ENHANCE BIG DATA VERACITY CLASSIFICATION USING NEUTROSOPHIC LOGIC AND NEURAL NETWORK

PROF. H.C INYIAMA, (*Professor of Electronics and Computer Engineering)*
NNAMDI AZIKIWE UNIVERSITY, AWKA NIGERIA

ANAZIA ELUEMUNOR KIZITO (*BSc, MSc Computer Science)*
DELTA STATE POLYTECHNIC, OZORO, NIGERIA.

Contact: kizitoanazia@gmail.com and kaymax07@yahoo.com

ABSTRACT

The benefits of the Internet cannot be over emphasized but not without some limitations which has hindered its full utilization. The Internet was brought to the limelight since the emergence of web 2.0, which contains huge volume of heterogeneous data collections that are usually generated, managed, assessed, and stored in high velocity that is referred to as Big Data. Big Data is made up of structured, semi structured and unstructured data whose biggest challenge is being able to analyze and give a better classification which has to deal with Big Data Veracity. The Online Social Network (OSN) forms a large percentage of the Big Data community with examples as Twitter platform and email messages. Twitter and email contents are good OSN platforms that provides a veritable platform for users to interact and share their views and comments about any topic of interest or discourse irrespective of its source and authenticity. A publicly available, verified and credible datasets obtained from Sanders Twitter Datasets, Email-Spambase Datasets and Smsspam Collection Datasets were used for the design of the proposed system. A hybrid Methodology which is a combination of Object Oriented System Analysis and Design Methods (OOAD) and Prototyping was adopted while Java and WEKA were used as Programming Language and Machine Learning Toolkit respectively. Accuracy, Precision, Recall Rate, F-Measure etc were used as performance metrics to determine its performance of the new system.

Key Words: *Big Data, Veracity, Neutrosophic, Machine Learning and Neural Networks*

Introduction

Since the advent of web 2.0, computing has witness a paradigm shift in the generation, processing, analysing and management of data and information due to its complex nature, speed of generation and high volume. There has been an influx of these huge and complex data over the internet by some data major players who try to keep in touch with their clients in real-time but not without some challenges like how to organise, manipulate and analyse these large chunks of data which is to be securely delivered through the internet and reach its destination unaltered or stolen by a third party. Also of great importance is the availability, authenticity, usefulness, durability and persistence of these data and that of the intended output which is information. All of these have been great concern to people that have to deal with them one way or the other. Due to its nature, size and rate of sending/retrieval, probably that is why they are referred to a "Big Data".

The definitions of big data depending on the nature of its generation, processing, analysing and management but not undermining its size, constitute and rate of assessment. 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 was reported as high-volume, high-velocity and/or variety information assets that demand cost-effective, innovative forms of information processing that enables enhanced insight, decision making and process automation (Duog, 2013). Big data are seen as data that exceeds the processing capacity of conventional database systems because they are too big, moves too fast and does not fit the structures of our usual database architecture. To gain value from this data, alternative means of processing and managing them must be taken.

Big data are hardly handled and managed by the traditional database nor the SQL queries of relational database management system because they are unlike the usual data format which are document/other text based files, rather it includes structured, semi-structured and unstructured data. They are better handled, processed and stored by software projects which supports the distributed processing of massive and complex data that are scattered in the form of clusters in millions of servers with the implementation of NoSQL. NoSQL is derived from "Not only SQL", which means that it allows not only regular SQL queries to be executed as proposed by oracle and other database management solution companies. Examples of these open source software are MapReduce and Firebase from Google, Mongo DB, Dynamo DB, Hadoop from Apache etc. The Open Source software is designed to support their processing from a single server to thousands of machines, with a very high degree of fault tolerance. Big Data are being generated, owned and managed by Individuals, Government, Communication Companies, Multinational and other major Techno-entrepreneurial players like Facebook, Apple, Google, Amazon, Microsoft etc.

Doug, (2013) 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. Veracity of big data can also be 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 (Pendyal, V, 2018). 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.

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.

Instructions and procedures are generated in the form of algorithms into huge volume of datasets that are processed as information and problem solving which are based on laid down rules. It should be stated clearly that those rules used by machine learning are created through learning of algorithm and not through specified computer programs generated by programmers at every new step. In a computer program driven application, users are required to write programs step by step throughout the application but in machine learning, it creates computer instructions which it will learn from the data without going through every new step of the program. By this laid down process, it is understood that computers/machines can do new jobs correctly from all it has learnt previously from stored data without adding new set of programs manually. 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). Machine Learning can be liken to a 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. Machine Learning can generally be grouped as supervised and unsupervised learning.

Supervised Machine Learning Algorithms

Supervised Machine Learning Algorithms can 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 Algorithms are 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.

Examples of machine learning techniques are Neural Network, Navies Bayes, Support Vector Machine, Logistic Regression, K-Means Clustering, K-Nearest Neighbour.

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). Neural Network can also be seen as is 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 (Haykin, 2009)

It has created a lot of breakthrough in Machine Learning like research such as speech recognition, computer vision and text processing. Some of the learning tools and skills in neural network are Weka kit, Scikit-Learn, Thread, Tensorflow, deeplearning4J etc. The diagram below shows the structure of a Neural Network

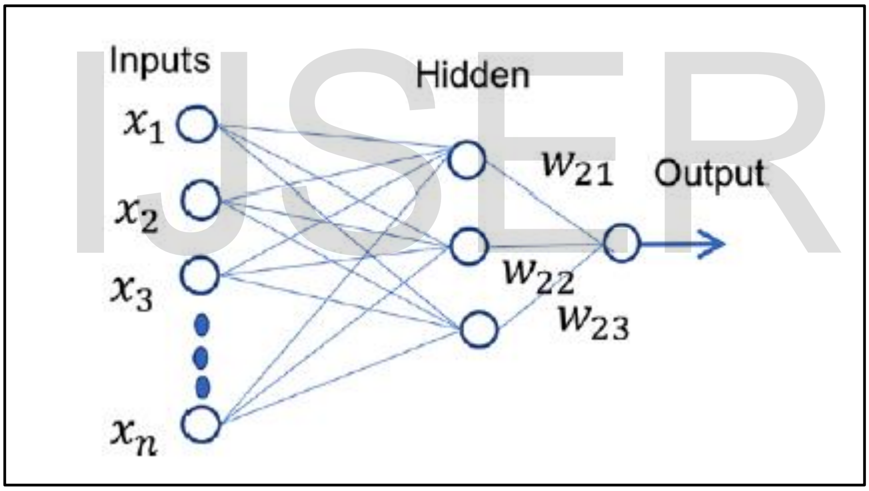

Figure 1: Structure of the Neural Network.

In this work, our Neural Network Classifier will make use of a Feed-Forward Neural Network approach that employs Back Propagation Algorithm and sigmoid function. 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 jth node in the hidden layer is expressed as;

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

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.

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

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;

$$\Delta_k = t_k - O_k \qquad \textbf{(3)}$$

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

$$\delta_k = \Delta_k O_k (1 - O_k) \qquad \textbf{(4)}$$

where the $O_k(1\text{-}O_k)$ 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:

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

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

where $\Delta w_{j,k}$ is the change in the weight between nodes j and k, l is the learning rate. In equation (6), 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.

Neutrosophic Logic

The term Neutrosophic Logic was derived from the word Neutrosophy which was introduce by Florentin Samarandche as a new branch of philosophy that deals with the origin, nature and scope of neutralities, as well as their interactions with different ideation spectra Umberto, (2007). There are so many new theories that have been formulated based on the laws of Neutrosophy like Neutrosophic Logic, Set Theory, Neutrosophic Set, Neutrosophic probability, Neutrosophic statistics etc which is generally an expansion of the Classical Logic (Binary Logic) and Fuzzy Logic. According to Samarandche (1995), Neutrosophic Logic represents an alternative to the existing logic as a mathematical model of uncertainty, vagueness, ambiguity, imprecision, undefined, unknown, incompleteness, inconsistency, redundancy and contradiction. In a Neutrosophic set unlike the classical logic and fuzzy logic that is made up of only Truth-membership (TA) and Falsity-membership (FA), it has Truth-membership (TA), Indeterminacy-membership (IA) and Falsity-membership (FA). TA(x), IA(x) and FA(x) are real standard or non-standard subsets of] 0-,1+[. That is;

$$TA: X \rightarrow]0\text{-},1\text{+}[\quad (7)$$

$$IA: X \rightarrow]0\text{-},1\text{+}[\quad (8)$$

$$FA: X \rightarrow]0\text{-},1\text{+}[\quad (9)$$

There is no restriction on the sum of TA(x), IA(x) and FA(x), so 0- ≤sup TA(x) + sup IA(x) + sup FA(x) ≤3+.

The Complement of a Neutrosophic set A is denoted by c(A) and is defined by;

$$Tc(A)(x) = \{1+\} – TA(x) \quad (10)$$

$$Ic(A)(x) = \{1+\} – IA(x) \quad (11)$$

$$Fc(A)(x) = \{1+\} – FA(x) \quad (12)$$

for all x in X.

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.

$$: p(X/Y) = \frac{p(X/Y)*p(Y)}{p(X)} \qquad (13)$$

where X is the features of a dataset class with the following features; $x_1, x_2, x_3, x_4 \ldots\ldots\ldots\ldots x_n$
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.

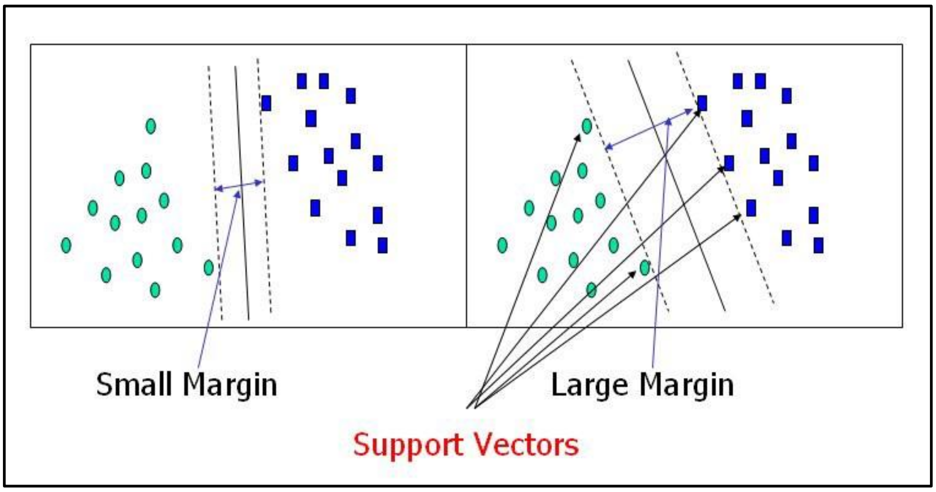

Figure 2 : Structure of Support Vector Machine

Review of Related Literatures

Big Data research basically is data driven in almost every discipline and field which makes it focus on methodological innovation or prioritizes the application of big data on people disease, geography, urban studies etc (Liu, et al, 2015). In trying to have an in-depth understanding of the proposed research the need arose to explore the strength and weakness of previously done work considering its methodologies and techniques on the related areas, the following works were reviewed.

Normala *et al*, (2015) did a systematic review on the profiling of Digital News Portal for Big Data Veracity. It was on manipulative journalism that tried to determine a portal that is bias or neutral in news reporting. They generated dataset from digital news contents and applied Concept Matrix with Classification Analysis. Their results showed that there are purpose of truth and issues related to manipulative writing but unable to eliminate noise and outliers completely.

According to Lukolanova, *et al* (2015), they proposed to manage content variation by way of quantifying the level of content objectivity, truthfulness and credibility (OTC) and expression variation using Rubin, (2006) and (2007) methodology. They further argued that quantification of subjectivity, deception and implausibility (SDI) reduces doubts in textual data content. Their dataset was obtained mainly from blog website. In their result, they brought forward a new method of determining veracity index through the combination of the three proposed dimension, OTC and concluded that it provides a useful way of assessing systematic variation in big data quality (veracity) across dataset with textual information.

In the work of Wang, (1998) a methodology known as Total Data Quality Management (TDQM) was employed to assess the veracity in a pool of generated data set. He was able to handle a wide range of measures that gave a detailed explanation of dataset though they are subjective and context dependent.

Pipino *et al.*, 2002 gave a broad range of mathematical models that can be used to determine the quality of a dataset. It has some drawbacks like having some data quality matrices that are context dependent and too many mathematical procedures.

Loshin, 2001 used a model known as Cost-Effect of Low Data Quality (COLDQ) which was based on the four matrices of data quality; accuracy, completeness, consistency and timelines combined with the thermo Nullity of Values. He places the cost of production of data quality higher than the authentication of the quality of the output of the processed data which is not good for data veracity assessment.

Crone, (2016) based his work on the methodology of Heterogeneous Data Quality (HDQ). He obtained his dataset from the insurance company in other to improve the risk assessment contents within the company. The result brought out from his work can be use to formalize the analysis of new data source that will replace the previous method in other to make it simpler for decision makers to evaluate and compare different source of data. It was noted that the methodology can only work if it converts semistructured/unstructured to structure before it can carry the quality assessment procedures.

In the work of Andrew H. Tapia et al (2013), they collected their dataset from tweets considering the reaction of twitter users before and after a major natural disaster or terrorism to see how social media platform is disseminating information in areas that are unreachable during the emergency process. It was concluded that information from the social media may not provide the needed information because of sentiments using Boston Marathon Bombing as a case study.

Prashanth, (2015) based his work on newly formulated veracity indices that are different from OTC as propagated by Lukoinov and Rubin, (2013). His new indices are Topic Diffusion, Geographic Dispersion and Spam Rate using Classification Analysis. The dataset used are tweets from different major oil companies and it was proved that the veracity of a topic depends on the veracity of contributing tweets though they were able to validate only two indices; Topic Diffusion, and Spam Rate and could not do that of geographic spread which makes it inappropriate for tweet from different geographical location.

Sanger *et al*, (2014) based their work on the Reputation-Based Trust (RBT) establishment of veracity of big data and proposed two-demission emerging from the combination of the “Big Data for Trust” and “Trust in Big Data”. It was discovered that with the computation of trust being reliant on trustworthy, trust in big data is a requirement for pursuant assessment and also that assessment of trust relies on the correct trust computation mechanism. The research fails to establish whether big data for trust in big data has a priority.

Kathleen & Fred, (2013), employed Topic Analysis in other to have a better knowledge of tweet contents and how two types of topics; informational topic and emotional topic affects users. They conducted their work using tweets as dataset and result showed that twitter data is biased and does not show the real level of veracity.

Zheng *et al*, (2015) carried out a content-based and user-based features using SVM algorithm to detect spam on OSN. A total of 30,116 users and more than 16 million messages were extracted from Sina Weibo OSN platform which was examined manually into spammer and non-spammer. This was further subjected to the SVM based system able to produce an excellent performance rate when compared to other single learning algorithm approach but not good as the combined methods though the system has an output whose computational and retraining rate was poor due to the use of SVM.

Ansari *et al*, (2011) introduce a new logic known as Netrosophic logic in the medical domain and by extension making fuzzy logic more powerful by employing Neutrosophic theories. They opined that the combination of fuzzy logic and Neutrosophic logic will improve intelligence of expert system in the cases of emergency or crisis.

Cheng *et al*, (2011) proposed a novel image segmentation approach based on Neutrosophic C-means clustering and indeterminacy filtering was combined with variety of experiments which determined the performance of the new system. The result showed that the proposed algorithm has a better performance quantitatively and qualitatively.

Swait *et al*, (2013) suggested the use of Neutrosophic logic for modeling real world uncertainties which will help in talking the conflicting attributes of the information captured. It produces a result that will be more generalized and indeterminacy tolerant in its working compared to fuzzy logic models though its result varies according to the nature of the control problems that it is meant to handle.

Kavitha *et al*, (2012) introduced a new technology for intrusion detection using Neutrosophic Logic Classifier which is an extended Fuzzy Logic. The system was tested with KDD 99 dataset and an improvised generic algorithm was adopted in order to detect the potential rules for performing a better classification. From their result, it was noticed that there is increase in detection rate and reduction in false alarm rate when compared to a fuzzy-based system.

Suman and Sankar (2015) proposed a system known as Fuzzy Granular Social Network –Model (FGSN) that is based on granular computing concept and fuzzy neighbour techniques to present a homogenous representation of social network. They evaluated the entropy and energy of the system to determine the uncertainties involved in the process and the fuzziness in the relationship of the actors. Their model showed a better classification in target set selection and community detection review.

Vadivukarassi et al, (2017) extracted raw tweets via twitter's API which was preprocessed using the Natural Language Toolkit based on some know keywords into positive and negative polarities. They used Chi Test and NB in selecting, training, testing the best features and also evaluating the sentimental polarities. It was noticed that higher the number of the features, the higher the accuracy of the selected features and the system had a better accuracy level more than the baseline model though the system failed to handle indeterminacy while the assumption of the shape of data distribution may have affected the veracity index.

Preety and Dahiya (2015) employed Support Vector Machine and Naïve Bayes approach on datasets collected from Twitter API. Their methodology was categorised into four modules of User's Interface, Log-Pre-processing, Features Clustering and Training/Testing modules. It was observed that this hybrid method gave a better accuracy and running time than using Support Vector Machine or Naïve Bayes separately.

In the work of Prabowo, R. and Mike, T. (2009), it was observed that their combinational approach had improvement on classification effectiveness as regards micro and macro-average fi. They used combined method of Support Vector Machine and Rule-Based approach on the review of movie and product contents.

Methodology Adopted

The methodology adopted in this work is the Hybrid of Object Oriented System Analysis and Design Methods (OOAD) and prototyping The aim of this research work is to carry out an enhanced big data veracity classification using Neutrosophic Logic and Neural Network on 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, Neural Network and Neutrosophic Logic/Neural Network) and their classification using performance metrics like Accuracy, Precision, Recall Value and F-Measure.

The diagram below shows the architectural structure of the proposed system showing it various components.

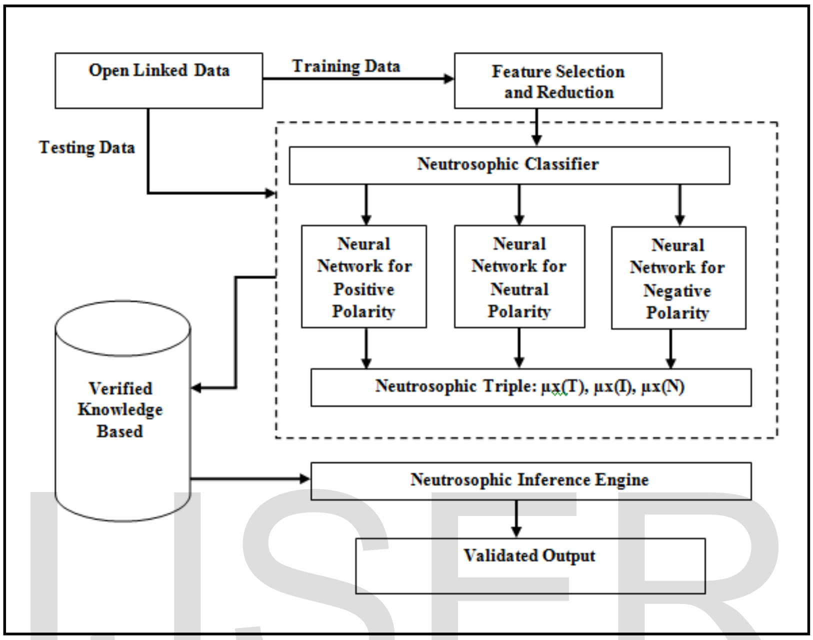
Open Linked Data
Training Data
Feature Selection and Reduction
Testing Data
Neutrosophic Classifier
Neural Network for Positive Polarity
Neural Network for Neutral Polarity
Neural Network for Negative Polarity
Neutrosophic Triple: µx(T), µx(I), µx(N)
Verified Knowledge Based
Neutrosophic Inference Engine
Validated Output

Figure 3: Architecture of the Proposed System

Data Collection

In this research work made use of data from online repository suitable for research in big data community in view of evaluating and validating our proposed model for determining the Veracity in Big Data (Twitter and email app), three big data datasets samples were used such Sanders Twitter datasets, Email-Spambase datasets and Smsspam collection datasets.

Dataset 1: Sanders –Twitter

The twitter sentiment corpus created by sanders analytics consists of 5513 hand classified tweets (however, 400 tweets missing due to the floating scripts created by the company). Each tweet was classified with respect to one of four different topics. This free dataset is for training and testing sentiment analysis algorithm and it is made up of positive, neutral and negative polarities.

Dataset 2: Email-Spambase

This is a collection of spam datasets that was generated from UCI (University of California, Irvine) machine learning data bank by Mark Hopkins, Erik Reeber, George Forman and Jaap Suermondt in Hewlett-Packard Labs. Email-Spambase contains 4601 instances and 58 attributes (57) continuous input attribute and 1 nominal class label target attribute. They are classified as positive, neutral and negative polarities.

Dataset 3: Smsspam Collection Dataset.

The SMS Spam Collection is a set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged according to being ham (legitimate) or spam. The files contain one message per line. Each line is composed by two columns: v1 contains the label (ham or spam) and v2 contains the raw text.

System Implementation

The process of big data Sentiment Analysis is purely a Natural language processing procedure which is one of the most complex part of data mining. It involves a combination of manual and automated processes of classifying opinions into polarity using machine learning approaches which is done with Java Programming Language and WEKA as the machine learning language. The machine learning algorithms (Naive Bayes and Support Vector Machine, Neural Network and neutrosophic/Neural Network,) will be implemented on three datasets (Sanders Twitter Datasets, Email-Spambase dataset and Smsspam Collection Datasets) and evaluates its classification using performance metrics like Accuracy, Precision, Recall Value and F-Measure.

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

i. Data Collection/Data Gathering
ii. Data Pre-Processing and Vectorization
iii. Data Training and Learning
iv. Data Testing and Classification

Algorithm of the Proposed System

Pre-processing Phase:

Step 1*: Generation of initial features from the datasets*

Step 2: *Perform String to word vector of the dataset*

Step 3: *Set attribute indices to the training set*

Step 4: *Set the minimum term frequency for the vector*

Step 5: *Tokenize the vectorised dataset*

Step 6: *Set the stemmer to LovinsStemmer*

Step 7: *Set StopwordsHandlers using Rainbow ()*

Step 8: *SetWordsToKeep to 100000*

Step 9: *Set output words count to be true*

Step 10*: Do a IDFT transform of the data*

Step 11*: Do a TFT Transform on the data*

Neural Network Training Phase:

Step 12: *The values of selected features in Step 11 should*
be fed into neural Network (Class I, Class II and
Class III NN).

Step 13: *Calculate the* ***Error=Target-Output*** *for both the Class I and Class II which is denoted*
by $\mathbf{e}^t$, $\mathbf{e}^i$
and $\mathbf{e}^f$ *for all feature vectors –feature (N) for every instance.*

Step 14: *Calculate the local gradient for nodes in each network.*

Step 15: *Calculate the hidden error of the network;*

Step 16: *Adjust the weights of the network using the learning rule until Learning is complete.*

Veracity Index:

Step 17: *Get the degree of belief in Class I denoted as* ***Pr(Class 1/X).***

Step 18: *Get the degree of belief in Class II denoted as* ***Pr(Class II/X).***

Step 19: *Get the degree of belief in Class III denoted as* ***Pr(Class III/X)***

Step 20: *Calculate Confusability Measurement = 1-|****Pr(Class 1/X)- Pr(Class II/X)|***

Step 21: *Determine the threshold CM from the validation data.*

Testing Phase:

Step 22*: Test the system putting Veracity Index and the complexity measurement from the test*
data into
consideration.

Training/Testing of the System

This type of testing comes along with the training of the selected datasets. In using the Naïve Bayes and Support Vector Machine algorithms, it made use of Cross Validation methods. Cross-Validation is defined as an approach used to evaluate a Machine Learning models by way of dividing the available datasets into folds (subsets) where part of the folds is used for training while the remaining folds is used for testing in different learning algorithms say about 60:40 percent in order to detect over-fitting or under-fitting. 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 and Neutrosophic Logic/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 neural network has a hidden layers of 3, learning rate of 0.1, momentum of 0.2 and epoch of 4.

Result and Discussion of Output

This gives the overall performance summary of the three learning algorithm on the three datasets and weighted (average) performance rate which is shown in the table and different graphical representations below.

Table 1: Summary Results of all the Datasets

Machine Learning Algorithm	**Overall System Performance Evaluation**			
NB Training Result	**Accuracy**	**Precision**	**Recall Value**	**F-Measure**
Sander Twitter Dataset	50.641%	0.510	0.506	0.507
Email-Spambase	62.000%	0.728	0.620	0.557
Smsspam Collection	91.917%	0.930	0.919	0.923
Total	204.558%	2.168	2.045	1.987
Average	**68.186%**	**0.7226**	**0.681**	**0.662**
SVM Training Result				
Sander Twitter Dataset	52.564%	0.554	0.526	0.520
Email-Spambase	62.000%	0.786	0.620	0.558
Smsspam Collection	96.849%	0.968	0.968	0.967
Total	211.413%	2.308	2.114	2.045
Average	**70.471%**	**0.769**	**0.706**	**0.6816**
NN Training Result				
Sander Twitter Dataset	83.9744%	0.864	0.840	0.841
Email-Spambase	85%	0.820	0.850	0.793
Smsspam Collection	98.0822%	0.981	0.981	0.980
Total	267.0566%	2.665	2.671	2.614
Average	**89.0188%**	**0.888**	**0.8903**	**0.8713**
NL/NN				
Sander Twitter Dataset	87.179%	0.885	0.872	0.873
Email-Spambase	100.000%	1.000	1.000	1.000
Smsspam Collection	99.315%	0.993	0.992	0.993
Total	286.494%	2.878	2.865	2.866
Average	**95.498%**	**0.959**	**0.955**	**0.962**

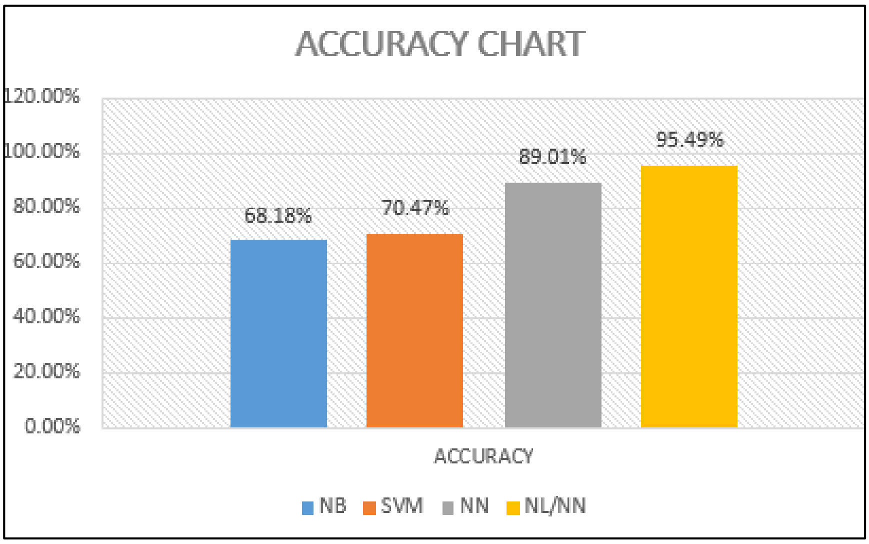

Figure 4: Average Accuracy Performance of all the Datasets

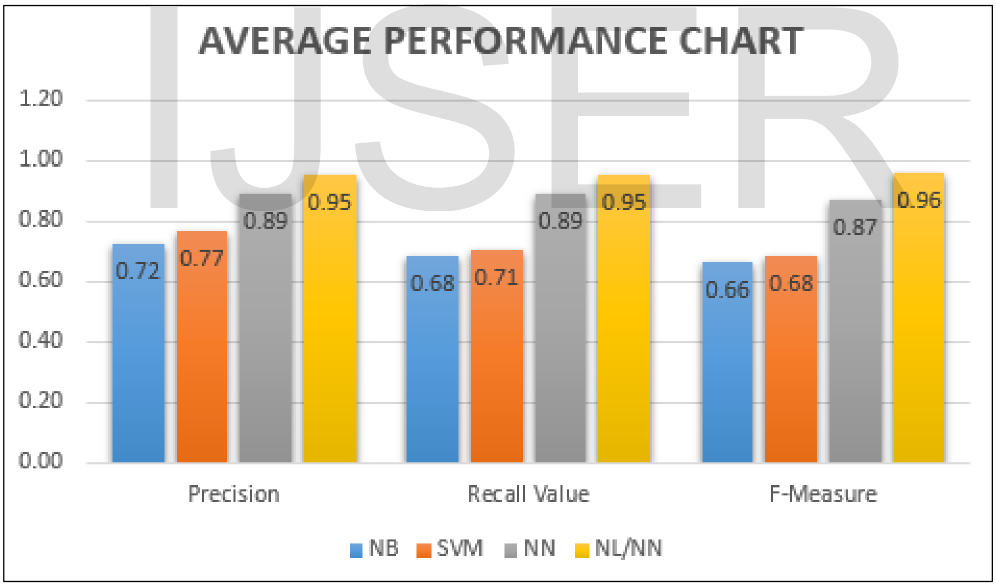

Figure 5: Average Precision, Recall Value and F-Measure of all the Datasets

From the analysis of the results shown in the table 1 and figures 4 & 5 above, it was clear that Netrosophic logic/Neural Network had a better performance than the other machine learning algorithms (Naïve Bayes and Support Vector Machine and Neural Network) though performance varied on different datasets. On the Sander Twitter datasets it has its accuracy rate of 50.64%, 52.564% and 87.179% on Naïve Bayes, Support Vector Machine, Neural Network and Neutrosophic Logic/Neural Network respectively, on Email-Spambase datasets, it has 62%, 62% and 100% using Naïve Bayes, Support Vector Machine and Neutrosophic Logic/Neural Network Learning Algorithms respectively while the Smsspam collection datasets produced 91.97%, 96.84% and 99.315% with Naïve Bayes, Support Vector Machine and Neutrosophic Logic/Neural Network respectively. The average Accuracy performance by the machine learning algorithms (Naïve Bayes, Support Vector Machine, Neural Network and Neutrosophic Logic/Neural Network) on all the datasets (Sander Twitter datasets, email-Spambase datasets and Smsspam collection datasets) are 68.186%, 70.471%, 89.018% and 95.498% respectively. The same level of performance were replicated considering metrics like precision, recall value, f-measure shown in table and graphs in figure 4 & 5.

Conclusion and Further Works

Three publicly and notable datasets suitable for big data analytic (Sander Twitter Dataset, Email-Spambase Datasets and Smsspam Collection Datasets) were considered in this research work and employed supervised machine algorithm (Naïve Bayes, Support Vector Machine, Neural Network and Neutrosophic Logic/Neural Network). From the analysis of the results shown in table 1 and figure 4 & 5 above, it was clear that Neutrosophic/Neural Network had a better classification accuracy than Naïve Bayes and Support Vector Machine and Neural Network. Neutrosophic Logic/Neural Network had average performance of 95.498%, 0959, 0.955 and 0.962 as it Accuracy, Precision, Recall Value and F-Measure respectively across all datasets, Naïve Bayes had 68.186, 0.7226, 0.681 and 0.662 as it Accuracy, Precision, Recall Value and F-Measure respectively across all datasets, Support Vector Machine had 70.471%, 0.769, 0.706 and 0.681 as it Accuracy, Precision, Recall Value and F-Measure respectively across all datasets while Neural Network had 89.018%, 0.888, 0.890 and 0.8713 as it Accuracy, Precision, Recall Value and F-Measure respectively across all datasets. From the result above, it can be concluded that an enhanced (hybrid) learning algorithm had better big data veracity classification than a single learning algorithm.

In this research work, the datasets used were basically text based but it should be understood that big data does not contain text only. Other formats like images, audio, video and spatial contents were totally ignored because they were not part of the datasets used, I will recommend that further works should be carried out in this area. Finally, it will be recalled that we used WEKA as our machine learning toolkit, I will suggest that comparative analysis should be done on several machine learning toolkits to see the one with a better classification.

References

Ansari, A.Q. Swati, A, and Biswas, R (2011). A Proposal for Applicability of Neutrosophic Set Theory Medical Artificial Intelligent, International Journal of Computer Applications (0975 – 8887), Volume 27– No.5.

Akshay, J. (2007). A Framework for Modelling Inuence, Opinions and Structure in Social Media, In Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, Vancouver, BC.

Cheng, H.D. Guo, Y. and Zhang, Y. (2011). A Novel Image Segmentation Approach Based on Neutrosophic C-Means Clustering and Indeterminacy Filtering, New mathematics and Natural Computation vol. 7, no.1 (2011) 155-171.

Doug, L. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety, Application Delivery Strategies, Meta Group Publishers, USA.

Durant, K. and Smith. M. (2006). Mining Sentiment Classification from Political Web Logs", In Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2006), Philadelphia, 2006.

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

Haykin, S. (2009). Learning Machines, Third Edition, Pearson Prentice Publishers, McMaster University, Hamilton, Ontario, Canada.

Kathleen, M. C. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose, Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.

Kavitha B, Karthikeyan S., Maybell P. S. (2012). An Ensemble Design of Intrusion Detection System for Handling Uncertainty Using Neutrosophic Logic Classifier, Knowledge-Based Systems. ACM Digital Library by the Association for Computing Machinery 28 (2012) 88-96.

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 Elsevier B.V.

Loshin, D., (2012). Data Governance and Quality: Data Reuse vs. Data Repurposing. http://dataqualitybook.com.

Lukoianova, T. and Rubin, L. V. (2014) Veracity Roadmap: Is Big Data Objective, Truthful and Credible?" Advances in Classification Research Online, vol. 24, pp. 4—15.

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

Normala, B. C. E. Iskandar, B. I. Fatimah, S. Lilly, S. A. and Ali, M. (2015). A Systematic Review on the Profiling of Digital News Portal for Big Data Veracity, procedia computer science, vol. 72, Elsevier.

Okpoko O.E., (2018). Neutrosophic-Based Decision Support System for Diagnosing Confusable Diseases. PhD thesis, Computer Science Department, University of Port Harcourt, Nigeria.

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

Prabowo, R. and Mike T. (2009). Sentiment Analysis: A Combined Approach, Journal of Informetrics, Vol.3. Issue 2.

Prashanth, K. (2015). A Case Study on Veracity in Twitter Data Using Oil Company Related Tweets, MSc Thesis Submitted to the Faculty of the Graduate College of the Oklahoma State University in partial fulfilment of the requirements for the Degree of Master of Science.

Preety, K. and Sunny, D. (2015). Sentiment Analysis Using SVM and Naïve Bayes Algorithm, International Journal of Computer Science and Mobile Computing, Vol.4 Issue.9.

Umberto, R. (2007).Neutrosophic Logics: Prospects and Problems, journal of Fuzzy Sets and Systems 159 Science Direct, Elsevier

Sanger, J. Richthammer, C, Pernul, G. (2014). Trust and Big Data: A Roadmap for Research, 25th International Workshop on Database and Expert System Application, Published by IEEE, Munich, Germany.

Swati, A, Ansari, A.Q. and Biswas, (2015).Neutrosophication of Fuzzy Models, Conference: IEEE Workshop on Computational Intelligence: Theories, Application and Future Directions, At IIT Kanpur.
Tagliaferri, L. (2017). An Introduction to Machine Learning, Digital Ocean.

Vadivukarassi, M. Puviarasan, N. and Aruna, P. (2017). Sentimental Analysis of Tweets Using Naive Bayes Algorithm, World Applied Sciences Journal 35 (1): 54-59, ISSN 1818-4952.

Wang, R.Y. (1998). A Product Perspective on Total Data Quality Management'', Communications of the ACM, Vol. 41 No. 2.

Zheng, X. Zeng, Z. Chen, Z. Yu, Y. Rong, C (2015). Detecting Spammer on Online Social Network, Journal of Neurocomputing vol. 159, Elsevier.