International Journal of Engineering Research and Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 22, Issue 4 (April 2026), PP 147-152

An Analysis of English Proficiency Among Arts College Students Using Data Mining Approach

J. Muthukumar, *Student of UG, Ayya Nadar Janaki Ammal College, Sivakasi, India*,
Dr. A. Dharmarajan, *Associate Professor, Ayya Nadar Janaki Ammal College, Sivakasi, India.*

Abstract— *This study examines the underlying factors contributing to the difficulties faced by Arts and Science college students in the Virudhunagar district in acquiring English proficiency and effective communication skills. A researcher-designed questionnaire was administered to assess students' grammar knowledge, vocabulary strength and speaking abilities. The collected responses were transformed into a structured dataset, upon which clustering techniques were applied to categorize students based on the primary causes of their limited English proficiency. Subsequently, classification methods were used within these clusters to identify the specific factors influencing each group. The findings reveal that the majority of students struggle due to an inadequate foundation in grammar and insufficient opportunities to practice spoken English. Additionally, limited exposure to English in their daily environment further diminishes their confidence. The study emphasizes the need for guiding students not only toward academic achievement but also toward developing strong English communication skills through continuous practice, enhanced grammar instruction and activity-based learning approaches.*

Keywords: *English Proficiency, Data Mining, Clustering, Classification, Machine Learning*

Date of Submission: 28-03-2026 Date of acceptance: 08-04-2026

I. INTRODUCTION

English proficiency has become a crucial requirement for academic achievement, career growth and active participation in today's globalized world. In India, especially within rural and semi-urban regions, English functions not only as a subject but also as a gateway to wider educational and professional opportunities. However, many undergraduate students in Arts and Science colleges struggle to attain effective communication skills, particularly in districts like Virudhunagar, where a large number of learners come from Tamil-medium backgrounds and have limited exposure to English outside the classroom.

The gap between classroom learning and practical communication needs continues to widen due to several persistent challenges. Students often lack a strong foundation in grammar, possess restricted vocabulary and receive minimal opportunities to practice speaking English. Traditional teaching approaches that emphasize rote learning and exam preparation further limit students' linguistic growth, leading to reduced confidence and fragmented language abilities. Consequently, these barriers affect not only their academic performance but also their employability and real-world communication competence.

In this context, it becomes essential to examine the underlying factors contributing to low English proficiency and to identify meaningful patterns that can guide targeted interventions. The use of data-driven methods such as clustering and classification enables a more systematic analysis of student performance by grouping learners based on shared difficulties and determining key influences within each group. This study adopts these analytical techniques to evaluate grammar knowledge, vocabulary strength and speaking skills among Arts and Science students in the Virudhunagar district. The findings aim to support educators and institutions in designing effective strategies that promote continuous practice, activity-based learning and enhanced instructional support for improved English communication skills.

II. THE METHODOLOGY

This study employed a quantitative research design supported by data-driven analytical techniques to investigate the factors affecting English proficiency among Arts and Science college students in the Virudhunagar district. The methodology consisted of four major phases: instrument design, data collection, dataset construction and analytical processing using clustering and classification methods.

A. Research Design and Participants

The study adopted a descriptive and analytical approach. Participants were undergraduate students from selected Arts and Science colleges located in the Virudhunagar district of Tamil Nadu. A purposive sampling method was

used to include students from diverse academic streams to ensure variability in English language exposure and learning backgrounds. The final sample size was determined based on student availability and voluntary participation.

B. Instrument Development

A researcher-designed questionnaire was created to assess three core components of English proficiency:

1. **Grammar Knowledge** – items related to sentence structure, tenses, parts of speech, and error identification.
2. **Vocabulary Strength** – items measuring word meaning, usage, synonyms, antonyms and contextual understanding.
3. **Speaking Ability** – evaluation of confidence, fluency, pronunciation and practice opportunities.

The questionnaire was validated through expert review to ensure clarity, content relevance and alignment with language proficiency indicators.

C. Data Collection Procedure

Students were provided with the questionnaire in a controlled environment to minimize external influence. The Participants were assured of confidentiality. The completed responses were collected, checked for completeness and prepared for digital entry.

D. Dataset Construction

All responses were transformed into a structured dataset suitable for computational analysis. Each questionnaire item was encoded numerically to represent students' performance levels. Derived variables were created to summarize grammar accuracy, vocabulary strength and speaking confidence scores. The final dataset served as input for the clustering and classification processes. Table presents the structure of the dataset used for analysis, including the key fields and their descriptions.

Column Name	Purpose
Student_ID	Each student unique ID
Student_Type	Rural / College
Question_ID	Q1, Q2, Q3…
Question	Actual question text
Correct_Answer	Expected correct answer
Student_Answer	Student written answer
Is_Correct	Yes / No
Mistake_Type	Grammar / Vocabulary / Tense / Spelling / Comprehension
Sub_Mistake	Article error / Verb form / Word meaning
Score	0 or 1 (or 2 marks)

Table 1. Structure of the Dataset

E. Clustering Analysis

To identify underlying patterns and group learners based on common difficulties, unsupervised machine-learning techniques were applied. The clustering process involved:

- **Data normalization** to ensure uniform scaling.
- **Feature selection** to retain variables most reflective of language proficiency.
- **Application of clustering algorithms** such as K- Means or hierarchical clustering to segment students into groups based on their performance.

The output clusters revealed distinct categories of learners experiencing specific barriers in grammar, vocabulary or speaking skills.

F. Classification Analysis

Following clustering, supervised classification methods were implemented to determine the dominant factors influencing each student group. Algorithms such as Decision Trees, Random Forest, or Naïve Bayes were applied

to:

- Analyze feature importance within each cluster.
- Identify predictors contributing to low proficiency.
- Validate the consistency of identified patterns across the dataset.

This two-stage approach ensured both unsupervised discovery of learner categories and supervised confirmation of contributing factors.

G. Ethical Considerations

Participation was voluntary and students provided informed consent. No personal identifiers were collected and all data were used strictly for research purposes.

III. RELATED WORKS

The information about Previous research on English proficiency and communication skill challenges has been widely explored through error analysis and linguistic studies, focusing on grammar, vocabulary and writing difficulties faced by EFL learners. Foundational work by Corder [4] emphasized the importance of learner errors in understanding language acquisition, which laid the theoretical basis for later grammatical error analysis studies. Several researchers analysed students' written outputs to identify recurring grammatical and lexical issues using systematic approaches.

Studies such as Agustin and Wulandari [1], Ayar [2], El Mahdy [5], Floranti and Adiantika [6], Huda and Wuda [8], Noor Azizah et al. [12], and Le Thi Trung Dinh [11] conducted detailed grammatical error analysis on EFL learners' writing samples. These works primarily focused on identifying error types related to tense usage, subject–verb agreement, article usage and sentence structure. Their findings highlighted that grammar and vocabulary deficiencies significantly affect students' overall English proficiency and communication skills.

Several studies applied computational and data-driven techniques to analyse linguistic patterns. Berzak et al. [3] introduced typology-driven estimation of grammatical error distributions in ESL learners, demonstrating how predictive models can be used to understand error tendencies. Similarly, Han et al. [7] and Lee and Seneff [10] explored automatic detection of grammatical errors in non-native English writing and speech, indicating the usefulness of machine learning– based classification techniques in identifying learner difficulties.

Recent works also focused on lexical, semantic and word- formation errors. Shawqi and Sultan [14] examined errors in word formation processes, while Rabinovich et al. [13] analyzed semantic infelicity detection in L2 English using automated techniques. These studies showed that linguistic challenges extend beyond grammar into deeper semantic and lexical levels, reinforcing the need for advanced analytical methods.

Although many earlier works relied on manual error analysis, recent educational analytics research has increasingly adopted clustering and classification techniques to group learners based on proficiency levels. Machine learning approaches such as K- Means clustering and classification models have been recognized as effective tools for identifying low-performing student groups and predicting learning difficulties. Studies in learner corpus analysis, including large-scale datasets such as the work reported in BMC Medical Education [15], demonstrated how pattern-based analysis can reveal high- frequency errors and performance gaps among learners.

Overall, these studies collectively support the application of clustering and classification algorithms for analysing English proficiency challenges. While traditional studies focused on grammatical and lexical error identification, recent data-driven approaches enable systematic grouping of learners and prediction of communication skill barriers. The insights gained from these works provide a strong foundation for applying machine learning techniques to analyse English language proficiency issues among Arts and Science college students.

IV. EXPERIMENTAL RESULTS

This section presents the key findings obtained from clustering and classification experiments performed on the English Wrong Answers dataset. The results highlight the patterns in student errors, the effectiveness of clustering and the predictive capability of the Naive Bayes classifier. The most relevant visual outputs are included to demonstrate the behaviour of the dataset and the performance of the applied algorithms.

A. Mistake Type Distribution

Figure 1 illustrates the overall distribution of mistake types made by students. Among all categories, *auxiliary/modal mistakes* were found to be the most frequent, followed by *other errors* and *wrong verbal forms*. *Spelling mistakes* and *tense mistakes* appeared comparatively less frequently. This uneven distribution is significant because it directly influences the accuracy and recall of the machine learning model.

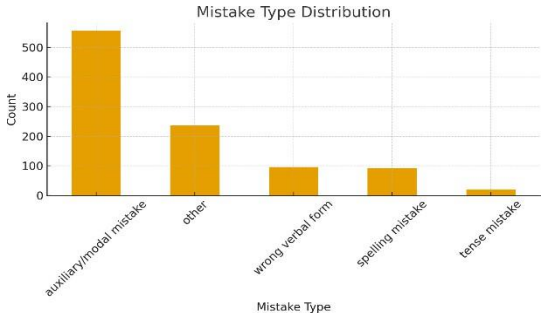

Fig. 1. Mistake Type Distribution

B. Clustering Results

K-means clustering was applied with **k = 5** (equal to the number of mistake types). The clusters formed successfully grouped similar error patterns, allowing meaningful interpretation of student weaknesses. Figure 2 displays the cluster distribution for the *auxiliary/modal* mistake category, which also represents the largest cluster. This demonstrates that clustering can effectively separate commonly occurring linguistic errors, making it useful for identifying dominant problem areas among learners.

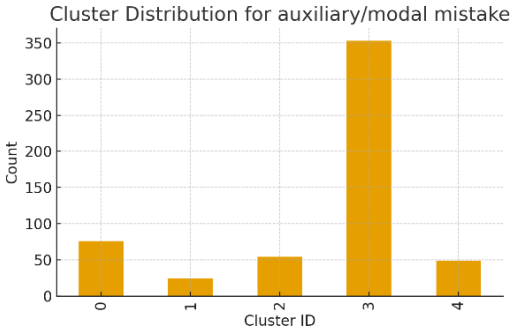

Fig. 2. Cluster Distribution for Auxiliary/Modal Mistakes

C. Classification Performance

The TF–IDF features combined with the Multinomial Naïve Bayes classifier achieved an overall accuracy of **70%**. To understand the model's performance across categories, a confusion matrix was generated.

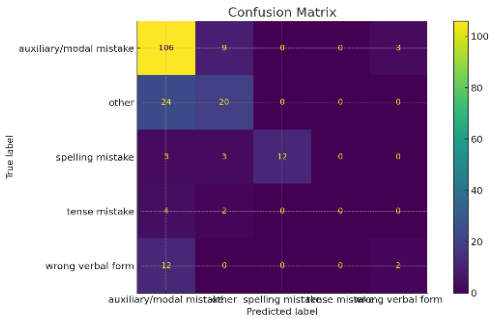

Fig. 3. Confusion Matrix for Mistake Classification

Figure 3 shows that the model performs strongly in predicting *auxiliary/modal* and *spelling mistakes*, while categories with fewer samples, such as *tense mistakes*, have lower recall. This reflects the influence of dataset imbalance on classifier performance.

D. Cluster Visualization Using PCA

To visualize how well the clusters were formed, PCA was used to project high-dimensional TF–IDF vectors into two dimensions.

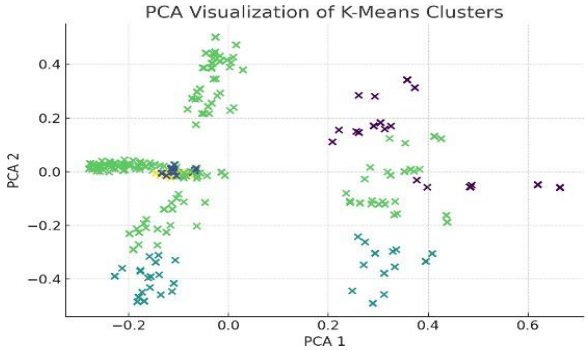

Fig. 4. PCA Visualization of K-Means Clusters

Figure 4 shows the PCA scatter plot where points are colored according to their K-means cluster assignments. Although some clusters partially overlap due to linguistic similarity, the plot clearly shows separated regions representing distinct error categories. This confirms that K-means successfully captured structural patterns in the dataset.

E. Summary of Findings

The experimental results indicate that:

- The dataset contains strong imbalance, with auxiliary/modal mistakes dominating the error distribution.
- K-means clustering effectively separated mistake categories into meaningful groups.
- The Naïve Bayes classifier achieved **70%** accuracy, with strong performance for frequent categories.
- PCA visualization confirmed the validity of the cluster structure.

These results demonstrate that machine learning can effectively analyze student English errors and support the design of targeted learning interventions.

V. CONCLUSION

This study analyzed the major factors contributing to low English proficiency among Arts and Science college students in the Virudhunagar district through machine learning techniques, including TF–IDF vectorization, K-Means clustering and the Multinomial Naive Bayes classifier. The experimental results revealed clear patterns in student errors, with auxiliary/modal mistakes being the most dominant and the classifier achieving a 70% accuracy rate. These findings demonstrate that inadequate grammar foundation, limited vocabulary development and insufficient exposure to spoken English significantly affect students' communication skills. By providing data-driven insights into linguistic weaknesses, this research highlights the usefulness of clustering and classification methods in supporting targeted instructional strategies and offers a foundation for developing improved English learning interventions and future educational applications.

REFERENCES

[1] Agustin, R., & Wulandari, S., "The Analysis of Grammatical Errors on Students' Essay Writing by Using Grammarly", *Jurnal Pendidikan Bahasa Inggris Proficiency*, 2022.

[2] Ayar, Zülal, "Error Analysis of Turkish Learners' English Paragraphs from Lexical and Grammatical Aspects", *ELT Research Journal*, 2020.

[3] Berzak, Y., Reichart, R., & Katz, B., "Contrastive Analysis with Predictive Power: Typology-Driven Estimation of Grammatical Error Distributions in ESL", *arXiv preprint*, 2016.

[4] Corder, S. P., "The Significance of Learner's Errors", *IRAL – International Review of Applied Linguistics in Language Teaching*, 1967.

[5] El Mahdy, F. M., "An Error Analysis of the Grammatical Errors of Egyptian EFL Learners and a Suggested Program for Enhancing Their Grammatical Competence", *Transcultural Journal of Humanities and Social Sciences*, 2023.

[6] Floranti, Astri Dwi, & Adiantika, Hanif Nurcholish, "Grammatical Error Performances in Indonesia EFL Learners' Writing", *Indonesian Journal of English Language Teaching and Applied Linguistics (IJELTAL)*, 2024/2025.

[7] Han, N. R., Chodorow, M., & Leacock, C., "Detecting Errors in English Article Usage by Non-Native Speakers", *Natural Language Engineering*, 2006.

[8] Huda, T., & Wuda, W., "Grammatical Errors Analysis on EFL Learners' Writing: A Case Study at Junior High Islamic Boarding School in Jember", *Journey: Journal of English Language and Pedagogy*, 2019.
[9] Kanwal, A., "Analysis of Errors in Written English EFL Learners: Evidence from Mixed Method", *Arab World English Journal (AWEJ)*, 2025.
[10] Lee, John, & Seneff, Stephanie, "An Analysis of Grammatical Errors in Non-Native Speech in English", *MIT CSAIL Technical Report / Conference Paper*, 2008.
[11] Le Thi Trung Dinh, "Grammatical Error Analysis of EFL Learners' English Writing Samples: The Case of Vietnamese Pre-Intermediate Students", *International Journal of TESOL & Education*, 2025.
[12] Noor Azizah, S. H., Baa, S., & Arham, M., "Analysis of EFL Students' Grammatical Errors in Writing Literature Review", *Tamaddun (Life & Language Journal)*, 2024.
[13] Rabinovich, Ella, Watson, Julia, Beekhuizen, Barend, & Stevenson, Suzanne, "Say Anything: Automatic Semantic Infelicity Detection in L2 English Indefinite Pronouns", *arXiv preprint*, 2019.
[14] Shawqi, Aisha Sonay Muhammed, & Sultan, Amra Ibrahim, "Error Analysis of EFL Students in Word Formation Process", *Journal of Tikrit University for Humanities*, 2024.
[15] "Analysis of High-Frequency Errors and Linguistic Patterns in EFL Medical Students' English Writing: Insights from a Learner Corpus", *BMC Medical Education*, 2024.