RESEARCH ARTICLE OPEN ACCESS

An Analysis of Machine Learning Techniques for Verifying and Improving Veracity in Big Data

[(1)]Dr. Anazia Eluemunor Kizito [(2)]Dr. Okpako Abugo Ejaita

[(1)]Lecturer, Delta State University of Science and Technology, Ozoro, Nigeria.
[(1)]email: kaymax07@yahoo.com

[(2)]Lecturer, Edwin Clark University, Kiagbodo, Nigeria.
[(2)]email: okpako.ejaita@ edwinclark university.edu.ng

Abstract:

Veracity is one of the characteristics of big data that deals with the quality, trustworthiness, accuracy, authenticity, truthfulness, provenance, unbiasness, and completeness of big data. It is observed that one of the major problems associated with big data at point of generation and storage, is usually itspoor veracity rate. Veracity problems in big data are due to activities like wilful falsity, domain negligence, value misrepresentation, deliberate biasness, inaccuracy in device measurement, computing errors, hacking, social engineering and other security breaches. In order to reduce and manage these veracity problems in big data so many methods and techniques have been proposed. In this research work, Machine Learning Techniques and Algorithms is proposed as the means of improving and verifying the Veracity in big data. We understudied the various techniques, algorithms, advantages and disadvantages of Machine Learning Techniques applicationsof improving and verifying the Veracity in big data classification.

Key Words: ***Big Data, Veracity, Machine Learning, Supervised, Semi-Supervised and Unsupervised, Reinforced Learning***.

Introduction

Computing recently has witnessed a surge in data generation which was as a result of advancement in new technologies, communication devices, applications like internet of things, smart devices, social media and new trends of doing businesses. It was recorded that Facebook had around 42 likes in every minute, Twitter had about3.5 lakhs tweets per minute, on YouTube, it got nearly 300 hours of data being generated per minute and other massive amount of data that are also generated through other social media platforms [1]. According to [2], Facebook only had 2.38 billion monthly active subscribers in the first quarter of 2019 and generated the sum of 55.8 billion US Dollars in the year 2018. These huge chunk of data sets with variant constituents generated in high velocity are referred to as big data. The term big data was first coined by Roger Magoulas of O'Reilly media when he was trying to describe a large amount of data, in variant formats that are generated in a very high speed and cannot be processed by traditional means of data processing. It is seen as the frontier of a firm's ability to store, process, and access all the data it needs to operate effectively, make decisions, reduce risks, and serve customers [3]. It was reported as High-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation [4].Big data are usually in huge volume, moves too fast, or doesn't fit the structures ofthe traditional database architectures and to gain value from this data, you must choose an alternative way to process it.Reference [4] had the first classification of characteristics of big data into Volume, Velocity and Variety after proper consideration of its complexity, size and speed of processing, difficulty of being managed by conventional database systems and inherent benefits, which is referred to as the 3Vs of big data. After further research, IBM introduced the fourth V which represents Veracity of data and other Vs were later proposed to handle emerging problems from the use of Big Data. Till date, the Big Data is attributed with more than 45Vs which includes Value, Validity, Variability, Volatility and Visualization etc.The question at this point is how far can users trust and relay on big data in the decision-making and problem solving? This brings in the fourth V of big data, known as "Veracity". Veracity of big data has to deal with the quality, trustworthiness and unbiased nature of big data. On a general note, it encompasses data inconsistency, data incompleteness, data freshness and timeliness, data uncertainty, error in data, provenance in data, fake data/information, security issues etc. Arguably, veracity is one of the less researched Vs of the big data but the onus of producing good result and making the right decision for business lies on the veracity of the big data used. This is so because since the axiom of computing says garbage in, garbage out it can be extended as "good data begets good information" high quality data is likely to produce high quality information while low quality data will surely produce poor information. It has been observed that one of the uphill tasks in the analysis and use of big data is determining and using a data of high quality.

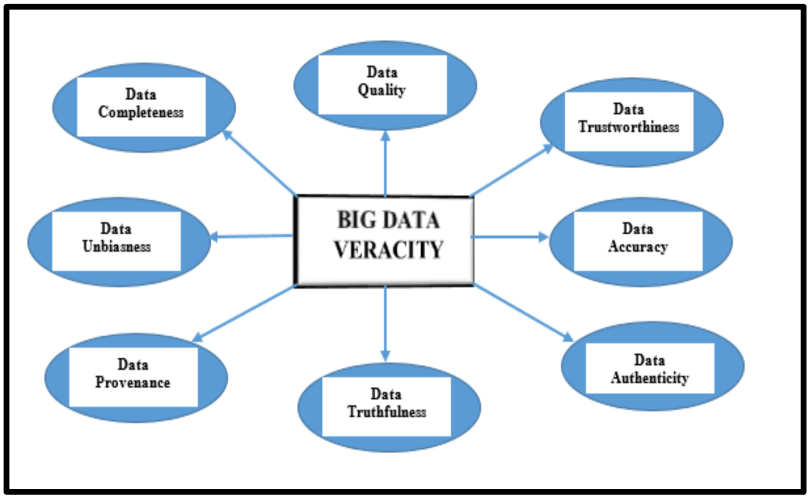

Figure 1: Components of Big Data Veracity

In trying to get it right with the veracity of big data, the choice of method of improving and verifying its truthfulness and quality of the dataset is sacrosanct and it should be handled with the utmost care and expertise. The earlier method of processing data with RDBMS is no longer practicable probably because of the complex nature of big data that is why more intelligent and interactive systems like machine learning, change detection, optimization techniques, natural language processing, formal methods, fuzzy logic, collaborative filtering, similarity measurements, blockchain technology amongst others.In this research work, it will be limited to only machine learning based approach to improving and verifying the veracity in big data. Machine learning (ML) is defined as a branch of Artificial Intelligence (AI) whose fundamental principal is that by giving access to the right data, machines can learn and solve problems on their own with little or no human interference. It works by leveraging complex mathematical and statistical that is capable of performing independently intellectual tasks that have been traditionally solved by human beings [5].

Statement of the Problem

The big data trend has changed the way both formal and non-formal businesses are done by making business approaches better, smarter, more intelligent and it has also improved the decision making, profit and participation of business with increase in growth rate. Some of the major drawbacks of Big Data Analytics is the problem of big data uncertainty, incompleteness, quality, biasness and ambiguity and this has to do with big data veracity. Veracity of big data is defined as the underlying accuracy or its lacks, of the data in question, specifically imparting the ability to derive actionable belief and value on the data [6]. The problem of big data veracity is on the increase with the penetration of internet and proliferation of hand-held and smart devices. These poor veracity rate in big data can be viewed from the angles of real-world dynamism, wilful falsity, domain negligence, value misrepresentation, deliberate biasness, inaccuracy in device measurement, computing errors, hacking, social engineering and other security breaches. These chunks of big data are from millions of emails, tweets, Google search, online transactions, likes/comments, blog posts/checks, spatial documents, other multimedia formats etc. Among many methods used in solving big data veracity problems, we are proposing the use of machine learning techniques/algorithms to improve and verify the veracity in big data in order to harness its full potentials.

Review of Related Literature

With so many underlying potentials of big data as proposed by data scientists, one of the best ways to tap into these proposed benefits is by improving and verifying the veracity of these huge datasets known as big data. Big data veracity is an important part of making big data meaningful and giving it its desired quality.Big data veracity refers to the uncertainty of available data; in such a case, quality and accuracy are difficult to control [7]. Researchers and data scientists has proposed several mechanisms and techniques to measure the veracity of big data, some of the techniques depend on information provenance as veracity ontology model [8]. Theveracity of big data is a crucial aspect of making sense and getting value out of big data and one of the most reliable and proven methods of doing that is machine learning techniques [6]. He further stated that when the ground truth is not reliable, even the best quality model on top of it will not be able to perform which machine learning techniques has helped in establishing the ground truth and building a reliable model.

It was proposed by [9] that machine learning techniques (supervised, semi supervised, unsupervised or reinforced) has proven to give the best classification of tweets (big data) during emergency, stating the true conditions of victims compared to other methods. Be that as it may, recent explosions in client-produced content on social sites are introducing unique difficulties in capturing, examining and translating printed content since information is scattered, confused, and divided but using machine learning algorithm will help in solving these problems of classification [10].

It was opined by [11] that from a technical point of view, identified machine learning, lexicon-based, statistical and rule-based approaches of verifying the veracity in big data but machine learning methods still has the best model and a better output.Big data veracity verification is usually performed by using techniques like natural language processing (NLP), machine learning, text mining/information theory and coding semantic approach but machine learning algorithms has advantages like continuous improvement, fully automation-based, trends and patterns identification, general applications andefficient handling of data etc[12]. Thehybrid methods also performed well and obtained reasonable classification accuracy scores, since they were able to take advantage of both machine learning classifiers and lexicon-based Twitter sentiment-analysis approaches [13].Despite all other methods proposed for big data veracity assessment, machine learning techniques stood out to be one of the most reliable and frequently used in assessing the veracity in big data [14].It was stated by [35] that process of analysing or classifying Online Social Media platforms contents into different polarities or classes like positive, neutral and negative is known as Sentiment Analysis.

Machine Learning

The word Machine Learning was coined by Arthur Samuel in the year 1959 and he defined it as a computer field that uses statistical methods to give computer system the ability to learn with data without being explicitly programmed. It is also seen as a branch of Artificial intelligence (AI) whose objective is to understand the structure of data and fit it into models that can be understood and utilized by people [15]. It makes computer to train on data inputs and use statistical analysis in order to produce result values that falls within required range.

Classification of Machine Learning Techniques

According to [15], Machine Learning is classified as Supervised Machine Learning, Unsupervised Machine Learning, Semi-supervised Machine Learning and Reinforcement Machine Learning Techniques.

(a) Supervised Machine Learning Techniques

Supervised Machine Learning Techniquesapply what has been learnt in the past to new data using labeled examples to predict future events. It starts with the analysis of a known training data set, which produces an inferred function to make predictions about the output values [15]. The system is able to provide targets for any new input after sufficient training. The algorithms in supervised machine learning can also compare its output with the correct, intended output and find errors in order to modify the model accordingly. Examples of algorithms of supervised machine learning techniquesare Nearest Neighbour, Naïve Bayes, Decision Tree, Support Vector Machine, Linear Regression and Neural Network, Logistic Regression, Random Forest etc.

(b) Unsupervised Machine Learning Techniques

Unsupervised Machine Learning Techniquesare used when the information used to train is neither classified nor labeled. Unsupervised learning studies how a system can infer a function to describe a hidden structure from unlabeled data [17]. The system does not figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data. K-Means, Hierarchical Cluster

Analysis, Expectation Maximization, Principal Component Analysis, T-Distributed Stochastic Neighbour Embedding are the examples of algorithms of unsupervised learning Techniques.

(c) Semi-Supervised Machine Learning Techniques

This type of machine learningtechniquesis a mid-point betweensupervised and unsupervised learning, because they use both labelled and unlabelled data for training mostly a small amount of labelled data and a large amount of unlabelled data[17]. Systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labelled data requires skilled and relevant resources in order to train it or learn from it. Otherwise, acquiring unlabelled data generally does not require additional resources. Examples of semi-supervised learning algorithms is the Deep Belief Neural Network.

(d) Reinforcement Machine Learning Techniques

Reinforcement Learning is actually different from supervised, unsupervised and semi supervised types of machine learning techniques. The Reinforcement Learning technique is designed to learn and become intelligent from its previous mistakes that is to say it is meant to learn based on the axiom that "experience is the best teacher". Reinforcement Learning is also referred to as the trial and error machine learning method because from the previous mistakes made, the machine can learn and understand the causes of those mistakes and try to avoid such whenencountered again [16].

A typical example of Reinforcement Learning is the learning process of human being most especially babies. If they touch fire by accident or knowingly, they will feel the pain, and they will never touch fire again in their entire life unless it is an accident. In Reinforcement Learning, the learning system is referred to as an agent. This system must learn by itself, which is the best strategy, known as a policy, to get the most positive reward over time.

The steps normally found in reinforcement learning is grouped into six steps.

i. Observe
ii. Select an action using the policy
iii. Do the action
iv. Get positive or negative rewards
v. Update the policy by analysing the rewards
vi. Repeat the same process until an optimal policy is obtained.

Many robots learn how to walk by implementing reinforcement learning. Example of algorithmsof reinforced machine learning technique is Q-learning algorithm and Temporalalgorithm.The diagram below shows the various steps followed in machine learning techniques.

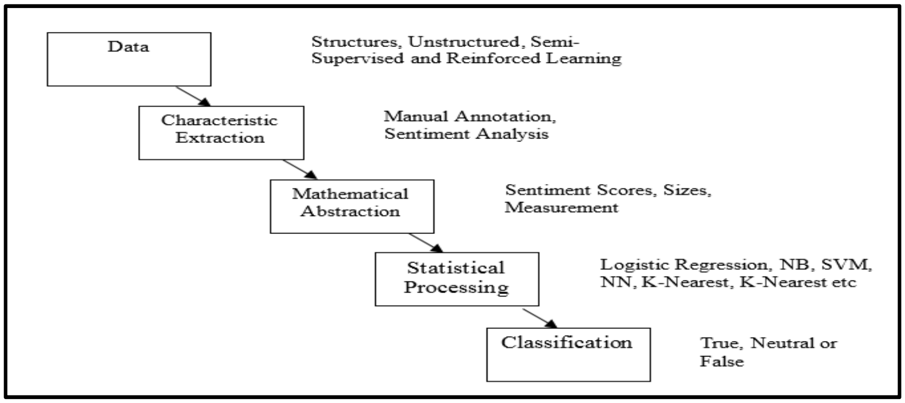
Data
Structures, Unstructured, Semi-Supervised and Reinforced Learning
Characteristic Extraction
Manual Annotation, Sentiment Analysis
Mathematical Abstraction
Sentiment Scores, Sizes, Measurement
Statistical Processing
Logistic Regression, NB, SVM, NN, K-Nearest, K-Nearest etc
Classification
True, Neutral or False

Figure 2: A general structure of Machine Learning Techniques

(a) Supervised Learning

(i) K-Nearest Neighbour Algorithm

K-Nearest Neighbour usually referred to as KNN is a learning algorithm that can be used to solve both classification and regression problems. It stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classify a data point based on how its neighbours are classified and assumes that similar things exist in close proximity. KNN determines similarity between points also known as distance, proximity, or closeness with some mathematical formula like Euclidean Function, Manhattan Function, Murkowski Function, Hammering Function etc[16].

The steps to implement KNN Algorithm is summarized below.

i. Load the data.
ii. Initialize K to your chosen number of neighbours.
iii. For each example in the data.
iv. Calculate the distance between the query example and the current example from the data.
v. Add the distance and the index of the example to an ordered collection.
vi. Sort the ordered collection of distances and indices in ascending order.
vii. Pick the first K entries from the sorted collection.
viii. Get the labels of the selected K entries.
ix. If regression, return the mean of the K labels
x. If classification, return the mode of the K labels

(ii) Naïve BayesAlgorithm

According to [18], it is a probabilistic based learning algorithm that is used in machine learning for different types of task classifications and predications that has its roots on a statistical theorem known as Bayes theorem created by Rev. Thomas Bayes (1702–61). The name Naïve is used because it assumes the features that go into the model is independent of each other. It implies that changing the value of one feature, does not directly influence or change the value of any of the other features used in the algorithm. Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. It assumes that the presence of one particular feature does not affect the other. Naïve Bayes Algorithm is used in spam filtering, classifying documents, sentiment prediction etc. It can be further divided into three types; Multinomial Naive Bayes, Bernoulli Naive Bayes and Gaussian Naive Bayes.

The equation below expresses Bayes' Theorem.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Where;

P is the symbol to denote probability.
P(A | B) = The probability of event A (hypothesis) occurring given that B (evidence) has occurred.
P(B | A) = The probability of the event B (evidence) occurring given that A (hypothesis) has occurred.
P(A) = The probability of event B (hypothesis) occurring.
P(B) = The probability of event A (evidence) occurring.

When using Naive Bayes algorithm to calculate the probability of an event, the following steps are taken;

i. Calculate the prior probability for given class labels
ii. Find Likelihood probability with each attribute for each class
iii. Put these value in Bayes Formula and calculate posterior probability.
iv. See which class has a higher probability, given the input belongs to the higher probability class.

(iii) Decision TreeAlgorithm

Decision Tree Algorithm is a supervised learning algorithm that can be used for solving regression and classification problems by creating a training model that can be used to predict the class or value of the target variable [19].It is defined as an upside-down tree that makes decisions based on the conditions present in the data. Decision Trees is based on learning simple decision rules inferred from prior data (training data) starting from the root node (parent) to the leaf node (children). It classifies the examples by sorting them down the tree from the root to some leaf/terminal node, with the leaf/terminal node providing the classification of the example. Each node in the tree acts as a test case for some attribute, and each edge descending from the node corresponds to the possible answers to the test case. This process is recursive in nature and is repeated for every sub-tree rooted to the new node.

Below is a decision tree for weather forecast to determine whether to go out or not.

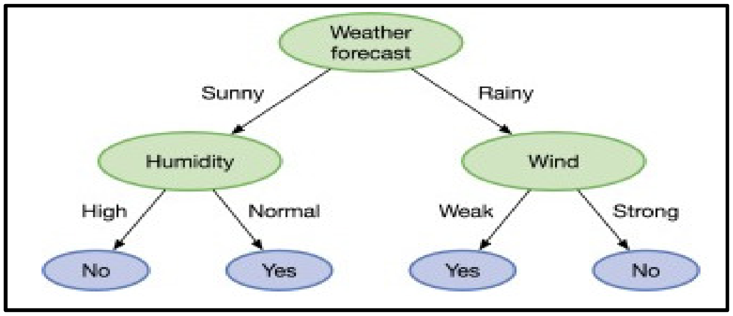

Figure 3: Structure of a Prediction Output of a Decision Tree Algorithm

The following assumptions can be made when the implementing decision tree;

i. Discretization of continuous variables is required.
ii. The data taken for training should be wholly considered as root.
iii. Distribution of records is done in a recursive manner on the basis of attribute values.

(iv) Support Vector MachineAlgorithm

Support Vector Machine (SVM) is a linear model for handling classification and regression problems that can solve linear and non-linear problems. SVM algorithm creates a line or a hyperplane which separates the data into different classes of into positive and negative classes [6]. The SVM algorithm indicates the points closest to the line from the classes and these points are called Support Vectors. Support Vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. The distance between the line and the support vectors are computed which is known as the “Margin” and the goal of SVM is to maximize the margin. Support Vector Machine uses kernel functions to model its classifier. The diagram below show how SVM algorithm classifiers datasets

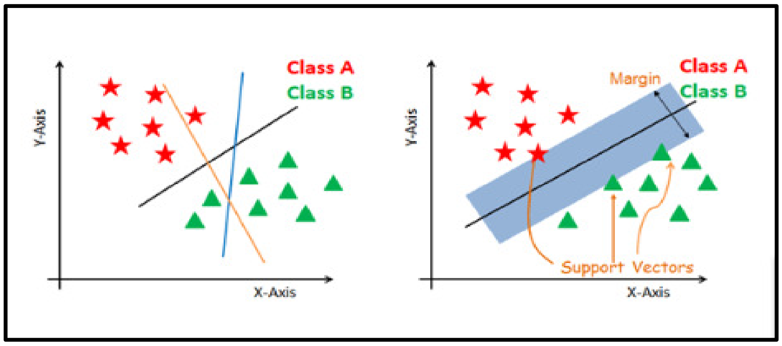

Figure 4: SVM Classifier

SVM searches for the maximum marginal hyperplane using the following steps:

i. Generate hyperplanes which segregates the classes in the best way. Left-hand side figure showing three hyperplanes black, blue and orange. Here, the blue and orange have higher classification error, but the black is separating the two classes correctly.
ii. Select the right hyperplane with the maximum segregation from either the nearest data points as shown in the right-hand side figure

(v) Linear Regression Algorithm

Linear Regression is a statistical analysis that tends to show a relationship between two variables or points by considering the various data points and plots a trend line. It models the relationship between two variables by fitting a linear equation to observed data [20].Linear Regression considers one variable to be an explanatory variable, and the other is considered to be a dependent variable. The variable that is to be predicted is called the dependent variable or outcome variable while the variable used to predict the other variable's value is called the independent variable or predictor variable. It can create a predictive model on apparently random data, showing trends in data, such as how students will pass in an examination considering their course work performance, the rate of consumption of a particular goods considering the number of people that are involved in the consumption of the product. The steps below are used to implement Linear Regression Algorithm;

i. Calculate Mean and Variance.
ii. Calculate Covariance.
iii. Estimate Coefficients.
iv. Make Predictions.
v. Predict Insurance.

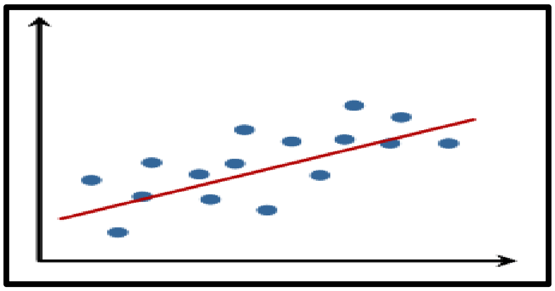

Figure 5: A Graph Plotted Using Linear Regression Algorithm

(vi) Neural Network Algorithm

Neural Network Algorithmis inspired by the working of neurons in human brain. The complex workings of the biological neuron are modelled through sophisticated abstraction that is used to solve real-world problems across all disciplines. Neural Network is seen as a massively parallel distributed processor made up of simple processing units that has a natural propensity for storing experimental knowledge and making it available for working with the fundamental of in knowledge acquisition from its environment and the interneuron connection strength known as synaptic weight which is used to store the acquired knowledge [22]. Each neuron can be thought of as a node and interconnection between them is an edge, which has weight associated with it, which represents how the two neurons which are connected can interact with each other [21 The neurons work in such a way that they can imitate or live in each other if there is an edge between two of them. When two neurons; A and B that has a weight w are interconnected to each other and adequate simulation is generated between two of them, it therefore means that a signal can be initiated between two of them which is dependent on the weight w and the nature of signal generated. Also the signal sent can either be positive or negative which can be sent not only to A and B but to all other interconnected nodes. The features or characteristics of the dataset used in training are used as input and it combines with the weight w to produce the output. The output can be made more accurate by introducing more neurons between the input and an output layer to generate the output classification as true or false, which means mores connection begets more processing. These function doses not receive input from neither external source nor output to the external source and thus referred to as “Hidden Layer”. Weight is a measurement of input connection strength that can be modified in response to various training dataset and according to a network specified topology or through its learning rule [21].

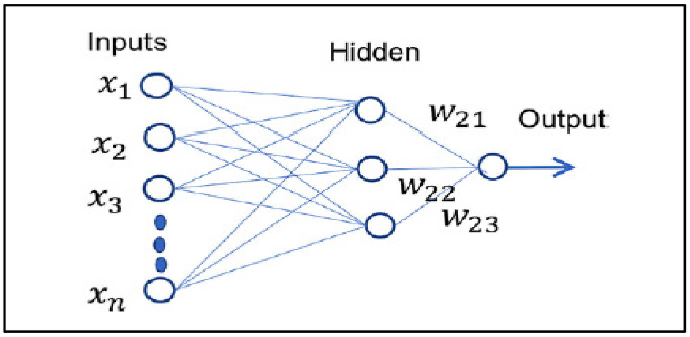

Figure 6: A Neural Network with Input and Output Layers.

The training process consists of the following steps using;

i. Forward Propagation: Take the inputs, multiply by the weights (just use random numbers as weights) and pass the result through a Sigmoid function to calculate the neuron's output. The Sigmoid function is used to normalise the result between 0 and 1.

ii. Backward Propagation: Calculate the error i.e. the difference between the actual output and the expected output. Depending on the error, adjust the weights by multiplying the error with the input and again with the gradient of the Sigmoid curve.

iii. Repeat the whole process as much as possible in order to improve the training and better the result.

(vii) Logistic RegressionAlgorithm

Logistic Regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. According to [23], the name Logistic Regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No while multinomial logistic regression is usually reserved for the case when the dependent variable has three or more unique values, such as True, Neutral and Negative. Although the type of data used for the dependent variable is different from that of multiple regression but the practical use of the procedure is similar. Many users feel that logistic regression is more versatile and better suited for modelling most situations than its discriminant analysis. It provides confidence intervals on predicted values, and provides ROC curves to help determine the best cut off point for classification. It allows users to validate results by automatically classifying rows that are not used during the analysis.

The steps below can be followed to implement Logistic Regression by Stochastic Gradient Descent.

i. Calculate Prediction. The process begins by assigning 0.0 to each coefficient and calculating the probability of the first training instance that belongs to class 0.

ii. Calculate New Coefficients.

iii. Repeat the Process.

iv. Make Predictions.

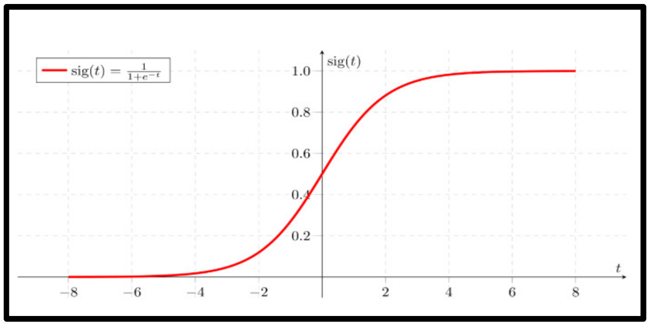

Figure 7: Logistic Regression Analysis Graph

Logistic Function is used to transform this score into a probability and the parametric statistical model adopted is known as logistic regression while the graph that has S shape is commonly referred to as Sigmoid Functions [6].

(viii) Random Forest Algorithm

Random Forest Algorithmbuilds multiple decision trees and merges them together to get a more accurate and stable prediction [24]. One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally produces a better model. Therefore, in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. Trees can be made more random by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

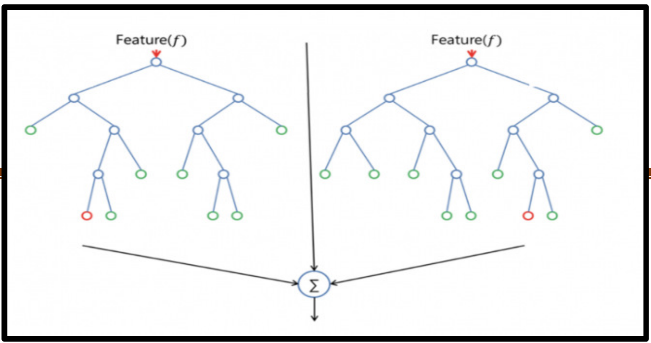

Figure 8: Random Forest Algorithm graph

The steps below indicates how the Random Forest Algorithm works.

i. Pick N random records from the dataset.
ii. Build a decision tree based on these N records.
iii. Choose the number of trees you want in your algorithm and repeat steps 1 and 2.
iv. In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output).

(b) Unsupervised Learning.

(i) K-Means Algorithm

K-Means clustering is a type of unsupervised learning that handles clustering problems when unlabelled or uncategorized data are used[25]. The aim of K-means clustering algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of the K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

i. The centroids of the K clusters, which can be used to label new data
ii. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering is used to find and analyse the groups that have formed organically. Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents. The diagram below shows how data are represented in a graph using K-means clustering algorithm.

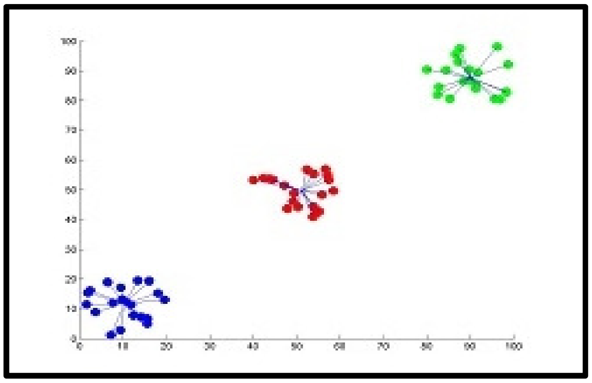

Figure 9: K-Means Algorithm

The way K-means algorithm works is as follows:

i. Specify number of clusters *K*.
ii. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
iii. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

iii a. Compute the sum of the squared distance between data points and all centroids.

iii b. Assign each data point to the closest cluster (centroid).

iii c. Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

(ii) Hierarchical Cluster Analysis Algorithm

Hierarchical Cluster Analysis, is an algorithm that puts similar objects into groups called clusters. It involves creating clusters that have a predetermined ordering from top to bottom. The endpoint is a set of clusters or groups, where each cluster is distinct from the other clusters, and the objects within each cluster are broadly similar to each other.There are two types of hierarchical clustering, Agglomerative Hierarchical Clustering and Divisive Hierarchical Clustering[26].

Agglomerative Hierarchical Clustering
It is also known as AGNES (Agglomerative Nesting). It works in a bottom-up manner. That is, each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes). This procedure is iterated until all points are member of just one single big cluster (root).

Divisive Hierarchical Clustering
It is also known as DIANA (Divise Analysis) and it works in a top-down manner. The algorithm is an inverse order of AGNES. It begins with the root, in which all objects are included in a single cluster. At each step of iteration, the most heterogeneous cluster is divided into two. The process is iterated until all objects are in their own cluster. The result of Agglomerative hierarchical clustering and Divisive hierarchical clustering is a tree which can be plotted as a Dendrogram as shown below.

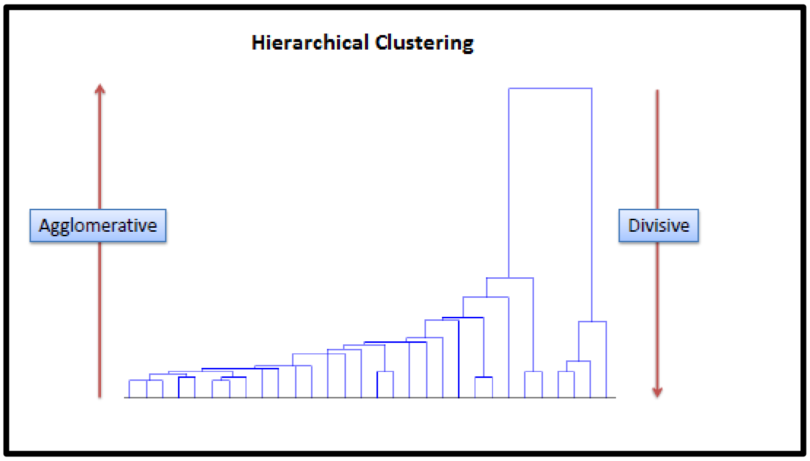

Figure 10: A Dendrogram from Hierarchical Cluster Analysis Algorithm

Steps to Perform Hierarchical Clustering

i. Assign all the points to an individual cluster.
ii. Identify the smallest distance in the proximity matrix and merge the points with the smallest distance and update the proximity matrix:
iii. Repeat step 2 until only a single cluster is left.

(iii) Expectation Maximization Algorithm

ExpectationMaximization (EM) Algorithm is an iterative method to find (local) maximum like hood of parameter in a statistical models where the model depends on unlabelled or unobserved latent variables[33]. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likehood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E* step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.
Expectation Maximization algorithm can be summarized as shown below;

i. Given a set of incomplete data, consider a set of starting parameters.

ii. Expectation step (E – step): Using the observed available data of the dataset, estimate (guess) the values of the missing data.
iii. Maximization step (M – step): Complete data generated after the expectation (E) step is used in order to update the parameters.
iv. Repeat step 2 and step 3 until convergence.

(iv) Principal Component Analysis Algorithm

Principal Component Analysis (PCA) Algorithmis a multivariate statistical procedure that summarizes information content in large data tables by means of a smaller set known as summary indices that can be easily visualized and analysed[28]PCA also helps to identify correlations between data points, such as whether there is a correlation between a particular data and the other. Principal Component Analysis forms the basis of multivariate data analysis based on projection methods. The most important use of Principal Component Analysis is to represent a multivariate data table as smaller set of variables in order to observe trends, jumps, clusters and outliers. Principal Component Analysis statistically finds lines, planes and hyper-planes in the K-dimensional space that approximate the data as well as possible in the least squares sense. A line or plane that is the least squares approximation of a set of data points makes the variance of the coordinates on the line or plane as large as possible as shown in the graph below.

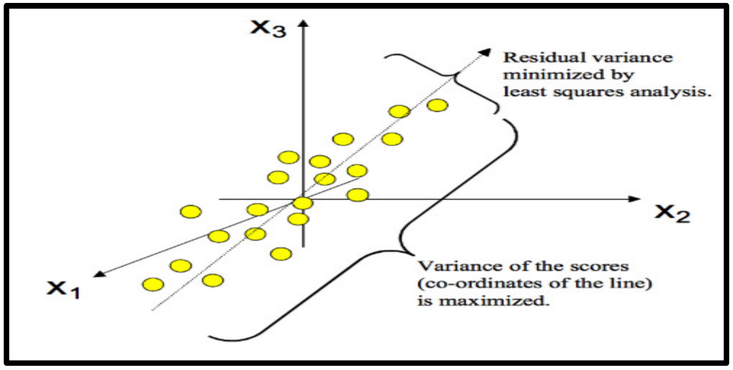

Figure 11: A Graph of Principal Component Analysis Algorithm

Steps to Perform Hierarchical Clustering

i. First, assign all the points to an individual cluster.
ii. Identify the smallest distance in the proximity matrix and merge the points with the smallest distance. Then update the proximity matrix.

Repeat step 2 until only a single cluster is left.

(v) T-Distributed Stochastic Neighbour Embedding Algorithm

T-Distributed Stochastic Neighbour Embedding (t-SNE) is an unsupervised and non-linear learning method that is used for exploration and visualizing high-dimensional data which was developed by Laurens van der Maatens and Geoffrey Hinton in 2008. Though it is similar to PCA but while PCA is used in preserving large pairwise distances, T-Distributed Stochastic Neighbour Embedding is used in preserving only small pairwise distances or local similarities to maximize variance[28]. It is extensively applied in image processing, NLP, genomic data and speech processing. The diagram below shows how the output of T-Distributed Stochastic Neighbour Embedding is displayed in a graphical manner.

Steps to implement T-Distributed Stochastic Neighbour Embedding Algorithm

i. Measure similarities between points in the high dimensional space.
ii. Use a Student t-distribution with one degree of freedom, which is also known as the Cauchy distribution.
iii. Let the set of probabilities from the low-dimensional space (Qij) to reflect those of the high dimensional space (Pij) as best as possible.

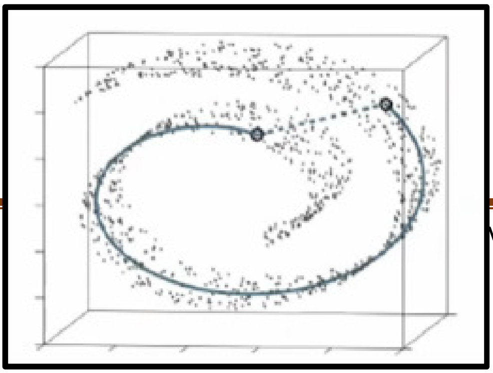

Figure 12: A T-Distributed Stochastic Neighbour Embedding Algorithm Graph

(c) *Semi-Supervised Learning* Techniques

(i) Deep Belief Neural Network Algorithms

This is a type of Deep Neural Network that comprisesof multiple layers of latent variables, with connections between the layers but not between units within each layer. It uses probabilities and unsupervised learning to produce outputs. Each layer in Deep Belief Networks learns the entire input and work globally to regulate each layer orderly while in convolutional neural networks, the first layers only filter inputs for basic features, such as edges, and the later layers recombine all the simple patterns found by the previous layers[29].The diagram below shows a Schematic overview of a Deep Belief Neural Network with arrows representing directed connections in the graphical model.

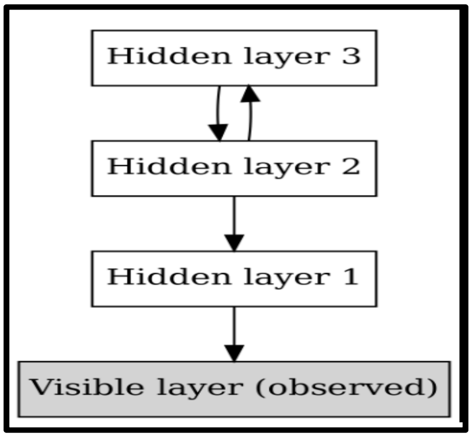

Figure 13: The Different Layers Shown in a Deep Belief Neural NetworkAlgorithms

Steps to implement Deep Belief Neural Network training using Greedy Layer Wise approach;

i. First layer is trained from the training data greedily, while all other layers are frozen and this is known as the positive phase**.**

ii. Reconstruct the visible units using negative phase which is similar to the positive phase.

iii. Update all associated weights. L is regarded as the learning rate that will be multiplied by the difference between the positive and negative phase values and add to the initial value of the weight.

(d) Reinforcement Machine Learning Algorithms

(i) Q-learning Algorithm

Q-learning is regarded as an off policy or free model reinforcement learning algorithm which seeks to find the best action to take given the current state[30]. The "Q" stands for quality, in this case it represents how useful a given action is in gaining some future reward. It is considered as an off-policy or free model because the q-learning function learns from actions that are outside the current policy, like taking random actions, and therefore a policy is not needed. More specifically, q-learning seeks to learn a policy that maximizes the total reward. The steps bellows summarizes how to implement Q-learning Algorithm.

i. First, initialize the Q function to some arbitrary value.
ii. Take an action from a state using epsilon-greedy policy and move it to the new state.
iii. Update the Q value of a previous state by following the update rule.
iv. Repeat steps ii and iii until it gets to the terminal state.

(ii) Temporal DifferenceAlgorithm.

Temporal Difference (TD) Learning is an approach on how to predict a quantity that depends on future values of a given signal [27]. It is a Reinforcement Learning Algorithm in which the learner uses the difference between the expected and actual rewards received in the different states of the world visited in a given episode to revise the expected value of the states. The acronym TD is derived from its changes or differences that occurs during predictions over successive time steps to drive the learning process. The prediction at any given time step is updated to bring it closer to the prediction of the same quantity at the next time step.

The steps involved in the TD prediction algorithm are summarized as follows;

i. First, initialize V(S) to 0 or some arbitrary value.
ii. Then, begin the episode and for every step in the episode, perform an action A in the state S and receive a reward R and move to the next state (s).
iii. Update the value of the previous state using the TD update rule.
iv. Repeat steps 2 and 3 until we reach the terminal state.

Advantages of Machine Learning

(i) Continuous Improvement

In using Machine Learning algorithms to verify big data veracity, it provides room for continuous improvement of the model because of its capability of learning from the datasets used. Whenever a new dataset is provided, the model's accuracy and efficiency to make decisions improve with subsequent trainings. The accuracy of finding associated products or recommendation engine improves with this huge amount of training data available.

(ii) It is Automation-Based

Machine Learning model are always known for reducing workload and time because nearly everything in the working of the model are automated which makes it very good for improving and verifying big data veracity. This process makes the algorithm to carry out the difficult tasks in the model because of the reliability of machine learning algorithms.

(iii) Trends and Patterns Identification

Machine learning generally are good in trends and patterns identification because its algorithms are very suitable for various classification and regression problems. Theyare used for the analysis buying patterns, search trends of customers and products prediction.

(iv) GeneralApplications

Machine Learning algorithms are used in nearly every industry these days, for example from Defence to Education. Companies generate profits, cut costs, automate, predict the future, analyze trends and patterns from the past data, and many more. Applications like GPS Tracking for traffic, Email spam filtering, text prediction, spell check and correction, etc are a few used widely these days.

(v) Efficient Handling of Data

Machine Learning algorithms are very efficient in Data Science applications which makes it better for big data analytic. It can handle any type of big data format like structured, semi structured and unstructured type of big data.

(b) Disadvantages of Machine Learning

(i) Time-consuming

Machine Learning models are usually used to process high volume of big datasets in veracity verification/improvement and in doing this, the model takes time in training and learning the datasets which may require additional resources for computing.

(ii) Highly Error-Prone

Because machine learning models handles huge amount of datasets, they are usually prone to errors bearing in mind that "Garbage In Garbage Out". So time has to be taken in order to make sure that datasets sent in are very accurate to minimize error during training and learning.

(iii) Data Acquisition

Veracity has to do with the improvement or verification of the quality of big datasets mostly in large volume of bogus and incorrect data. No doubt acquiring the needed data for machine learning training and learning will always put the burden of cost on the developer.

(iv) Choice of Machine Learning Algorithm

It will be observed that they are many machine learning algorithms for big data improvement and verification. Most times data scientists are faced with the problem of choosing the right algorithm for a particular task. So one of the major drawbacks of machine learning is choosing the most efficient and effective algorithm for a particular task because if care is not taken an inappropriate algorithm may be chosen and it will lead to wrong output, high cost, waste of time and resources.

Conclusion and Recommendation

With the increasing demand of big data-driven applications, there is need to improve the quality and ascertain its source. This will help in attaining full utilization of the potentials of big data in decision making, scientific and other applications. Though the procedure of improving and verifying the quality of big data which is referred to as big data "Veracity" has always been an uphill task for data users and data scientists. From our findings, it is obvious that machine learning techniques will be a better method of improving and verifying the veracity of big data which will yield the desired goals in big data technology. For further studies, we will recommend the use of hybrid techniques (combination of two or more machine learning algorithms) and other methods of improving and verifying the veracity of big data.

References

[1] S. Kotiand S.V Seeri, "A Survey on Big Data Issues and Challenges", IOSR Journal of Computer Engineering, 2017), 19(5).

[2] Statista, "Social Media Revenue, In Statista", The Statistics Portal,2019, December, Retrived June, 2020.

[3] M. Gualtieri, "Big Data Predictive Analytics Solutions",The Forrester Wave, USA,2015.

[4] L. Doug, "3D Data Management: Controlling Data Volume, Velocity and Variety, Application Delivery Strategies", Meta Group Publishers, USA, 2001.

[5] S. Marsland,"Machine Learning: An Algorithmic Perspective", CRC Press,2015

[6] V. Pendyala, "Veracity In Big Data: Machine Learning And Other Approaches To Verifying Truthfulness", Apress Publishers, San Jose, California, USA, 2018.

[7] M. Al-Jepooriand Z. A. Al-Khanjari, "Framework for Handling Data Veracity in Big Data" International Journal of Computer Science and Software Engineering (IJCSSE), 2018 Volume 7, Issue 6, www.IJCSSE.org

[8] B. Grégoire, E. C. Amparo, R. Matthew and S. Alfonso, "Representing, Proving and Sharing Trustworthiness of Web Resources Using Veracity" In International Conference on Knowledge Engineering and Knowledge Management, Springer Publishers, 2010.

[9] A. Kaur, "Analyzing Twitter Feeds to Facilitate Crises Informatics and Disaster Response During Mass Emergencies", Dissertation M.Sc. in Computing (Data Analytics), TU Dublin,2019.

[10] A. M. Kaplan and M. Haenlein, “Users of the World, Unite! The Challenges and Opportunities of Social Media”, Business Horizons, 2010. Vol. 53, No. 1.

[11] D’A., Alessia, F. Fernando,Patrizia, G. andG. Tiziana, “Approaches, Tools and Applications for Sentiment Analysis Implementation”, International Journal of Computer Applications, 2015, (0975 – 8887) Volume 125 – No.3.

[12] B. Supriya, V. Moralwar, N. Sachin N. and H. Deshmuk, “Different Approaches of Sentiment Analysis”, International Journal of Computer Sciences and Engineering Open Access,2015). Vol 3 Issue 3.

[13] A. Abdullah and Z. K. Mohammad, “A Study on Sentiment Analysis Techniques of Twitter Data”, International Journal of Advanced Computer Science and Applications, (IJACSA),2019.Vol. 10, No. 2.

[14] G. L Marianela, B. Joel, F. Ulrik, R. Magnus, T. Edward, V. Stefan, and V. Vladimir, “Veracity Assessment of Online Data”, Decision Support Systems, 2020, 129, 113132.

[15] L. Tagliaferri, “An Introduction to Machine Learning”, Digital Ocean, 2017.

[16] B.K. Muhammad, and B.M.B Eihab, “Machine Learning: Algorithms and Applications”, Tylor and Francis Group Publishers, LLC,2017..

[17] T. M. Mitchell, The Discipline of Machine Learning, Machine Learning Department technical Report, Pittsburgh, PA: Carnegie Mellon University, 2006.

[18] B. Daniel, “Bayes’ Theorem and Naive Bayes Classifier”, Data Science Laboratory Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo, Japan,2019.

[19] H. P. Harsh, and P. Purvi, “Study and Analysis of Decision Tree Based Classification Algorithms”, International Journal of Computer Science and Engineering, 2018, Vol. 6, Issue 10.

[20] O. D. Oyerinde, and P. A. Chia, “Learning Analytics Approach Using Multiple Linear Regression”, International Journal of Computer Applications, 2017, Vol. 157, No. 4.

[21] O.E. Okpako,“Neutrosophic-Based Decision Support System for Diagnosing Confusable Diseases”. PhD thesis, Computer Science Department, University of Port Harcourt, Nigeria, 2018.

[22] S. Haykin,,“Learning Machines”, Third Edition, Pearson Prentice Publishers, McMaster University, Hamilton, Ontario, Canada, 2009.

[23] E. F. Runyi“The Application of Binary Logistic Regression Analysis on Staff Performance Appraisal”, Science Journal of Applied Mathematics and Statistics, 2017.

[24] Niklas D. A “Complete Guide to the Random Forest Algorithm”, Markov Solution, Data Science Publication, 2019.

[25] N. Sajid, and W. Aishan, “Study and Implementing K-Means Clustering Algorithm on English Text and Techniques to Find the Optimal Value of K International Journal of Computer Applications, 2018, Vol. 182, No. 31.

[26] R. Yogita, and R. Harish, “A Study of Hierarchical Clustering Algorithm”, International Journal of Information and Computation Technology, 2013.

[27] A.G. Barto, Dept. Of Computer Science, University Of Massachuestts - Amherst, 2018.

[28] K. Sasan, M.A. Shahidan, A.M. Azizah, Z.Mazdak, and H. Alireza, An Overview of Principal Component Analysis, Journal of Signal and Information Processing, 2013,Vol. 4.

[29] L.R. Nicole and Yoshua, B. Representational Power of Restricted Boltzmann Machines and Deep Belief NetworksNeural Computation Journals,(2008). Vol. 20 Iss**.** 6.

[30] J. Beakcheol, K. Myeonghwi, H. Gaspard, and W. K. Jong, "Q-Learning Algorithms: A Comprehensive Classification and Applications, IEEE ACCESS Publication, 2019, DOI 10.1109/.2019.294122

[31] W. Adigwe and K.E. Anazia, "Sentiment Analysis Using Neural Network" International Journal of Trend in Research and Development, IJTRD, 2020, Volume 7(1).

[32] A. Abirami, and V. Gayathri, "A Survey on Sentiment Analysis Methods and Approach", in Advanced Computing (ICoAC), Eighth International Conference IEEE, 2017.

[33] P. Wikanda, and B. Kamon, "Parameter Estimation of the Crack Lifetime Distribution Using The Expectation-Maximization Algorithm", Far East Journal of Mathematical Sciences (FJMS),2017. Vol. 102, No. 5.