IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 8, August 2021

DOI 10.17148/IJARCCE.2021.10820

# An Experimental Comparison of Classification Tools for Fake News Detection

**Ekemini Anietie Johnson[1], Jude Alphonsus Inyangetoh[2] , Mfon Okpu Esang[3]**

Department of Computer Science Federal Polytechnic Ukana Akwa Ibom State, Nigeria [1]

Department of Statistics Federal Polytechnic Ukana Akwa Ibom State, Nigeria[2]

Department of Computer Science Federal Polytechnic Ukana Akwa Ibom State, Nigeria [3]

**Abstract**: Fake news in the media is not new. It has been with us since the development of the earliest writing systems. Fake news have caused a lot of damage to humanity and hence the need to detect it. The term "fake news" is not new but detecting it quickly has really been a problem. This study used random forest and decision tree algorithms on a dataset containing both fake and real news to do classification. The software used for the experiment was Weka and the result generated show that random forest correctly classified instance is 100% and incorrectly classified instance is 0% while the decision tree correctly classified instance is 93.6364% and incorrectly classified instance is 6.3636%. The results is a proof that random forest algorithm is a better classification tool as compared to decision tree.

**Keywords:** Fake news, Random Forest, Decision Tree, Algorithm, tool, Classification.

## 1. INTRODUCTION

The term "fake news" is not new. Contemporary discourse, particularly media coverage, seems to define fake news as referring to viral posts based on fictitious accounts made to look like news reports (Edson C. Tandoc Jr., Zheng Wei Lim and Richard Ling 2017). One of the most commonly accepted definitions by the research community is that "Fake news is a news article that is intentionally and verifiably false" (Allcott and Gentzkow, 2017). A recent study defined fake news to be news articles that are intentionally and verifiably false, and could mislead readers" (Allcott and Gentzkow 2017). Two main motivations underlie the production of fake news: financial and ideological. On one hand, outrageous and fake stories that go viral precisely because they are outrageous provide content producers with clicks that are convertible to advertising revenue. On the other hand, other fake news providers produce fake news to promote particular ideas or people that they favor, often by discrediting others (Allcott and Gentzkow 2017). . According to Wikipedia, fake news is: "a type of yellow journalism or propaganda that consists of deliberate misinformation or hoaxes spread via traditional print and broadcast news media or online social media". Fake news have caused a lot of damage to humanity and hence the need to detect it.

Machine learning methodologies have proven to be of great practical value in a variety of application domains in situations where it is impractical to manually extract information from data. Machine learning can be used to do prediction and it is capable of working with partial input data (Ivan Makhotina et. al. 20121). Machine learning is said to be a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence (Aliyuda and Howell, 2019). Machine learning explores the construction and study of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data driven predictions or decisions, rather than following strictly static program instructions. Machine learning is closely related to and often overlaps with computational statistics; a discipline that also specializes in prediction making. It has strong ties to mathematical optimization, which deliver methods, theory and application domains to the field. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms is infeasible. The data driven approach allows retrieving non trivial dependencies and building powerful predictive models from historical data.

In this study we are going to train a news data obtained from https://www.kaggle.com/clmentbisal-news-dataset?select=true.csv using the random forest and decision tree algorithms and compare their outputs.

## 2. LITERATURE REVIEW

Xichen Zhang, Ali A. GhorbaniIn (2019) carried out a survey and presented a comprehensive overview of the finding to date relating to fake news. They characterized the negative impact of online fake news, and the state-of-the-art in detection methods. Many of these rely on identifying features of the users, content, and context that indicate

  

DOI 10.17148/IJARCCE.2021.10820

misinformation. They also studied existing datasets that have been used for classifying fake news. Finally, they proposed promising research directions for online fake news analysis.

Edson C. Tandoc Jr., Zheng Wei Lim and Richard Ling (2017) did a review of how previous studies have defined and operationalized the term "fake news." An examination of 34 academic articles that used the term "fake news" between 2003 and 2017 resulted in a typology of types of fake news: news satire, news parody, fabrication, manipulation, advertising, and propaganda.

Joao Pedro Baptista and Anabela Gradim (2020) did a review that aimed to see why fake news is widely shared on social media and why some people believe it. The presentation of its structure (from the images chosen, the format of the titles and the language used in the text) can explain the reasons for going viral and what factors are associated with the belief in fake news. They showed that fake news explores all possible aspects to attract the reader's attention, from the formation of the title to the language used throughout the body of the text. The review showed that fake news continues to be widely shared and consumed because that is the main objective of its creators.

Claire Wardle (2017) in a research titled "Fighting Fake News" tried to explore the ongoing efforts to define fake news and discuss the viability and desirability of possible solutions. The discussion encompassed attempts to identify the particular harm associated with fake news; This workshop was meant to be a first step towards encouraging interdisciplinary conversation and work on these issues.

Iftikhar Ahmad ,Muhammad Yousaf, Suhail Yousaf , and Muhammad Ovais Ahmad (2020) detected fake news using Machine Learning Ensemble Methods. They proposed to use machine learning ensemble approach for automated classification of news articles. Their study explored different textual properties that can be used to distinguish fake contents from real. By using those properties, they trained a combination of different machine learning algorithms using various ensemble methods and evaluated their performance on 4 real world datasets. Experimental evaluation confirms the superior performance of the proposed ensemble learner approach in comparison to individual learners.

Oberiri Destiny Apuke , Bahiyah Omar(2020) aimed at understanding the effects of fake news spreading in Nigeria, the reasons for fake news sharing among social media users, and eventually propose preventive measures (i.e. awareness strategies) to combat the proliferation of fake news in Nigeria. Some grave implications of fake news sharing were identified such as death, conflict escalation, political hostility, and societal panic. Meanwhile, people were motivated to share news mainly because of their civil obligation to inform others and provide advice or warning. These motivations, together with other contextual reasons such as media control, interpersonal trust and youth unemployment, had led to fake news proliferation in Nigeria. The study adopted a documentary research method to generate the information necessary to investigate fake news spread in Nigeria. A total of 265 articles were drawn from Google Scholar search and after a close examination, only 20 articles were included for analysis. The researchers were of the opinion that Social media users should be constantly informed through adequate advertisements, workshops, conferences, and other forms of sensitization, about the consequences of fake news sharing, how to spot and differentiate fake news with made-up news and why it is imperative to be self-aware before forwarding any message.

Xichen Zhang, Ali A. GhorbaniIn (2019) carried out a survey and presented a comprehensive overview of the finding to date relating to fake news. They characterized the negative impact of online fake news, and the state-of-the-art in detection methods. Many of these rely on identifying features of the users, content, and context that indicate misinformation. They also studied existing datasets that have been used for classifying fake news. Finally, they proposed promising research directions for online fake news analysis.

Gulizar Haciyakupoglu, Jennifer Yang Hui, V. S. Suguna, Dymples Leong, and Muhammad Faizal Bin Abdul Rahman (2018) the study suggested that the following mearsures be used to prevent fake news.

- Pre-emptive measures that are focused on an issue (i.e., elections) and supplemented by continuous collaborative engagements with the industry, non-governmental sector and regional fora;
- Immediate measures that comprise an agile crisis communications plan and fact-checking initiatives; and
- Long-term measures that strengthen social resilience through media literacy, inculcation of social norms on responsible information sharing, and defining the responsibilities of technology companies.
- Going forward, a multi-pronged strategy that comprises both legislation and non-legislative measures given that each have their challenges would form a more sustainable bulwark against fake news.

Iftikhar Ahmad ,Muhammad Yousaf, Suhail Yousaf , and Muhammad Ovais Ahmad (2020) detected fake news using Machine Learning Ensemble Methods. They proposed to use machine learning ensemble approach for automated classification of news articles. Their study explored different textual properties that can be used to distinguish fake contents from real. By using those properties, they trained a combination of different machine learning algorithms using various ensemble methods and evaluated their performance on 4 real world datasets. Experimental evaluation confirms the superior performance of the proposed ensemble learner approach in comparison to individual learners.

Marianna Isaakidou, Emmanouil Zoulias and Marianna Diomidous (2021) aimed at introducing an Artificial Intelligence approach, the Decision Trees algorithm to identify fake news on the COVID-19. This was necessary because they claimed that European Commission has conducted a survey about "Fake News" through EU citizens to

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 8, August 2021

DOI 10.17148/IJARCCE.2021.10820

estimate the awareness and people behaviour related to the appearance of fake news and disinformation on electronics and the findings are quite worrying, since about 40% come across fake news daily and 85% evaluate fake news as a problem.

### 2.1 The design

The design comprises of database of news, data preprocessing, random forest module, decision tree module, model evaluation and comparison of random forest and decision tree as shown in Figure 1.

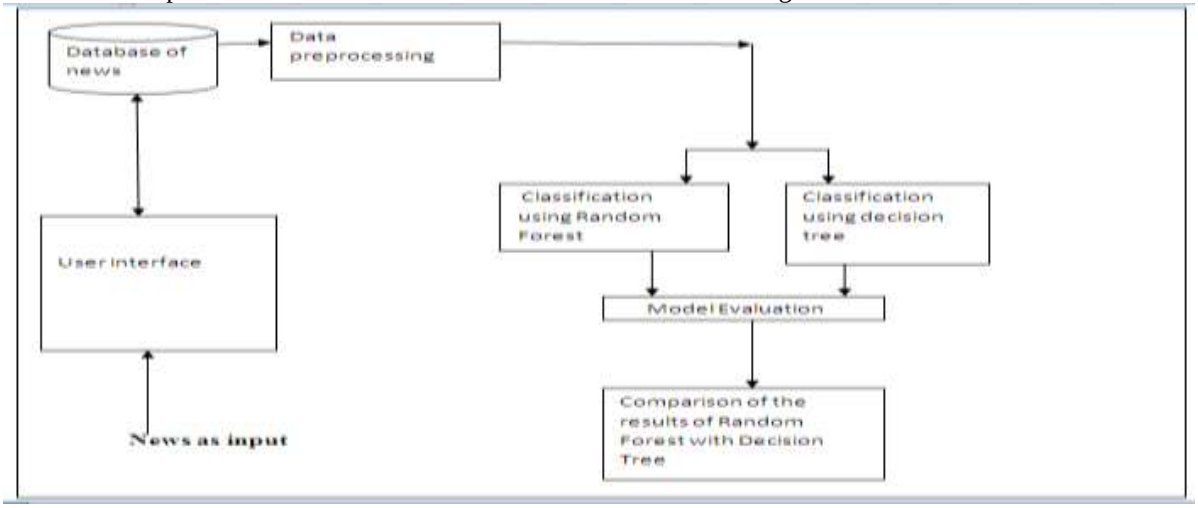


**Figure 1: A Frame Work for an Experimental Comparison of Classification Tools for Fake News Detection.**

### 2.2 The database

Database of news consist of one hundred and ten (110) news and the characteristics of fake news.

### 2.3 Data Pre Processing

Here the data from the database is preprocessed so as to be in a format that can be processed by the computer

### 2.4 Random Forest module

The random forest module takes the preprocessed data and does classification of the news into fake and real news.

Random Forest (RF). Random forest (RF) is an advanced form of decision trees (DT) which is also a supervised learning model. RF consists of large number of decision trees working individually to predict an outcome of a class where the final prediction is based on a class that received majority votes. The error rate is low in random forest as compared to other models, due to low correlation among trees (Gregorutti, B. Michel, and P. Saint-Pierre 2017).

According to Gregorutti and Saint-Pierre (2017), the random Forest package optionally produces two pieces of information: a measure of the importance of the predictor variables, and a measure of the internal structure of the data (the proximity of different data points to one another). The random forest algorithm is as follows:

i. Take a random sample of size N with replacement from the data.
ii. Take a random sample without replacement of the predictors.
iii. Construct the first CART partition of the data.
iv. Repeat Step 2 for each subsequent split until the tree is as large as desired. Do not prune.
v. Repeat Steps 1–4 a large number of times.

### 2.5 Decision Tree.

The decision tree module takes the preprocessed data and does classification of the news into fake and real news using decision tree algorithm. The decision tree algorithm follows these steps:

- Step-1: Begin the tree with the root node, say S, which contains the complete dataset.
- Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).
- Step-3: Divide the S into subsets that contains possible values for the best attributes.
- Step-4: Generate the decision tree node, which contains the best attribute.
- Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 8, August 2021

DOI 10.17148/IJARCCE.2021.10820

While implementing a Decision tree, the main issue arises on how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called Attribute selection measure or ASM. By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are:

- Information Gain
- Gini Index

i. $\mathbf{Information\ Gain} = \mathbf{Entropy}(S) - [(\mathbf{Weighted\ Avg}) * \mathbf{Entropy}(\mathbf{each\ feature})]$ **eqn1**

Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

$$\mathbf{Entropy}(s) = -\mathbf{P}(\mathbf{yes})\mathbf{log2\ P}(\mathbf{yes}) - \mathbf{P}(\mathbf{no})\ \mathbf{log2\ P}(\mathbf{no}) \qquad \textbf{eqn 2}$$

Where,

- S= Total number of samples
- P(yes)= probability of yes
- P(no)= probability of no.

**ii.** Gini Index:

- Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.
- An attribute with the low Gini index should be preferred as compared to the high Gini index.
- It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.
- Gini index can be calculated using the below formula:

$$\mathbf{Gini\ Index} = 1 - \sum \mathbf{jPj2} \qquad \textbf{eqn 3}$$

## 3.0 PERFORMANCE METRIC

To evaluate the performance of algorithms, different metrics are used. Most of them are based on the confusion matrix. Confusion matrix is a tabular representation of a classification model performance on the test set, which consists of four parameters: true positive, false positive, true negative, and false negative.

### 3.1 Accuracy

Accuracy is often the most used metric representing the percentage of correctly predicted observations, either true or false. To calculate the accuracy of a model performance, the following equation can be used:

$$\mathbf{Accuracy} = (\mathbf{TP} + \mathbf{TN}) / (\mathbf{TP} + \mathbf{TN} + \mathbf{FP} + \mathbf{FN}). \qquad \textbf{eqn 4}$$

Most Times, high accuracy value represents a good model, but considering the fact that here a classification model is trained, an article that was predicted as true while it was actually false (false positive) can have negative consequences; similarly, if an article was predicted as false while it contained factual data, this can create trust issues. Therefore, other metrics can be used that take into account the incorrectly classified observation, i.e., precision, recall, and F1-score.

### 3.2 Recall.

Recall represents the total number of positive classifications out of true class. In our case, it represents the number of news classified as true out of the total number of true news.

**Table 1: confusion matrix**

| | Predicted true | Predicted false |
|---|---|---|
| Actual true | True positive (TP) | False negative (FN) |
| Actual false | False positive (FP) | True negative (TN) |

$$\mathbf{Recall} = \mathbf{TP} / (\mathbf{TP} + \mathbf{FN}). \qquad \textbf{eqn 5.}$$

### 3.3 Precision.

Precision score represents the ratio of true positives to all events predicted as true

$$\mathbf{Precision} = \mathbf{TP} / (\mathbf{TP} + \mathbf{FP}). \qquad \textbf{eqn 6.}$$

  

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 8, August 2021

DOI 10.17148/IJARCCE.2021.10820

### 3.4 F1-Score.

F1-score represents the trade-off between precision and recall. It calculates the harmonic mean between each of the two. Thus, it takes both the false positive and the false negative observations into account. F1-score can be calculated using the following formula:

$$\mathbf{F1-score} = 2\,((\mathbf{Precision} \times \mathbf{Recall})\,/(\,\mathbf{Precision} + \mathbf{Recall})) \qquad \textbf{eqn 7}$$

## 4.0 THE EXPERIMENT AND RESULT

The experiment was carried out with Weka software as the computation and analysis tool. Figure 2, 3,4, 5, 6 and 7 depict the data capturing and processing environment in a GUI of Weka for Random Forest and Decision Tree**.** The classification is done in two ways: (i) classification using random forest and (ii) classification using decision tree.


**Figure 2:Weka GUI chooser**

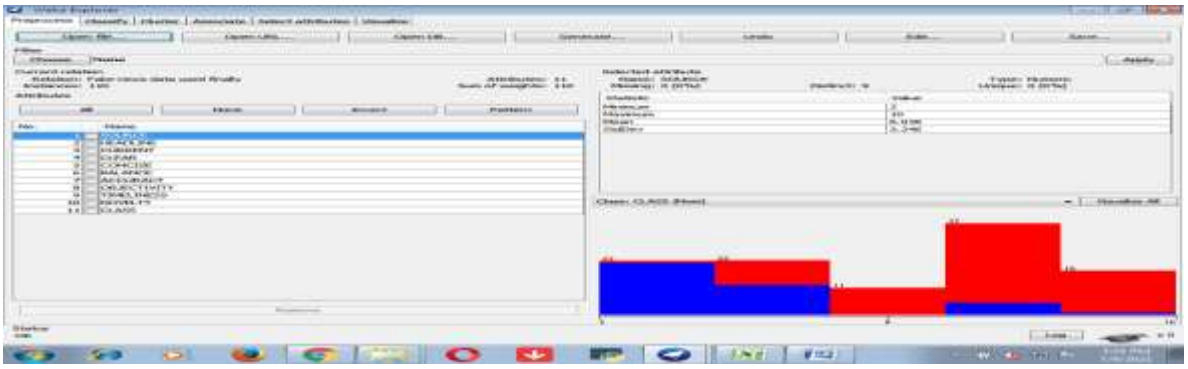

**Figure 3: Graphical Representation of the ratio of fake news to Genuine news in the dataset used**

**Figure 4: Weka Explorer showing classification output using Random Forest**

IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 8, August 2021

DOI 10.17148/IJARCCE.2021.10820

**Figure 5: Weka Explorer showing classification output using Random Forest (contd)**

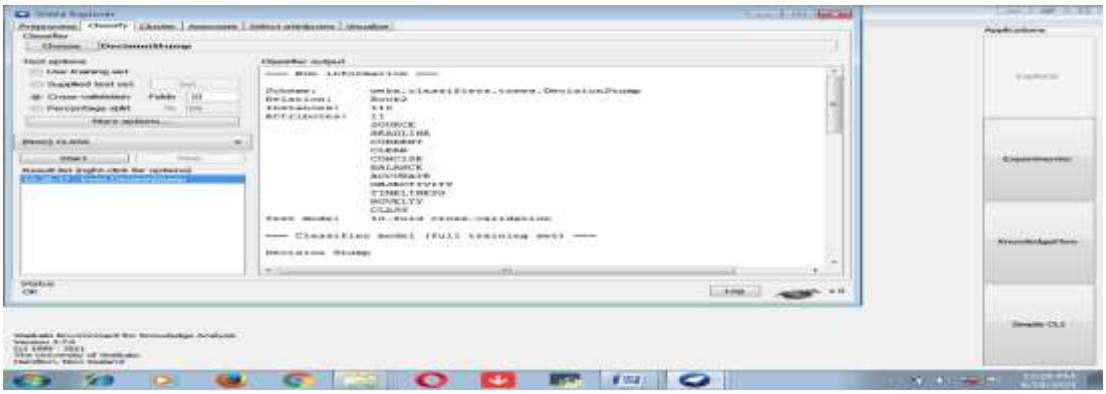

**Figure 6: Weka Explorer showing classification output using Decision tree**

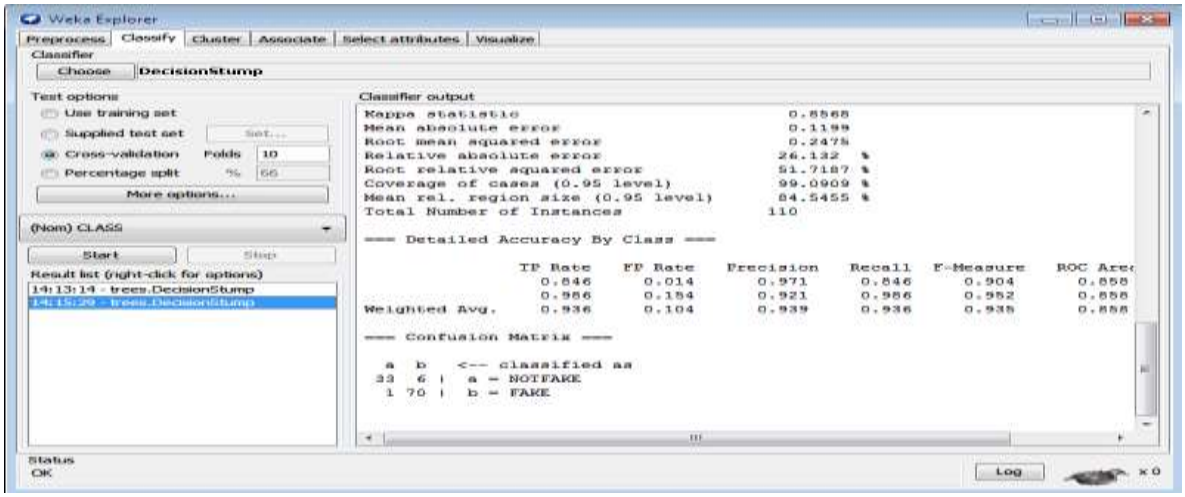


**Figure 7: Weka Explorer showing classification output using Decision tree (contd)**

### 4.1 Results

The random forest result shows a confusion matrix with True Positive (TP) of 39, False Negative (FN) of 0, False Positive (FP) of 0 and True Negative of 71. With the confusion matrix, the performance measures are computed using the formulas explain in section 3 and the following results obtained:

- Accuracy = 1.
- Recall = 1.
- Precision = 1.
- F1- score = 1.

The correctly classified instance is 100% and incorrectly classified instance is 0%.

  

DOI 10.17148/IJARCCE.2021.10820

The decision tree result show a confusion matrix with True Positive (TP) of 33, False Negative (FN) of 6, False Positive (FP) of 1 and True Negative of 70. With the confusion matrix, the performance measures are computed using the formulas explain in section 3 and the following results obtained:

- Accuracy = 0.9364.
- Recall = 0.8462.
- Precision = 0.9706.
- F1- score = 0.9042.

The correctly classified instance is 93.6364% and incorrectly classified instance is 6.3636%.

## 5. CONCLUSION AND RECOMMENDATION FOR FUTURE WORK

The results obtained show that Random Forest is a better classification tool with correctly classified instance of 100% and incorrectly classified instance of 0% as compared to the decision tree with correctly classified instance of 93.6364% and incorrectly classified instance of 6.3636%.

It is recommended that future studies be carried out in the area of fake news prevention so that fake news after being detected can be blocked from gaining access into the society.

## REFERENCES

[1]. Allcott T and Gentzkow, G,.(2017). Social Media and Fake News in the 2016 Election Journal of Economic Perspectives Vol. 31, No. 2, Spring (pp. 211-36)

[2]. Aliyuda K., Howell J., (2019). Machine learning algorithm for estimating oil recovery factor using a combination of engineering and stratigraphic dependent parameters. Interpretation SE151{SE159. doi:https://doi.org/10.1190/INT-2018-0211.1.

[3]. Bahzad Taha Jijo and Adnan Mohsin Abdulazeez (2021) Classification Based on Decision Tree Algorithm for Machine Learning. Journal of applied science and technology trends, Vol.02, No.01, pp. 20 – 28 (2021) ISSN: 2708-0757

[4]. Breiman, L., Friedman, J., Olshen, R., and Stone, C., (1984) Classification and Regression Trees, Springer, Berlin, Germany.

[5]. Claire Wardle (2017) Fighting Fake News: Workshop Report hosted by The Information Society Project The Floyd Abrams Institute for Freedom of Expression.

[6]. Edson C. Tandoc Jr., Zheng Wei Lim and Richard Ling (2017) DEFINING "FAKE NEWS" A typology of scholarly definitions. Digital Journalism, https://doi.org/10.1080/21670811.2017.1360143

[7]. Gregorutti, B. Michel, and P. Saint-Pierre (2017) "Correlation and variable importance in Random forests," Statistics and Computing, vol. 27, no. 3, pp. 659–678.

[8]. Gulizar Haciyakupoglu, Jennifer Yang Hui, V. S. Suguna, Dymples Leong, and Muhammad Faizal Bin Abdul Rahman (2018). Countering Fake News a Survey of Recent Global Initiatives. Policy Report S. Rajaratnam School of international studies.

[9]. Iftikhar Ahmad ,Muhammad Yousaf, Suhail Yousaf , and Muhammad Ovais Ahmad (2020) Fake News Detection Using Machine Learning Ensemble Methods Hindawi Complexity Volume 2020, Article ID 8885861, 11 pages https://doi.org/10.1155/2020/8885861

[10]. Ivan Makhotina, Denis Orlova, Dmitry Koroteeva, Evgeny Burnaeva, Aram Karapetyanb, Dmitry Antonenko (2021 ). Machine learning for recovery factor estimation of an oil reservoir: a tool for de-risking at a hydrocarbon asset evaluation. Journal of Petroleum Science and Engineering 184 106513.

[11]. Julie Posetti, Cherilyn Ireton, Claire Wardle, Hossein Derakhshan, Alice Matthews, Magda Abu-Fadil, Tom Trewinnard, Fergus Bell, Alexios MantzarlisJOURNALISM(2018). 'FAKE NEWS' & DISINFORMATION Handbook for Journalism Education and Training UNESCO Series on Journalism Education.

[12]. João Pedro Baptista and Anabela Gradim (2020) Understanding Fake News Consumption: A Review. Soc. Sci. 2020, 9, 185; doi:10.3390/socsci9100185

[13]. Oberiri Destiny Apuke and Bahiyah Omar(2020) Fake News Proliferation in Nigeria: Consequences, Motivations, and Prevention Through Awareness Strategies

[14]. Xichen Zhang⋆ , Ali A. Ghorbani (2021) An overview of online fake news: Characterization, detection, and discussion. Canadian Institute for Cybersecurity (CIC), Faculty of Computer Science, University of New Brunswick (UNB), Fredericton, NB E3B 5A3, Canada

[15]. Marianna Isaakidou, Emmanouil Zoulias and Marianna Diomidous (2021) Machine Learning to Identify Fake News for COVID-19 European Federation for Medical Informatics (EFMI) and IOS Press.