University of Ibadan
Journal of Science and Logics in ICT Research

An Experimental Study on the Effectiveness of Random Forest (RF) Algorithm in Predicting Website Trustworthiness

[1]Alaba, O. B., [2]Ogunsanwo, G. O., [3]Aiyelokun, O. O. and [4]Abass, O. A.

[1,2]Tai Solarin University, Nigeria [3]University of Ibadan, Nigeria [4]Tai Solarin College of Education, Nigeria
[1]alabaob@tasued.edu.ng, [2]ogunsanwogo@tasued.edu.ng. [3]aiyelokuntobi@gmail.com, [4]abassoa@tasce.edu.ng

Abstract

Web users frequently depend on presentation and layout of a website for evaluating the trustworthiness of information contained therein. This can be disguised by the pervasive availability of professionally designed templates making the web information seem trustworthy regardless of its actual quality or source. As a result, web users are liable to arrive at false conclusions about the trustworthiness of the information available to them. This study seeks to improve the credibility of websites by assessing the effectiveness of Random Forest (RF) algorithm in predicting web trustworthiness. Dataset used entails scrapped data of nine thousand, five hundred and forty (9,540) websites collected from the training set and raw web files provided by Kaggle. The variables used in predicting web trustworthiness were average daily visitors, child safety, average daily page view, privacy. The dataset used was divided into two groups with a ratio of 80% to 20%. The 80% of the data was used for training of models, while the remaining 20% was used for the testing (validation). The experiment was performed using Sklearn Python library. The result showed that RF model was able to achieve an absolute precision, recall and F-measure of 1 for each class of website trustworthiness. The experimental study revealed that RF is effective in predicting web trustworthiness on the bases of average daily visitors, child safety, average daily page view, privacy, and traffic rank, while privacy and child safety were the most important input features for the model.

Keywords: Random Forest, trustworthiness, website, machines learning algorithm

1. Introduction

According to Sule [26], web content has become the main attraction in the digital world. The web contents are very expedient, whether in the case of valuable material and entertainment being accessible to internet users or to online advertisement of services and product. Whichever case, the web is an essential commodity [26] and the scale of users on the Web in terms of human interaction and communication, has grown very fast [5]. Trust is an essential issue that must be considered when internet users consume data [19]. This is principally factual with regard to the web, which has abundant information but characterized by lack of quality control that enables incorrect or low quality information to be published [19]. In addition, ordinary Web users tend to base their decisions on whether to trust web information on trial-and-error factors that are mainly based on surface level characteristics of the web page, that is user interface design [12]. Such characteristics are easily disguised, and web users can arrive at the wrong conclusions about the trustworthiness of the information they consume. Therefore, the importance of assessing the trustworthiness of websites cannot be overemphasized. This justifies the need for the study.

Machine learning (ML) is one of the most rapidly developing techniques, and it can help computers to address the problems by learning through experience [13]. Due to rapid advances of computer technology and intelligent technology globally, intelligent machine identification skills have been well developed [8]. ML emphasizes on the development of computer programs that can change when exposed to a new set of data. It is the aspect of

Alaba, O. B., Ogunsanwo, G. O., Aiyelokun, O. O. and Abass, O. A. (2021). An Experimental Study on the Effectiveness of Random Forest (RF) Algorithm in Predicting Website Trustworthiness. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)*, Vol. 6 No. 2, pp. 1 - 9

data testing that automates analytical model building, based on algorithms that repeatedly learn from data. ML enhances computers to discover hidden information without writing programs to know where such information is located. ML today is not like ML of the past because of the level of advancement in computing technologies.

ML is the ability of a machine to perform better at a given task, using its previous experience [26]. It is that segment of Artificial Intelligence (AI), which aims at making computers learn similar to humans. It is the automation of a learning process and learning is equivalent to the construction of rules based on observations of environmental states and transitions [26]. The repetitive aspect of ML is important because as models are exposed to new data, they are able to independently adapt. In ML, previous computations serve as medium of learning in order to generate reliable, repeatable decisions and results. While many ML algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data is a recent development [1]. Some widely advertised examples of ML applications are self-driving Google car, online recommendation offers such as those from Amazon and Netflix, Fraud detection and weather prediction. However, little or no work has been done to measure the effectiveness of Random Forest (RF) algorithm to predict website trustworthiness.

This study employed the use of Random Forest algorithm for predicting web trustworthiness based on average daily visitors, child safety, average daily page view, privacy, and traffic rank. Random forest was used due to its robustness to noise and overfitting compared to other ML counterpart [4, 22]. A model with overfitting problem will generally have poor predictive performance and will not generalize well [4].

1. Literature Review

2.1 Website Trustworthiness

the depth to which those topics are focused upon. A social network is a platform that allows its users to obtain services and share their experiences [7]. Based on such feedback gathered, a data processing center (DPC) can

Trust is an essential factor in the process of consuming data; for Web and Semantic Web environments, which are decentralized and have little control over publishing quality [19]. The field of information quality research comes up with tools and methods that can be used to analyzing the quality of Web data and its data sources [18]. In particular, it describes a number of quality criteria to help in assessing the quality of information. For instance, Tate [27], defined five quality values for their value-model, namely: accuracy (the data should be error-free); comprehensiveness (the completeness of the coverage of a particular subject or discipline); currency (how recent the data is); reliability (the consistency of the quality of the system and its output over time); and validity (the correctness of the information).

According to Rieh and Belkin [21], more studies were conducted on the judgment of information quality and authority by scholars when interacting with information on the web, their study collected data based on the scholars' actual searching behaviours and then concluded six major categories of criteria for evaluating information quality and authority. The six major criteria consist of characteristics of information objects, characteristics of sources, users' own personal experiences, situation, ranking in search output, and general assumptions (e.g. a salesman will always over-state the positive aspects of their product and omit any negatives, so do not trust everything they say) [18].

Similarly, Taylor [28] proposed information quality criteria for web resources as: authority is the degree to which a person or organization is perceived as having the required knowledge to provide information on a given subject area; accuracy is the degree to which the information is accurate and free from errors; objectivity is the degree to which the material conveys neutral facts or information (i.e. the facts are not influenced by personal feelings or other biases); currency is the degree to which the material or information is up-to-date; and coverage is the scope of topics and

provide quality ratings for different services, which can further give suggestions for new users. Both reliability and trustworthiness of the feedback on the side of the users must be checked. In terms of feedback and decision systems, the management of trust and reputation has become a great challenge. Therefore, Kim *et. al.* [14] and Du *et. al.* [9] proposed many trustworthiness evaluation mechanisms for social networks in order to ensure that visitors of the websites eventually become customers and the website presented creditably to establish trust.

2.2 Random Forest Algorithm

RF is a collection or ensemble of Classification and Regression Trees (CART) trained on datasets of the same size as training set, called *bootstraps*, created from a random resampling on the training set itself [3], [23]. Once a tree is constructed, a set of bootstraps, which do not include any particular record from the original dataset (*out-of-bag* (OOB)) or known seen samples), is used as test set [23]. The error margin for the classification of all the validation data is the generalization error OOB estimate. Breiman [2] disclosed by experimental proof that, for the "bagged classifiers, the OOB error" is precise as using a validation set of the similar size as the calibration set. Consequently, adopting the OOB approximation eliminates the necessity for a discrete test set [23]. To categorize new input data, each of the individual CART tree votes for one class and thereby the forest predicts the class that gets the plurality of votes (colored branches in Figure 1).

Figure 1 shows the illustration of a random forest construct that is superimposed on a coronal slice of the MNI 152 (Montreal Neurological Institute) standard template. RF trails unique rules for tree growing, tree composition, self-testing and post-processing. It is resistant to computational complexity, and is rated more reliable in the presence of noise and in very high-dimensional datasets parameters than other ML approaches. Burlutskiy [5], Caruana and Niculescu-Mizil [6], and Menze et.al. [16] stated that the significance of variable is an underlying feature selection performed by RF usi3ng a random subspace methodology. This is assessed by the Gini impurity criterion index. The Gini index assesses the prediction ability of variables in regression or classification by adopting the principle of impurity reduction [25]. The index is non-parametric which does not depend on data belonging to a specific type of distribution. For a binary split (white circles in Figure 1), the Gini index of a node *n* is calculated as follows:

$$\text{Gini}(n) = 1 - \sum_{j=1}^{2} (p_j)^2 \qquad (1)$$

where p_j is the relative frequency of class *j* in the node *n*.

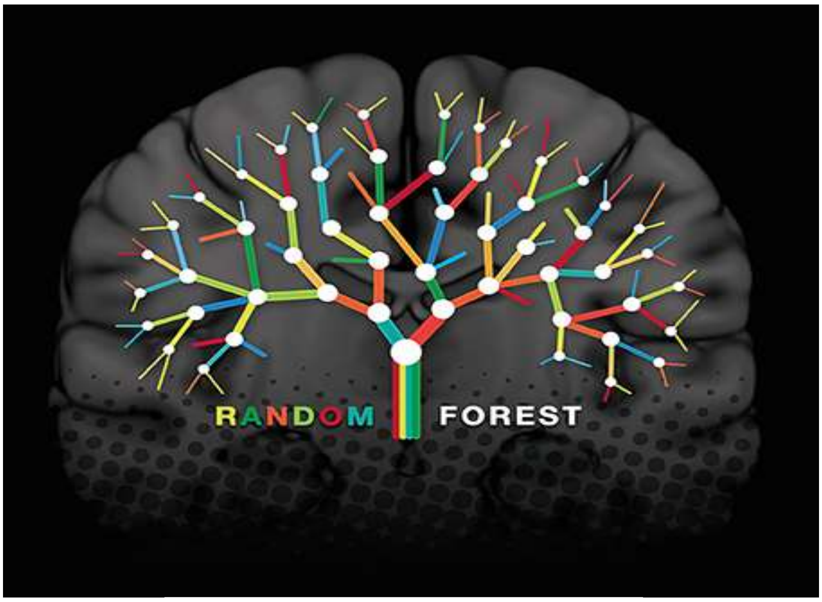


Figure 1: Random Forest Structure

Splitting a binary node enhances the improvement in the Gini index which should be maximized. This implies that a low Gini (i.e., a greater decrease in Gini) means that a certain predictor feature plays a greater role in partitioning the data into the two classes. Thus, the Gini index can be used to rank the importance of features for a classification problem.

3. Methodology

3.1 Data Collection

Dataset used in the study entails scrapped data of nine thousand, five hundred and forty (9,540) websites collected from application gathered from the training set and raw web files provided by Kaggle. The dataset consists of information about each website such as average daily visitors, child safety, average daily page view, privacy, and traffic rank. The collected data were cleaned and preprocessed, while attribute selection processes were further performed on the data to identify the input variables for the developed model.

3.2 Description of Dataset

Table 1 shows the dataset that was used for RF model development. As depicted, the study employed one output variable or response variable (trustworthiness of website), and five input variables or predictors; which include the average daily visitors, child safety, average daily page view, privacy, and traffic rank. The dataset used consist of both class and numeric data type; while the variable with class data type were categorized into excellent, good, poor, very poor.

3.3 Model Training and Validating

The dataset used was divided into two groups with a ratio of 80% to 20%. 80% of the data was used for training of models, while the remaining 20% was used for the testing (validation) as shown in Figure 2.

The experiment was performed using Sklearn Python library. The Random Forest Regressor class of the sklearn.ensemble library was used to predict Website Trustworthiness using n_estimator = 10.

This parameter defines the number of tree used as shown in Figure 3. Random forest was used due to its robustness to noise and overfitting compared to other ML counterpart [4], [22]. A model with overfitting problem will generally have poor predictive performance and will not generalize well [11].

3.4 Performance Evaluation of Models

The performance of the models developed was assessed using the Precision, Recall, F-measure, Accuracy, Macro-average measure and Weighted-average. In addition, the confusion matrix, which shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data, was used.

The matrix is NxN, where N is the number of target values (classes) and the performance of the models is commonly evaluated using the data in the matrix.

Table 1: Description of Dataset used in Model Development

S/N	Variable	Variable Type	Classification	Data Type
1	Trustworthiness	Output	Excellent, Good, Poor, Very Poor	Class
2	Average Daily Visitors	Input	Non	Numeric
3	Child Safety	Input	Excellent, Good, Poor, Very Poor	Class
4	Average Daily Page View	Input	Non	Numeric
5	Privacy	Input	Excellent, Good, Poor, Very Poor	Class
6	Traffic Rank	Input	Non	Numeric

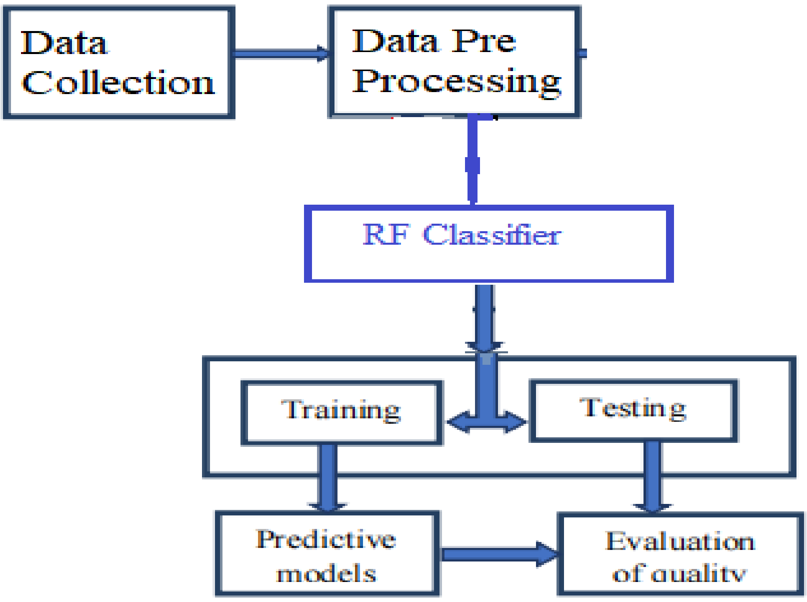


Figure 2: Framework of the Model

```
c:\users\envy 15\miniconda3\envs\tensorflow\lib\site-packages\sklearn\ensemble\fore
ue of n_estimators will change from 10 in version 0.20 to 100 in 0.22.
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)

Out[10]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=None, max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=10,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)
```

Figure 3: RF Model

4. Results and Discussion

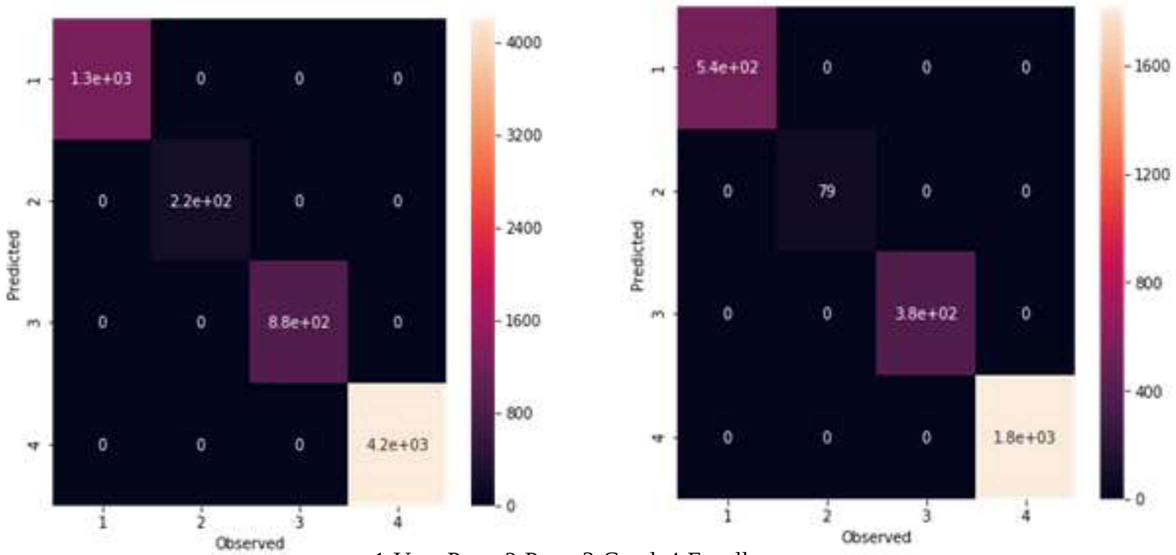
4.1 Model Performance at Training and Testing Phases

The performance of the RF model at training and testing phase is presented in Table 2. The table shows that at the training phase, RF model was able to achieve an absolute precision, recall and F-measure of 1 for each class of website trustworthiness. The table also shows that at the training phase RF model had respective score of 1 for the Marco-average and Weighted-average, implying that the model has 100% accuracy when calibrating it. At the testing phase the RF was found to be effective in predicting website trustworthiness based on unseen dataset (test dataset). As shown in Table 2, the RF was able to achieve 100% accuracy in precision, recall and F1-score of 1 for each class of website trustworthiness which constitutes very poor, poor, good and excellent. The loss of mis-classification of the model at training and testing phase was further assessed using the confusion matrix presented in Figure 4 which shows that in comparison with the frequency values presented in Table 2, all the class of trustworthiness was correctly classified.

Table 2: Model Performance at Calibration and Validation State

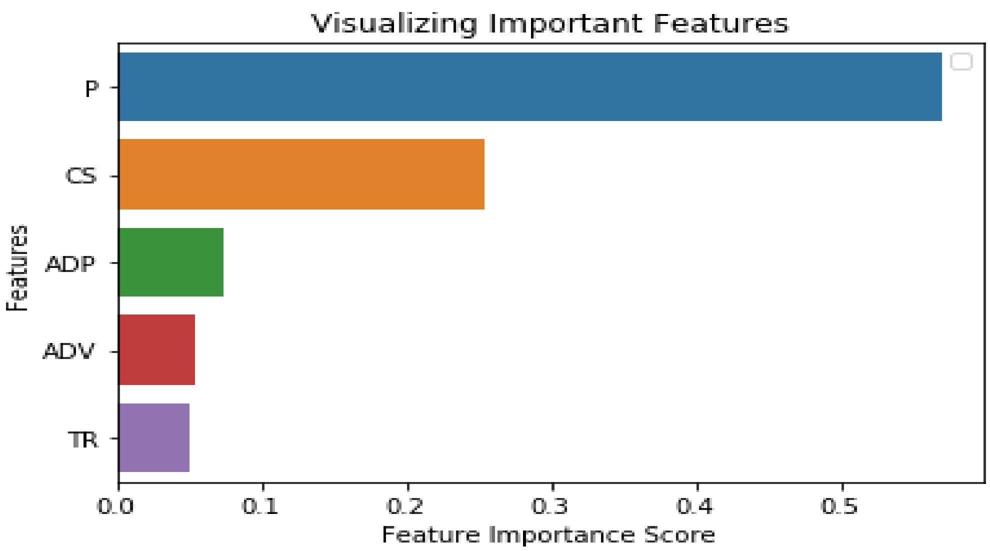
	Training Phase				Testing Phase			
	precision	recall	f1-score	Frequency	precision	recall	f1-score	Frequency
Very Poor	1.0	1.0	1.0	1267	1.0	1.0	1.0	540
Poor	1.0	1.0	1.0	220	1.0	1.0	1.0	79
Good	1.0	1.0	1.0	885	1.0	1.0	1.0	381
Excellent	1.0	1.0	1.0	4205	1.0	1.0	1.0	1819
Accuracy			100%				100%	
Macro avg	1.0	1.0	1.0		1.0	1.0	1.0	
Weighted avg	1.0	1.0	1.0		1.0	1.0	1.0	

.



1-Very Poor, 2-Poor, 3-Good, 4-Excellent

Figure 4: Confusion Matrix for Training Phase (a) and Testing Phase (b)



P-Privacy, CS-Child Safety, ADV-Average Daily, ADP- Average Page View Visitors, TR-Traffic Rank

Figure 5: Feature Importance Score

4.2 Analysis of Feature Importance

The level of importance of the input features were further assessed using feature importance scores as depicted in Figure 5. The figure shows that privacy is the more important factor with feature importance score of about 0.6, followed by child safety with feature importance score of about 0.25. The number of average page view visitors was consider next to be important in predicting trustworthiness of a website with a low feature importance score of about 0.8. Other factors that were ranked to be less important were the average daily visitors and the traffic rank with importance feature score of less than 0.05.

4.3 Discussion of Findings

There is still limitation in the understanding of the extent by which people interact with each other, and the kind of information that can be deduced from their activities over the Web. Flanagin and Metzger [10] unraveled that the use of traditional sources like books, newspapers, interpersonal contacts or even television by people has drastically reduced; as there is now strong tendency that people will turn mostly to Internet when searching for information.

However, even more importantly, the study concludes that people perceived searching for information as the most important functionality of the Internet [20]. Although, searching for information on the Internet is quick, easy, and free, the trustworthiness of many content shared via Internet might be dubious [20] and people usually do not seek for verification of information found on the Internet in other sources [10], [17], [24], [15].

Most of the time, web users depend on descriptive criteria such as information presentation and layout [12], for evaluating trustworthiness of information, which can be disguised by the pervasive availability of professionally designed templates making the web information seem trustworthy regardless of its actual quality or source. Consequently, web users might give false conclusions about the trustworthiness of information.

This study builds on this gap by assessing the effectiveness of RF as a potential ML algorithm for predicting web trustworthiness based on average daily visitors, child safety, average daily page view, privacy, and traffic rank which can be used to ensure credibility of any website. The study unraveled that RF was able to achieve 100% accuracy at both training and testing phase, implying that RF is effective for predicting web trustworthiness based on average daily visitors, child safety, average daily page view, privacy, and traffic rank. The study further presents new insights on the most important features that have high tendency of affecting the predictive performance of RF with respect to trustworthiness of websites. It was established in the study that privacy and child safety are more likely to determine the variations of trustworthiness of websites.

4. Conclusion

The web has become a ubiquitous environment for human interaction, communication, and data sharing. As a result, large amount of data are produced which can be utilized by building predictive models of user behavior in order to support business decisions. Since data mining and predictive analytics have become one of the key features of many security initiatives developed to monitor internet activities; the present study endeavors to assess the effectiveness of RF in predicting the trustworthiness of websites on the bases of average daily visitors, child safety, average daily page view, privacy, and traffic rank. Facts emerging from the study have established that RF algorithm has high predictive accuracy, having been able to achieve 100% accuracy at both training and testing phase, while privacy and child safety are the most important input features for the model.

References

[1] Alexander, Y. S. & Bridget, R. S. (2019). How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Published by IOP Publishing Ltd.

[2] Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi: 10.1007/BF00058655

[3] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Boca Raton, FL: CRC press.

[4] Boinee, P., Angelis. A., & Foresti G, (2006): Meta Random Forest, *International Journal of Computational Intelligence, 2*(3), 138-147.

[5] Burlutskiy, N. (2017). Prediction Of User Behaviour On The Web. A PhD thesis submitted to the University of Brighton.

[6] Caruana, R., & Niculescu-Mizil, A. (2006). "An empirical comparison of supervised learning algorithms," in *23rd International Conference on Machine Learning*, (Pittsburgh, PA: ACM Press), 161–168.

[7] Ceriani, L., & Verme, P. (2012). The origins of the gini index: extracts from variabilità e mutabilità (1912) by Corrado Gini. *Journal of Economic Inequality 10*(3), 421– 443. doi: 10.1007/s10888-011-9188-x

[8] Du, J., Jiang, C., Chen, K., Ren, Y., & Poor, H. V. (2018). Community structured evolutionary game for privacy protection in social networks in IEEE Transaction on Information Forensics and Security, 13(3), 574–589.

[9] Du J., Gelenbe E., Jiang C., Zhang H., Ren, Y., & Poor, H. V. (2018). Peer Prediction Based Trustworthiness Evaluation and Trustworthy Service Rating in Social Networks IEEE Transactions on Information Forensics and Security DOI: 0.1109/TIFS.2018.2883000

[10] Flanagin A.J. & Metzger M. (2000). Perceptions of Internet information credibility. *Journalism & Mass Communication Quarterly*, *77,*3, 515-540.

[11] Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences, 44*(1), 1–12. [Crossref], [PubMed], [Google Scholar]

[12] Fogg, B. J., Soohoo C., Danielson, D. R., Marable, L., Stanford, J., & Tauber, E. R. (2003). How do users evaluate the credibility of Web sites?: a study with over 2,500 participants", Proceedings of the 2003 conference on Designing for user experiences, 1–15.

[13] Jordan, M. and Mitchell, T. (2015). Machine learning: Trends, perspectives, and prospects. Science 2015, 349, 255–260.

[14] Kim Y., E.-W. Jhee, J. Choe, J.- Choi S., and Shin Y,(2015) A measurement model for trustworthiness of information on social network services," in 2015 Int. Conf. Inform. Networking (ICOIN). Cambodia, Jan. 2015, pp. 437–438.

[15] Lucassen, T. (2013). *Trust in online information*. Universiteit Twente. https://doi.org/10.3990/1.9789036534857.

[16] Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W. & Hamphrecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics 10*:213. doi: 10.1186/1471-2105-10-213

[17] Metzger, M. J. (2007). Making Sense of Credibility on the Web: Models for Evaluating Online Information and Recommendations for Future Research. *Journal of the American Society for Information Science and Technology, 58*(13), 2078-2091.

[18] Pattanaphanchai, J., O'Hara, K., Hall, W. (2013). Trustworthiness criteria for supporting users to assess the

credibility of web information Conference Proceedings of the 22nd International Conference on World Wide Web companion.
DOI: 10.1145/2487788.2488132

[19] Pattanaphanchai, J. (2014). Trustworthiness of Web Information Evaluation Framework. PhD thesis submitted to the Faculty of Physical Sciences and Engineering, University of Southampton.

[20] Rafalak, M., Abramczuk, K. & Wierzbicki, A. (2014). Incredible: Is (Almost) All Web Content Trustworthy? Analysis of Psychological Factors Related to Website Credibility Evaluation. the International World Wide Web Conference Committee (IW3C2).

[21] Rieh, S. Y. & Belkin, N. J. (2000) Interaction on the Web: Scholars' Judgment of Information Quality and Cognitive Authority Proceedings of the 63rd Annual Meeting of the ASIS, 2000, 25-38.

[22] Robnik M. S. (2004): Improving Random Forests in J F Boulicaut et al. (eds): Machine Learning, ECML 2004 Proceedings, Springer, Berlin.

[23] Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. *Front Aging Neurosci*. 9:329.
doi: 10.3389/fnagi.2017.00329

[24] Sparks, J. R. & Rapp, D. N. (2011). Readers' Reliance on Source Credibility in the Service of comprehension. *Journal of Experimental Psychology*, *37*(1), 230–247.

[25] Strobl, C., Boulesteix, A. L. & Augustin, T. (2007). Unbiased split selection for classification trees based on the gini index. *Comput. Stat. Data Anal.* 52, 483–501.
doi: 10.1016/j.csda.2006.12.030.

[26] Sule, M. D. (2003). Machine Language Techniques for Conversational Agents. Retrieved May 28, 2021 from https://digital.library.unt.edu.

[27] Tate, M. A. (2018). Web Wisdom: How to Evaluate and Create Information Quality on the Web (3rd ed.). CRC Press. https://doi.org/10.1201/b22397.

[28] Taylor, R. S. (2007). Value-Added Processes in the Information Life Cycle. *Journal of The American Society for Information Science and Technology*, *33*(5). DOI:10.1002/asi.4630330517.