Informationization Design Of College English Corpus Using Deep Learning Algorithms

Cheng Lin* and Ruixue Li

School of Foreign Languages, Jiujing University, Jiujiang Jiangxi, 332005, China

Corresponding author. E-mail: forrestlincheng@163.com

Received: Feb. 21, 2026; Accepted: Mar. 27, 2026

This study proposes the ISSO-SNN model for semantic readability assessment in a college English corpus, addressing limitations of traditional methods. Using NLP preprocessing and TF-IDF features, the model leverages a Siamese network optimized with ISSO for improved prediction. The study proposes an Intelligent Shuffled Shepherd Optimized Siamese Neural Network (ISSO-SNN) model for semantic readability assessment in a college English corpus. Implemented in Python 3.11, it achieved high performance (accuracy 0.98, precision 0.97, recall 0.97 , F1 0.98), supporting effective and personalized English learning.

Keywords: Natural Language Processing (NLP); Intelligent Shuffled Shepherd optimized Siamese Neural Network (ISSO-SNN); deep learning (DL); English Corpus

http://dx.doi.org/10.6180/jase.202609_32.023

1. Introduction

Language acquisition has advanced with digital technology and deep learning (DL), enabling effective analysis of collegiate English corpora-large collections of academic texts [1–3]. DL and NLP techniques identify grammatical and semantic patterns, transforming static corpora into dynamic systems that support real-time analysis and personalized learning [4–6]. Informationization integrates AI to digitize, organize, and analyze language data, improving educational insights beyond traditional rule-based methods [7, 8].

The proposed ISSO-SNN framework enhances readability assessment by combining semantic similarity modeling with optimized convergence, overcoming limitations of conventional models in handling context and complex academic language. DL-based techniques such as sentiment analysis, speech recognition, and machine translation further improve language evaluation and curriculum design [9–12].

This research introduces the ISSO algorithm to optimize Siamese Neural Networks, enabling efficient corpus analysis, accurate readability prediction, and adaptive, data-driven language learning in higher education.

Advanced Siamese Neural Network Architecture: Enhances readability assessment by capturing semantic relationships between sentences.

Comprehensive NLP Preprocessing: Includes tokenization, stop-word removal, and normalization of corpus data.

Effective Feature Extraction Using TF-IDF Identifies important words representing sentence structure and language features.

Adaptability to Complex Linguistic Patterns: Effectively analyzes diverse and challenging language patterns beyond traditional readability methods.

Research design: Part 2 presents the relevant literature, Part 3 describes the methodology, Part 4 presents the results, Part 5 discusses the findings, and Part 6 concludes the study.

Materials and methods

Previous studies on English language learning and assessment show that deep learning (DL) and corpus-based methods enhance vocabulary teaching, speech assessment, and machineassisted translation. DL and visual recognition achieve over 90% accuracy in online vocabulary learning and support pronunciation assessment [13, 14]. Low-resource Chinese-English translation and clinical guideline evaluation for international students have also benefited from DL [15, 16]. CNNs improve vocabulary detection [17], while NLP-based chatbots address university queries [18]. Multilingual code-mixing, mobile DL systems, and multimodal DL models aid vocabulary, lexical, and semantic learning [19–22]. Corpus analysis using GBC and CAI models support grammar pattern identification and corpus-based instruction [23, 24].

Problem statement

Despite advances in deep learning and NLP, traditional readability tools still struggle to capture meaning and adapt to different contexts. Studies show that corpora help students learn vocabulary and language rules independently [13], and a CNN-based model has improved online vocabulary detection without splitting words [17]. These limitations highlight the need for improved methods, leading to the proposed ISSO-SNN model for better readability assessment in college English corpora.

Methodology

The Informationization Design of the College English Corpus follows a systematic approach. Data are collected from BBC articles and preprocessed using NLP techniques such as tokenization and stop-word removal. Important linguistic features are extracted using TF-IDF. The proposed ISSO-SNN model combines a Siamese Neural Network with Shuffled Shepherd

Optimization to efficiently evaluate sentence readability, improving accuracy and convergence, as shown in Figure 1.

Corpus: A vast and methodically arranged collection of natural language data, typically in digital format, is called a corpus. It is employed to examine linguistic structures, usage trends, and other noteworthy aspects. Corpora are essential resources for empirical research and the creation of computer models in the fields of linguistics, NLP and language instruction.

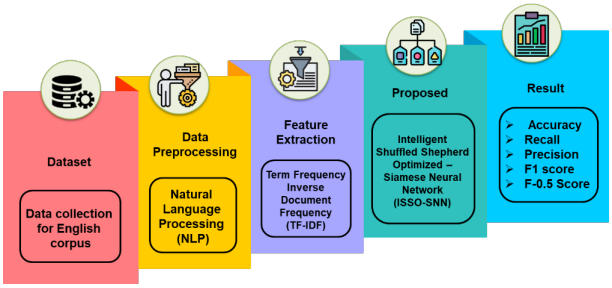

Fig. 1. Overall flow for the research

2.1 Data Collection

The dataset comprises BBC News articles collected from Kaggle via RSS feeds, including fields like title, pubDate, guide, link, and description. Articles are organized into sentence-level samples with readability labels assigned through standard scoring and validated via tokenization and stop-word removal. The data is split into training, validation, and testing sets for reliable evaluation of the ISSO-SNN model. RSS feeds were gathered using Requests_html and BeautifulSoup, providing well-structured, linguistically rich text suitable for college-level English readability analysis. BBC News text provides semi-formal, structured language similar to academic English while acknowledging its limitations in representing domain-specific discourse, yet still offering a diverse and practical foundation for readability modeling.[1]

2.2 Preprocessed using the Natural Language Processing (NLP)

Tokenize: Tokenization, essential for deep learning-based English corpora, converts text into manageable units for processing and analysis. Tools like NLTK, TextBlob, MBSP, and Pattern segment text into sentences using PunktSentence Tokenizer and into words with TreebankWordTokenizer, WordPunctTokenizer, PunctWordTokenizer, or Whitespace Tokenizer, ensuring effective DL model training for linguistic pattern analysis.

Stop word removal: Stop-word removal eliminates common, less meaningful words (e.g., "the," "in") to improve deep learning model performance in NLP tasks like text analysis, summarization, and question-answering. Tools like NLTK provide predefined English stopword lists for this process.

2.3 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF assigns weights to words based on their importance and serves as input to the ISSOSNN pipeline. Sentences are

[1] https://www.kaggle.com/datasets/gpreda/bbc-news

converted into TF-IDF vectors, processed by parallel subnetworks to generate semantic embeddings, while the ISSO algorithm optimizes network weights for improved readability prediction. TF captures a word's relevance within a document, combined with IDF to rank terms across the corpus. Quantifying how often a given word appears in the context of a document, higher values of TF mean important words in the text are defined by Equation (1).

$$TF_{s,c} = \frac{e_{s,c}}{\max\{e_{s',c} : s' \in c\}} \quad (1)$$

Where $e_{s,c}$ signifies that a phrase or term s occurs many times in document C. Conversely, the IDF determines a word or term's infrequency and importance across all texts. The definition of the IDF is in Equation (2). The high IDF value denotes a word that is rarely used in all documents.

$$IDF_{s,C} = \log \frac{C}{\{c \in C : s \in c\}} \quad (2)$$

Here the logarithmic scale IDF_s, are the number of documents that comprise the word or phrase s separated by the entire number of documents C. The TF-IDF weighting is defined in Equation (3). When a term or word appears frequently in a document but is included in a few papers, its weighting increases.

$$TF - IDF = TF_{s,c} \times IDF_{s,C} \quad (3)$$

2.4 The Intelligent Shuffled Shepherd optimized Siamese Neural Network (ISSO-SNN)

The ISSO-SNN uses Siamese networks to measure text similarity and improve readability analysis. ISSO optimization with OBL variants and adaptive step sizes enhances performance and avoids local minima. It also supports personalized learning and resource recommendations, as shown in Figure 2.

2.4.1 Siamese Neural Network (SNN)

Siamese Neural Networks convert sentences into embeddings for semantic comparison and perform better than traditional similarity methods. They use pair and triplet models with different loss functions for tasks like classification and plagiarism detection. The ISSO-SNN model improves performance using optimized parameters and training settings for stable results. There are two popular loss functions in Equation (4).

$$\text{Loss} = (Z)\left(-\log\left(Z_{\text{pred}}\right)\right) + (1 - Z)\left(-\log\left(1 - Z_{\text{pred}}\right)\right) \quad (4)$$

Where, Z is the label (1 if texts are similar, 0 otherwise), Z_{pred} is the predicted similarity of the contrastive loss in Equation (5).

Algorithm 1. ISSO-SNN

```
function InitializePopulation
    population ← ∅
    for each solution in population do
        OX ← OppositeSolution(solution)
        QRX ← QuasiReflectionSolution(solution)
        QOX ← QuasiOppositeSolution(solution)
        SUX ← SelectSolution(OX, QRX, QOX)
        Append SUX to population
    return population
function UpdateStepsize(population)
    for each solution in population do
        if random() < 0.8 then
            new_shepherd ← ComputeNewShepherd(solution)
            UpdateSolution(solution, new_shepherd)
function TrainSiameseNetwork(training_data)
    for each (anchor, positive, negative) do
        loss ← ComputeTripletLoss(anchor, positive, negative)
        UpdateNetworkWeights(loss)
function Main
    population ← InitializePopulation
    for iteration = 1 to max_iterations do
        UpdateStepsize(population)
        TrainSiameseNetwork(training_data)
        if iteration mod update_interval = 0 then
            RecommendResources
function OppositeSolution(solution)
    return solution_max + solution_min - solution
function QuasiReflectionSolution(solution)
    MID ← (solution_max+solution_min)/2
    return MID + (MID − solution) · random()
function QuasiOppositeSolution(solution)
    MID ← (solution_max+solution_min)/2
    return MID + (MID− OppositeSolution(solution)) · random()
function SelectSolution(OX, QRX, QOX)
    return OX if condition else QOX
function ComputeTripletLoss(anchor, positive, negative)
    return max(d(a, p) − d(a, n) + α, 0)
function EuclideanDistance(a, b)
    return √(∑(a_i − b_i)^2)
```

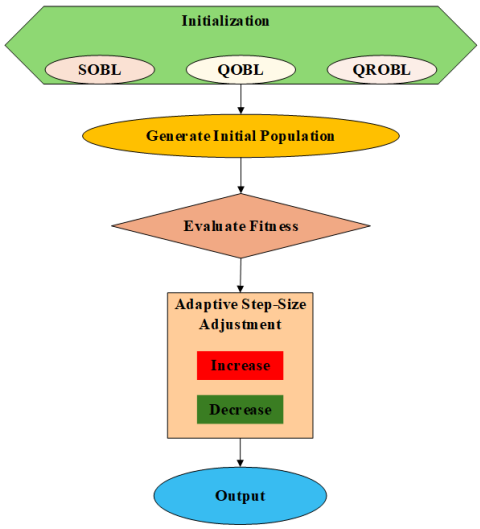

Fig. 2. ISSO Framework with OBL Variants and Adaptive Step-Size Mechanism

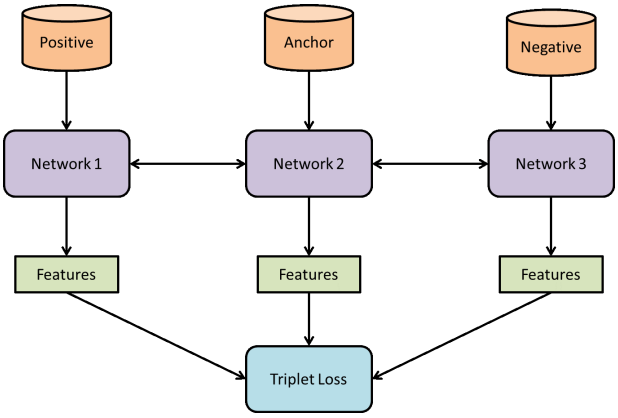

Fig. 3. Siamese neural network structure with a triplet of components as the input

$$\text{Loss} = Z * C^2 + (1 - Z) * \max(\alpha - C, 0)^2 \tag{5}$$

In triplet-based SNNs, Z is the label, C is the Euclidean distance between outputs, and α is a margin separating nearby and distant samples. Each triplet compares an anchor text, a similar (positive) text, and a dissimilar (negative) text. These networks maximize similarity evaluations inside the corpus by utilizing Triplet Loss in Equation (6).

$$K(B, OM) = \max\left(\|f(B) - f(O)\|^2 - \|f(B) - f(M)\|^2 + \alpha, 0\right) \tag{6}$$

Where B is the anchor text, O is the positive text, M is the negative text, α is the margin parameter, and f ()represents the text embeddings.

Recommender systems (RSs) filter large linguistic corpora to deliver personalized learning, using content-based analysis of keywords, grammar, and semantics to match resources with learner profiles. Key components include learner profiling, content analysis, and preferencebased filtering show in Figure 3.

Collaborative filtering provides personalized recommendations by analyzing student interactions, using model-based prediction or similarity-based matching to suggest relevant learning materials.

2.4.2 Intelligent Shuffled Shepherd Optimization (ISSO)

The ISSO algorithm enhances deep learning-based collegiate English corpora using OppositeBased Learning (OBL) and its variants-SOBL, QOBL, and QROBL-to optimize model initialization. SOBL strengthens global search, QOBL balances exploration and efficiency, and QROBL improves convergence to avoid local minima, improving corpus quality and linguistic analysis. The following is the definition of the opposite approach in Equation (7):

$$OX_r = w_{r,\max} + w_{r,\min} - w_r R = 1, 2, \ldots, m \,\text{Var} \tag{7}$$

As a result, OX is the opposite of the approach below consideration, the other variation of OBL is created by using the opposite of the measured approach.

The Quasi-Reflection resultin QRX of the measured approach wis referred to as a random strategy, which is created among the r then central point $MID_r = (w_{r,\max} + w_{r,\min})/2.QRX$ for the r_{th} design factor is as follows in Equation (8).

$$QRX_r = MID_r + (MID_r - w_r) \times rand \tag{8}$$

Equation (9) is used to determine the quasi-opposite approachQOX of the measured approach r. It is defined as a random approach that is formed among OX and the central point.

$$QOX_r = MID_r + (MID_r - OX_r) \times \text{ rand} \tag{9}$$

The remedy that is completely contrary to Equation (10) is used to define the SUX of the measured approach r.

$$\text{SUX}_r = \begin{cases} OX_r + (w_{r,\max} - OX_r) \times \text{rand}\, OX_r > \text{MID}_r \\ w_{r,\min} + (OX_r - w_{r,\min}) \times \text{ randotherwise} \end{cases} \tag{10}$$

Equations (7-10) generate $5 \times$ mT members for the population, calculating four solutions for each random outcome. The ISSO algorithm selects mT associates from the sorted population and uses parameters like population size, iteration limits, and adaptive step-size, converging when

solutions stabilize or reach maximum iterations. A limitation is its tendency to get stuck in local minima due to low population diversity.

This is addressed by adjusting the step-size: if a randomly generated value is below 0.8 , a new step-size is computed, improving exploration and optimization (see Equation 11).

$$w_{j,i,r}^{\text{newshepherd}} = V\Big(\text{Mean}_{i,r} - \text{Std}_{i,r} - \sigma_{j,r}, \text{Mean}_{i,r} + \text{Std}_{i,r} + \sigma_{i,r}\Big) \quad (11)$$

Where V is the function that yields a random integer derived from an ongoing random delivery with lesser and higher bounds given by $\text{Mean}_{i,r} - \text{Std}_{i,r} - \text{sigma}_{j,r}$ and $\text{Mean}_{i,r} + \text{Std}_{i,r} + \text{sigma}_{i,r}$ Mean $_{i,r}$, and Std $_{i,r}$ is the standard deviation and mean of the r_{th} variable in the i_{th} herd; $\text{sigma}_{i,r}$ is a variable that, when the entire population joins to the designated value in Equation (12), helps the statistically regenerated stepsize to function effectively.

$$\text{sigma}_{i,r} = \begin{cases} 0.01 \times (w_{r,\max} - w_{r,\min}) \\ \quad \text{if std}_{i,r} < 0.01 \times (w_{r,\max} - w_{r,\min}) \\ 0 \quad \text{otherwise} \end{cases} \quad (12)$$

The framework combines Siamese Neural Networks with ISSO optimization, enhancing accuracy and reliability in readability assessment and recommendations. By avoiding local minima and speeding convergence, ISSO-SNN offers an efficient solution for analyzing large English corpora.

3.Results and discussion

The ISSO-SNN model, implemented in Python 3.11 on a Windows 10 system, predicts readability scores in a college English corpus. The scatter plot of actual vs. predicted scores shows low variance and few outliers, indicating consistent performance with minor errors, demonstrating its effectiveness for language resource recommendations. show in Figure 4

The ISSO-SNN model outperforms BERT [25], H-ERDC [25], and ANN [26] in readability assessment, achieving faster convergence, lower memory usage, and higher accuracy and F scores. It combines TF-IDF, ISSO optimization, and Siamese networks for efficient corpus processing and improved recommendations, with robustness validated through cross-validation and ablation studies as Table 1 and Figure 5.

Precision: Precision measures the accuracy of retrieved relevant content. ISSO-SNN achieved 0.97, outperforming BERT (0.90) and ANN (0.95), demonstrating strong recommendation capability.

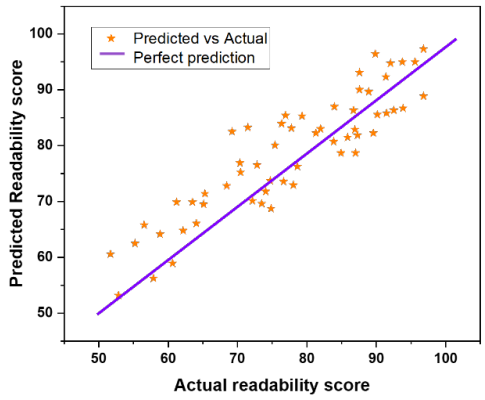

Fig. 4. Scatter Plot Visualization of Readability Score Predictions in ISSO-SNN Model.

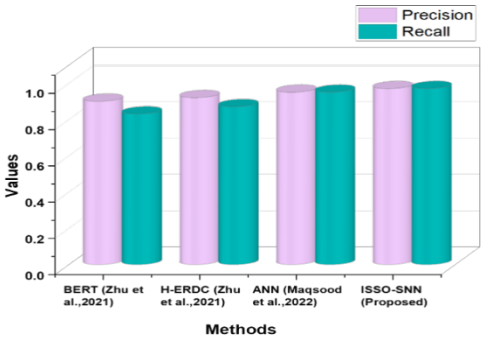

Fig. 5. Comparison of Recall and Precision Using Various Methods

Recall: Recall measures the ability to capture relevant patterns. ISSO-SNN achieved 0.97, outperforming BERT (0.83) and H-ERDC (0.87), ensuring more comprehensive corpus retrieval for accurate recommendations.

Table 1. Comparison of DL Techniques for College English Corpus Analysis in terms of Precision and F1 Score

Method	Precision	Recall
BERT [25]	0.90	0.83
H-ERDC [25]	0.92	0.87
ANN [26]	0.95	0.95
ISSO-SNN (Proposed)	0.97	0.97

Accuracy: Accuracy measures overall prediction correctness. ISSO-SNN achieved 0.98 accuracy, outperforming ANN (0.96), improving semantic pattern detection and personalized recommendations.

F1 score: The F1 score balances precision and recall for unbalanced datasets. ISSO-SNN achieved a higher F1

score of 0.98 than ANN, indicating superior accuracy and reliability in corpus analysis and content recommendation in Table 2 and Figure 6.

Table 2. Comparison of Deep Learning Techniques for College English Corpus Analysis in Terms of Accuracy and F1 Score

Method	Accuracy	F1 score
ANN [26]	0.96	0.95
ISSO-SNN (Proposed)	0.98	0.98

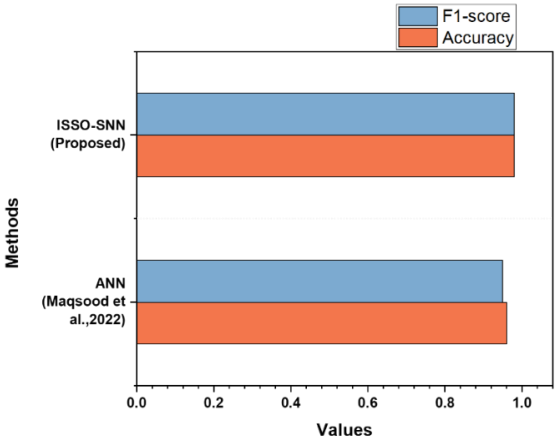

Fig. 6. Comparison of F1 score and Accuracy Using Various Methods

F-0.5 score: The F0.5 score, which emphasizes precision over recall, highlights ISSO-SNN's improved accuracy in identifying relevant content. It outperformed BERT (0.86) and H-ERDC (0.89) with a score of 0.97 , demonstrating higher precision and effectiveness, as shown in Table 3 and Figure 7.

Table 3. Comparison of DL Techniques for College English Corpus Analysis in Terms of F0.5 score

Method	F-0.5 score
BERT [25]	0.86
H-ERDC [25]	0.89
ISSO-SNN (Proposed)	0.97

Discussion

This research develops an ISSO-SNN-based system for accurate sentence readability assessment in higher-education English corpora, supporting real-time adaptive learning and personalized recommendations. It overcomes limitations of BERT [25], H-ERDC [27], and ANN by improving

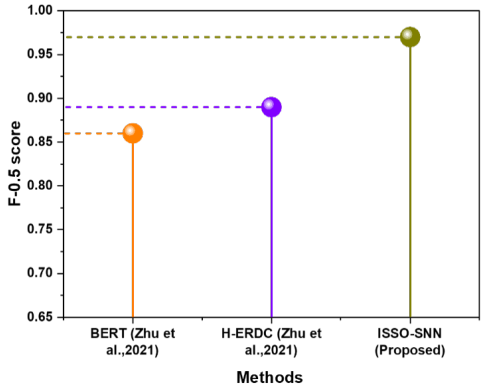

Fig. 7. Comparison of F-0.5 score Using Various Methods

semantic learning, reducing computational cost, and enabling faster, more accurate analysis of complex academic texts, as supported by prior studies [21, 22].

2. 4.Conclusion

The ISSO-SNN model uses TF-IDF, NLP preprocessing, and a Siamese network optimized with ISSO to assess sentence readability in a college English corpus. It achieves high performance (accuracy 0.98 , precision 0.97 , recall 0.97 , F1 0.98), improving semantic analysis and adaptive recommendations, though evaluation on the BBC dataset limits multilingual generalization.

Limitation and Future Scope: A limitation is the model's reliance on a specific English corpus, limiting coverage of diverse language patterns. Future work can expand datasets with complex, multilingual, and varied texts, and incorporate real-time adaptation with user input.

Acknowledgments

Declarations

Funding:

This work is supported by the Humanities and Social Sciences Research Project of Jiangxi Province (YY21107), titled "Investigation of register characteristics of Chinese English learners and their academic writing ability."

Conflicts of interests:

Authors do not have any conflicts.

Data availability statement:

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Clinical trial number:

Not applicable.

Ethics:

Not applicable.

Consent to participate:

Not applicable.

Consent to publish declarations:

Not applicable.

Code availability:

Not applicable.

Authors' contributions:

Cheng Lin, Ruixue Li is responsible for designing the framework, analyzing the performance, validating the results, and writing the article.

References

[1] A. Lusta, Ö. Demirel, and B. Mohammadzadeh, (2023) *"Language corpus and data driven learning (DDL) in language classrooms: A systematic review"* **Heliyon** ***9***(12): e22731. DOI: 10.1016/j.heliyon.2023.e22731.

[2] J. Panwar. *"Techniques for text classification and text generation: Enhanced online sexism detection and template driven Wikipedia article generation"*. (phdthesis). International Institute of Information Technology Hyderabad, India, 2024.

[3] I. H. Sarker, (2021) *"Deep learning: A comprehensive overview of techniques, taxonomy, applications, and research directions"* **SN Computer Science** ***2***(6): 420. DOI: 10.1007/s42979-021-00815-1.

[4] T. R. Lee and S. C. Mouritsen, (2021) *"The corpus and the critics"* **The University of Chicago Law Review** ***88***(2): 275–366.

[5] J. Kozal, M. Leś, P. Zyblewski, P. Ksieniewicz, and M. Woźniak, (2022) *"Lifelong learning natural language processing approach for multilingual data classification"* **arXiv preprint *arXiv:2206.11867***: DOI: 10 . 48550 / arXiv.2206.11867.

[6] J. Zhu, C. Zhu, and S. B. Tsai, (2021) *"Construction and analysis of intelligent English teaching model assisted by personalized virtual corpus by big data analysis"* **Mathematical Problems in Engineering *2021***: DOI: 10.1155/2021/5374832.

[7] S. Vernim, M. Krauel, and G. Reinhart, (2021) *"Identification of digitization trends and use cases in assembly"* **Procedia CIRP *97***: 136–141. DOI: 10.1016/j.procir.2020.05.215.

[8] N. L. Rane, O. Kaya, and J. Rane. "Artificial intelligence, machine learning, and deep learning applications in smart and sustainable industry transformation". In: *Artificial Intelligence, Machine Learning, and Deep Learning for Sustainable Industry*. ***5***. Deep Science Publishing, 2024, 2–29. DOI: 10.70593/978-81-981271-8-1_2.

[9] H. M. Nasir, N. M. A. Brahin, F. E. M. Sani, M. S. Mispan, and N. H. Abd Wahab, (2023) *"AI educational mobile app using deep learning approach"* **JOIV: International Journal on Informatics Visualization** ***7***(3): 952–958. DOI: 10.30630/joiv.7.3.1247.

[10] S. Chauhan, R. Kumar, S. Saxena, A. Kaur, and P. Daniel, (2024) *"Semsyn: Semantic-syntactic similarity-based automatic machine translation evaluation metric"* **IETE Journal of Research** ***70***(4): 3823–3834. DOI: 10.1080/03772063.2023.2195819.

[11] C. Zhai and S. Wibowo, (2023) *"A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university"* **Computers and Education: Artificial Intelligence** ***4***: 100134. DOI: 10.1016/j.caeai.2023.100134.

[12] R. M. Horst. *"Higher education executives and data-driven decision making: A phenomenological study"*. (phdthesis). Concordia University, Oregon, 2020.

[13] J. Cui, (2020) *"Application of deep learning and target visual detection in English vocabulary online teaching"* **Journal of Intelligent & Fuzzy Systems** ***39***(4): 5535–5545. DOI: 10.3233/JIFS-189035.

[14] F. Jiao, J. Song, X. Zhao, P. Zhao, and R. Wang, (2021) *"A spoken English teaching system based on speech recognition and machine learning"* **International Journal of Emerging Technologies in Learning (iJET)** ***16***(14): 68–82. DOI: 10.3991/ijet.v16i14.24049.

[15] A. D. Yacoub, S. Slim, and A. Aboutabl, (2024) *"A survey of sentiment analysis and sarcasm detection: Challenges, techniques, and trends"* **International journal of electrical and computer engineering systems** ***15***(1): 69–78. DOI: 10.32985/ijeces.15.1.7.

[16] L. Lin, J. Liu, X. Zhang, and X. Liang, (2021) *"Automatic translation of spoken English based on an improved machine learning algorithm"* **Journal of Intelligent & Fuzzy Systems *40***(2): 2385–2395. DOI: 10.3233/JIFS-189234.

[17] M. Ji, Y. Liu, M. Zhao, Z. Lyu, B. Zhang, X. Luo, Y. Li, and Y. Zhong, (2021) *"Use of machine learning algorithms to predict the understandability of health education materials: Development and evaluation study"* **JMIR Medical Informatics *9***(5): e28413. DOI: 10.2196/28413.

[18] J. Wu and B. Chen, (2020) *"English vocabulary online teaching based on machine learning recognition and target visual detection"* **Journal of Intelligent & Fuzzy Systems *39***(2): 1745–1756. DOI: 10.3233/JIFS-179948.

[19] R. Guha, (2021) *"Designing a chat-bot for college information using information retrieval and automatic text summarization techniques"* **Current Chinese Computer Science *1***(1): 42–51. DOI: 10.2174/2665997201999201022191540.

[20] G. I. Ahmad, J. Singla, A. Anis, A. A. Reshi, and A. A. Salameh, (2022) *"Machine learning techniques for sentiment analysis of code-mixed and switched Indian social media text corpus: A comprehensive review"* **International Journal of Advanced Computer Science and Applications *13***(2): DOI: 10.14569/IJACSA.2022.0130254.

[21] N. Shen. "A deep learning approach of English vocabulary for mobile platform". In: *2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*. IEEE, 2021, 463–466. DOI: 10.1109/ICMTMA52658.2021.00106.

[22] L. Diao and P. Hu, (2021) *"Deep learning and multimodal target recognition of complex and ambiguous words in an automated English learning system"* **Journal of Intelligent & Fuzzy Systems *40***(4): 7147–7158. DOI: 10.3233/JIFS-189543.

[23] D. Wang, J. Su, and H. Yu, (2020) *"Feature extraction and analysis of natural language processing for deep learning English language"* **IEEE Access *8***: 46335–46345. DOI: 10.1109/ACCESS.2020.2974101.

[24] M. Alhamami, (2022) *"Google Books corpus and designing English for specific purposes materials"* **Journal on English as a Foreign Language *12***(2): 421–457. DOI: 10.23971/jefl.v12i2.4254.

[25] S. Maqsood, A. Shahid, M. T. Afzal, M. Roman, Z. Khan, Z. Nawaz, and M. H. Aziz, (2022) *"Assessing English language sentences readability using machine learning models"* **PeerJ Computer Science *8***: e818. DOI: 10.7717/peerj-cs.818.

[26] Y. Jiang, (2023) *"The application of corpus linguistics under the computer-assisted instruction model in college English teaching"* **Revista Ibérica de Sistemas e Tecnologias de Informação** (E55): 475–484.

[27] J. Zhu, X. Shi, and S. Zhang, (2021) *"Machine learning-based grammar error detection method in English composition"* **Scientific Programming *2021***(1): 4213791. DOI: 10.1155/2021/4213791.