Global Scientific JOURNALS

GSJ: Volume 8, Issue 2, February 2020, Online: ISSN 2320-9186
www.globalscientificjournal.com

DR. A.E OKPAKO (*BSc, MSc, PhD Computer Science)*
EDWIN CLARK UNIVERSITY, KIAGBADO, NIGERIA.

ANAZIA ELUEMUNOR KIZITO *BSc, (MSc Computer Science)*
DELTA STATE POLYTECHNIC, OZORO, NIGERIA.
Contact: kizitoanazia@gmail.com and kaymax07@yahoo.com

MACHINE LEARNING BASED BIG DATA CLASSIFICATION

ABSTRACT

In recent times, the generation of data has changed from the simple data format that can be analysed and classified by the traditional method of Structured Query Language (SQL) or Relational Database Management System (RDBMS) to a more complex and massive formats that are generated in a very high speed. This new form of data collection is known as Big Data. Machine Learning which is defined as a computer field that uses statistical methods to give computer system the ability to learn with data without being explicitly programmed is one of the new methods proven to handle big data management and classification. In this research work, we explored the powers of supervised machine learning algorithms like Naïve Bayes, Support Vector and Neural Network to classify big data contents obtained from the Online Social Media platform. The methodology used is the prototype approach while JAVA and WEKA were used as the programming language and machine learning tool-kit respectively. Considering the results from the performance metrics like Accuracy, Precision, Recall Rate and F- Measure, it was observed that Neural Network has a better performance when compared with Naïve Bayes and Support Vector.

Key Word: ***Big Data, Machine Learning, Naïve Bayes, Support Vector and Neural Network***

Introduction

Recent research has indicated that there is increase in need for big data and its applications which is as a result of its variant forms of generation, processing and management that has helped in breaking new grounds in problems solving that were difficult to solve in the past. Big data is defined as datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyse (Manayika,*et al*, 2011). 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 decision, reduce risks, and serve customers (Gultieri, 2015). Big data users have looked beyond its drawbacks and focused on its potentials as a revolution that will transform how we live, work and think by creating knowledge discovery and better decision making (Liu *et al*, 2015).

Characteristics of Big Data

The characteristics of big data are often viewed from the dimensional perspective of its inherent benefits and challenges (Manyika*et al*, 2011). These characteristics are used to define big data as well as to

determine its benefits and drawbacks. Below are some of the major characteristics though in recent times, research has proposed not less than 45 characteristics of big data which include Volume, Velocity, Variety, Veracity, Value, Validity, volatility among others.

i. Volume

The Volume signifies the vast amount of data that are generated and stored every second which consist of structured, semi-structured and unstructured data. They are known to come in large and complex quantity due to the various numbers of users that are involved in the generation and management of such data.

ii. Velocity

Velocity represents the speed at which the data are generated, analysed, stored and retrieved. The data can be described as fleetingly increasing quantity of unstructured data from ever-growing amount of sources streams across the Internet (Karapetyan, 2012). They provide real-time services in other to meet up with their demand and also to make updating of transaction possible.

iii. Variety

The Variety of big data has to deal with the diverse combination of data that makes up the Big Data. Data come from different data sources and also in various format such as transaction and log data from various applications, structured database table, semi-structured data such as XML data, unstructured data such as text, images, video streams, audio statement, and more.

iv. Veracity

The quality and accuracy of data captured can differ in many regards which is referred to as the Veracity of data. It should be noted at this point that "good data begets good information" hence the source of the information must be of good repute. As other properties of big data improve, its complexity also increases and its reliability and trust drops which is the veracity in question. So the veracity of big data is one of the few properties that are affected negatively when others are improved. Providing satisfactory answers to the above questions will determine the rate of confidence or trustworthiness on the data set by users which is the veracity of the data set.

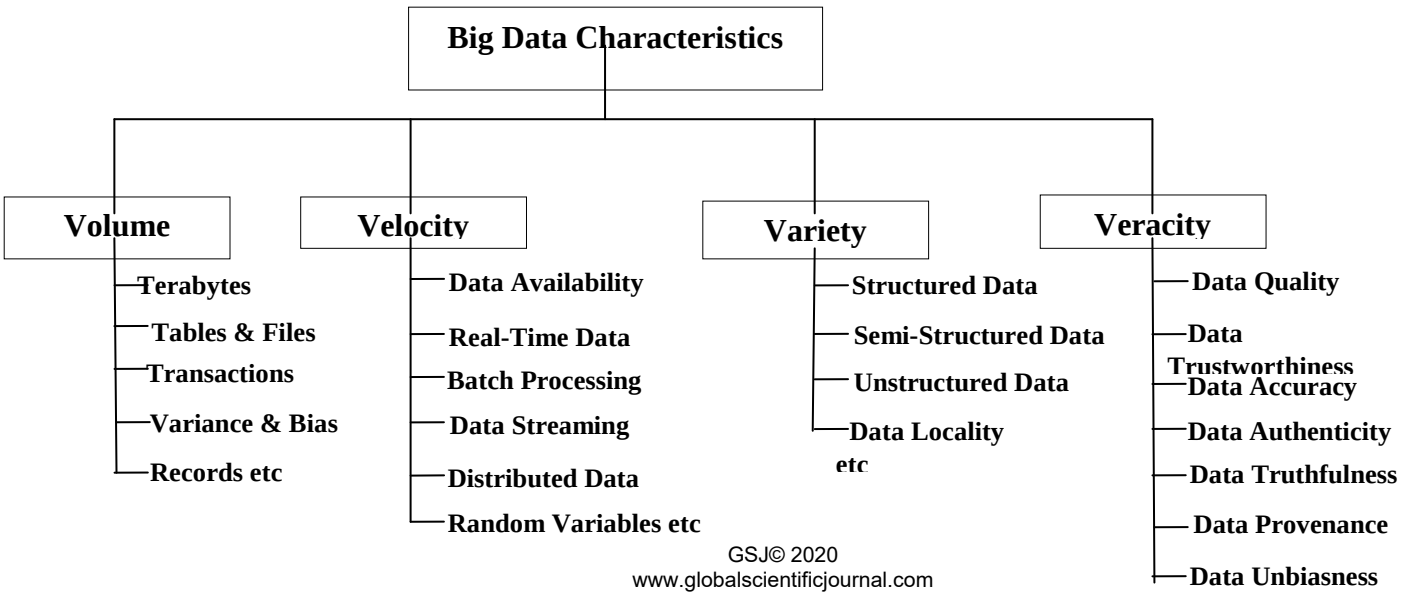

Figure 1: Diagram showing the first 4 Vs of Big Data

Big Data Technology

In the time past, major players in database management generate transactional structured data using traditional relational database and transactional query processing for the information extraction but currently the technologies are evolving (Raghavendra*et al*, 2011). To perform the investigations on the whole data using distributed computing and storage technologies such as MapReduce, distributed file systems and in-memory computing with highly optimized capabilities for different business and scientific purposes. With the increase in data storage capacity, data generating, analysing and processing tools, the analysis of data can be carried out in real time or close to real time, acting on full data sets rather than on the summarized elements, leveraging tools and technologies enough to address the issue. In addition, the number of options to interpret and analyse the data has also increased, with the use of various visualization technologies. Big data technology can be view under the following sub heading; Open Source Platform, Commercial Framework, File System and Tools.

Since traditional data processing has not yielded the needed results in the processing of big data despite its enormous potentials, other methods likes Machine Learning, Change Detection, Optimization, Natural Language Processing, Formal Method, Fuzzy Logic, Collaborative Filtering, Similarity Measurement, Blockchain Technology has been proposed as the means to unravel the gains of big data but machine learning has stood out among them Pendyla (2018). In this our research work, machine learning techniques will be used in the classification of big data datasets which is basically Naïve Bayes, Support Vector Machine and Neural Network.

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. Machine learning is 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 (Tagliaferri, 2017). It makes computer to train on data inputs and use statistical analysis in order to produce result values that falls within required range.Machine learning is based on the idea of giving “training data” to a “learning algorithm” that will make the learning algorithm generate a new set of rules based on inferences from the data. (Internet Society, 2017).The primary aim of machine learning is to allow the computer to learn on its own with little or no human intervention. It focuses on the development of computer program that can access data and use it to learn for themselves (Andrea *et al*, 2017). Machine Learning as situation where a computer is said to learn from experience E with respect to some task T and

some performance measure P, its performance on T as measured by P, improves with experience E. examples of machine learning techniques are Neural Network, Navies Bayes, Support Vector Machine, Logistic Regression, K-Means Clustering, K-Nearest Neighbour

Classification of Machine Learning

Machine Learning are classified as supervised machine learning, unsupervised machine learning, Semi-supervised machine learning algorithms and Reinforcement machine learning algorithms.

Supervised Machine Learning Algorithms

Supervised Machine Learning Algorithmscan apply what has been learned 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. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

Unsupervised Machine Learning Algorithms

Unsupervised Machine Learning Algorithmsare used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

Semi-Supervised Machine Learning Algorithms

This type of algorithm is 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. 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 / learn from it. Otherwise, acquiring unlabelled data generally doesn't require additional resources.

Reinforcement Machine Learning Algorithms

Reinforcement machine learning algorithmsis a learning technique that interacts with its environment by producing actions and it's discover errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. It supports machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize its performance.

There are many types of machine learning algorithms like Neural Network, Naïve Bayes, Support Vector Machine, K-Nearest Neighbour, K-Means etc.

2.7 Neural Network

This is a machine learning that is 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. The working algorithm of the brain neurons simulates where data are trained to handle problem in that manner. Neural Network is a computational model of ML that is based on the way biological neural network in the human brain process information (Ujjwalkarn, 2016). The diagram below shows the structure of a Neural Network

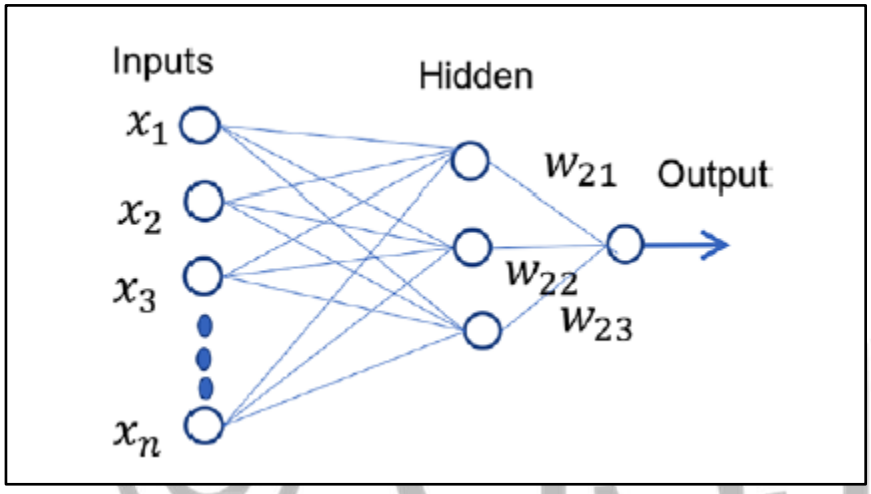

Figure 2: Structure of the Neural Network.

In this work, our Neural Network Classifier will make use of a Feed-Forward Neural Network approach that employs Back Propagation Algorithm. They are implemented using the mathematical expression shown below;

When a given training method is fed to the input layer, the weighted sum of the input to the j^{th} node in the hidden layer is expressed as;

$$Net_j = \sum w_{i,j} x_j + \theta_j \qquad (1)$$

Equation (1) is used to calculate the aggregate input to the neuron. The θ_j term is the weighted value from a bias node that always has an output value of 1. If any input pattern has zero values, the Neural Network could not be trained without a bias node. To decide whether a neuron should fire, Sigmoid function is used as the activation function and its result is used to calculate the neuron's output, and becomes the input value for the neurons in the next layer connected to it.

$$O_j = x_k = \frac{1}{1 + e^{-Net_j}} \quad (2)$$

Equations (1) and (2) are used to determine the output value for node k in the output layer. Let the actual activation value of the output node k be O_k, and the expected target output for node k be t_k the difference between the actual output and the expected output
is given by;

$$\Delta_k = t_k - O_k \quad (3)$$

The error signal for node k in the output layer can be calculated as

$$\delta_k = \Delta_k O_k (1 - O_k) \quad (4)$$

where the $O_k(1-O_k)$ term is the derivative of the Sigmoid function. With the delta rule, the change in the weight connecting input node j and output node k is proportional to the error at node k multiplied by the activation of node j. The formulas used to modify the weight $w_{j,k}$ between the output node, k and the node j is:

$$\Delta w_{j,k} = l_r \delta_k x_k \quad (5)$$

$$w_{j,k} = w_{j,k} + \Delta w_{j,k} \quad (6)$$

where $\Delta w_{j,k}$ is the change in the weight between nodes j and k, l is the learning rate. In equation (4.12), it was observed that the x_k variable is the input value to the node k and the same value as the output from node j.

Naïve Bayes Algorithm

According to Akshay (2007), Naïve Bayes Algorithm 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 Y happening, given that X has occurred. Here, X is the evidence and Y 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 in three types; Multinomial Naive Bayes, Bernoulli Naive Bayes and Gaussian Naive Bayes.

$$: p(X/Y) = \frac{p(X/Y) * p(Y)}{p(X)} \quad (7)$$

where X is the features of a dataset class with the following features; x_1, x_2, x_3, x_4……..x_n
and Y are dataset classes like Positive, Indeterminacy (Neutral) and Negative.
$p(X, Y)$ = the joint probability of a dataset with the given features is either Positive, Indeterminacy (Neutral) and Negative.
$p(X|Y)$ = probability of the dataset having features X given that the dataset is either Positive, Indeterminacy (Neutral) and Negative.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a linear model for handling classification and regression problems that can solve linear and non-linear problems (Durant and Smith, 2006). SVM algorithm creates a line or a hyperplane which separates the data into different classes of either positive and negative or positive, negative and neutral classes. 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 SVM and its margins are shown in the diagram below.

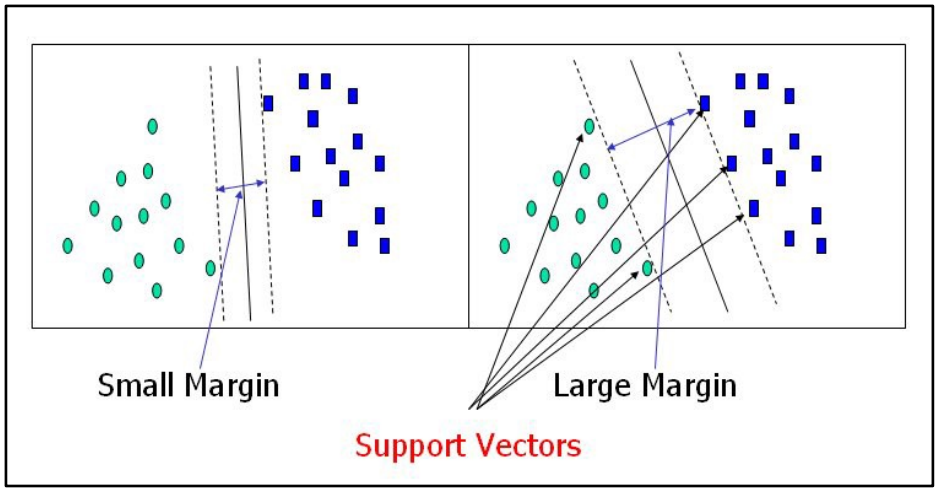

Figure 3: SVM diagram

REVIEW OF RELATED LITERATURE

Adiska, *et al*, (2016), employed Content Analysis on Data Quality Management (DQM) to analysis the quality of dataset collected from financial organisations and ways to improve them. They concluded their findings that there are no fundamentally new data quality problems in big data when compared with data sets from traditional form. They proposed ten antecedents of big data quality (veracity) which includes data, technology, people, process, procedure, organization and external expects. Also on their finding is that data quality is central issues in the values creation from big data.

Sindhu& Hedge (2017). In their research work, dataset of sex and age was generated from patients suffering from a disease know as Amyotrophic Lateral Sclerosis (ALS) which was used to find the missing value by applying Modified Moving Average (MMA) and compared with the earlier result obtained using Standard Deviation from an Auto-Regressive Tree (ART) and a better result using MMA. It was proposed that MMA had a superior result because this work was based on non-linear dataset and it is believed that ART will work better on linear dataset. It was suggested that projecting a methodology that can work well in both linear and non-linear dataset will be better off in the assessment veracity indices of big data.

Vosoughi, (2014), proposed the combination of Computational and Predictive Methodology to investigate the nature of rumours surrounding real world issues in Twitter and Reddit using the April 2013 Boston marathon bombings as a case study which will be analysed using natural languages processing and network algorithm. They focused on getting the right information for user but deterring the veracity index of the "assumed right information"

Liu *et al*, 2015. They proposed that big data research should closely follow good scientific practice to provide reliable and scientific "stories" as well as explore and develop techniques and methods to mitigate or rectify those big "errors brought by big data"

Fu &Chau, (2013), used random sampling approach to collect dataset from microbloggers from their services providers in other to analyse the profile and the pattern of usage of about 29,998 microblogger. From the regression carried out on the dataset, it was put forward that volume of followers is a key determinant in creating original microblogger's post, reposting, messages being reposted and receiving comments. It also notes that gender difference and regional disparities in the use of microblogger in china are also observed. From the research, it was noticed that there are many account with little or no posts which are unnoticed by the microblogger though it was unable to classify these accounts.

Robin *et al*, (2016) cross-compared sets of unrelated data flow of human movement from their homes locations to shopping mail generated from telecom service providers, commercial consumer survey and geo-tagged social media (twitter) messages. In their conclusion, they stressed the importance of verification in flow of data derived from new sources and demonstrated ways of actualizing it. One of the limitations of their work is that it is more of spatial based research.

Agarwal *et al*, (2016) generated dataset from twitter (tweets) social media platform through Crowdsourcing which was analyzed using sentiment analysis. The analysis was further evaluated using a Bayesian predictor trained with a trinomial function which produced a better accuracy level when compared with previous techniques. They concluded that using sentiment analysis on crowdsourced datasets is a better

method of improving the big data veracity but the veracity index of the system was actually in doubt due to the fact that their dataset was crowdsourced.

The work of Bo Pang *et al* (2002) was not based on classification of topics rather on general sentiment of viewer on the movies. They conducted a sentiment analysis on the datasets from the movies viewers which was further trained using machine learning methods (Naives-Bayes (NB), Maximum Entropy Classification (MEC) and support vector machine (SVM). They concluded that standard machine learning methods perform better than human produced baselines but did not perform so well on sentiment classification as on tradition topic based on classification.

Celikyilmaz*et al,* (2010) employed the probabilistic model on sentiment analysis on tweets generated datasets and they were able to classify them into polar and non-polar tweets. In their results, they obtained better F-scores that are relatively better than classification baseline that uses a probabilistic model-based sentiment analysis on tweets messages.

Peddinti et al (2011), focused on sentiment analysis using domain adaptation whose datasets were generated from other domain. They result showed that algorithm based model called expectation maximization (EM) and Rocchio SVM produced a better accuracy level than other methods used in the previous research.

Luneva*et al.,* (2016) employed fuzzy logic on the analysis of OSN user's influence and authorship on a group of messages. They analysed the reviews on products/services and users communication designed interface based on the robotic technology. The designed system was an analytic tools and it made of use of web observatories datasets.

Dragoni*et al.,* (2015) had a probabilistic based concepts that was modelled using fuzzy logic. Their work threw more light into the uncertainties that has to do with fuzzy logic and its application areas. It was a hybrid approach of linguistic concept of SenticNet and WordNet that was managed by a graph-propagation algorithm using Blitzer dataset. The system presented an enhanced performance and work rate as against other systems using the same dataset.

Rashma*et al.,* (2015) proposed a hybrid system of NB and fuzzy logic which will create a mechanism to understand the entire opinion about certain product/service from customers and producers. The system used SentiWordnet dataset to obtain the values of customers and manufacture's opinions and linguistic verification. The fuzzy rule applied helped to identify the varying degree of values to express the level of vagueness and imprecision on customer and manufactures information.

Mary, A. J. J. and Arockiam, L. (2017) applied fuzzy logic concepts in considering the objectivity of sentences and determination of the polarity review. They opined that the polarity of sentence objectivity contained in the aspects will help in enhancing the accuracy of the aspect polarity to improve customer's decisions. They demonstrated their work using a review on Samsung Galaxy Note 7.

Methodology Adopted

The methodology adopted in this work is the Prototyping. Software prototype is an executable model of the proposed software system. It must be producible with significantly less effort than the planned product (Gustav and Gunther, 1996). It is a model or a standard for other things of same category. It must be readily modifiable and extensible and also must have all the features near the targeted system. In many fields, there is uncertainty as to whether a new design will actually do what is expected or desired. New designs often have unexpected problems. A prototype is often used as part of the product design process to allow engineers and designers to explore the designs alternatives, test theories and confirm performances prior to starting the production of a new product.

Theaim of this research work is to classify big data datasets obtained from Twitter Sander datasets, Email-Spambase datasets and Smsspam Collection datasets into three polarities of positive, neutral and negative using machine learning algorithms (Naive Bayes, Support Vector Machine and Neural Network) and draw a comparison of the algorithm with a better classifying Accuracy rate and other performance metrics like Precision, Recall Value and F-Measure

System Implementation

The implementation of this system is done with Java Programming Language and WEKA as the machine learning language. Three machine learning algorithms (Neural Network, Naive Bayes and Support Vector Machine) will be implemented on three datasets (Sanders Twitter Datasets, Email-Spambase dataset and Smsspam Collection Datasets which a comparison will be drawn using a performance metrics like accuracy, precision, recall value and f-measure.

In order to achieve our objectives of proposed system, the entire system design is divided into the following phases;

(i) Data Collection/Data Gathering

This has to deal with choosing the particular dataset that will be used for the design though there are so many examples of big data dataset available online but there is need to choose the one that will be suitable for the nature of your research. As discussed in the previous chapter, after due considerations, Sanders Twitter datasets, Spam Email dataset and Smsspam collection datasets were chosen, downloaded and analysed in this work.

(ii) Data Pre-Processing and Vectorization

These Natural Language Processing models which involves text transformation through algorithms deal with large amount of text to perform classification correctly mostly at the backend. These datasets are usually in a raw and unprocessed state and will be difficult for WEKA to handle and process, so there is need for the datasets to be brought to a state that conforms to WEKA's data representation format. It is unfortunate that at present, java programming language has not developed the application for these dataset

cleaning which includes stop-words, hash-tags, symbols, numeric, slang words, twitter address, quotation marks, lowercases, etc so data pre-processing for this research are done manually.

It should also be noted that these datasets are originally in CSV (Comma Separated Values) format but cannot be processed in this format, they are converted to a processible format known as ARFF (Attribute-Relation File Format) though this processing is handled in an automate procedure by WEKA and it is known as Vectorization. This stage comes immediately after the initial pre-processing of data that involves data cleaning manually. Vectorization can be defined as the process of transforming raw data into feature vectors. In text classification, each term, phrase or character can be represented as a feature. A feature is a measurable property about a tweet text or email spam in the dataset. The reason for Vectorization of the data strings i.e. to convert sequences of text into attributes with number and categorical values is performed in the hope of finding the best feature vector for a learning classifier. The feature vectors are generated from the existing data by transforming the raw data into machine readable units, called, 'feature nuggets' which essentially carry significant information about the data and its characteristics. These feature vectors are able to 'learn' certain aspects about the dataset that will be utilized during machine classification. Vectorization can be achieved in the following different ways; Count Vectors, Term-Frequency Inverse-Document-Frequency (TF-IDF), Word Embeddings, Text or Natural Language Processing based and Topic Model based features. Everything about Vectorization can be summarized as the conversion from string to word vector format which was conveniently handled by WEKA using topic models as attribute and instances features.

(iii) **Data Training and Learning**

Having completed the process of pre-processing and vectorization on the dataset, the datasets are in the ready state to be trained and learn the features of the datasets using Neutrosophic-Based Neural Network, Naïve Bayes, Support Vector Machine and Neural Network approaches. In the training process, the model tries to understand the features of a particular dataset which serves as an input as regards its attributes selection and instance representation. The feature extractor transfers the text input into a feature vector which classifies its polarity. The datasets are usually divided into two parts like a ratio of 70% to 30% for training and learning respectively at the choice of the programmer. After the attributes and instance are selected, the training data is represented in terms of the attributes. Training can be seen as learning procedure in which prediction of test data will be made. During the training, the presence of each attribute value in each of the polarity (negative, positive, and neutral) is determined.

(iv) Data Testing and Classification

This is the process of verifying all the possible combinations and estimates how well a model will predict the desired or expected results by comparing the result of the training datasets with that of the testing datasets. If the expected result is far different from the output result, input can be adjusted and the model will be fine-tuned based on the results of the test data set. This is done by matching the attributes of the testing dataset with that of the trained dataset and the probabilities of each of the hypotheses are computed

from the attributes and the one the closest result is classified in that class. At first, the classifier is fed with the testing datasets to check the accuracy of the algorithm there after real-live datasets can be put into the classifier for the classification of tweets and spam emails considering all the classifying methods either as a single selection or bulk selection.

In implementing Naïve Bayes and Support Vector Machine algorithms, it made use of Cross Validation methods. We used the n-fold cross-validation method to perform cross-validation where we split the input data into n folds of data. Training was done on all the folds but not on one of the folds (n-1) which is then used for testing of the model. This process is repeated in n times, with a different fold reserved for testing and excluded from training each time.

In sander's twitter datasets, a total of 156 datasets (instances), in email-Spambase datasets, a total of 100 datasets (instances) and in Smsspam collection datasets, a total of 730 datasets (instances) were used and divided into n folds (subsets) where n folds is 10. As the training was going on, the 10 fold that was used for testing at the end of every training session in order to evaluate the performance of the learning algorithm. But in the case of Neural Network, a test file was created which was equally divided into the ratio of 90: 10 percent. Training was done with the 90% of the entire dataset selected while testing was done with the remaining 10% of the entire dataset that was not part of the training. The detail of the implementation is done in the diagram below;

Table 1: System Detailed Dataset Structure

MACHINE LEARNING ALGORITHMS	**DATASETS USED**		
	Sander's Twitter Dataset	Email Spam	Smsspam Collection
Neural Network	Input Neurons: The number of neuron is determined by the number of attributes for any datasets being considered. Hidden Layers: 3 Learning Rate: 0.2 Momentum: 0.2 Epoch:4	Input Neurons: The number of neuron is determined by the number of attributes for any datasets being considered. Hidden Layers: 3 Learning Rate: 0.1 Momentum: 0.2 Epoch:4	Input Neurons: The number of neuron is determined by the number of attributes for any datasets being considered. Hidden Layers: 3 Learning Rate: 0.1 Momentum: 0.2 Epoch:4
Naive Bayes	Cross Validation: 10 Random Number: 1	Cross Validation: 10 Random Number: 1	Cross Validation: 10 Random Number: 1
Support Vector Machine	Cross Validation: 10 Random Number: 1	Cross Validation: 10 Random Number: 1	Cross Validation: 10 Random Number: 1

Result and Discussion of Output.

This gives the overall performance summary of the three learning algorithm on the three datasets which is shown in shown in table 2.

Table 2: Performance Summary of Naïve Bayes on the Datasets

NAÏVE BAYES LEARNING ALGORITHM				
	Accuracy	**Precision**	**Recall Value**	**F-Measure**
Sander Twitter Datasets	50.641%	0.510	0.506	0.507
Email-Spambase Datasets	62.000%	0.728	0.620	0.557
Smsspams Collection Datasets	91.917%	0.930	0.919	0.923

From the performance metrics shown in the table 2 above, 50.641%, 62.000% and 91.991% were presented as the Accuracy Rateacross all the datasets with their respective Precision, Recall Value and F-Measure.

Table 3: Performance Summary of Support Vector Machine on the Datasets

SUPPORT VECTOR MACHINE Algorithm				
	Accuracy	**Precision**	**Recall Value**	**F-Measure**
Sander Twitter Datasets	52.564%	0.554	0.526	0.520
Email-Spambase Datasets	62.000%	0.786	0.620	0.558
Smsspams Collection Datasets	96.849%	0.968	0.968	0.967

Using SVM algorithms on all the datasets, it has Accuracy Rateof 52.564%, 62.000% and 96.849% respectively as shown in table 3 above which has their respective Precision, Recall Value and F-Measure.

Table4: Performance Summary of Support Vector Machine on the Datasets

NEURAL NETWORK LEARNING ALGORITHM

	Accuracy	Precision	Recall Value	F-Measure
Sander Twitter Datasets	83.974%	0.864	0.840	0.841
Email-Spambase Datasets	85.000%	0.820	0.850	0.793
Smsspams Collection Datasets	98.082%	0.981	0.981	0.980

From table 4 above, Neural Network presented Accuracy Rate of 83.9744%, 85.000% and 98.0822%respectively on all the datasets with corresponding Precision, Recall Value and F-Measure as shown in table 3.

Considering the general performance of the new model, it will be observed that training/test on Smsspam Collection Datasets (91.9178%, 96.8493% and 98.082%) had a better output compared to other datasets (Twitter Sanders Dataset and Email-Spambase Datasets) and also the accuracy level across all the datasets was better in Neural Network with an Accuracy Rate of 83.974%, 85.000% and 98.0822% across all datasets as compared to the results obtained withNaïve Bayes and Support Vector Machine as seen in the table3 (NB has 50.641%, 62.000% and 91.917% while SVM has 52.564%, 62.000% and 96.849%). The Total Average Accuracy Rate of the machine learning algorithms (NB, SVM and NN) on all the datasets can be summarized as 68.186%, 70.471% and 95.498% respectively. The figure 4 below shows the graphical representation of the Accuracy Rate of all the machine learning algorithms (NB, SVM and NN) on all the datasets.

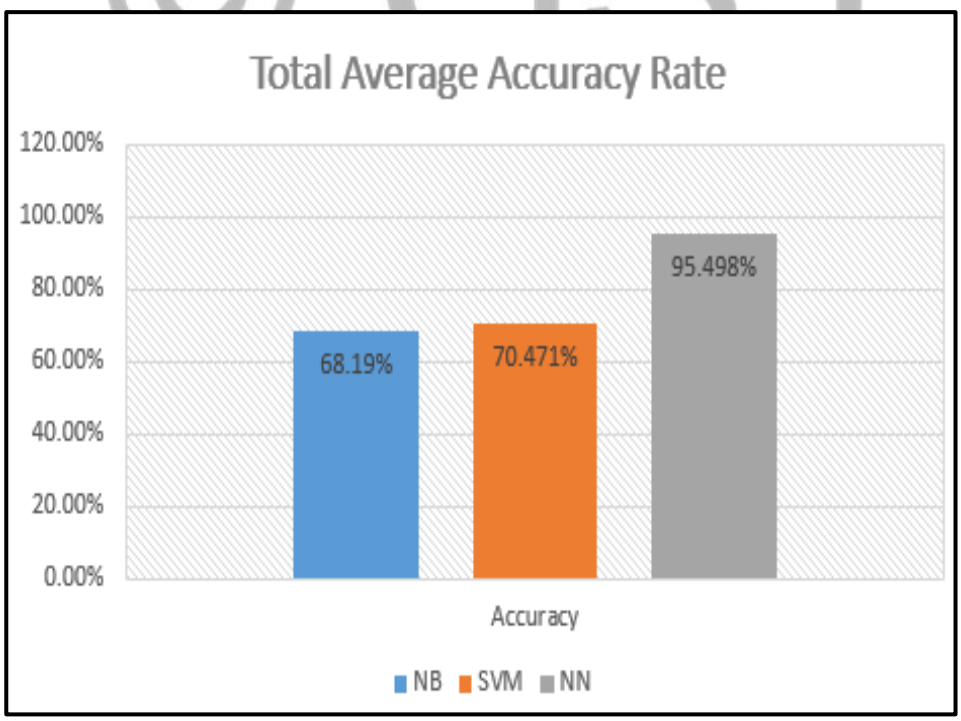

Figure 4: Total Average Accuracy Rate Performance

Conclusion and further works

In this research work, we had the task of classifying big data datasets obtained from Twitter Sander datasets, Email-Spambase datasets and Smsspam Collection datasets into three polarities of positive, neutral and negative using machine learning algorithms (Naive Bayes, Support Vector Machine and Neural Network)and draw a comparison of the algorithm with a better classifying Accuracy rate. From the results obtained, it is evident that Neural Network (83.974%, 85.000% and 98.0822%) had a better classification Accuracy Rate compared to Naïve Bayes and Support Vector Machine that had (50.641%, 62.000% and 91.917%) and (52.564%, 62.000% and 96.849%)respectively across all the three datasets which also reflected on other performance metric like Precision, Recall Rate and F-Measure as presented on the tables 2, 3, 4 and figure 4 above. Also it is observed that the Accuracy Rate improved on Smsspam collection dataset than Twitter Sander datasets and Email-Spambase datasets.

I will recommend that more machine learning algorithms as well as more datasets should involve in further works in order to determine how these algorithms will perform in a more critical model.

REFERENCE

Adiska, F. H. Joris, H. Agung, W. Haiko, V. and Marijn J. (2016). Antecedents of Big Data Quality an Empirical Examination in Financial Service Organizations. 4TH IEEE International Conference on Big Data, 4 Washington.

Agarwal, B. Ravikumar, A. and Saha, S. (2016). A Novel Approach to Big Data Veracity Using Crowdsourcing Techniques and Bayesian Predictors, https://www.researchgate.net/publication/308401450.

Celikyilmaz, A. Hakkani-Tur, D. and Feng, J. (2010). Probabilistic Model-Based Sentiment Analysis of Twitter Messages. Spoken Language Technology Workshop (SLT), IEEE.

Dragoni, M. Andrea, G. B. Tettamanzi and Celia Da Costa, P, (2015). Propagating and Aggregating Fuzzy Polarities for Concept-Level Sentiment Analysis, Cognitive Computation, Vol. 7, No. 2, pp. 186-197.

Fu and Chau (2013). Assessing Censorship on Micro blogs in China: DiscriminatoryKeyword Analysis and the Real-Name Registration Policy, Internet Computing, IEEE, 2013, v. 17 n. 3.

Gualtieri, M. (2015). Big Data Predictive Analytics Solutions, The Forrester Wave, USA.

Internet Society (2017).Artificial Intelligence and Machine Learning: Policy Paper Series for Artificial Intelligence and Technology.

Karapetyan, K. (2012). Big Data and its Opportunities for the Engineering Companies
Research Publication of Digiboost Project, Savonia University of Applied Sciences, 1st Edition, Tapio, Aaito.

Liu, J. Li, J. Li, W. and Wu, J. (2015). Rethinking Big Data: A Review on Data Quality and Usage Issues. ISPRS Journal of Photogrammetry and remote sensing, Published by Elservier B.V.

Luneva, E.E. Banokin, P.I. Yefremov, A.A. and Tiropanis, T. (2016). Method of Evaluation of Social Network User Sentiments Based on Fuzzy Logic, International Journal of Key Engineering Materials, Vol. 685, pp. 847-851.

Manyika, J. Chui, M. Brown, B. Bughin, J. Dobbs, R. Roxburgh, C. and Byers, A. H. (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity.

Mary, A. J. J. and Arockiam, L. (2017). Imputation of Missing Sentiment in the Aspect based Sentiment Analysis, Proceedings of International Conference on IoT, Data Science and Security, pp. 263-273.

Mckinsey Global Institute, International Journal of Intelligence Science, Vol. 5 No.3.

Pang, B. Lee, L. and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification Using Machine Learning Techniques. Proceedings of EMNLP.

Peddinti, V. M. K. and Chintalapoodi, P. (2011). Domain Adaptation in Sentiment Analysis of Twitter. Analyzing Microtext Workshop, AAAI,

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

Raghavendra, K. Pramod, K. K. Arun, A. Raghavendra, R. C. and Rajkumar, B. (2011). The Anatomy of Big Data Computing, Future Fellowship Project of the Australian Research Council and Melbourne Chindia Cloud Computing (MC3) Research Network.

Reshma, V and Ansamma, J. (2015). Aspect Based Summarization of Reviews Using Naive Bayesian Classifier and Fuzzy Logic, Proceedings of IEEE International Conference on Control, Communication and Computing, pp. 617-621.

Robin, L. Mark, B. Philip, C. and Martin, C. (2016).From Big Noise to Big Data: Toward the Verification of Large Data Sets for Understanding Regional Retail Flows'. Geographical Analysis 48, no. 1.

Tagliaferri, L. (2017). An Introduction to Machine Learning, Digital Ocean.
Vosoughi, S. (2014). Automatic Detection and Verification of Rumors on Twitter, a PhD thesis Submitted to the department of Program in Media Arts and Sciences, at the Massachusetts Institute of Technology.