Online harassment detection on online data science platforms optimized by metaheuristic

Nebojsa Bacanin[1]*, Milos Kabiljo[2], Lepa Babic[1], Vuk Gajic[1], Jelena Kaljevic[1], and Milos Dobrojevic[1]

Singidunum University, Belgrade, Serbia
nbacanin@singidunum.ac.rs*, lbabic@singidunum.ac.rs, vgajic@singidunum.ac.rs, jkaljevic@singidunum.ac.rs, mdobrojevic@singidunum.ac.rs
Department for Information Systems and Technologies, Belgrade Academy for Business and Arts Applied Studies, Belgrade, Serbia
milos.kabiljo@bpa.edu.rs

Abstract. Cyberbullying denotes one of the recent pervasive problems, mostly found on social networks, that poses a considerable challenge to keep safe and inclusive environment. It can lead to serious psychological problems for the victim. As one of possible responses, artificial intelligence emerged as a powerful option to identify cases of cyberbullying, and it has garnered considerable attention. This paper suggest using a combination of natural language processing, paired with machine learning XGBoost classifier tuned by an altered variant of the sine cosine metaheuristics to classify and identify the cases of cyberbullying in data collected from a variety of social networks including Kaggle, Twitter and Youtube. The obtained simulation outcomes suggest considerable potential of machine learning models to address this problem.

Keywords: Cyberbullying · Harassment detection · Machine learning · XGBoost · Swarm intelligence · metaheuristics optimization · sine cosine algorithm.

1 Introduction

Cyberbullying relies on the utilization of the electronic communication platforms, like social media, forums or messaging applications, for the purpose of harassment, intimidation or harming others. It comes in many different forms, where the most common are sending messages with harmful content, embarrassing photo or video materials (real or manipulated), and exclusion of the individuals from conversation. It can result in serious psychological issues for the victim, that include depression, anxiety, no self-esteem, and ultimately, suicidal thoughts and self-harming. In worst case scenarios, it can raise safety concerns, as it may escalate to physical violent behavior and threats.

Although detection of cyberbullying is essential, it poses several considerable challenges regarding its complex nature paired with the evolving methods used

N. Bacanin and H. Shaker (eds.), *Proceedings of the 2nd International Conference on Innovation in Information Technology and Business (ICIITB 2024)*,
Advances in Computer Science Research 113,
https://doi.org/10.2991/978-94-6463-482-2_9

by predators. One of the biggest issues is tremendous amount of data, where moderators are often not capable to follow all the comments on the platform. Another point is the variability in language, where predators are using sarcasm and slang that evolves with the culture, which makes extremely difficult for the automated systems to precisely identify harmful messages. Other challenges include contextual understanding, where some words and/or phrases may be used in different context, which is vital for the proper classification. Finally, predators are not only sending text messages, but also multimodal content including images, videos and audio files, that makes accurate detection even harder [24].

Artificial intelligence (AI) may be a potential solution for cyberbullying detection, as it can leverage a wide spectrum of methods from natural language processing (NLP), machine learning (ML) and data analysis [27]. AI models are capable of analyzing wast amount of textual data captured from social media posts and comments, and detect the linguistic patterns that are related to the cyberbullying. NLP methods like sentiment analysis can additionally aid in comprehending both content and context of comments. ML models, on the other side, may be trained to adopt problematic patterns and automatically perform classification of the new messages and comments as either cyberbullying or normal content. However, the biggest challenge of ML methods is the proper configuration of the hyperparameters' values, that is crucial for proper classification in every single classification problem. This task is considered NP-hard by nature, meaning that only stochastic methods are suitable to resolve it in the acceptable time.

Metaheuristics algorithms, a subcategory of stochastic methods, are renowned as potent optimizers, and may be utilized to select the most appropriate set of hyperparameters' merits of the observed model. Consequently, the model will be capable to perform classification better, and the performance level will be significantly better. Nevertheless, as elaborated by No free lunch theorem [63], single optimization method able to solve every optimization challenge does not exist, necessitating experimentation with multiple methods prior to selecting the adequate optimizer. This manuscript suggests an altered variant of the famous sine cosine algorithm (SCA) [43] for tuning the XGBoost model for this peculiar task.

2 Related works and background

Several recent research papers dealt with the application of AI methods to detect unacceptable behavior on the Internet and cyberbullying. Application of explainable AI to identify cyberbullying in social media messages was discussed by [26], by applying NLP paired with ML methods, with promising results. Deep learning methods were investigated in [27], while ML approaches were compared to transfer learning algorithms for social networks in [60]. Other notable recent publications also mostly deal with evaluation of different models on social media [1, 59, 16]. The rest of this section introduces technologies utilized in the experiments in this research.

2.1 TF-IDF

Term Frequency and Inverse Document Frequency (TF-IDF) [54] represents a numerical statistic approach utilized for data retrieval and text mining aiming to mark how significant a word is within a document with respect to the set of documents. It is frequently employed in NLP for text classification and keyword extraction over the set of documents. As discussed in [21], TF-IDF is based on a pair of concepts: TF and IDF. The calculation of TF and IDF is depicted by Eq 1 and Eq 2:

$$\mathrm{TF}(t,d) = \frac{\text{Count of term } t \text{ located in document } d}{\text{total number of words in document } d} \tag{1}$$

$$\mathrm{IDF}(t,D) = \log\left(\frac{\text{Number of documents in } D}{\text{Count of documents containing term } t+1}\right) \tag{2}$$

The overall TF-IDF calculation is explained by the Eq 3:

$$\text{TF-IDF}(t,d,D) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t,D) \tag{3}$$

These final values showcase the relative importance of every term inside document relative to the overall set of documents, where terms having larger scores have greater relevance to the specific observed document.

2.2 XGBoost

The XGBoost approach [19] makes use of an ensemble learning method relying on the decision trees to integrate predictions made by a collection of weak classifiers. Every tree in this model tackles the faults introduced by its predecessors. XGBoost effectiveness is relying on its reqularization techniques and powerful parallel processing capabilities. XGBoost model is excelling in predictions by utilizing complex relations between input features and target pattern. Moreover, incremental training approach is employed to refine the objective function. XGBoost has a considerable number of hyperparameters that must be tuned for each classification task, which is an NP-hard challenge.

The model must be trained in iterations in order to generate the best predictions. The XGBoost model's fitness function is given by Eq. 4

$$\mathrm{obj}(\Theta) \;=\; L(\theta) \;+\; \Omega(\Theta), \tag{4}$$

above, Θ describes the collection of XGBoost control variables, $L(\Theta)$ denotes the loss function, while $\Omega(\Theta)$ corresponds to the regularization term. The final variable is controlling how complex is the model. The loss function depends on the particular classification task.

$$\mathrm{L}(\Theta) \;=\; \textstyle\sum_i \,(y_i \;-\; \hat{y_i})^2, \tag{5}$$

where y_i denotes the forecasted value, and $\hat{y_i}$ corresponds to the predicted target for every round i.

$$L(\Theta) = \sum_i [\ y_i \ln\ (1 + e^{-\hat{y}_i})\ +\ (1 - y_i) \ln\ (1 + e^{\hat{y}_i})]. \tag{6}$$

The objective of this algorithm is differentiation among actual and anticipated scores. The overall loss function needs to be minimized to leverage classification.

2.3 Stochastic optimization

Taking inspiration from the thriving of animals in extensive swarms and their utilization of group behavior, swarm intelligence techniques exhibit impressive efficacy in cases where individual efforts are not adequate for task completion. This approach has achieved significant success in dealing with NP-hard problems.

Algorithms falling under the swarm intelligence umbrella have demonstrated impressive proficiency in addressing broad spectrum of real-world challenges. Some prominent samples of their practical implementations include applications in medicine [39, 65, 10, 34, 41, 33, 69, 23, 62, 52], identification of the credit card scams [30, 46]. Additionally, swarm methods achieved impressive efficiency in cloud computing tasks [48, 8], plant identification [14], electricity predictions [58, 5], wide spectrum of economic challenges [35, 55, 57, 12, 51, 47], audit opinion enhancement [61], identification of software bugs [71], feature selection [6, 17, 31, 68], variety of problems falling in the area of computer and network security [67, 2, 53, 37, 50, 15, 36, 11], pollution prediction and environmental observation, [38, 32, 7, 42], optimization of IoT and WSNs [66, 4] and overall optimization of ML algorithms [9, 56, 13, 22, 49, 70, 3, 23, 18].

3 Methods

3.1 Original sine cosine algorithm

The SCA metaheuristics tackles optimization challenges by employing principles galvanized by mathematics to a collection of agents. It initiates with arbitrary solutions, subjecting them to multiple assessments prior to refining them withr respect to the foundational principles guiding its operation [43].

During the exploration stage, the SCA method makes use of both exploration and exploitation strategies to discover promising areas of the search domain. While this stage progresses, arbitrary solutions undergo smaller alterations compared to the exploitation stage. The SCA utilizes two major equations representing this pair of phases:

$$x_{ij}^{t+1} = x_{ij}^t + r_1 * sin(r_2) * |r_3 P_j^t - x_{ij}^t|, \tag{7}$$

$$x_{ij}^{t+1} = x_{ij}^t + r_1 * cos(r_2) * |r_3 P_j^t - x_{ij}^t|, \tag{8}$$

above, t depicts the current round, i denotes the solution, j marks the dimension, while r_1, , r_2, and r_3 mark the random values. The i-th individual's position over j-th dimension within t-th round of execution is depicted as $x_{ij(t)}$. The target in j-th dimension is given as p_j, while $||$ marks the absolute value.

This pair of equations can be exhibited mathematically as follows:

$$x_{ij}^{t+1} = \begin{cases} x_{ij}^t + r_1 \sin(r_2) \cdot |r_3 P_j^t - x_{ij}^t|, & \text{if } r_4 < .5 \\ x_{ij}^t + r_1 \cos(r_2) \cdot |r_3 P_j^t - x_{ij}^t|, & \text{if } r_4 \geq .5 \end{cases} \tag{9}$$

above, r_4 marks random value in range $[0, 1]$, and r_1 parameter guides fluctuations of X_i in different directions. In practice, in case of $r_1 > 1$, Xi proceeds in the direction of P. Otherwise, it goes away from P. Moreover, r_1 also controls the balance betwixt the exploration and exploitation procedures. The r_2 parameter is indicating the amplitude of solution's motion. The r_3 parameter denotes a random weight value utilized to either emphasize $(r_3 > 1)$ or reduce $(r_3 < 1)$ its effect to decide the distance. Sine and cosine properties are employed to account exploitation and growing through higher dimensions, necessitating adaptive alterations to keep the balance.

3.2 Modified SCA

Although baseline SCA is regarded as very potent optimization method, it was noted that it has certain drawbacks when executing comprehensive test on standard CEC [29] evaluation function set. These flaws have been addressed in this research by hybridization with some concepts taken from the genetic algorithm (GA) [45]. The novel method was simply named genetically inspired SCA (GISCA).

During the execution of GISCA, during the first $T/2$ iterations, the worst solution in every round is replaced by the novel solution synthesized as follows: a random new individual is produced within the borders of the search space, and then combined with the random existing solution from the population with the uniform crossover inherited from GA.

During the last $T/2$ rounds, the worst solution in each round is deleted from the populace, and substituted by the individual synthesized with the uniform crossover betwixt two best-performing individuals. Since neither alteration introduces additional fitness function evaluations, the GISCA complexity is identical to the baseline SCA. The pseudocode of GISCA is depicted within Algorithm 1.

Algorithm 1 GISCA algorithm

```
Produce initial population P
while t < T do
    Assess the individuals within P based on their fitness
    for Every individual X belonging to P do
        Update solutions by employing baseline SCA search procedure
    end for
    if t < T/2 then
        Produce a novel solution
        Perform uniform crossover between this novel solution and the random agent
        from P
    else
        Create novel individual by utilizing uniform crossover over a pair of best-
        performing individuals
    end if
    Replace the worst-performing individual in P by the hybrid solution synthesized
    in the previous step
end while
return The best agent from P
```

4 Simulation setup

The simulations were revolving around the publicly open dataset, that may be accessed on `https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset` It is a collection of data taken from a variety of social networks including Kaggle, Twitter and YouTube, where each entry is comprised of text and label (either marked as bullying or not bullying). Different sorts of bullying are tracked, including hate speech, aggressive comments, insulting others and toxic messages. Dataset was separated to 70%/30% subsets that were used to train and test the model. XGBoost was chosen to execute the classifying task, and the hyperparameters' that were optimized accompanied by the search limits were learning rate $[.1, .9]$, minimum child weight $[1, 10]$, subsample $[.1, 1.0]$, colsample bytree $[.01, 1.00]$, max depth $[3, 10]$ and γ $[0, .8]$.

Eight algorithms were employed to perform optimization of the above-mentioned XGBoost hyperparameters, namely the suggested GISCA, elementary SCA, GA, firefly algorithm (FA) [64], artificial bee colony (ABC) [40], whale optimization algorithm (WOA) [44], Harris hawk's optimization (HHO) [25] and crayfish optimization algorithm (COA) [28]. Every metaheuristics algorithm was given the population N of 10 individuals, 10 iterations in each run (T), and 30 separate executions. All experiments were developed in Python, and common set of libraries was used, like scikit-learn, pandas, scipy and seaborn.

Since dataset is not balanced, Cohen's kappa indicator κ has been selected as fitness function necessitating maximization, which is defined as follows [20]:

$$\kappa = \frac{k_o - k_e}{1 - k_e} = 1 - \frac{1 - k_o}{1 - k_e} \tag{10}$$

here, k_o and k_e denote vectors that contain observed and actual results. Cohen's κ parameter accounts disbalance among classes, allowing it to yield stronger forecasts than accuracy, that could be misleading in such situation.

5 Experimental outcomes

The simulation results are summarized in Tables 1 and 2, regarding the fitness function (Cohen's κ value) and indicator function (classification error), where the best value in each class is marekd with bold font.

Table 1. Harassment detection objective function outcomes for each optimizer.

Approach	Best	Worst	Mean	Median	Sd	Variance
XG-GISCA	**.466874**	**.452175**	**.458387**	**.459136**	**.003849**	**1.48E-05**
XG-SCA	.457893	.438936	.447574	.447855	.005897	3.48E-05
XG-GA	.458372	.442895	.450005	.449114	.004066	1.65E-05
XG-FA	.454448	.433291	.446607	.449356	.007072	5.00E-05
XG-ABC	.451148	.425347	.441123	.442637	.007888	6.22E-05
XG-WOA	.457918	.435196	.448488	.448384	.005662	3.21E-05
XG-HHO	.464081	.435196	.448616	.448015	.009017	8.13E-05
XG-COA	.461683	.439908	.451150	.451561	.006629	4.39E-05

Table 2. Harassment detection indicator function outcomes for each optimizer.

Approach	Best	Worst	Mean	Median	Sd	Variance
XG-GISCA	**.222222**	**.225756**	**.225794**	**.225750**	**.001926**	**3.71E-06**
XG-SCA	.228395	.228395	.231085	.231702	.002252	5.07E-06
XG-GA	.228836	.233686	.231261	.232804	.003024	9.15E-06
XG-FA	.232363	.238536	.233995	.233466	.002682	7.20E-06
XG-ABC	.229718	.233245	.231526	.231041	.001865	3.48E-06
XG-WOA	.230600	.237213	.232716	.233025	.002644	6.99E-06
XG-HHO	.223104	.237213	.231437	.233025	.004190	1.76E-05
XG-COA	.225309	.234127	.229718	.229718	.002300	5.29E-06

The results show the supremacy of the introduced GISCA, that attained superior scores in every observed category for both fitness and indicator, over thirty separate executions of each algorithm. The second best outcome was attained by XGBoost tuned by HHO, and COA metaheuristics finished in third place regarding the best metric.

To visualize the performance of the regarded metaheuristics algorithms for this peculiar task, Violin plots of Cohen's κ coefficient, accompanied by the box plots of the error rate over thirty runs are given in Fig. 1.

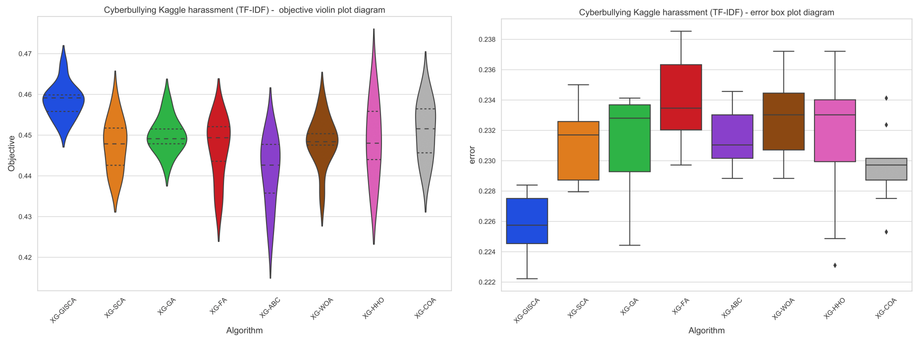

Fig. 1. Outcome distributions for each optimizer objective and indicator functions.

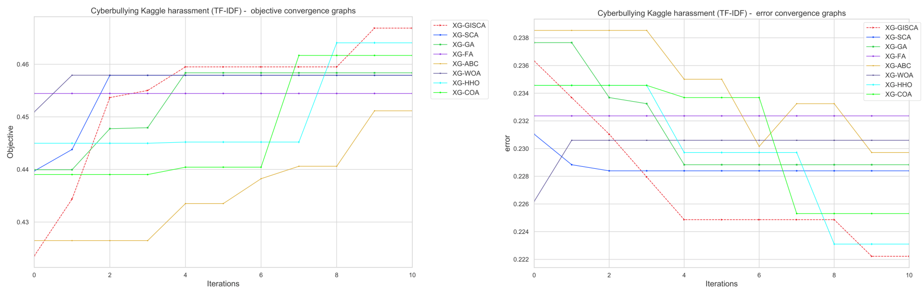

Fig. 2. Convergence rates for objective and indicator functions during optimization.

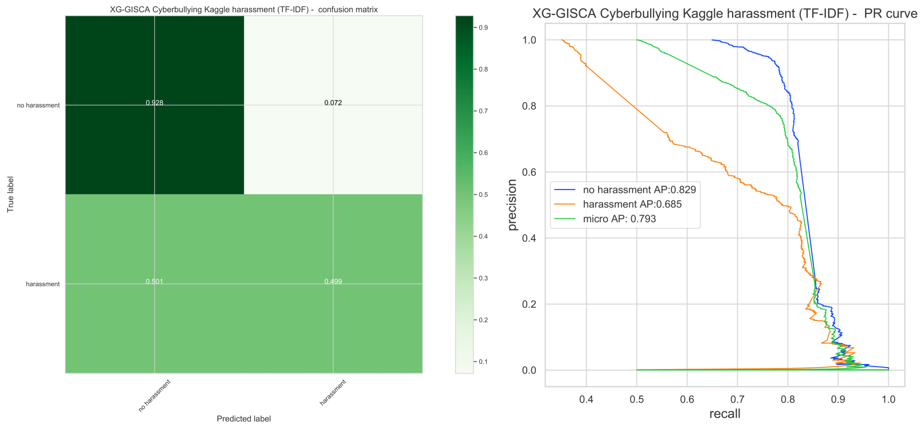

Fig. 3. Best performing model (optimized by GISCA) confusion matrix and PR curve.

Table 3. Harassment detection detailed metrics for each optimizer best constructed model.

Method	Metric	harassment	non-harassment	Accuracy	Macro avg.	Weighted avg.
XG-GISCA	prec.	.774504	.789264	**.777778**	**.781884**	**.779678**
	sens.	.928038	.499371	**.777778**	.713705	**.777778**
	f1-value	**.844348**	.611710	**.777778**	**.728029**	**.762802**
XG-SCA	prec.	**.775852**	.757914	.771605	.766883	.769564
	sens.	.911745	.511950	.771605	.711847	.771605
	f1-value	.838327	.611111	.771605	.724719	.758681
XG-GA	prec.	.777003	.752747	.771164	.764875	.768501
	sens.	.908350	.516981	.771164	.712666	.771164
	f1-value	.837559	.612975	.771164	.725267	.758836
XG-FA	prec.	.778892	.734266	.767637	.756579	.763249
	sens.	.896809	.528302	.767637	.712556	.767637
	f1-value	.833701	**.614484**	.767637	.724093	.756859
XG-ABC	prec.	.771689	.765504	.770282	.768597	.769521
	sens.	.917855	.496855	.770282	.707355	.770282
	f1-value	.838450	.602593	.770282	.720522	.755775
XG-WOA	prec.	.779412	.739437	.769400	.759424	.765399
	sens.	.899525	.528302	.769400	**.713913**	.769400
	f1-value	.835172	.616288	.769400	.725730	.758446
XG-HHO	prec.	.773318	**.789579**	.776896	.781449	.779018
	sens.	**.928717**	.495597	.776896	.712157	.776896
	f1-value	.843924	.608964	.776896	.726444	.761564
XG-COA	prec.	.774543	.775194	.774691	.774869	.774771
	sens.	.921249	.503145	.774691	.712197	.774691
	f1-value	.841550	.610221	.774691	.725886	.760463
	entries	1473	795			

Table 4. Parameter selections made by each optimizer for the respective best performing models.

Method	Learning Rate	Min Child W.	Subsample	Col by Tree	Max depth	Gamma
XG-GISCA	.900000	4.380805	.894683	.642173	9	.751739
XG-SCA	.768268	1.000000	.998044	1.000000	5	.000000
XG-GA	.900000	1.066000	1.000000	1.000000	5	.800000
XG-FA	.900000	3.493051	1.000000	1.000000	10	.800000
XG-ABC	.862100	1.000000	.884828	.741802	9	.438964
XG-WOA	.900000	1.450853	1.000000	1.000000	7	.800000
XG-HHO	.900000	5.335281	1.000000	.630253	8	.800000
XG-COA	.874073	1.000000	.921518	.501137	9	.447086

Moreover, Fig. 2 provides meaningful insight to the convergence rates of both fitness and indicator during the best run achieved by each regarded algorithm. Suggested GISCA attains superior convergence and avoids local optimums, which may hinder the outcomes in case of prematurely converging to less favourable areas, exhibited by other algorithms like WOA and baseline SCA.

The top-performing XGBoost models generated by every metaheuristics were analyzed in details in Table 3. GISCA attained superior accuracy of approximately 77.8%, followed by the HHO and COA. GISCA also attained the best results in the majority of the observed metrics, however, it must be said that other metaheuristics algorithms performed very well. Finally, to make the replication of the simulations easier, the best established collection of XGBoost parameter values by each algorithm is shown in Table 4. PR curve accompanied by the confusion matrix of XGBoost-GISCA classifier are presented in Fig. 3.

6 Conclusion

This study examined the ability of hybrid metaheuristics-XGBoost classifier to perform the cyberbullying classification problem. Cyberbullying is a considerable contemporary problem, that may affect the victims dearly. Proper classification of the content on social networks may help in faster identification of inappropriate behavior and prompt intervention. XGBoost classifier has been optimized with the help of a modified SCA metaheuristics, while the results have been validated against the scores of other potent algorithms. The suggested XGBoost-GISCA structure attained supreme accuracy of 77.8%.

The limitations of the study must also be emphasized here. The experiments are very computationally intensive, consequently a limited count for algorithms was evaluated, with reduced number of solutions in the population and rounds per run. Search space of each hyperparameter has also been constrained. Future works will aim to tackle these limitations, in case additional computing resources are obtained.

Acknowledgment

This research was supported by the Science Fund of the Republic of Serbia, Grant No. 7373, Characterizing crises-caused air pollution alternations using an artificial intelligence-based framework - crAIRsis and Grant No. 7502, Intelligent Multi-Agent Control and Optimization applied to Green Buildings and Environmental Monitoring Drone Swarms - ECOSwarm.

References

1. Joseph D Akinyemi, Ayodeji OJ Ibitoye, Christianah T Oyewale, and Olufade FW Onifade. Cyberbullying detection and classification in social media texts using machine learning techniques. In *International Conference on Computer Science, Engineering and Education Applications*, pages 440–449. Springer, 2023.
2. Nadheera AlHosni, Luka Jovanovic, Milos Antonijevic, Milos Bukumira, Miodrag Zivkovic, Ivana Strumberger, Joseph P Mani, and Nebojsa Bacanin. The xgboost model for network intrusion detection boosted by enhanced sine cosine algorithm. In *International Conference on Image Processing and Capsule Networks*, pages 213–228. Springer, 2022.
3. Nebojsa Bacanin, Milos Antonijevic, Nikola Vukobrat, Timea Bezdan, and Miodrag Zivkovic. Enhanced seagull optimization algorithm adapted for artificial neural network training. In *ICT with Intelligent Applications: Proceedings of ICTIS 2021, Volume 1*, pages 753–761. Springer, 2022.
4. Nebojsa Bacanin, Uros Arnaut, Miodrag Zivkovic, Timea Bezdan, and Tarik A Rashid. Energy efficient clustering in wireless sensor networks by opposition-based initialization bat algorithm. In *Computer Networks and Inventive Communication Technologies: Proceedings of Fourth ICCNCT 2021*, pages 1–16. Springer, 2022.
5. Nebojsa Bacanin, Luka Jovanovic, Miodrag Zivkovic, Venkatachalam Kandasamy, Milos Antonijevic, Muhammet Deveci, and Ivana Strumberger. Multivariate energy forecasting via metaheuristic tuned long-short term memory and gated recurrent unit neural networks. *Information Sciences*, 642:119122, 2023.
6. Nebojsa Bacanin, Aleksandar Petrovic, Miodrag Zivkovic, Timea Bezdan, and Milos Antonijevic. Feature selection in machine learning by hybrid sine cosine metaheuristics. In *International Conference on Advances in Computing and Data Sciences*, pages 604–616. Springer, 2021.

7. Nebojsa Bacanin, Marko Sarac, Nebojsa Budimirovic, Miodrag Zivkovic, Ahmad Ali AlZubi, and Ali Kashif Bashir. Smart wireless health care system using graph lstm pollution prediction and dragonfly node localization. *Sustainable Computing: Informatics and Systems*, 35:100711, 2022.
8. Nebojsa Bacanin, Vladimir Simic, Miodrag Zivkovic, Melfi Alrasheedi, and Aleksandar Petrovic. Cloud computing load prediction by decomposition reinforced attention long short-term memory network optimized by modified particle swarm optimization algorithm. *Annals of Operations Research*, pages 1–34, 2023.
9. Nebojsa Bacanin, Catalin Stoean, Miodrag Zivkovic, Dijana Jovanovic, Milos Antonijevic, and Djordje Mladenovic. Multi-swarm algorithm for extreme learning machine optimization. *Sensors*, 22(11):4204, 2022.
10. Nebojsa Bacanin, Miodrag Zivkovic, Fadi Al-Turjman, K Venkatachalam, Pavel Trojovskỳ, Ivana Strumberger, and Timea Bezdan. Hybridized sine cosine algorithm with convolutional neural networks dropout regularization application. *Scientific Reports*, 12(1):6302, 2022.
11. Nebojsa Bacanin, Miodrag Zivkovic, Milos Antonijevic, K Venkatachalam, Jinseok Lee, Yunyoung Nam, Marina Marjanovic, Ivana Strumberger, and Mohamed Abouhawwash. Addressing feature selection and extreme learning machine tuning by diversity-oriented social network search: an application for phishing websites detection. *Complex & Intelligent Systems*, 9(6):7269–7304, 2023.
12. Nebojsa Bacanin, Miodrag Zivkovic, Luka Jovanovic, Milica Ivanovic, and Tarik A Rashid. Training a multilayer perception for modeling stock price index predictions using modified whale optimization algorithm. In *Computational Vision and Bio-Inspired Computing: Proceedings of ICCVBIC 2021*, pages 415–430. Springer, 2022.
13. Nebojsa Bacanin, Miodrag Zivkovic, Mohamed Salb, Ivana Strumberger, and Amit Chhabra. Convolutional neural networks hyperparameters optimization using sine cosine algorithm. In *Sentimental Analysis and Deep Learning: Proceedings of ICSADL 2021*, pages 863–878. Springer, 2022.
14. Nebojsa Bacanin, Miodrag Zivkovic, Marko Sarac, Aleksandar Petrovic, Ivana Strumberger, Milos Antonijevic, Andrija Petrovic, and K Venkatachalam. A novel multiswarm firefly algorithm: An application for plant classification. In *International Conference on Intelligent and Fuzzy Systems*, pages 1007–1016. Springer, 2022.
15. Nebojsa Bacanin, Miodrag Zivkovic, Catalin Stoean, Milos Antonijevic, Stefana Janicijevic, Marko Sarac, and Ivana Strumberger. Application of natural language processing and machine learning boosted with swarm intelligence for spam email filtering. *Mathematics*, 10(22):4173, 2022.
16. Vimala Balakrisnan and Mohammed Kaity. Cyberbullying detection and machine learning: a systematic literature review. *Artificial Intelligence Review*, 56(Suppl 1):1375–1416, 2023.
17. Timea Bezdan, Dusan Cvetnic, Luka Gajic, Miodrag Zivkovic, Ivana Strumberger, and Nebojsa Bacanin. Feature selection by firefly algorithm with improved initialization strategy. In *7th conference on the engineering of computer based systems*, pages 1–8, 2021.
18. Milos Bukumira, Miodrag Zivkovic, Milos Antonijevic, Luka Jovanovic, Nebojsa Bacanin, and Tamara Zivkovic. The extreme gradient boosting method optimized by hybridized sine cosine metaheuristics for ship vessel classification. In *International Conference on Advances in Data-driven Computing and Intelligent Systems*, pages 255–270. Springer, 2023.

19. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.
20. Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.
21. Mamata Das, PJA Alphonse, et al. A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset. *arXiv preprint arXiv:2308.04037*, 2023.
22. Masa Gajevic, Nemanja Milutinovic, Jelena Krstovic, Luka Jovanovic, Marina Marjanovic, and Catalin Stoean. Artificial neural network tuning by improved sine cosine algorithm for healthcare 4.0. In *Proceedings of the 1st international conference on innovation in information technology and business (ICIITB 2022)*, volume 104, page 289. Springer Nature, 2023.
23. Luka Gajic, Dusan Cvetnic, Miodrag Zivkovic, Timea Bezdan, Nebojsa Bacanin, and Stefan Milosevic. Multi-layer perceptron training using hybridized bat algorithm. In *Computational Vision and Bio-Inspired Computing: ICCVBIC 2020*, pages 689–705. Springer, 2021.
24. Vincenzo Gattulli, Donato Impedovo, Giuseppe Pirlo, and Lucia Sarcinella. Cyber aggression and cyberbullying identification on social networks. In *ICPRAM*, pages 644–651, 2022.
25. Ali Asghar Heidari, Seyedali Mirjalili, Hossam Faris, Ibrahim Aljarah, Majdi Mafarja, and Huiling Chen. Harris hawks optimization: Algorithm and applications. *Future generation computer systems*, 97:849–872, 2019.
26. Mohammad Rafsun Islam, Ahmed Saleh Bataineh, and Mohammad Zulkernine. Detection of cyberbullying in social media texts using explainable artificial intelligence. In *International Conference on Ubiquitous Security*, pages 319–334. Springer, 2023.
27. Celestine Iwendi, Gautam Srivastava, Suleman Khan, and Praveen Kumar Reddy Maddikunta. Cyberbullying detection solutions based on deep learning architectures. *Multimedia Systems*, 29(3):1839–1852, 2023.
28. Heming Jia, Honghua Rao, Changsheng Wen, and Seyedali Mirjalili. Crayfish optimization algorithm. *Artificial Intelligence Review*, 56(Suppl 2):1919–1979, 2023.
29. Shouyong Jiang, Shengxiang Yang, Xin Yao, Kay Chen Tan, Marcus Kaiser, and Natalio Krasnogor. Benchmark functions for the cec'2018 competition on dynamic multiobjective optimization. Technical report, Newcastle University, 2018.
30. Dijana Jovanovic, Milos Antonijevic, Milos Stankovic, Miodrag Zivkovic, Marko Tanaskovic, and Nebojsa Bacanin. Tuning machine learning models using a group search firefly algorithm for credit card fraud detection. *Mathematics*, 10(13):2272, 2022.
31. Dijana Jovanovic, Marina Marjanovic, Milos Antonijevic, Miodrag Zivkovic, Nebojsa Budimirovic, and Nebojsa Bacanin. Feature selection by improved sand cat swarm optimizer for intrusion detection. In *2022 International Conference on Artificial Intelligence in Everything (AIE)*, pages 685–690. IEEE, 2022.
32. Gordana Jovanovic, Mirjana Perisic, Nebojsa Bacanin, Miodrag Zivkovic, Svetlana Stanisic, Ivana Strumberger, Filip Alimpic, and Andreja Stojic. Potential of coupling metaheuristics-optimized-xgboost and shap in revealing pahs environmental fate. *Toxics*, 11(4):394, 2023.
33. Luka Jovanovic, Nebojsa Bacanin, Miodrag Zivkovic, Milos Antonijevic, Aleksandar Petrovic, and Tamara Zivkovic. Anomaly detection in ecg using recurrent

networks optimized by modified metaheuristic algorithm. In *2023 31st Telecommunications Forum (TELFOR)*, pages 1–4. IEEE, 2023.
34. Luka Jovanovic, Marko Djuric, Miodrag Zivkovic, Dijana Jovanovic, Ivana Strumberger, Milos Antonijevic, Nebojsa Budimirovic, and Nebojsa Bacanin. Tuning xgboost by planet optimization algorithm: An application for diabetes classification. In *Proceedings of fourth international conference on communication, computing and electronics systems: ICCCES 2022*, pages 787–803. Springer, 2023.
35. Luka Jovanovic, Dejan Jovanovic, Nebojsa Bacanin, Ana Jovancai Stakic, Milos Antonijevic, Hesham Magd, Ravi Thirumalaisamy, and Miodrag Zivkovic. Multistep crude oil price prediction based on lstm approach tuned by salp swarm algorithm with disputation operator. *Sustainability*, 14(21):14616, 2022.
36. Luka Jovanovic, Dijana Jovanovic, Milos Antonijevic, Bosko Nikolic, Nebojsa Bacanin, Miodrag Zivkovic, and Ivana Strumberger. Improving phishing website detection using a hybrid two-level framework for feature selection and xgboost tuning. *Journal of Web Engineering*, 22(3):543–574, 2023.
37. Luka Jovanovic, Dijana Jovanovic, Milos Antonijevic, Miodrag Zivkovic, Nebojsa Budimirovic, Ivana Strumberger, and Nebojsa Bacanin. The xgboost tuning by improved firefly algorithm for network intrusion detection. In *2022 24th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 268–275. IEEE, 2022.
38. Luka Jovanovic, Gordana Jovanovic, Mirjana Perisic, Filip Alimpic, Svetlana Stanisic, Nebojsa Bacanin, Miodrag Zivkovic, and Andreja Stojic. The explainable potential of coupling metaheuristics-optimized-xgboost and shap in revealing vocs' environmental fate. *Atmosphere*, 14(1):109, 2023.
39. Luka Jovanovic, Miodrag Zivkovic, Milos Antonijevic, Dijana Jovanovic, Milica Ivanovic, and Hothefa Shaker Jassim. An emperor penguin optimizer application for medical diagnostics. In *2022 IEEE Zooming Innovation in Consumer Technologies Conference (ZINC)*, pages 191–196. IEEE, 2022.
40. Dervis Karaboga and Bahriye Basturk. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (abc) algorithm. *Journal of global optimization*, 39:459–471, 2007.
41. Stanislava Kozakijevic, Mohamed Salb, Ali Elsadai, Joseph Mani, Kanchana Devi, Anni Dasho Sharko, and Suresh Muthusamy. Seizure detection via time series classification using modified metaheuristic optimized recurrent networks. *Theoretical and Applied Computational Intelligence*, 1(1):82–94, 2023.
42. Joseph Mani, Hotefa Shaker, Luka Jovanovic, et al. Sunspot occurrence forecasting with metaheuristic optimized recurrent neural networks. *Theoretical and Applied Computational Intelligence*, 1(1):15–26, 2023.
43. Seyedali Mirjalili. Sca: a sine cosine algorithm for solving optimization problems. *Knowledge-based systems*, 96:120–133, 2016.
44. Seyedali Mirjalili and Andrew Lewis. The whale optimization algorithm. *Advances in engineering software*, 95:51–67, 2016.
45. Seyedali Mirjalili and Seyedali Mirjalili. Genetic algorithm. *Evolutionary algorithms and neural networks: Theory and applications*, pages 43–55, 2019.
46. Aleksandar Petrovic, Nebojsa Bacanin, Miodrag Zivkovic, Marina Marjanovic, Milos Antonijevic, and Ivana Strumberger. The adaboost approach tuned by firefly metaheuristics for fraud detection. In *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)*, pages 834–839. IEEE, 2022.
47. Aleksandar Petrovic, Luka Jovanovic, Miodrag Zivkovic, Nebojsa Bacanin, Nebojsa Budimirovic, and Marina Marjanovic. Forecasting bitcoin price by tuned long short

term memory model. In *1st International conference on innovation in information technology and business (ICIITB 2022)*, pages 187–202. Atlantis Press, 2023.
48. Bratislav Predić, Luka Jovanovic, Vladimir Simic, Nebojsa Bacanin, Miodrag Zivkovic, Petar Spalevic, Nebojsa Budimirovic, and Milos Dobrojevic. Cloud-load forecasting via decomposition-aided attention recurrent neural network tuned by modified particle swarm optimization. *Complex & Intelligent Systems*, pages 1–21, 2023.
49. Mohamed Salb, Nebojsa Bacanin, Miodrag Zivkovic, Milos Antonijevic, Marina Marjanovic, and Ivana Strumberger. Extreme learning machine tuning by original sine cosine algorithm. In *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)*, pages 143–148. IEEE, 2022.
50. Mohamed Salb, Luka Jovanovic, Nebojsa Bacanin, Milos Antonijevic, Miodrag Zivkovic, Nebojsa Budimirovic, and Laith Abualigah. Enhancing internet of things network security using hybrid cnn and xgboost model tuned via modified reptile search algorithm. *Applied Sciences*, 13(23):12687, 2023.
51. Mohamed Salb, Miodrag Zivkovic, Nebojsa Bacanin, Amit Chhabra, and M Suresh. Support vector machine performance improvements for cryptocurrency value forecasting by enhanced sine cosine algorithm. In *Computer Vision and Robotics: Proceedings of CVR 2021*, pages 527–536. Springer, 2022.
52. Marko Šarac, Nikola Pavlović, Nebojsa Bacanin, Fadi Al-Turjman, and Saša Adamović. Increasing privacy and security by integrating a blockchain secure interface into an iot device security gateway architecture. *Energy Reports*, 7:8075–8082, 2021.
53. Nikola Savanović, Ana Toskovic, Aleksandar Petrovic, Miodrag Zivkovic, Robertas Damaševičius, Luka Jovanovic, Nebojsa Bacanin, and Bosko Nikolic. Intrusion detection in healthcare 4.0 internet of things systems via metaheuristics optimized machine learning. *Sustainability*, 15(16):12563, 2023.
54. Julia Silge and David Robinson. Term frequency and inverse document frequency (tf-idf) using tidy data principles. *CRAN–R project. URL: https://cran. rproject. org/web/packages/tidytext/vignettes/tf_idf. html*, 2018.
55. Marko Stankovic, Nebojsa Bacanin, Miodrag Zivkovic, Luka Jovanovic, Joseph Mani, and Milos Antonijevic. Forecasting ethereum price by tuned long short-term memory model. In *2022 30th Telecommunications Forum (TELFOR)*, pages 1–4. IEEE, 2022.
56. Marko Stankovic, Jelena Gavrilovic, Dijana Jovanovic, Miodrag Zivkovic, Milos Antonijevic, Nebojsa Bacanin, and Milos Stankovic. Tuning multi-layer perceptron by hybridized arithmetic optimization algorithm for healthcare 4.0. *Procedia Computer Science*, 215:51–60, 2022.
57. Marko Stankovic, Luka Jovanovic, Nebojsa Bacanin, Miodrag Zivkovic, Milos Antonijevic, and Petar Bisevac. Tuned long short-term memory model for ethereum price forecasting through an arithmetic optimization algorithm. In *International Conference on Innovations in Bio-Inspired Computing and Applications*, pages 327–337. Springer, 2022.
58. Catalin Stoean, Miodrag Zivkovic, Aleksandra Bozovic, Nebojsa Bacanin, Roma Strulak-Wójcikiewicz, Milos Antonijevic, and Ruxandra Stoean. Metaheuristic-based hyperparameter tuning for recurrent deep learning: application to the prediction of solar energy generation. *Axioms*, 12(3):266, 2023.
59. Daniyar Sultan, Aigerim Toktarova, Ainur Zhumadillayeva, Sapargali Aldeshov, Shynar Mussiraliyeva, Gulbakhram Beissenova, Abay Tursynbayev, Gulmira Baenova, and Aigul Imanbayeva. Cyberbullying-related hate speech detection using shallow-to-deep learning. *Computers, Materials & Continua*, 75(1), 2023.

60. Teoh Hwai Teng and Kasturi Dewi Varathan. Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. *IEEE Access*, 2023.
61. Mihailo Todorovic, Nemanja Stanisic, Miodrag Zivkovic, Nebojsa Bacanin, Vladimir Simic, and Erfan Babaee Tirkolaee. Improving audit opinion prediction accuracy using metaheuristics-tuned xgboost algorithm with interpretable results through shap value analysis. *Applied Soft Computing*, 149:110955, 2023.
62. Milan Tuba and Nebojsa Bacanin. Jpeg quantization tables selection by the firefly algorithm. In *2014 International Conference on Multimedia Computing and Systems (ICMCS)*, pages 153–158. IEEE, 2014.
63. David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.
64. Xin-She Yang. Firefly algorithm, stochastic test functions and design optimisation. *International journal of bio-inspired computation*, 2(2):78–84, 2010.
65. Miodrag Zivkovic, Nebojsa Bacanin, Milos Antonijevic, Bosko Nikolic, Goran Kvascev, Marina Marjanovic, and Nikola Savanovic. Hybrid cnn and xgboost model tuned by modified arithmetic optimization algorithm for covid-19 early diagnostics from x-ray images. *Electronics*, 11(22):3798, 2022.
66. Miodrag Zivkovic, Nebojsa Bacanin, Tamara Zivkovic, Ivana Strumberger, Eva Tuba, and Milan Tuba. Enhanced grey wolf algorithm for energy efficient wireless sensor networks. In *2020 zooming innovation in consumer technologies conference (ZINC)*, pages 87–92. IEEE, 2020.
67. Miodrag Zivkovic, Luka Jovanovic, Milica Ivanovic, Nebojsa Bacanin, Ivana Strumberger, and P Mani Joseph. Xgboost hyperparameters tuning by fitness-dependent optimizer for network intrusion detection. In *Communication and intelligent systems: Proceedings of ICCIS 2021*, pages 947–962. Springer, 2022.
68. Miodrag Zivkovic, Luka Jovanovic, Milica Ivanovic, Aleksa Krdzic, Nebojsa Bacanin, and Ivana Strumberger. Feature selection using modified sine cosine algorithm with covid-19 dataset. In *Evolutionary computing and mobile sustainable networks: Proceedings of ICECMSN 2021*, pages 15–31. Springer, 2022.
69. Miodrag Zivkovic, Catalin Stoean, Aleksandar Petrovic, Nebojsa Bacanin, Ivana Strumberger, and Tamara Zivkovic. A novel method for covid-19 pandemic information fake news detection based on the arithmetic optimization algorithm. In *2021 23rd international symposium on symbolic and numeric algorithms for scientific computing (SYNASC)*, pages 259–266. IEEE, 2021.
70. Miodrag Zivkovic, Ana Vesic, Nebojsa Bacanin, Ivana Strumberger, Milos Antonijevic, Luka Jovanovic, and Marina Marjanovic. An improved animal migration optimization approach for extreme learning machine tuning. In *International Conference on Intelligent and Fuzzy Systems*, pages 3–13. Springer, 2022.
71. Tamara Zivkovic, Bosko Nikolic, Vladimir Simic, Dragan Pamucar, and Nebojsa Bacanin. Software defects prediction by metaheuristics tuned extreme gradient boosting and analysis based on shapley additive explanations. *Applied Soft Computing*, 146:110659, 2023.