Metaheuristic Optimized Ensemble Model for Classification of SMS Spam in Computer Networks

Mohamed Saber[1*], El-Sayed M. El-Kenawy[2], Abdelhameed Ibrahim[3], Marwa M. Eid[4], Abdelaziz A. Abdelhamid[5]

[1]Electronics and Communications Engineering Dep., Faculty of Engineering, Delta University for Science and Technology, Gamasa City, Mansoura, Egypt
[2]Department of Communications and Electronics, Delta Higher Institute of Engineering and Technology,

Mansoura, 35111, Egypt

[3]Computer Engineering and Control Systems Department, Faculty of Engineering, Mansoura University,

35516, Mansoura Egypt

[4]Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura, Egypt

[5]Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, 11566, Egypt
Emails: Mohamed.saber@deltauniv.edu.eg; skenawy@ieee.org; afai79@mans.edu.eg; mmm@ieee.org; abdelaziz@cis.asu.edu.eg

Abstract

By use of electronic communication, we are able to communicate a message to the recipient. In this digital age, a collaboration between several people is possible thanks to a variety of digital technologies. This interaction may take place in a variety of media formats, including but not limited to text, images, sound, and language. Today, a person's primary means of communication is their smart gadget, most commonly a cell phone. Spam is another side effect of our increasingly text-based modes of communication. We received a bunch of spam texts on our phones, and we know they're not from anyone we know. The vast majority of businesses nowadays use spam texts to advertise their wares, even when recipients have explicitly requested not to receive such messages. As a rule, there are many more spam emails than genuine ones. We apply text classification approaches to define short messaging service (SMS) and spam filtering in this study, which effectively categorizes messages. In this paper, we use "machine learning algorithms" and metaheuristic optimization to determine what percentage of incoming SMS messages are spam. This is why we used the optimized models to evaluate and contrast many classification strategies for gathering data.

Keywords: Support vector machine; Neural network; Computer Networks; Decision tree; Voting ensemble; Dipper throated optimization.

1. Introduction

With its roots in data mining, deep learning, and big data, data science is an interdisciplinary subject that uses experimental approaches, algorithms, processes, and systems to extract meaningful information and observation from large amounts of structured and unstructured data. Computer scientists employ a process called "data mining" to sift through large amounts of unstructured data in

DOI: https://doi.org/10.54216/IJWAC.060205
Received: September 11, 2022 Accepted: November 22, 2022

order to extract meaningful insights. And the system relied heavily on data for its interpretation. Classification, grouping, and other similar techniques are only some of the numerous available options. SMS stands for short message service. When communicating over SMS, each message can only be 160 characters long, and longer ones must be broken up into several shorter ones [1]. To make it possible for mobile phones to exchange brief text messages, it made advantage of existing communication standards. The government's goal is to keep up with the rapid pace of technological progress, and the number of texts sent has grown over the past several years [3]. SMS spam is more advanced than it formerly was. Due to the low cost of sending and receiving SMS messages, both consumers and service providers have abandoned the challenge and limited window of opportunity presented by mobile spam filtering software [2]. Compared to email, spam sent by text message is rather rare. Even while it accounts for around 1% of transcripts mailed to the US and 30% of typewritten correspondence mailed to modern Asia, this is still a significant number. According to the Telephone Consumer Protection Act of 2004, spam text messages were outlawed in the United States. Whoever receives unsolicited text messages from foreign nations already knows how to tag along with the guidance counselor in the direction of a trivial situation court. In China, the three largest movable cell phone hands have been working together since 2009 [3] on a unified approach to combat mobile unwanted messaging by imposing limits on the total number of messages of a certain format sent to a single user.

Here we provide a few examples of categorization algorithms' practical use in the real world. As a means of determining whether or not a communication is a spam, we employ classification algorithms. A training set containing the necessary objects must be available at this location [4]. The contents of an SMS consist of text messages. SMS spam class or SMS to a human being; those texts sent by individuals, meaning that these messages from human class or mobile phone, and spam messages often originate from organizations and firms to sell their products via adverts; this is why we use text classification in this study. With so many people carrying and using mobile phones and smartphones, it's no surprise that voice mail is a common form of communication. As a result of their often small size, SMS spam datasets are fewer in number compared to email filter spam datasets. Due to the compact nature of spam SMS, a system based on the filtering techniques generally used to combat spam in the email would be ineffective [5]. Email spam is less common than spam text messages (SMS) in several countries, including Korea. However, the reverse approach was used in the West, where email spam was more common due to its low cost than SMS spamming, as SMS spamming is both more expensive and produces fewer messages [3]. As much as half of all SMS messages received appear spammy on mobile devices [6]. It is for this reason that an SMS filtering system has to be as resource-efficient as possible, functioning in both primary and secondary memory on mobile devices. This analysis made use of real information, such as ham and spam. We employ many categorization methods, some of which have been used before and others that are novel, to provide data for further study and comparison.

2. Literature Review

As a new controversy, SMS spamming is widely expected to grow into a major poorly mannered or subject of concern in the future. Maintain a constant 360° surveillance arc over the connected work zone. By combining the KNN ordering technique with a bumpy set, Duan and Huang separate spam SMS from ham with a double run through a filter [9]. Despite the fact that this has been shown to cause errors in height accuracy, it is nonetheless shown as having been consumed as part of the haste with which sorting is conducted. Problems with script cataloging are linked to the unwanted passage through a filter structure, and a classifier is needed to group together similar versions written in different discussion brands. A few of the parts of this hierarchy are dualistic [7]. To begin with, it is a fresh way of gathering data because the ongoing group required a means to catalog texts. As an added bonus, the above may be sorted. High-caliber search terms are used in [8]. Moreover, every category has its own signature set of keywords, and each of those phrases must also be applicable to loads. The popularity of those keywords is then shared among the masses to form the categorization system. There are several grant-related feature collection instruments [9] since there are so many grants that need to be acquired. For the purpose of identifying SMS spam, it is necessary to place a test on a collection of cataloging strategies. Support Vector Machines (SVMs) [10] were reused as part of the most rudimentary approach to the problem of SMS spam exposure. This development effort at the proper entry layer [11] and the use of SMS spam straining categorization lead to new learnings. A unified SMS corpus consisting of 2,000 spam messages and 4,000 useful ones is now in the works. Standard two-octet creäte options include a random bytes-per-second rate and an octet [12-14]. Our work has led to a significant improvement in performance compared to other classifiers, making Bayes classifiers the preferred choice in many applications. There are a select few cellular

DOI: https://doi.org/10.54216/IJWAC.060205
Received: September 11, 2022 Accepted: November 22, 2022

machinists that have spent some time in the past preparing a demanding design, much as Open Mobile Alliance, to distinguish the outset as SMS spam. There are significant differences between the methods used to identify spam in electronic messages, which are highlighted by these performances[15-18].

3. Proposed Methodology

Now that we've gotten our research structured, we can utilize categorization techniques from machine learning. Before using our dataset, it must be cleaned up (preprocessed) to ensure accuracy. Next, we examine whether or not any of our attribute's values are blank. Once we use the appropriate algorithms, we can obtain reliable data in near-real time. Figure 3 shows the overall structure of our research.

A. Dataset

The kaggle.com spam SMS dataset is available for download. Out of the 611 records in this database, some are undoubtedly Spam and others are undoubtedly Ham as shown in Figure 1.

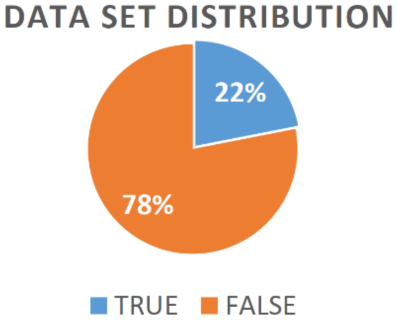

Figure 1: Distribution of the dataset samples

B. Formulating the dataset

CSV records are still the default for the dataset being consumed. Each of these LPs was given its own set of handwritten liner notes. There is just one additional field except for the integer id that defines the route, and that is the sticky label "is spam." A zip file including multiple script libraries and a memorandum constitutes the Spam SMS collection. We need to conduct some preprocessing so that the oblique extent classifier can make use of the data.

C. Data preprocessing

Not all classifiers have the same preprocessing strategy, and this affects the documents they require for training. Using this approach, we can determine whether or not there is a missing value in our dataset and then fill it in.

D. Training and Testing Sets

In the first, we were split into the files that self-control remained used to boat train our classifiers, and in the second, we were split into the files that were used to test those classifiers. We divided our data collection such that the most comprehensive part could be used right away to make crucial and difficult preparations as shown in Figure 2.

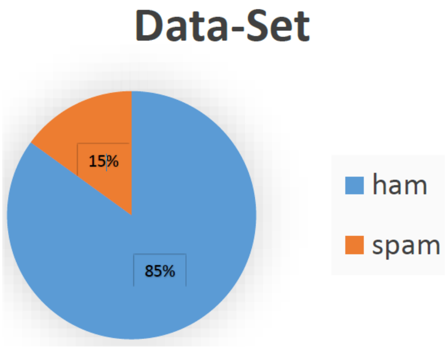

Figure 2: Dataset Split

DOI: https://doi.org/10.54216/IJWAC.060205
Received: September 11, 2022 Accepted: November 22, 2022

E. The Proposed Ensemble Model

The first shows the files that were used to boat-train our classifiers after we exercised some restraint and let go of the reins, and the third depicts the recommended process. There were three different types of basic classifiers used to create this diagram. Decision trees (DT), multilayer perceptron (MLP), and support vector machines are the aforementioned classifiers (SVM). When utilized in a voting ensemble model, these classifiers' votes are optimized using a DTO/PSO metaheuristic hybrid (PSO).

F. Support Vector Machines (SVM)

As a subset of supervised learning methods, support vector machines (SVMs) are commonly employed for the purpose of classification. To restate, it can predict the classification of unseen samples if given a set of training data that have previously been divided into two categories. Using a dataset with examples annotated in various ways, the SVM training approach creates a model for generating predictions. To help in the classification process, support vector machines (SVMs) create a hyperplane or a series of hyperplanes in a high-dimensional space. The big picture of the SVM method is seen in Figure 3.

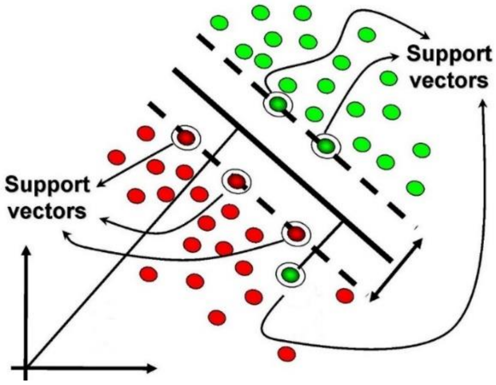

Figure 3: Structure of support vector machines.

G. Multilayer Perceptron (MLP)

A neural network may be defined as a group of interconnected nodes (neurons) that communicate with one another via connections (synapses). Due to their remarkable similarity to the actual thing, estimates are frequently recreated using artificial neural networks that are inspired by the human nervous system. Every artificial neural network consists of three distinct parts: an input layer, a hidden layer, and an output layer. A set of input layer nodes receive data and produce an activation function using this technique. The weighting layer, which is concealed from view, stands between the input and output layers, giving the former more influence over the latter. Ultimately, it is the output layer that provides the result. A multilayer neural network, like the one shown in Figure 4, has many interconnected nodes.

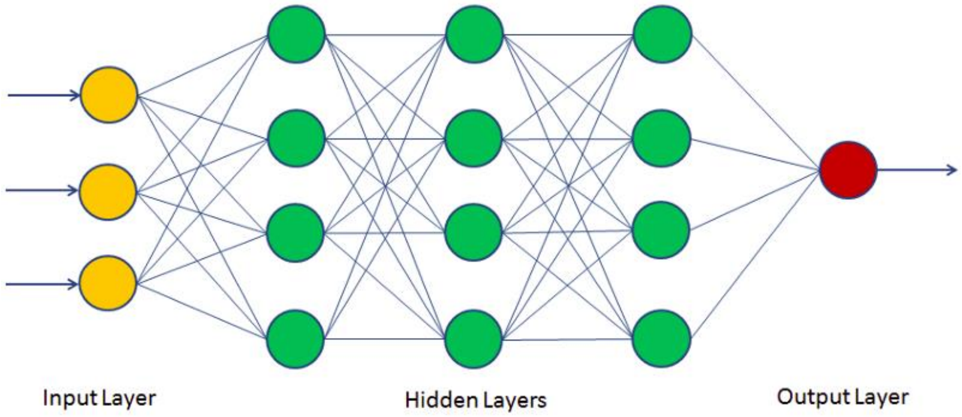

Figure 4: Structure of a multilayer neural network.

DOI: https://doi.org/10.54216/IJWAC.060205
Received: September 11, 2022 Accepted: November 22, 2022

H. Decision Trees (DT)

To find the best branching points in a decision tree, this method of issue-solving employs a greedy search. Once some data has been partitioned, the procedure is iterated from highest to lowest until all data fits into the designated containers. The likelihood that all points in the data set are clustered together is very sensitive to the complexity of the decision tree. Leaf nodes, or data points that correspond to a given class, can be more easily extracted from a short tree. While initially easy to achieve, maintaining this purity as a tree expands becomes increasingly difficult, and as a result, there is often inadequate information contained inside a particular subtree. Because of the dispersed nature of the data, overfitting frequently occurs. The preference for tiny trees in decision trees is in line with the notion of parsimony espoused by Occam's Razor, which warns against "entities being multiplied beyond necessity." The simplest explanation isn't necessarily the most appealing, so decision trees shouldn't get any more convoluted than they need to be. The pruning process serves both goals by removing forks that produce offspring with few desirable characteristics (reducing complexity and preventing overfitting). Following this, cross-validation may be used to assess the model's precision. The random forest approach is one technique to maintain the reliability of decision trees by ensuring that the trees in the ensemble are independent of one another, which allows the classifier to make more accurate predictions. In Figure 5 we see a schematic representation of a decision tree.

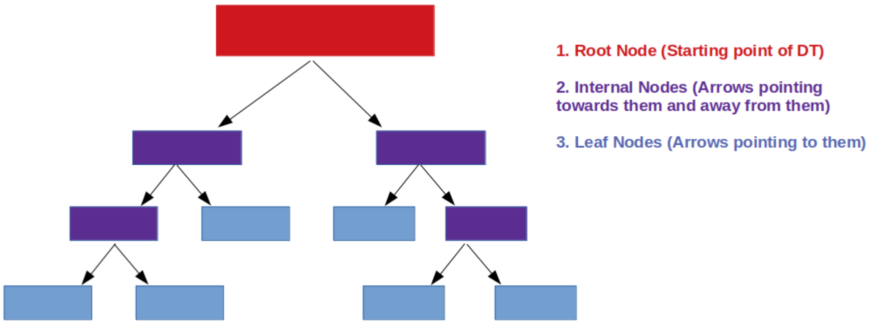
1. Root Node (Starting point of DT)
2. Internal Nodes (Arrows pointing towards them and away from them)
3. Leaf Nodes (Arrows pointing to them)

Figure 5: Structure of a decision tree.

I. Dipper Throated Optimization (DTO)

In computer science and mathematical optimization, metaheuristics are used to identify, create, or choose a heuristic (partial search algorithm) that, despite imperfect information or restricted processing power, may provide a good enough solution to an optimization issue. While it would be difficult to systematically enumerate or investigate every single conceivable solution, metaheuristics let us take a small but representative bite out of the enormous pie. Since various metaheuristics may be used in a variety of settings without extensive planning or a detailed understanding of the optimization problem at hand, they are frequently more flexible. In contrast to optimization algorithms and iterative techniques, the employment of metaheuristics does not guarantee the identification of a globally optimal solution for a given category of issues. As a result of the stochastic optimization used by many metaheuristics, the solution that is ultimately obtained might shift depending on the values of the random variables that are generated. Metaheuristics explore a large number of possible solutions, perhaps identifying excellent ones with less computational work than optimization algorithms, iterative techniques, or simple heuristics in combinatorial optimization. This means that they can be considered potential solutions to optimization problems. A great deal of literature has been produced on the subject. The bulk of articles written about metaheuristics discusses the author's personal experiences with implementing and testing these algorithms within the software, making them primarily experimental in nature. However, there are also some formal theoretical findings available, focusing on topics such as convergence and the possibility of obtaining the global optimum. There has been a proliferation of recently published articles proposing several metaheuristic techniques, each of which promises to be novel and beneficial in practice. The majority of the articles written on this subject have been of poor quality due to issues including vague language, insufficient explanation of key concepts, unreliable research techniques, and inadequate citation of prior work.

4. Results

To put machine learning (ML) models for tagging into action, we make use of a number of Python3 packages. There are a variety of useful Python packages, such as NumPy, SciPy, sci-kit-learn, Keras, pandas, and Matplotlib. It has been proven that Scikit-learn is the most trustworthy and user-friendly

DOI: https://doi.org/10.54216/IJWAC.060205
Received: September 11, 2022 Accepted: November 22, 2022

machine-learning library. Python packages NumPy, SciPy, and Matplotlib form the backbone of this bundle. An important metric for evaluating a classifier's efficacy is the confusion matrix, which summarizes the percentage of times that the training data resulted in incorrect predictions. In the case of a genuine positive, the observed value and the model's forecast are both correct (TP). In a genuine negative, both the actual and expected results are wrong (TN). As shown in Table 1, the proposed approach is compared to popular machine learning methods. The table below shows the enhanced accuracy, sensitivity, specificity, p-value, n-value, and F-score achieved by using the recommended optimal voting ensemble classifier.

Table 1: Classification results using the proposed method compared to other methods

	Accuracy	Sensitivity	Specificity	Pvalue	Nvalue	F-score
NN	0.9091	0.9091	0.9091	0.8475	0.9474	0.8772
SVM	0.8974	0.9091	0.8911	0.8197	0.9474	0.8621
DT	0.9032	0.9259	0.8911	0.8197	0.9574	0.8696
DTO+PSO	0.9524	0.9615	0.9502	0.8197	0.9906	0.8850

The statistical analysis presented in Table 2 shows the superiority of the proposed voting ensemble classifier. These results are better when the proposed optimized voting ensemble is employed.

Table 2: Statistical analysis of the results recorded by the proposed method

	NN	SVM	DT	DTO+PSO
Number of values	10	10	10	10
Minimum	0.9091	0.8874	0.9032	0.9524
25% Percentile	0.9091	0.8974	0.9032	0.9524
Median	0.9091	0.8974	0.9032	0.9524
75% Percentile	0.9091	0.8974	0.9032	0.9524
Maximum	0.9191	0.8974	0.9132	0.9524
Range	0.01	0.01	0.01	0
10% Percentile	0.9091	0.8884	0.9032	0.9524
90% Percentile	0.9181	0.8974	0.9122	0.9524
95% CI of median				
Actual confidence level	97.85%	97.85%	97.85%	97.85%
Lower confidence limit	0.9091	0.8974	0.9032	0.9524
Upper confidence limit	0.9091	0.8974	0.9032	0.9524
Mean	0.9101	0.8964	0.9042	0.9524
Std. Deviation	0.003162	0.003162	0.003162	0
Std. Error of Mean	0.001	0.001	0.001	0
Lower 95% CI of mean	0.9078	0.8942	0.902	0.9524
Upper 95% CI of mean	0.9124	0.8987	0.9065	0.9524
Coefficient of variation	0.3475%	0.3528%	0.3497%	0.000%
Geometric mean	0.9101	0.8964	0.9042	0.9524
Geometric SD factor	1.003	1.004	1.003	1
Lower 95% CI of geo. mean	0.9078	0.8942	0.902	0.9524
Upper 95% CI of geo. mean	0.9123	0.8987	0.9065	0.9524
Harmonic mean	0.9101	0.8964	0.9042	0.9524
Lower 95% CI of harm. mean	0.9078	0.8941	0.902	0.9524
Upper 95% CI of harm. mean	0.9123	0.8987	0.9065	0.9524

DOI: https://doi.org/10.54216/IJWAC.060205
Received: September 11, 2022 Accepted: November 22, 2022

Quadratic mean	0.9101	0.8964	0.9042	0.9524
Lower 95% CI of the quad. mean	0.9078	0.8942	0.902	0.9524
Upper 95% CI of the quad. mean	0.9124	0.8987	0.9065	0.9524
Skewness	3.162	-3.162	3.162	
Kurtosis	10	10	10	
Sum	9.101	8.964	9.042	9.524

Comparing the suggested method to the other methods is investigated using the Wilcoxon signed-rank test. Table 3 shows the outcomes of this analysis. It is demonstrated by the p-value in the table.

Table 3: Wilcoxon signed rank test of the recorded results of the proposed method

	NN	SVM	DT	DTO+PSO
Theoretical median	0	0	0	0
Actual median	0.9091	0.8974	0.9032	0.9524
Number of values	10	10	10	10
Wilcoxon Signed Rank Test				
Sum of signed ranks (W)	55	55	55	55
Sum of positive ranks	55	55	55	55
Sum of negative ranks	0	0	0	0
P value (two-tailed)	0.002	0.002	0.002	0.002
Exact or an estimate?	Exact	Exact	Exact	Exact
P value summary	**	**	**	**
Significant (alpha=0.05)?	Yes	Yes	Yes	Yes
How big is the discrepancy?				
Discrepancy	0.9091	0.8974	0.9032	0.9524

Figure 6 is a scatter plot displaying the improvement in classification accuracy over the starting models brought about by the suggested optimal voting ensemble classifier. This diagram illustrates how the improved efficiency of the suggested method.

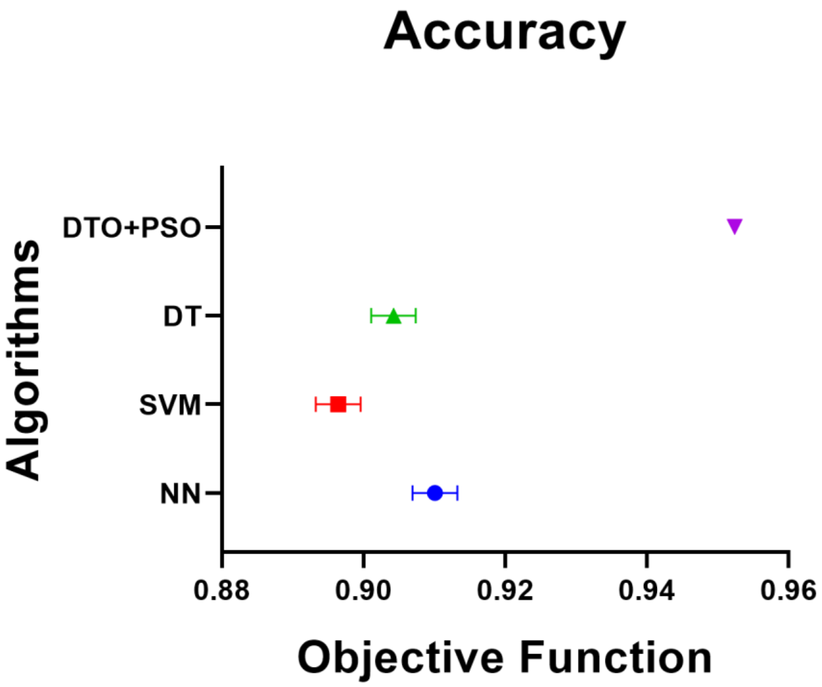

Figure 6: The accuracy of the proposed method compared to other methods

5. Conclusion

The evaluation of the machine learning methods used to identify spam SMS messages brings this article to a close. The focus of this work is on spam messages sent to mobile phone SMS systems. In this work, several classifiers were trained under supervision to discover the most accurate results possible while trying to identify spam emails. We experiment with a variety of machine-learning techniques and evaluate their efficacy. Among the classifiers we tested, Random forest and random tree Classifiers were found to have a perfect score of 100% accuracy. Content-linked documents have demonstrated significant improvements across the most popular classifier and achieved the greatest accuracy when used to classify textual information. It was established that the LMT and JRip perform well in traditional classification problems, with adequate results literally adjacent to trees. In practice, cutting-edge work and research yield helpful outcomes for the identification of spam SMS. There is a need for feature reduction methods and alternative stemming in future research.

Funding: “This research received no external funding”

Conflicts of Interest: “The authors declare no conflict of interest.”

References

[1] Hon, J., 2020. What’S The Difference Between SMS And MMS?. [online] Twigby Help & Support. Available at: <https://twigby.zendesk.com/hc/en-us/articles/115010624828-What-s-the-difference-between-SMS-and-MMS->.

[2] En.wikipedia.org. 2020. Mobile Phone Spam. [online] Available at: <https://en.wikipedia.org/wiki/Mobile_phone_spam> .

[3] SearchMobileComputing. 2020. What Is SMS Spam (Cell Phone Spam Or Short Messaging Service Spam)? - Definition From Whatis.Com. [online] Available at: <https://searchmobilecomputing.techtarget.com/definition/SMS-spam>.

[4] J. Han, M. Kamber. Data Mining Concepts and Techniques. by Elsevier inc., Ed: 2nd, 2006

[5] A. Tiago, Almeida , José María GómezAkebo Yamakami. Contributions to the Study of SMS Spam Filtering. University of Campinas, Sao Paulo, Brazil.

DOI: https://doi.org/10.54216/IJWAC.060205
Received: September 11, 2022 Accepted: November 22, 2022

[6] M. Bilal Junaid, Muddassar Farooq. Using Evolutionary Learning Classifiers To Do Mobile Spam (SMS) Filtering. National University of Computer & Emerging Sciences (NUCES) Islamabad, Pakistan.

[7] Inwhee Joe and Hyetaek Shim, "An SMS Spam Filtering System Using Support Vector Machine," Division of Computer Science and Engineering, Hanyang University, Seoul, 133-791 South Korea.

[8] Xu, Qian, Evan Wei Xiang, Qiang Yang, Jiachun Du, and Jieping Zhong. "Sms spam detection using noncontent features." IEEE Intelligent Systems 27, no. 6 (2012): 44-51. Yadav, K., Kumaraguru, P., Goyal, A., Gupta, A., and Naik, V. "SMSAssassin: Crowdsourcing driven mobile-based system for SMS spam filtering," Proceedings of the 12th Workshop on Mobile Computing Systems and Applications, ACM, 2011, pp. 1-6.

[9] Duan, L., Li, N., & Huang, L. (2009). “A new spam short message classification” 2009 First International Workshop on Education Technology and Computer Science, 168-171.

[10] Weka The University of Waikato, Weka 3: Data Mining Software in Java, viewed on 2011 September 14.

[11] Mccallum, A., & Nigam, K. (1998). “A comparison of event models for naive Bayes text classification”. AAAI-98 Workshop on 'Learning for Text Categorization'

[12] Bayesian Network Classifiers in Weka, viewed on 2011 September 14.

[13] Llora, Xavier, and Josep M. Garrell (2001) Evolution of decision trees, edn., Forth Catalan Conference on Artificial Intelligence (CCIA2001).

[14] B. G. Becker. Visualizing Decision Table Classifiers. pages 102- 105, IEEE (1998).

[15] El-Kenawy, El-Sayed M., Seyedali Mirjalili, Fawaz Alassery, Yu-Dong Zhang, Marwa Metwally Eid, Shady Y. El-Mashad, Bandar Abdullah Aloyaydi, Abdelhameed Ibrahim, and Abdelaziz A. Abdelhamid. "Novel Meta-Heuristic Algorithm for Feature Selection, Unconstrained Functions and Engineering Problems." IEEE Access 10 (2022): 40536-40555.

[16] El-kenawy, El-Sayed M., Marwa M. Eid, and Abdelhameed Ibrahim. "Anemia estimation for covid-19 patients using a machine learning model." Journal of Computer Science and Information Systems 17, no. 11 (2021): 2535-1451.

[17] El-Kenawy, El-Sayed M., Marwa Eid, and Alshimaa H. Ismail. "A New Model for Measuring Customer Utility Trust in Online Auctions." International Journal of Computer Applications 975: 8887.

[18] El-kenawy, El-Sayed M., Hattan F. Abutarboush, Ali Wagdy Mohamed, and Abdelhameed Ibrahim. "Advance artificial intelligence technique for designing double T-shaped monopole antenna." CMC-COMPUTERS MATERIALS & CONTINUA 69, no. 3 (2021): 2983-2995.

DOI: https://doi.org/10.54216/IJWAC.060205
Received: September 11, 2022 Accepted: November 22, 2022