IJWR

# *Focused Crawler for Event Detection Using Metaheuristic Algorithms and Knowledge Extraction*

Hossein Moradi[a], Fatemeh Azimzadeh*[b]

[a]Department of Computer Engineering, University of Science and Culture, Tehran, Iran; Hoseinmoradi9776@gmail.com

[b]SID (Scientific Information Database), ACECR, Tehran, Iran; f.azimzadeh@gmail.com

***ABSTRACT***

**The surge in internet usage has sparked new demands. Historically, specialized web crawlers were devised to retrieve pages pertaining to specific subjects. However, contemporary needs such as event identification and extraction have gained significance. Conventional web crawlers prove inadequate for these tasks, necessitating exploration of novel techniques for event identification, extraction, and utilization. This study presents an innovative approach for detecting and extracting events using the Whale Optimization Algorithm (WOA) for feature extraction and classification. By integrating this method with machine learning algorithms, the proposed technique exhibits improvements in experiments, including decreased execution time and enhancements in metrics such as Root Mean Square Error (RMSE) and accuracy score. Comparative analysis reveals that the proposed method outperformed alternative models. Nevertheless, when tested across various data models and datasets, the WOA model consistently demonstrated superior performance, albeit exhibiting reduced evaluation metrics for Wikipedia text data.**

***Keywords****— Knowledge Extraction, Focused Crawler, Whale Optimization Algorithm (WOA), Feature Selection, Event Detection.*

## 1. Introduction

In today's highly competitive world, data has emerged as a valuable asset. The sheer volume of data underscores the importance of transforming this data into useful information and knowledge. This challenge is evident in various domains, including search engines, digital business, news dissemination management, and many daily needs. One of the common applications in this area is search engines. The ultimate goal of a user employing a search engine is to receive information relevant to their needs [1]. To this end, crawlers in search engines attempt to identify pages related to the user's needs. They then extract the links within the documents and repeat this process for the obtained links [2]. In a focused crawler, the crawler seeks to gather documents on one or several specific topics. Using a relevance function, the focused crawler analyzes the links and estimates their relevance to the desired topic, determining the priority of the links for crawling accordingly [3].

When a significant event occurs, many users try to find the most up-to-date information about it. However, there is no systematic method for collecting and archiving event information. As a result, a focused crawling system is required to gather web data on key events [4]. In the past, focused crawling was primarily applied to topic-based crawling (i.e., collecting web pages on a specific subject or domain). However, traditional focused crawling approaches no longer suffice for extracting, identifying, and utilizing events to meet user needs. Therefore, we need different models for event identification through event-based focused crawlers.

The main challenge associated with focused crawlers is the accurate classification of web pages based on the given topic due to the unstructured nature of data on web pages. The primary goal of this paper is to present an improved focused crawler model based on web page classification. A focused web crawler always uses different algorithms and

http://dx.doi.org/10.22133/ijwr.2024.454772.1215

Citation H. Moradi, F. Azimzadeh, " Focused Crawler for Event Detection Using Metaheuristic Algorithms and Knowledge Extraction", *International Journal of Web Research*, vol.6, no.2,pp.143-150, 2023, doi: http://dx.doi.org/ 10.22133/ijwr.2024.454772.1215.

*Coressponding Author

Article History: Received: 28 September 2023 ; Revised: 24 Desember 2023; Accepted: 28 Desember 2023.

features to identify relevant pages. In computer science and mathematical optimization, metaheuristic methods are high-level procedures designed to find, generate, adjust, or select a heuristic mode.In this research, we aim to perform knowledge extraction using important page features combined with page classification. Therefore, we propose a focused crawling method based on machine learning, assisted by the Whale Optimization Algorithm (WOA). In this approach, the relevance of a page's content to an event is calculated for each page. On the one hand, page classification is considered based on the event, and on the other hand, other important features on the event page are utilized. For the first part (knowledge extraction process), the probability of the page being in a topic-related category is calculated. For the second part (topic classification), the content similarity [5] with the desired event is computed based on important features on the page. By combining these two computed values, the page's relevance to the event is determined.

The next sections of this paper review the latest related research. Chapter 3 describes the research methodology, Chapter 4 compares the proposed method with other methods, and Chapter 5 discusses the obtained results and offers suggestions for further research.

## 2. Relate Works

A web crawler is based on the idea of extracting information from the World Wide Web by analyzing various web pages and other resources. A focused crawler provides an initial clue that can help in the search direction.

Machine learning-based focused crawling approaches utilize text classification algorithms [6, 7] to build a model from training data. The focused crawler then employs this model to assess the relevance of unvisited web pages. This model enhances the classifier's performance by incorporating specific knowledge and online feedback. Rena and Barth [8] applied reinforcement learning to tackle the issue of focused crawling. Another reinforcement learning algorithm, temporal difference learning, utilizes a function to gauge the importance of web pages in directing towards future relevant web pages. The method offers increased classification accuracy as a benefit, with increased complexity being considered a drawback.

Subsequent research [9] adopted a reinforcement learning framework and enhanced its performance through incremental online learning. In this algorithm, each new URL is classified, and its features are utilized to update the class features. They refine the supervised learning algorithm based on the new training data. This approach eliminates the data bias present in the test data, ensuring that unseen URLs not present in the training data emerge during the testing phase. The precise updates can be viewed as a positive aspect of this method, while widespread complexity and low speed are its drawbacks.

Dong et al. [10] noted in their paper that since a new page can only be discovered by following a new outbound link (OL), predicting new links is highly effective in practice. In the literature, many feature designs for predicting changes on the web have been proposed. In their work, they provide a structured analysis of this problem using new outbound links as the prediction target. Specifically, they unify previous feature designs into a classification arrangement in two dimensions: static versus dynamic features, and features of a page versus features of its surrounding network. This classification is complemented with new features (mainly dynamic network features) and identifies the best predictors for new outbound links. Thus, most informative features are the recent history of new outbound links within the page itself and related content pages. As a result, a new model called "Look Back, Look Around" (LBLA) is proposed, which only uses these features.

Wang and YU [11] considered that existing focused crawlers are weak in the case of crawling pages with specific subjects, they proposed a learning-based focused crawler using a URL knowledge base. The proposed method considered the relevance of the page heading which was evaluated with the content of the linked page, text associated with the link, and information encoded in the URL. The URL content is also learned and updated iteratively and continuously. Within the crawler, they implement a crawling mechanism based on a combination of textual evaluation and relationship exploration between webpages, which decreases computational complexity and avoids the locality problem of crawling. Experimental results show that the proposed algorithm achieves better precision.

This overview captures the essence of the major research efforts and approaches in the domain of focused web crawling, highlighting the evolution from basic heuristic methods to sophisticated machine learning and reinforcement learning techniques. The subsequent sections will delve deeper into the methodology and performance comparison of the proposed approach against existing methods.

Rajeev et al. [12] also proposed a novel and efficient method for enhancing keyword sets. This method focuses on the automatic classification of web pages used to determine their relevance. Optimized keyword weights based on an efficient metaheuristic are employed. Feature extraction based on Term Frequency (TF) and keyword weight optimization using the Random Search Algorithm (RSA) are utilized in event-focused web crawling.

Gradient descent, a popular algorithm for achieving optimization, has the advantage of smooth subdifferentiable and differentiable fitness functions, making it suitable for large data optimization. This algorithm focuses on optimizing the keyword set, and if a better keyword set is identified, the resultant documents returned can be even more relevant to user queries. Support Vector Machines (SVM) classifiers are used for this task. Experimental results demonstrated that the proposed method outperforms other methods, including Particle Swarm Optimization (PSO)-based weight optimization solutions. The proposed RSA weight optimization showed a 5.8% improvement over PSO, indicating its ability to handle large data volumes. However, a drawback of this method is its lower accuracy on some datasets due to the use of SVM.

Overall, various algorithms based on collective intelligence behavior have been proposed and introduced, such as Grey Wolf Optimization [13], Lion Ant Algorithm[14], Whale Optimization Algorithm (WOA) [15], Bacterial Foraging Optimization Algorithm [16], Social Spider Algorithm [17], and Needlefish Optimization Algorithm. These algorithms have numerous applications in fields such as knowledge discovery, machine learning techniques, and data mining, such as association rules [18]. They can play a significant role in improving artificial neural networks as weight selectors and bias enhancers, enhancing SVM parameters [19], and discovering association rules. However, to increase their accuracy, the population size and the number of iterations need to be sufficiently large to reduce errors in calculations.

The Whale Optimization Algorithm (WOA) is one of the optimization algorithms based on swarm intelligence, inspired by the hunting behavior of humpback whales. The specific features of this algorithm that make it preferable to other methods include:

1. ***Simplicity and Efficiency***: The WOA is simple and easy to understand, and its implementation is easier compared to many other complex optimization algorithms.
2. ***Fast Convergence:*** Due to the use of techniques such as encircling and spiral updating, the WOA has a high convergence speed, quickly moving towards the best solutions.
3. ***Balance Between Local and Global Search***: The WOA uses various techniques to maintain a balance between local search (exploitation) and global search (exploration), improving the accuracy and quality of results.
4. ***Ability to Handle Nonlinear and Complex Problems***: The WOA performs well when dealing with nonlinear and complex problems and can find global optima [8].

The main challenge of metaheuristic algorithms is their execution time, especially as the population size increases to enhance accuracy. In the proposed method, an embedding operation using linear regression and the Whale Optimization Algorithm is employed to address the challenges of classification and knowledge extraction.

This section has reviewed significant research contributions and the evolution of focused crawling techniques, highlighting the transition from heuristic methods to advanced machine learning and metaheuristic approaches. The next sections will detail the research methodology, the proposed method's advantages, and its performance comparison with existing techniques.

## 3. Proposed Method

The utilization of metaheuristic algorithms in this research is driven by their capacity to replicate the finest and most distinctive attributes of nature. Our proposed technique incorporates the Whale Optimization Algorithm (WOA), which adeptly balances global exploration and local exploitation through two primary stages:

- Bubble Encasing Stage: This serves as local search, progressively moving towards the best local solutions.
- Prey Search Stage: This functions as global search, dispersing solutions to pinpoint the optimal points in the search space.

This trait enables the WOA algorithm to discover superior solutions in intricate and multi-dimensional spaces, thereby enhancing the model's accuracy and dependability. The algorithm's adaptability and tailoring to the WOA optimization algorithm are well-suited for various optimization challenges. This adaptability is facilitated by dynamic parameters and automatic adjustment mechanisms, enabling the algorithm to approach optimal points more effectively and reduce the root mean square error. The swift stability and convergence of the WOA optimization algorithm are typically characterized by rapid convergence and high stability owing to intelligent search mechanisms and updates. This capability allows the algorithm to attain optimal outcomes in less time and enhance evaluation metrics such as accuracy, precision, recall, and program execution time.

In the proposed approach, linear regression is employed for the imputation process. Linear regression, recognized for its easily interpretable behavior, high performance on sparse data, and speed, is one of the simplest machine learning algorithms. Alternate string matching algorithms will be utilized for keyword adaptation, while the WOA optimization algorithm will be leveraged for

classification and knowledge extraction. The methodology involves page classification based on events and the utilization of other significant features on the event page. The probability of a page belonging to a topic-related class is computed for the first part, while content similarity with the desired event is calculated based on important features on the page for the second part. The amalgamation of these calculated values determines the page's relevance to the event.

### 3.1. Focused Crawling Method

The general stages of the proposed method are as follows:

***Step 1:*** Selection of the desired event, date, and location of the event (events are considered as topics and represented with a list of keywords).

***Step 2:*** Extraction of web pages related to the selected event which performed by the concentrated crawler.

***Step 3:*** Keyword matching with the extracted set of URLs.

***Step 4:*** The concentrated crawler estimates the content relevance of the web page to the selected event.

***Step 5:*** Scoring of web pages and URLs based on relevance to the event.

The method used in steps 1, 2, and 3, which involve data preprocessing operations, keyword analysis, and data dimensionality reduction, will be handled by the linear regression algorithm and the alternating string algorithm.

Steps 4 and 5 will also utilize the WOL algorithm for classification and feature selection.

In the first and second steps, to implement the calculation of the content similarity score of a page with a topic, a method based on the following steps is proposed:

- Initial pages from which the crawler starts crawling (seed pages) are considered as input.
- Outgoing links of pages are extracted to form a set of pages related to an event.
- Important features of the page such as page title, URL text string, anchor text, and inheritance feature are extracted for calculating the page's content similarity.

By combining the calculated probability of the page's existence in the topic-related category with the content similarity score of the target event, the degree of relevance of the page to the event, namely step three, which is keyword matching with the extracted pages, will be obtained.

### 3.2. WOA Algorithm

WOA has the ability to avoid local optima and obtain global optimal solutions. It also has less computational overhead. These advantages make WOA an appropriate algorithm for solving various optimization problems, constrained or unconstrained, for practical applications without structural modifications to the algorithm. The proposed model is inspired by the behavior of whale pods, and in this algorithm, modeling of spiral-up or double-doubling movement is used to search the problem space and find optimal solutions. This behavior makes the problem space better traversed to find the optimal solution. In the WOA optimization algorithm, the most appropriate whale population location is considered as the current optimal point, and fish shoal aggregation is considered around this point so that other whales move around this point using spiral or circular movements. The following equations can be used for rotational or circular movements in the optimization algorithm [23]:

In the Equ (1) and Equ (2) ; represents the distance vector (indicating the working space or the web environment for the crawler), where the position of a whale, denoted as (which refers to the focused web crawler navigating through the web), is measured from the optimal position (which denotes the optimal state where the crawler can match the search criteria with the retrieved instances). The optimal position is the best state resulting from establishing relevant indices. Additionally, and are random and decreasing coefficients based on the algorithm iterations.

$$\vec{D} = \left|\vec{C}.\vec{X}^*(t) - \vec{X}(t)\right| \quad (1)$$

$$\vec{X}(t+1) = \vec{X}^*(t) - \vec{A}\vec{D} \quad (2)$$

In the Whale Optimization Algorithm, Equ (3) can be used to perform the spiral movement[20]:

In the Equ(3), represents the distance between the position of a whale and the optimal whale position, is a constant, and is a uniformly distributed random number within the interval [-1,1]. A significant advantage of the Whale Optimization Algorithm (WOA) is its higher accuracy compared to well-known algorithms such as the Genetic Algorithm (GA) and Particle Swarm Optimization (PSO).

$$\vec{X}(t+1) = \overrightarrow{D'} + e^{bl}.cos\,(2\pi l) + \vec{X}^*(t) \quad (3)$$

To perform classification and knowledge extraction, the WOA is employed. In this stage, the TF-IDF model and word embedding are used for feature extraction of textual information. After this step, feature grouping is performed. During the clustering phase, all extracted features are grouped based on their similarities into different clusters. Clustering is referred to as an unsupervised learning process that identifies hidden patterns in sample data. It is also used to identify meaningful features for accurate recommendations. Finally, a network model

based on the Whale Optimization Algorithm is used to select the feature vectors for the final recommendation. In the WOA-based model, feature labels along with cues are required to generate the recommended list.

## 4. Results and Evaluation

This section reviews some of the most significant research conducted in this fieldLeft-aligned headings should be used. They need to be numbered. "2. Headings and

### 4.1. Simulation Environment

The proposed method was implemented using the Python software environment, version 3.9.

### 4.2. Dataset

To evaluate the proposed method using real-world data, two sets of datasets were utilized:

1. Financial datasets: FORD, GMC, and TESLA from Yahoo Finance (2010-2023), and AAPL, IBM, MSFT, and WMT from Google Finance.
2. Terrorism event datasets: Data on terrorist events obtained from Wikipedia.

### 4.3. Evaluation Metrics

In this section considered some different criteria to evaluating the proposed method in compare the other methods.

***Root Mean Squared Error (RMSE):***

RMSE represents the standard deviation of the residuals (prediction errors) and is calculated using Equ (4):

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \left(\frac{|y_t - \widehat{y_t}|}{|y_t|}\right)^2} \qquad (4)$$

***Accuracy (Acc):***

Accuracy is defined as the ratio of correctly predicted classes (including both error-prone and error-free) to the total number of classes, and is calculated using the Equ (5):

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \qquad (5)$$

***Precision (Pre):***

Precision indicates the number of error-prone classes that are correctly predicted by the model as error-prone. Equ (6) show the Precision formula:

$$Precision = \frac{TP}{TP+FP} \qquad (6)$$

***Recall (Rec):***

Recall is the number of error-prone classes that are predicted as error-prone by the model, and is calculated using the Equ (7):

$$Recall = \frac{TP}{TP+FN} \qquad (7)$$

***F1 Score:***

The F1 score takes both precision and recall into account to calculate accuracy, and can be interpreted as a weighted average of precision and recall, which show in the Equ (8).

$$F1 - measure = 2 * \frac{Recall * Precision}{Recall + Precision} \qquad (8)$$

### 4.4. Evaluation

To evaluate the proposed method, it is compared with ANN, GWO, and Bi-LSTM methods. GWO (Grey Wolf Optimizer) is a metaheuristic optimization algorithm inspired by the leadership hierarchy and hunting mechanism of grey wolves in nature. It simulates the natural behavior of grey wolves during the hunting process.

Artificial Neural Networks (ANNs) are a branch of machine learning models constructed using the principles of neural organization found in biological neural networks, which constitute the brains of animals. An ANN is made up of interconnected units or nodes called artificial neurons, which model the neurons in the brain.

Bi-LSTM is a type of Recurrent Neural Network (RNN) that processes sequential data in both forward and backward directions. It combines the power of LSTM with bidirectional processing, allowing the model to capture both past and future context of the input sequence.

Due to the relative similarity in the working principles of these methods with WOA, they are used for comparison in this section.

Three different testing approaches are used in this evaluation. The first test uses datasets, the second test is performed in real-time, and the third test involves a new domain focused on information from crawling Wikipedia for terrorist events.

***First Experiment:***

Evaluation of the ANN method as shown in Tables 1., it is evident that ANN achieves the highest value in the Precision factor, followed by F1 Score and Accuracy in subsequent positions. In the evaluation of the GWO method, GWO attains the highest value in the Accuracy factor, followed by Precision and F1 Score. The Bi-LSTM method demonstrates that it achieves the highest value in the

F1 Score factor, followed by Precision, Accuracy, Recall, and RMSE.

Regarding the proposed method, WOA, it is clear that WOA attains the highest value in the F1 Score factor, followed by Precision, Accuracy, and Recall. The four factors—Accuracy, Precision, Recall, F1 Score and RMSE, show improved values in this method compared to the other methods.

Figure 1. show that WOA method achieve the best results.

***Second Experiment:***

In the second experiment, as show in Table 2. it is evident from the analysis of the ANN method that the ANN method has assigned the highest value to the Accuracy factor in computational metrics, and subsequently, F1 Score and Recall have been ranked next. Investigation of the GWO method reveals that the GWO method has allocated the highest value to the Accuracy factor, followed by Recall and F1 Score in subsequent ranks. Additionally, the execution time of this method is higher compared to other methods. The value of the Root Mean Square Error (RMSE) in this method has also increased compared to the first experiment. Examination of the Bi-LSTM method indicates that the Bi-LSTM method has allocated the highest value to the Accuracy factor, followed by Precision, F1 Score, and Recall in subsequent ranks. The RMSE has also enhanced in this experiment compared to the previous one. The execution time of this method is more optimal compared to the GWO method and higher than ANN. Analysis of the proposed method also shows that the proposed WOA method has improved in the factors of F1 Score, Precision, Accuracy, Recall, and RMSE compared to other methods. The accuracy, precision, and recall rates have increased, and the systematic error rate has decreased compared to other tested methods. Additionally, the execution time of this method has shown improvement compared to other methods.

In Figure 2. Based on the provided table comparing the performance of four methods—Artificial Neural Network (ANN), Grey Wolf Optimizer (GWO), Bidirectional Long Short-Term Memory (Bi-LSTM), and Whale Optimization Algorithm (WOA)—across metrics such as Execution Time, Precision and Root Mean Square Error (RMSE) achieved better results.

Table 1. Results Comparison in the First Experiment.

| *Methods* | *RMSE* | *F1 Score* | *Recall* | *Precision* | *Accuracy* |
|---|---|---|---|---|---|
| ANN | 0/91 | 0/826 | 0/789 | 0/862 | 0/793 |
| GWO | 0/78 | 0/719 | 0/696 | 0/731 | 0/746 |
| Bi-LSTM | 0/36 | 0/92 | 0/876 | 0/899 | 0/893 |
| WOA | 0/18 | 0/953 | 0/919 | 0/946 | 0/943 |

Table 2. Results Comparison in the Second Experiment

| *Methods* | *RMSE* | *F1 Score* | *Recall* | *Precision* | *Accuracy* | *ExecutionTime* |
|---|---|---|---|---|---|---|
| ANN | 0/93 | 0/466 | 0/466 | 0/165 | 0/59 | 0/033 |
| GWO | 0/944 | 0/372 | 0/372 | 0/158 | 0/80 | 0/05 |
| Bi-LSTM | 0/844 | 0/424 | 0/324 | 0/666 | 0/77 | 0/034 |
| WOA | 0/787 | 0/212 | 0/397 | 0/649 | 0/85 | 0/0330 |

***Third Experiment :***

AS shown in the Table 3, the investigation of the ANN method indicates that ANN, in terms of Accuracy factor, has assigned the highest value to itself in the computational metrics, followed by F1 Score and Recall in subsequent ranks. Examination of the GWO method reveals that GWO has assigned the highest value to itself in the Accuracy factor, followed by Recall and F1 Score in subsequent ranks. Moreover, the execution time of this method has allocated a higher value to itself compared to other methods. The value of the mean square error has also increased in this method compared to the first experiment. Analysis of the Bi-LSTM method indicates that Bi-LSTM, in terms of Accuracy factor, has assigned the highest value to itself, followed by Precision, F1 Score, and Recall in subsequent ranks. The mean square error has also increased in this experiment compared to the previous one. The execution time of this method is also more optimal compared to the GWO method and higher than ANN. Examination of the proposed method indicates that

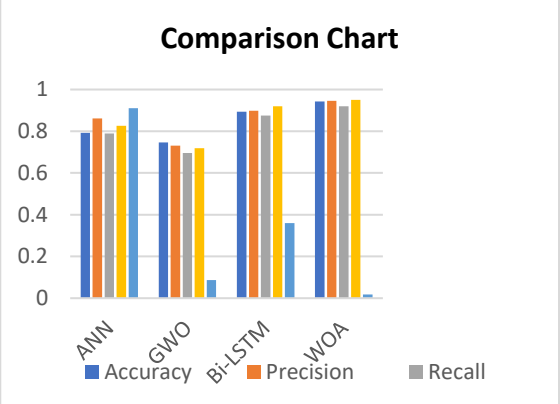


Figure. 1. Results Comparison in the First Experiment

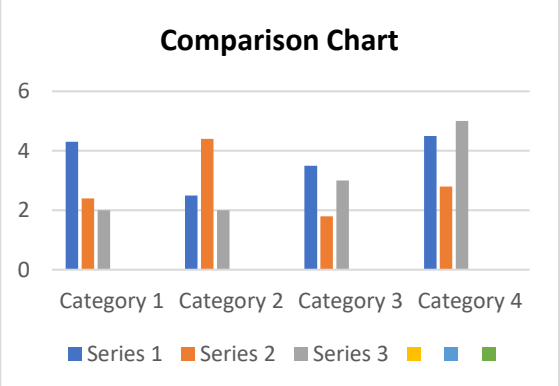


Figure. 2. Results Comparison in the Second Experiment

Table 3. Results Comparison in the Third Experiment

| *Methods* | *RMSE* | *F1 Score* | *Recall* | *Precision* | *Accuracy* | *Execution Time* |
|---|---|---|---|---|---|---|
| ANN | 1.2274 | 0.4872 | 0.3462 | 0.8218 | 0.4010 | 0.9440 |
| GWO | 1.2783 | 0.5042 | 0.3462 | 0.9275 | 0.3685 | 167.75 |
| Bi-LSTM | 1.4137 | 0.5142 | 0.3462 | 0.9992 | 0.3464 | 0.9152 |
| WOA | 1.1950 | 0.4706 | 0.3462 | 0.7347 | 0.4278 | 0.2413 |

the proposed method has improved in terms of F1 Score, Precision, Accuracy, Recall, and RMSE compared to other methods. Additionally, the execution time of this method has improved compared to other methods.

In Figure 3. Based on the provided table comparing the performance of four methods—Artificial Neural Network (ANN), Grey Wolf Optimizer (GWO), Bidirectional Long Short-Term Memory (Bi-LSTM), and Whale Optimization Algorithm (WOA)—across metrics such as Execution Time, Accuracy, Recall, F1 Score, and Root Mean Square Error (RMSE), achieved the better result, except Precision which, Bi-LSTM take the better result.

## 5. Conclusion and Research Recommendations

In this paper, a model for web page retrieval using a focused crawler based on the Whale Optimization Algorithm (WOA) was utilized. By employing this model on two different datasets, better results were achieved, indicating that this approach performs better regardless of the dataset. This crawler was tasked with loading and collecting pages focused on a specific topic.

Various machine learning methods were evaluated in this study based on six evaluation criteria: execution time, accuracy, precision, recall, F1 Score, and Root Mean Square Error (RMSE). The obtained results demonstrated that our proposed method, WOA, outperformed others with the lowest execution time (0.2413 seconds), highest accuracy (0.4278), and the lowest RMSE value (1.195), thus being recognized as the best evaluation method. These results underscore the significant superiority of WOA compared to other examined methods.

However, methods such as ANN, GWO, and Bi-LSTM couldn't compete fully with the performance of WOA. Although Bi-LSTM showed good performance in some criteria such as precision (0.9992) and F1 Score (0.5142), due to its lower accuracy and higher RMSE compared to WOA, its overall efficiency was lower. GWO, with a significantly longer execution time (167.75 seconds) and weaker performance in other criteria, also failed to be competitive.

In general, the results of this study confirm that our proposed method, WOA, is introduced as the optimal method due to its balanced combination of

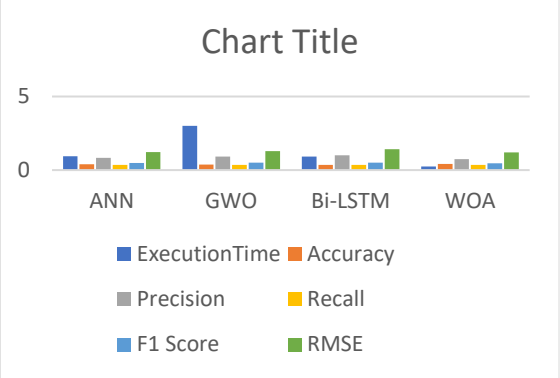


Figure. 3. Results Comparison in the Third Experiment

execution time, accuracy, and the lowest RMSE value. This superiority highlights WOA's capability in providing optimal and efficient results in various practical fields.

From the above comparison analysis, it is evident that the proposed method, owing to the optimal optimization method, was able to achieve better results in the mentioned criteria compared to other methods. Indeed, WOA's ability to avoid local optima and obtain a general optimal solution makes it a suitable algorithm for solving various optimization problems, limited or unlimited, for practical applications. A significant improvement in RMSE compared to ANN and GWO methods was observed. The reasons for these improvements include the advantages of employing the WOA method (simple structure, fewer operator requirements, rapid convergence, and faster discovery stages) and its combination with linear regression embedding, as well as preprocessing operations and removal of unrelated pages, leading to improved execution time, accuracy, and error score.

As a result, we believe that the proposed system can be considered as the preferred system for retrieving web pages relevant to queries in search engines and web content management. Other classification methods such as SVM can be considered as future research topics due to their potential to classify other web pages based on features, input diversity, and changes in membership functions.

## Declarations

### *Funding*

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

### *Authors' contributions*

HM: Study design, creation of models, formal techniques to analyze the study data and experiments, evaluation of the results, drafting the manuscript FA:Study design, creation of models,

formal techniques to analyze the study data and experiments, evaluation of the results, revision of the manuscript

### *Conflict of interest*

The author declares that no conflicts of exist.

## References


[1] B. Oliveira and C. Teixeira Lopes, "The Evolution of Web Search User Interfaces-An Archaeological Analysis of Google Search Engine Result Pages," in Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, 2023, pp. 55-68. https://doi.org/10.1145/3576840.3578320.

[2] Q. Liu, R. Yahyapour, H. Liu and Y. Hu, "A novel combining method of dynamic and static web crawler with parallel computing," Multimedia Tools and Applications, pp. 1-22, 2024. https://doi.org/10.1007/s11042-023-17925-y.

[3] N. Kumar and D. Aggarwal, "LEARNING-based focused WEB crawler," IETE Journal of Research, vol. 69, no. 4, pp. 2037-2045, 2023. https://doi.org/10.1080/03772063.2021.1885312.

[4] H. Wu and D. Hou, "A Focused Event Crawler with Temporal Intent," Applied Sciences, vol. 13, no. 7, pp. 4149, 2023. https://doi.org/10.3390/app13074149.

[5] J. Xu, B. Wang, Q. Peng and W. Li, "Key-frame reference selection for error resilient video coding using low-delay hierarchical coding structure," Signal, Image and Video Processing, vol. 18, no. 1, pp. 215-222, 2024. https://doi.org/10.1007/s11760-023-02742-5.

[6] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys (CSUR), vol. 34, no. 1, pp. 1-47, 2002. https://doi.org/10.1145/505282.505283.

[7] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in Mining Text Data, MA: Springer, 2012, pp. 163-222. https://doi.org/10.1007/978-1-4614-3223-4_6.

[8] J. D. M. Rennie and A. McCallum, "Using reinforcement learning to spider the web efficiently," in Proceedings of the 16th International Conference on Machine Learning (ICML'99), 1999, pp. 335-343.

[9] M. U. Demirezen and T. S. Navruz, "Lambda Architecture-Based Big Data System for Large-Scale Targeted Social Engineering Email Detection," International Journal of Information Security Science, vol. 12, no. 3, pp. 29-59, 2023. https://doi.org/10.55859/ijiss.1338813.

[10] T. K. N. Dang, D. Bucur, B. Atil, G. Pitel, F. Ruis, H. Kadkhodaei, and Litvak, "Look back, look around: A systematic analysis of effective predictors for new outlinks in focused Web crawling," Knowledge-Based Systems, vol. 260, p. 110126, 2023. https://doi.org/10.1016/j.knosys.2022.110126.

[11] W. Wei and U. Lihua, "UCrawler: A learning-based web crawler using a URL knowledge base," Journal of Computational Methods in Sciences and Engineering, vol. 21, no. 2, pp. 461-474, 2021.

[12] S. Rajiv and C. Navaneethan, "Keyword weight optimization using gradient strategies in event focused web crawling," Pattern Recognition Letters, vol. 142, pp. 3-10, 2021. https://doi.org/10.1016/j.patrec.2020.12.003.

[13] D. Ho, E. Liang, X. Chen, I. Stoica and P. Abbeel, "Population based augmentation: Efficient learning of augmentation policy schedules," in International conference on machine learning, 2019, pp. 2731-2741. https://proceedings.mlr.press/v97/ho19b.html.

[14] E. Rios, L. S. Ochi, C. Boeres, V. N. Coelho, I. M. Coelho and R. Farias, "Exploring parallel multi-GPU local search strategies in a metaheuristic framework," Journal of Parallel and Distributed Computing, vol. 111, pp. 39-55, 2018. https://doi.org/10.1016/j.jpdc.2017.06.011.

[15] H. Faris, I. Aljarah, M. A. Al-Betar and S. Mirjalili, "Grey wolf optimizer: a review of recent variants and applications," Neural Computing and Applications, vol. 30, pp. 413-435, 2018. https://doi.org/10.1007/s00521-017-3272-5.

[16] A. A. Heidari, H. Faris, S. Mirjalili, I. Aljarah and M. Mafarja, "Ant lion optimizer: theory, literature review, and application in multi-layer perceptron neural networks," Nature-Inspired Optimizers: Theories, Literature Reviews and Applications, pp. 23-46, 2020. https://doi.org/10.1007/978-3-030-12127-3_3.

[17] M. Abd Elaziz and S. Mirjalili, "A hyper-heuristic for improving the initial population of whale optimization algorithm," Knowledge-Based Systems, vol. 172, pp. 42-63, 2019. https://doi.org/10.1016/j.knosys.2019.02.010.

[18] R. Nagpal, P. Singh and B. P. Garg, "Concurrent bacterial foraging with emotional intelligence for global optimization," International Journal of Information Technology, vol. 11, pp. 313-320, 2019. https://doi.org/10.1007/s41870-018-0215-z.

[19] S. Aggarwal, P. Chatterjee, R. P. Bhagat, K. K. Purbey and S. J. Nanda, "A social spider optimization algorithm with chaotic initialization for robust clustering," Procedia Computer Science, vol. 143, pp. 450-457, 2018. https://doi.org/10.1016/j.procs.2018.10.417.

[20] S. Mirjalili and S. M. Saremi, "Whale optimization algorithm: theory, literature review, and application in designing photonic crystal filters," Nature-Inspired Optimizers: Theories, Literature Reviews and Applications, pp. 219-238, 2020. https://doi.org/10.1007/978-3-030-12127-3_13.

[21] S. Mirjalili, Evolutionary Algorithms and Neural Networks, Cham: Springer, 2019, pp. 43-60.

[22] D. C. Montgomery, E. A. Peck and G. G. Vining, Introduction to Linear Regression Analysis, John Wiley & Sons, 2012.

[23] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in Mining Text Data, Boston, MA: Springer, 2012, pp. 163-22. https://doi.org/10.1007/978-1-4614-3223-4_6..


**Hossein Moradi** is a master's student in Computer Engineering - Software at the University of Science and Culture. His research interests include information retrieval, feature extraction, and focused crawling. During his master's studies, he has participated in various research projects.

**Fatemeh Azimzadeh** received a Ph.D. degree in Information Technology from University Putra Malaysia in 2012. Currently, she is an assistant professor in ACECR, Tehran, Iran. She is also the director of SID (Scientific Information Database) in Iran. Her research interests include information retrieval and Natural Language Processing..