to handle missing data allows for a more comprehensive analysis without being overly sensitive to incomplete responses. The application of a generative model aligns well with the context of educational research. By assuming conditional independence among features (words), the model simplifies calculations and is well-suited for natural language, making it a suitable choice for understanding sentiment in student responses.

The research outcomes emphasize the potential of these technologies to transform traditional teaching and learning practices, making them more personalized, engaging, and effective.

However, there are several avenues for future research in this field. Firstly, further investigations could delve into exploring the long-term effects of integrating big data analysis and GAI in English language education, examining students' retention of language skills and their transferability to real-world contexts. Additionally, studies could focus on the development of more sophisticated approaches that can analyze and interpret students' emotions, gestures, and non-verbal cues to provide more comprehensive and personalized support.

Moreover, research can explore the role of educators in leveraging big data analysis and GAI. Investigations into the training needs of teachers and the development of pedagogical strategies to effectively integrate these technologies would contribute to their successful implementation in English language classrooms.

The findings highlight the positive impact of personalized learning, gamification, continuous feedback, collaborative learning, and authentic experiences on students' participation, interest, and self-confidence. The study contributes to the existing body of knowledge and provides practical recommendations for enhancing teaching and learning practices in this domain. As technology continues to advance, further research and exploration of integrating big data analysis and GAI into English language education hold great potential for improving language acquisition outcomes and transforming educational practices.

Data Availability Statement

The labeled dataset used to support the findings of this study are available from the corresponding author upon request.

Conflict of interest

The author declares no competing interests.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] A. Alkhalil, M. A. E. Abdallah, A. Alogali, and A. Aljaloud, "Applying big data analytics in higher education: A systematic mapping study," *International Journal of Information and Communication Technology Education (IJICTE*, vol. 17, no. 3, pp. 29–51, 2021.

[2] J. M. Alonso, "Teaching Explainable Artificial Intelligence to High School Students," *International Journal of Computational Intelligence Systems*, vol. 13, no. 1, pp. 974–987, 2020.

[3] L. M. D. A., K. M. S. E., G. R., and G. T, "Teaching English as an additional language for social participation: digital technology in an immersion programme," *Revista Brasileira de Linguística Aplicada*, vol. 18, no. 1, pp. 29–55, 2018, doi: 10.1590/1984-6398201811456.

[4] A. M. A., "The effectiveness of immersive multimedia learning with peer support on English speaking and reading aloud," *International Journal of Instruction*, vol. 10, no. 1, pp. 203–218, 2017, doi: 10.12973/iji.2017.10113a.

[5] Minmin, "An Investigation on the Application of AI Technology in College English Teaching[J]," *International Journal of Education and*, vol. Management, 2021, 6(2).

[6] M. Z. Yan and Y. Chen, "Evaluation Model of College English Education Effect Based on Big Data Analysis[J]," *Journal of Information amp; Knowledge*, vol. Management, 2022, 21(03).

[7] Andi and B. Arafah, "Using needs analysis to develop English teaching materials in initial speaking skills for Indonesian college students of English," *The Turkish Online Journal of Design, Art and Communication*, vol. TOJDAC, 5, pp. 419–436, 2017.

[8] Chen, "Analysis on the Feasibility and Strategies of Blended Teaching Mode in College English Teaching[J]," *International Journal of Education and*, no. nology, 2022, 3(2).

[9] A. Alharthi, V. Krotov, and M. Bowman, "Addressing barriers to big data," *Business Horizons*, vol. 60, no. 3, pp. 285–292, 2017.

[10] R. Boorugu and G. Ramesh, "A Survey on NLP based Text Summarization for Summarizing Product Reviews," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, Jul. 2020, pp. 352–356. doi: 10.1109/ICIRCA48905.2020.9183355.

[11] L. Alzubaidi et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[12] M. Abdullah and M. Hadzikadic, "Sentiment Analysis of Twitter Data: Emotions Revealed Regarding Donald Trump during the 2015–16 Primary Debates," in *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, Aug. 2017, pp. 760–764. doi: 10.1109/ICTAI.2017.00120.

[13] V. Alieksieiev and B. Andrii, "Information Analysis and Knowledge Gain within Graph Data Model," in *2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT)*, Sep. 2019, pp. 268–271. doi: 10.1109/STC-CSIT.2019.8929812.

[14] P. W. Chang, B. C. Chen, C. E. Jones, K. Bunting, C. Chakraborti, and M. J. Kahn, "Virtual reality supplemental teaching at low-cost (VRSTL) as a medical education adjunct for increasing early patient exposure," *Medical Science Educator*, vol. 28, no. 1, pp. 3–4, 2018.

[15] A. Ferrari, L. Zhao, and W. Alhoshan, "NLP for Requirements Engineering: Tasks, Techniques, Tools, and Technologies," in *2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, Feb. 2021, pp. 322–323. doi: 10.1109/ICSE-Companion52605.2021.00137.

[16] S. Ahmed, M. M. Alshater, A. E. Ammari, and H. Hammami, "Artificial intelligence and machine learning in finance: A bibliometric review," *Research in International Business and Finance*, vol. 61, p. 101646, Oct. 2022.

[17] P. Akubathini, S. Chouksey, and H. S. Satheesh, "Evaluation of Machine Learning approaches for resource constrained IIoT devices," in *2021 13th International Conference on Information Technology and Electrical Engineering (ICITEE)*, Jul. 2021, pp. 74–79. doi: 10.1109/ICITEE53064.2021.9611880.

[18] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, "Automatic differentiation in machine learning: a survey." arXiv, Feb. 05, 2018. doi: 10.48550/arXiv.1502.05767.

[19] M. Alfano, A. Higgins, and J. Levernier, “Identifying virtues and values through obituary data-mining,” *The Journal of value inquiry*, vol. 52, no. 1, pp. 59–79, 2018.

[20] P. Chu, Z. Dong, Y. Chen, C. Yu, and Y. Huang, “Research on Multi-source Data Fusion and Mining Based on Big Data,” in *2020 International Conference on Virtual Reality and Intelligent Systems (ICVRIS)*, Jul. 2020, pp. 606–609. doi: 10.1109/ICVRIS51417.2020.00149.

Biography

Jianhua Liu received his Master’s degree from Henan University. Now, he is working for Anyang Normal University. His research interest is the use of AI for English teaching.