[1]Fang Yu

Research and Application of Key Vocabulary Extraction Algorithm in English Vocabulary Learning

Journal of Electrical Systems

Abstract: - An effective key vocabulary extraction is significant for English Language Learning, assisting the learners to achieve proficiency in English. However, the process of vocabulary extraction is often challenging and multifaceted, leading to inaccurate vocabulary acquisition, which reduces the learner's proficiency. To address these issues, we presented a novel key vocabulary extraction using the combination of Radial Basis Function Neural Network (RBFNN) and Emperor Penguin Optimizer (EPO). The primary objective of this research is to identify the most relevant vocabulary from the huge English language corpus to assist learners in improving their proficiency levels and learning process. The proposed work commences with the collection of English corpus and the collected database undergoes pre-processing steps like tokenization, stop word removal, and stemming to improve the efficiency of subsequent analysis. Then, the developed EPO-RBFNN was designed to extract the most informative and relevant features from the pre-processed database. The RBFNN module was trained using the pre-processed database to learn and capture the semantic patterns and interconnections within the corpus, while the EPO was employed for selecting the vocabulary sequence considering their relevance and importance in English learning. The proposed framework was implemented in the Python tool, and the results are evaluated in terms of accuracy, precision, recall, and f-measure. Furthermore, a comparative assessment was made with existing vocabulary extraction methodologies to validate the outcomes. The comparative analysis showcases that the proposed strategy outperformed the conventional models, making it a suitable solution for key vocabulary extraction.

Keywords: Radial Basis Functional Neural Network, Heap-based Optimizer, English Language Learning, Key Vocabulary Extraction

1. INTRODUCTION

In recent years, the data management of education resources has transitioned progressively towards automated systems. The evolution of web-based educational resources significantly enhanced the data management level and service capabilities [1]. Vocabulary is one of the fundamental resources of the English language, hence teaching vocabulary is one of the key ways of teaching the English language [2]. However, the increasing dependence on automated data management in the educational field demands an advanced technique for extracting and capturing key vocabulary effectively for enhancing English learning. Vocabulary extraction is the process of identifying the most important vocabulary, which enhances the process of English learning [3]. The conventional process of vocabulary extraction depends on the categorization of words done by teachers or trainers, which is time-consuming and personalized. This creates a demand for automatic vocabulary extraction approaches for effectively identifying key vocabularies from the large corpus data.

The automatic algorithms use advanced techniques like natural language processing (NLP), artificial intelligence, data analytics, etc., to process the large corpus of data and extract words, that are more relevant and significant for English language learning [4]. These techniques analyze the textual data available online and extract the key vocabulary. Also, it extracts and identifies domain-specific terms relevant to different educational disciplines [5]. By automating this identification process aids in allocating time and resources more effectively in online English learning. In addition, it offers more personalized learning experiences and improves the English proficiency level of individual learners [6]. Many researches are conducted using these advanced techniques for extracting keywords and phrases to improve the proficiency level of English learners.

In [9], a study developed a supervised algorithm for automatic keyword extraction from a single document. In this algorithm, the text was modeled as a complex network, and the feature sequence was constructed by capturing select node properties from it. Further, the node properties are examined using an unsupervised algorithm, and a graph-assisted keyword extraction technique was deployed for differentiating keywords and

[1] School of Software, Dalian University of Foreign Languages, Dalian 116000, China

*Corresponding author's e-mail: fangyucencen@163.com

non-keywords. In addition, the training sequence was constructed from the feature sequence by allocating a label to each candidate keyword based on whether the candidate is listed as a gold-standard keyword or not. This system utilizes two public databases from the scientific domain for training and is validated using three unknown scientific corpora. However, this system is dependent on domain and language. In [10], the author developed a vocabulary extraction algorithm for improving English vocabulary online teaching using the application of deep learning and a target visual identification algorithm. The deep learning approach uses non-linear mapping functions to determine high-level abstract representation of textual data. By capturing attributes from data, the identifiable elements in the textual information are tracked and extracted. The implementation outcomes suggest that it offers 90% accuracy in extracting key vocabulary, and also assists the student in improving the English vocabulary knowledge. However, the acceleration of the training process is still a concern.

In [12], an innovative feature extraction algorithm using deep learning to capture the relevance of words in the English corpus. This study develops a multi-modal neural system for analyzing the features in the corpus. The primary concern of this research is to resolve the challenges in vocabulary segmentation by analyzing the semantic interconnections within the document. This study uses BI-GRU (Bidirectional Gated Recurrent Unit) for English word segmentation and uses the CRF (Conditional Random Field) model to annotate sentences in sequence. The implementation results suggest that it improved the document processing speed by 1.94 times compared to the existing techniques and enhanced the effectiveness of word segmentation processing. Although various studies are developed for extracting the keywords or key vocabularies from the large document corpus, they face challenges like domain dependency, high computational complexity, large training time, inaccurate extraction, etc. To resolve these issues, we developed an innovative hybrid key vocabulary extraction algorithm using a combination of deep learning and meta-heuristic optimization algorithms. The utilization of deep learning captures the semantic relations within the textual data in the corpus, while the optimization algorithm enables to selection of the most relevant and optimal vocabulary sequence for enhancing English vocabulary learning.

The enduring sections of the article are organized as follows: section 2 presents the system model and its problem statement, section 3 illustrates the developed strategy, section 4 depicts the results of the developed work, and section 5 provides the conclusion of the research.

2. SYSTEM MODEL AND PROBLEM STATEMENT

English vocabulary learning is the basic aspect of improving language proficiency. Key vocabulary extraction involves identifying the most significant and relevant vocabulary from a large textual corpus, assisting the learners to focus on the most important words for learning. Typically, the trainers or teachers identify the keywords or vocabulary. This type of vocabulary selection is more personalized, and may adversely influence the knowledge of the learners. Hence, the automation of this process is essential for identifying and prioritizing key vocabularies for learners. The general system model comprises an English corpus, pre-processing module, key vocabulary extraction module, and user interface. Initially, the collected English corpus data was pre-processed using the NLP steps to improve the efficiency of further analysis. The vocabulary extraction algorithm module analyzes the contextual relevance, semantic relations, and correlations to identify words significant for understanding the document. Finally, these vocabularies are ranked based on their relevance to the text. Currently, advanced techniques such as artificial intelligence, deep learning, and machine learning algorithms are employed for automating the vocabulary extraction process. However, handling large volumes of English corpus is difficult and complex, making them ineffective. Moreover, training these models is complex and they need fine-tuning. Also, the selection of an optimal sequence of vocabularies is still a concern because of the diverse nature of language and the varying needs of individual learners. To resolve these challenges, we proposed an innovative algorithm to automate the extraction of key vocabulary. The proposed strategy combines the efficiency of deep learning and a meta-heuristic optimization algorithm for optimally selecting the key vocabulary sequence from the large English document corpus. Through this research, we aim to assist English vocabulary learners by identifying and prioritizing the most relevant and significant words to enhance their learning process.

3. PROPOSED EPO-RBFNN FOR KEY VOCABULARY EXTRACTION

This study proposed an innovative algorithm for extracting key vocabulary from the large English corpus database to enhance the proficiency level of English. The developed algorithm combines the Radial Basis Functional Neural Network (RBFNN) and Emperor Penguin Optimizer (EPO) for extracting the key vocabulary. The proposed methodology follows three steps: data collection, pre-processing, and vocabulary extraction. In the data collection phase, we accumulate the English corpus dataset and train the system. In the pre-processing phase, the collected dataset undergoes steps like tokenization, stop word removal, and stemming to enhance the quality of the dataset for further analysis. In the vocabulary extraction phase, we integrate the proposed EPO-RBFNN for extracting and capturing the most relevant and important vocabulary. The RBFNN was trained using the dataset to understand the correlation and significance of the vocabularies within the corpus, while the EPO was employed for selecting the most relevant vocabularies by analyzing the learned RBFNN outcomes. Figure 1 presents the architecture of the developed methodology.

3.1 Data collection

The proposed study commences with the collection of the English language corpus database. This database includes textual information written in English, which can be collected from diverse sources through assessment or from literature. The presented study utilized 100 English literature randomly selected from the British Academic Written English (BAWE) Corpus [13]. This corpus database contains 672 keywords, and the content of the literature is presented in Table 1.

Table 1: Content of the input English database

Subject	Number of documents
Military	6
Electrical engineering	9
Chemical engineering	7
Biology	12
Power engineering	10
Chemical engineering	7
Natural science	15
Mechanical engineering	11
Water conservancy project	8
Mineral Engineering	10
Architecture science	12

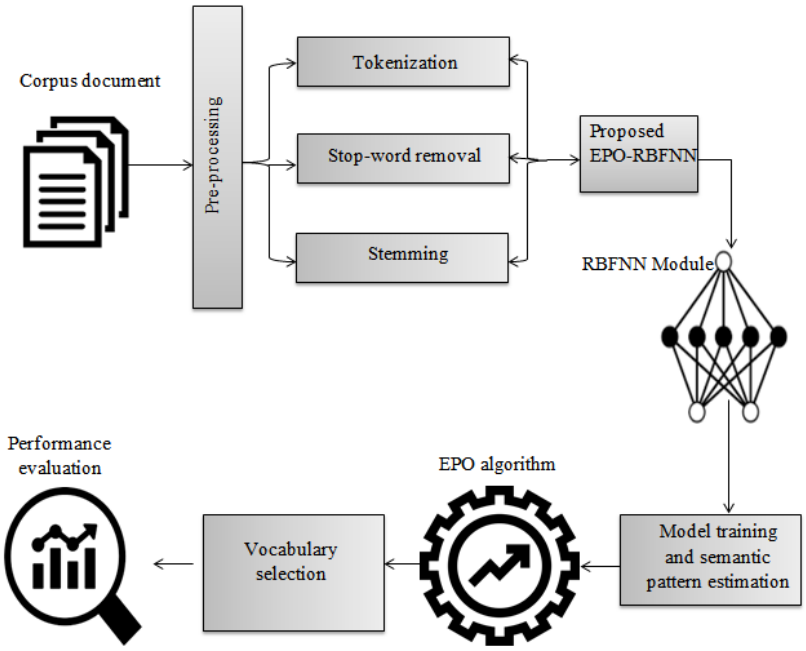

Figure 1: Architecture of the proposed framework

3.2 Pre-processing

In NLP, pre-processing includes three major steps namely, tokenization, stop-word removal, and stemming, which are significant for improving the quality of the raw text data.

Tokenization:

Tokenization indicates the process of splitting the text into a sequence of meaningful pieces, which are commonly referred to as tokens. This step enables to standardization and structure of the textual corpus data and makes them effective and reliable for further analysis. Depending upon the task, we can divide the textual data into words or subwords, which assists in analyzing the intricate patterns and semantic relations within the corpus.

Stop-word removal:

Stop-word removal is one of the common steps in NLP in which the stop words in the English language including "a," "the," "are," "is," etc., are removed from the tokenized dataset. This step helps to identify the words, which are not relevant and significant for English vocabulary learning, making the extraction algorithm more effective. Moreover, the removal of the stop-words makes the database more concise; thereby reducing the computational time and improving the processing speed of the system.

Stemming:

In NLP, stemming defines the process of minimizing a word to its root word (stem) by eliminating the suffixes or prefixes from the words. This process enables the transformation of the word to its common form, minimizing the dimensionality of the database and enhancing the accuracy of data analysis for key vocabulary extraction.

The utilization of these steps in data pre-processing standardizes the textual corpus data, and makes them more effective and reliable for further analysis.

3.3 Radial Basis Function Neural Network

RBFNN is an artificial neural network commonly deployed for approximation problems. The unique feature of this network is its universal approximation and quick learning ability. This is a simple feed-forward neural

network with three important layers. Each layer performs a certain function depending upon the task. In the proposed work, we utilized RBFNN for understanding the semantic interconnections within the pre-processed corpus database for extracting the most significant vocabularies. The first layer is the input layer, which accepts the input of the network. The second layer is the hidden layer, which contains numerous RBF non-linear activation units for performing the above-mentioned task. The final layer is the output layer, which produces the final results of the network. The non-linear activation functions in RBFNN are typically executed as Gaussian functions. The structure of the RBFNN is presented in Figure 2.

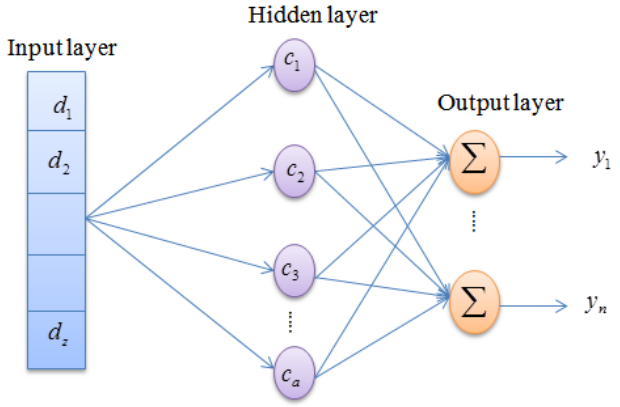

Figure 2: Structure of RBFNN

In the developed work, the input layer accepts the pre-processed corpus database as input. The neuron in the input layer transforms the input data into a suitable format for further analysis. The hidden layer consists of several neurons, which apply a non-linear transformation to the input data using an RBF (Gaussian function). This layer analyzes the input database and converts it into a high-dimensional data sequence in which the semantic interconnections within the input corpus are more apparent. Each neuron in the hidden layer evaluates its output considering the similarity between the input data sequence and its associated center (Gaussian functions). The outcome of the a^{th} activation function in the hidden layer is determined based on the distance between the input pattern and the center and is expressed in Eqn. (1).

$$G_a\left(\|d - c_a\|\right) = \exp\left(-\frac{\|d - c_a\|^2}{2w_b^2}\right) \quad (1)$$

Where G_a represents the Gaussian activation function, $\|d - c_a\|$ indicates the Euclidean distance between the input data d and the center c_a, and w_b denotes the width of the hidden neuron. This output provides the correlations and semantic interrelationships within the corpus data, and it is fed into the output layer of the network. Each neuron in this layer combines the outputs of the neurons in the hidden layer to produce the final output, which indicates the feature vector representing the significance of each vocabulary within the corpus. The final output of the network is represented in Eqn. (2).

$$y_n = \sum_{b=1}^{m} w_{gbn} G_b(d) \quad (2)$$

Where y_n defines the output of the nth neuron in the output layer, w_{gbn} denotes the weight connecting the b^{th} neuron in the hidden layer to the nth neuron in the output layer, and m represents the number of neurons. At each iteration, RBFNN is trained to produce the feature vector by iteratively adjusting its parameters such as the width of the neuron, weight, center, etc. The training of the system is terminated when the calculation error is

less than 0.01 or when it reaches the maximum iterations. The calculation error is determined in terms of Mean Square Error, and it is minimized by using the backpropagation algorithm.

3.4 Emperor Penguin Optimizer

The EPO is a nature-inspired algorithm developed based on the huddling characteristics of emperor penguins (EPs) for solving optimization problems. Generally, the EPs travel in rafts/colonies for foraging, and this unique feature found in these animals during foraging is defined as huddling characteristics. The major objective of this algorithm is to determine an effective mover from the swarm. In the presented study, we applied EPO to select the key vocabulary from the English corpus for improving English vocabulary learning. Here the EPs define the sequence of vocabularies, and the objective is to select the vocabulary with greater correlation score. A correlation score indicates the significance of the word to improve English vocabulary learning. If the correlation score is high for a word, then the word will be selected, and vice versa. The formula for calculating the correlation score is expressed in Eqn. (3).

$$C_s = \frac{k\left(\sum uv\right) - \left(\sum u\right)\left(\sum v\right)}{\sqrt{\left[k\sum u^2 - \left(\sum u\right)^2\right]\left[k\sum v^2 - \left(\sum v\right)^2\right]}} \quad (3)$$

Where u indicates the frequency of the word, k represents the number of words, and v denotes the effectiveness of the word in improving English vocabulary learning. The optimal selection of vocabulary commences with the random initialization of the feature vector, which contains relative vocabularies. Each parameter v in the vector indicates a vocabulary, and it is initialized using Eqn. (4).

$$K_v(r) = \{v_1, v_2, v_3, ..., v_r\} \quad (4)$$

Where r defines the population size. After initialization, the fitness value of each vocabulary is determined based on the pre-defined objective function (correlation score). For each vocabulary in the feature vector, the fitness value was determined. The fitness evaluation is represented in Eqn. (5).

$$f(v_i) = C_s(v_i) \quad (5)$$

Then, determine the vocabulary with the highest correlation based on the fitness value. In the EPO algorithm, before position updation, the EPs normally huddle together to preserve temperature to protect them from collisions. In the developed work, each vocabulary term tends to huddle together to preserve its significance in enhancing English vocabulary learning. For this reason, it uses two vectors $\vec{X}$ $\vec{H}$ which are expressed in Eqn. (6), and (7).

$$\vec{X} = \left\{m_p \times \left(s_p + f(v)\right) \times Rand(\ )\right\} - s_p \quad (6)$$

$$\vec{H} = Rand(\ ) \quad (7)$$

Where m_p indicates the movement parameter, and s_p denotes the significance profile. Then, the significance of the vocabulary was updated to determine the optimal sequence of vocabulary for English vocabulary learning. This updation is influenced by the movement direction and distance from the optimal solution, and it is mathematically represented in Eqn. (8).

$$v(t+1) = v(t) - m_d \,.\, d_{os} \quad (8)$$

Where m_d denotes the movement direction, and d_{os} defines the distance from the optimal solution. After position updation, the fitness value of each vocabulary was upgraded. Finally, the vocabulary terms with a fitness value greater than 0.5 are selected for English vocabulary learning. This optimization process is an

iterative mechanism and it continues until maximum iteration count or convergence is reached. The step-by-step procedure of the developed key vocabulary extraction algorithm is illustrated in Algorithm 1.

Algorithm 1: EPO-RBFNN
Start {
Input: English corpus;
Output: Extracted vocabularies
Initialize the input corpus database;
Pre-process the database;
Design RBFNN {
Initialize RBFNN parameters (weights, center, hidden neuron, width);
Determine semantic relations within the corpus using the non-linear activation function in Eqn. (1);
Determine the feature vector using Eqn. (2);
EPO {
Initialize EPO parameters (population size, maximum iteration count);
Initialize the feature vector;
Define objective function;
while (t< maximum iteration)
Calculate the fitness value for each vocabulary term;
Find the best solution based on fitness value;
for each vocabulary term:
Generate huddle behavior;
Update current position using Eqn. (8)
end for
Calculate fitness for each vocabulary term;
Update current iteration count (t++);
end while
Select vocabulary with a fitness value greater than 0.5;
}End

4. RESULTS AND DISCUSSION

In this study, we proposed a hybrid algorithm for effective and optimal extraction of key vocabulary from the large English corpus. The proposed algorithm was modeled in the Python language, version 3.8, and the experimental computer environment is a Windows 10 operating system, CPU for Intel four core, memory 8 G, and hard disk 1T. The experimental results are examined in terms of precision, recall, and computational efficiency.

4.1 Performance analysis

In this section, the training and testing performances of the proposed algorithm were assessed in terms of accuracy, and loss. Initially, the English corpus database was split into 80:20 ratios for training and testing purposes. The accuracy measures how well the proposed algorithm learns the semantic relations within the corpus, while the loss defines the incorrect extractions made by the system. Figure 3 presents the training and testing performances of the proposed system.

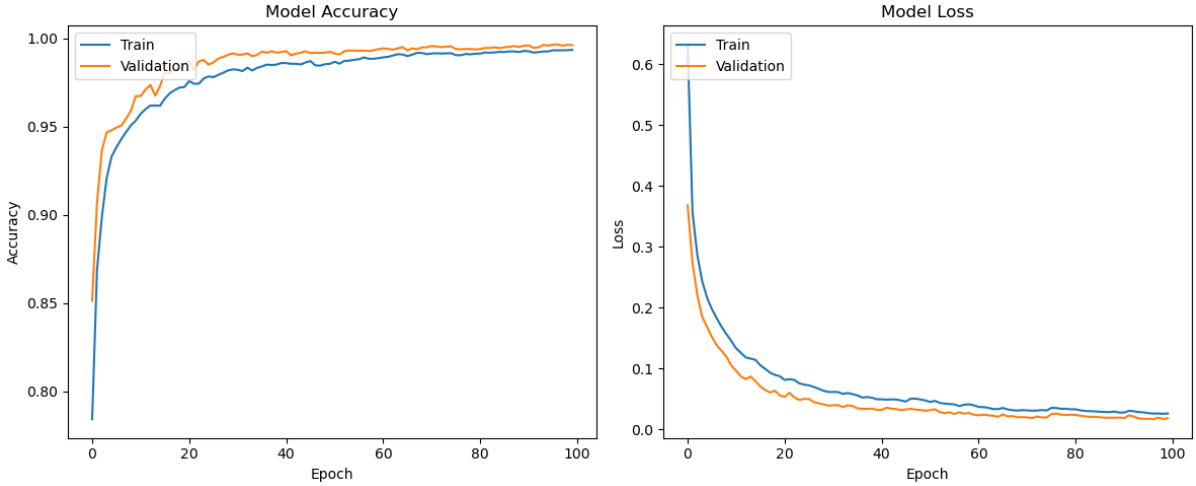

Figure 3: Training and testing performances

The training accuracy measures how effectively the developed algorithm understands the relevance, semantic relations, and correlations within the train corpus set, and the high accuracy of 0.95 depicts that the developed algorithm accurately learns the relations. On the other hand, the designed strategy achieved 0.94 accuracy in the testing phase, which depicts that it can generalize well on unseen corpus databases. This makes it more effective and reliable for real-world vocabulary extraction scenarios. Consequently, we assessed the loss performances to validate how precisely the developed algorithm extracts the key vocabularies. The presented algorithm obtained minimum training and testing losses of 0.08, and 0.09, illustrating that it precisely extracts the key vocabularies for improving English learning proficiency. From this evaluation, it is clear that the developed algorithm quickly and accurately learns the semantic relations within the corpus, and generalizes well on the unseen corpus, making them more effective and reliable for real-world scenarios.

Comparative assessment

In this section, we compare the results of the proposed strategy with the conventional models to validate its effectiveness and robustness in key vocabulary extraction. The existing techniques used for comparative analysis include Convolutional Neural Network (CNN) [14], Deep Neural Network (DNN) [15], Web Embedded System with Machine Learning (WES-ML) [16], and Unsupervised Algorithm with Graph-assisted keyword extraction (UA-GKE) [11], and BI-GRU [12]. The parameters used for comparative analysis include precision, recall, and computational time.

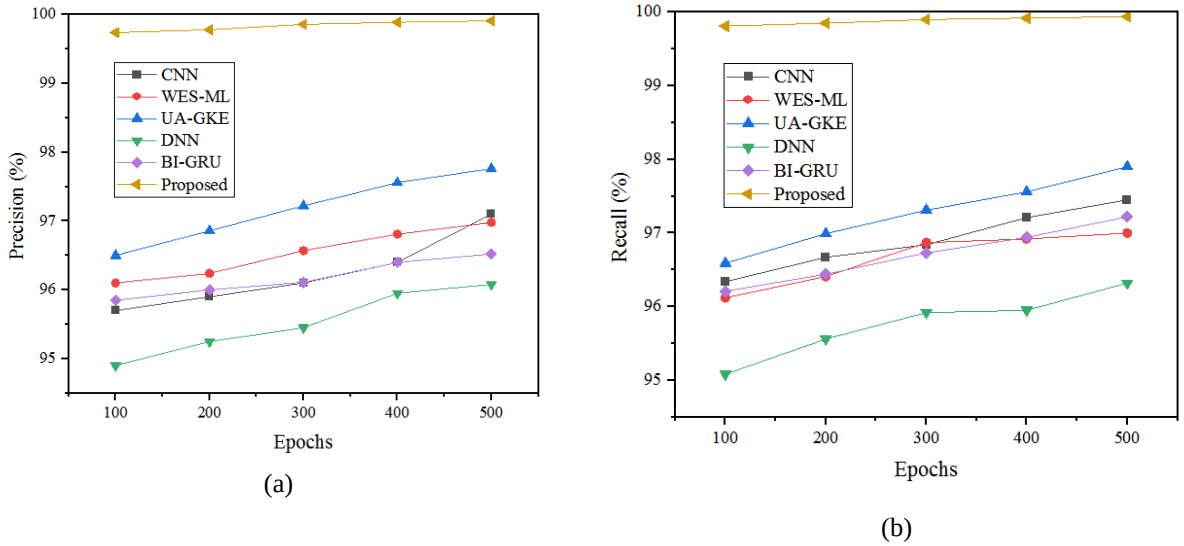

Figure 4: Comparative analysis: (a) Precision, (b) Recall

The precision metric quantifies the accuracy of the extracted key vocabularies compared to a reference set of true positives. It indicates the proportion of correctly identified key vocabularies among the total extracted. Figure 4 (a) presents the comparative evaluation of precision. Here, the precision value of the models are evaluated by increasing the epoch count from 100 to 500. The existing techniques such as CNN, WES-ML, UA-GKE, DNN, and BI-GRU obtained an average precision of 96.1%, 96.57%, 97.22%, 95.45%, and 96.11%, respectively, while the proposed EPO-RBFNN obtained higher precision of 99.86%. The improved precision depicts that the developed strategy accurately extracts the key vocabularies from the corpus database. Moreover, the improvement of precision over epochs signifies that the presented algorithm extracts more relevant and important vocabulary, highlighting its reliability and effectiveness. Recall measures the completeness of the extracted key vocabularies. It quantifies the ratio of true key vocabularies that were correctly identified by the algorithm among all true key vocabularies present in the dataset. Figure 4 (b) presents the comparative evaluation of recall. Here, the recall value of the models are evaluated by increasing the epoch count from 100 to 500. The existing techniques such as CNN, WES-ML, UA-GKE, DNN, and BI-GRU obtained an average recall rates of 96.84%, 96.87%, 97.31%, 95.92%, and 96.73%, respectively, while the proposed EPO-RBFNN obtained higher recall of 99.90%. The increased recall rate manifest that the developed the vocabularies extracted by the developed algorithm are correct and matches with the true key vocabularies in the database. This illustrates that the combination of RBFNN and EPO precisely extracts the key vocabularies. Moreover, the increase of recall over the epoch count highlights its scalability and robustness, making the designed system effective for real-time vocabulary extraction.

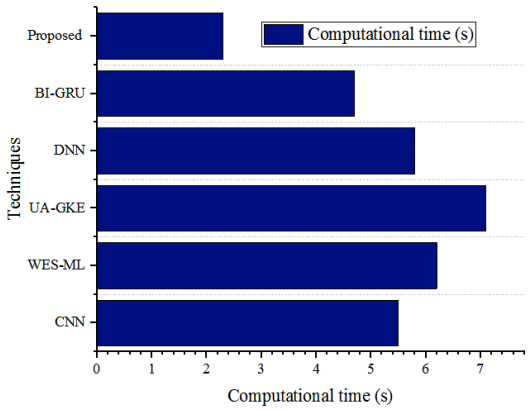

Figure 5: Comparison of computational time

Consequently, we compared and equated the computational efficiency of the system with the conventional techniques such as CNN, WES-ML, UA-GKE, DNN, and BI-GRU. The computational time measures how quickly the developed strategy processes the corpus database and extracts the key vocabularies. The comparison of computational time is presented in Figure 5. The above-stated conventional models obtained computational time of 5.5s, 6.2s, 7.1s, 5.8s, and 4.7s, while the developed algorithm obtained minimum computational time of 2.3s. The minimum computational time manifests that the presented algorithm quickly understands the patterns and extracts the key vocabularies

From this evaluation, it is clear that the presented strategy is accurate and reliable in extracting the key vocabularies than the conventional models. These improved performances manifest that the developed algorithm extracts the vocabularies accurately, assisting the learners with optimal words sequence and improving their learning proficiency.

5. CONCLUSION

This study proposed a distinct hybrid key vocabulary extraction algorithm using the combination of deep learning and meta-heuristic optimization algorithms. The developed strategy integrates the Emperor Penguin Optimizer into the Radial Basis Functional Neural Network for optimally selecting the most relevant and important vocabulary sequences from the large English corpus. The developed strategy utilizes the publicly available document corpus dataset as input, and the corpus was pre-processed to improve the efficiency of the subsequent analysis. Further, the developed EPO-RBFNN was applied for extracting the key vocabularies by examining the semantic relations and patterns within the corpus. The RBFNN was trained to understand the semantic interconnections within the database, while the EPO was employed for selecting the optimal vocabulary sequence based on the fitness value calculated based on its correlation score. The developed algorithm was implemented in Python tool, and the experimental results suggest that it achieved higher precision of 99.86%, and greater recall rate of 99.90%. Also, the proposed algorithm obtained minimum computational time of 2.3s, highlighting its computational efficiency. Moreover, we made a comparative study with the conventional algorithms like CNN, WES-ML, UA-GKE, DNN, and BI-GRU and it manifest that the performances like precision and recall are enhanced in the developed strategy by 2.64%, and 2.59%. These improved outcomes validates its effectiveness and reliability over the conventional models, making it robust and optimal for real-time vocabulary extraction for enhancing the English vocabulary proficiency of the learners.

Acknowledgements:

This work was supported by the project of “Research on the Innovation of ESP Vocabulary Teaching Model Assisted by Corpus” (2022), Dalian University of Foreign Languages.

REFERENCES

[1] Meirbekov, Akylbek, Inga Maslova, and Zemfira Gallyamova. "Digital education tools for critical thinking development." Thinking Skills and Creativity 44 (2022): 101023.

[2] Vu, Duy Van, and Elke Peters. "Vocabulary in English language learning, teaching, and testing in Vietnam: A review." Education Sciences 11.9 (2021): 563.

[3] Ghalebi, Rezvan, Firooz Sadighi, and Mohammad Sadegh Bagheri. "Vocabulary learning strategies: A comparative study of EFL learners." Cogent Psychology 7.1 (2020): 1824306.

[4] Egger, Roman, and Enes Gokce. "Natural language processing (NLP): An introduction: Making sense of textual data." Applied data science in tourism: Interdisciplinary approaches, methodologies, and applications. Cham: Springer International Publishing, 2022. 307-334.

[5] Brack, Arthur, et al. "Domain-independent extraction of scientific concepts from research articles." European Conference on Information Retrieval. Cham: Springer International Publishing, 2020.

[6] Anis, Muneeba. "Leveraging Artificial Intelligence for Inclusive English Language Teaching: Strategies and Implications for Learner Diversity." Journal of Multidisciplinary Educational Research 12.6 (2023).

[7] Mishra, Siba, and Arpit Sharma. "Automatic word embeddings-based glossary term extraction from large-sized software requirements." Requirements Engineering: Foundation for Software Quality: 26th International Working Conference, REFSQ 2020, Pisa, Italy, March 24–27, 2020, Proceedings 26. Springer International Publishing, 2020.

[8] Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. Information Sciences, 509, 257-289.

[9] Zhou, N., Shi, W., Liang, R., & Zhong, N. (2022). Textrank keyword extraction algorithm using word vector clustering based on rough data-deduction. Computational Intelligence and Neuroscience, 2022.

[10] Duari, Swagata, and Vasudha Bhatnagar. "Complex network based supervised keyword extractor." Expert Systems with Applications 140 (2020): 112876.

[11] Cui, Jinying. "Application of deep learning and target visual detection in english vocabulary online teaching." Journal of Intelligent & Fuzzy Systems 39.4 (2020): 5535-5545.

[12] Wang, Dongyang, Junli Su, and Hongbin Yu. "Feature extraction and analysis of natural language processing for deep learning English language." IEEE Access 8 (2020): 46335-46345.

[13] Li, Jinye. "A comparative study of keyword extraction algorithms for English texts." Journal of Intelligent Systems 30.1 (2021): 808-815.

[14] Khatun, Rubaya, and Arup Sarkar. "Deep-KeywordNet: automated english keyword extraction in documents using deep keyword network based ranking." Multimedia Tools and Applications (2024): 1-33.

[15] Wang, Hailin, Ke Qin, Rufai Yusuf Zakari, Guoming Lu, and Jin Yin. "Deep neural network-based relation extraction: an overview." Neural Computing and Applications (2022): 1-21.

[16] Han, Rui, and Yanlin Yin. "Application of web embedded system and machine learning in English corpus vocabulary recognition." Microprocessors and microsystems 80 (2021): 103634.