**CLASSIFICATION OF NIGERIAN NEWS HEADLINES USING SOME SELECTED MACHINE LEARNING ALGORITHMS**

**A PROJECT SUBMITTED**

**BY**

**ABIODUN LATIFAT OLUWADAMILOLA**

**CSC/2017/2002**

**UNDER THE SUPERVISION OF**

**DR T.M. FAGBOLA**

**IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF**

**BACHELOR OF SCIENCE DEGREE (BSc)**

**IN**

**COMPUTER SCIENCE**

**FEDERAL UNIVERSITY OYE-EKITI**

**AUGUST, 2021**

**CHAPTER ONE**

**INTRODUCTION**

* 1. **Background to the Study**

According to Collins dictionary, a news headline is the title of the Newspaper story or the summary of the whole news. It is important because it helps the reader to identify the nature of the news so that they don’t go through the whole news to understand what it is about. Online distribution of news stories has become a standard conduct of media sources. News headlines are vital in both capturing readers’ attention and in influencing their online reading experience of news because in most cases, people limit their reading to headlines only, without clicking on a link to the full news (Holsanova*et al*., 2006). However, in a national context, the need for News headlines classification is eminent to understand the hidden trend regarding issues of national interest such as kidnap, rape, corruption, child abuse, crime and so on.

Classification is a challenging task in the field of text mining as it requires pre-processing steps to prepare the textual data into organized form which is initially available in the unorganized form. The different steps involved in news headline classification are collection of news, pre-processing of collected news, feature extraction, different classification techniques to classify news and evaluating performance measure for different classification technique. (Nabamita Deb *et al*., 2020).

News article Classification is a recently growing area of interest in text classification because of its associated multiple matching categories (Fagbola, Thakur and Olugbara, 2019). Basically, classification is a process of categorizing a given set of data into classes; it can be on either structured or unstructured data. The classes are often referred to as target, label or categories. In this study, the Classification of Nigeria News Headlines is treated as a binary classification problem. The news headlines are thereby classified into X<Y, Z categories using a classifier or ensemble of classifiers. The classifier, in this case, needs training dataset to understand how the given input variables are related to the class. Once the classifier is trained accurately, it is then used by inputting a different dataset from testing set, and then allowed to predict the member of the class based on what it has learned from the training set.

In this study, the performance of each machine learning algorithm such as Naïve Bayes, Support Vector Machine (SVM) and decision tree will be evaluated the for Classification of Nigeria News Headline and also in terms of accuracy, time and ability to predict correctly. The use of other machines such as F1 score, specificity and sensitivity measures are adopted to evaluate these algorithms. The Machine learning algorithms that will be considered in this study include: Logistic Regression, Naïve Bayes Classifier, Support Vector Machine (SVM) and Decision Tree.

* 1. **Statement of the Problem**

Despite the absolute importance of headline features on news production and utilization, little or no attention has been paid to the technical analysis and classification of the Nigerian News headlines in recent times. An attempt in this direction can produce better understanding of the changing trends in sentiments, frequency and emotions regarding key societal issues expressed on Nigerian News headlines. Hence, classification of Nigerian News headlines is pertinent towards pattern discovery of hidden insights regarding issues of national interest. By implication, it could assist the government to determine and evaluate the extent to which certain policies have impacted on some issues of national interest and able to make a more informed decision.

Improvements during the classification phase of Nigerian news headlines is being sought to avoid misclassification. However, each algorithm is important at its own place but the right choice of classifier must be made after taking a look at its advantages as well as disadvantages listed with it. It can be seen that proper choice of classifier and making right decision can boost up your entire news headline classification accuracy (Mazhar, Shehzad and Muhammad, 2014)

**1.3 Aim and Objectives**

**1.3.1** **Aim**

The aim of this project is to apply some selected machine learning algorithms (Support Vector Machine, Naïve Bayes, Logistic Regression and Decision Tree) to classify Nigerian News headlines.

**1.3.2** **Objectives**

The specific objectives are to:

i. develop temporal trend pattern of key terms of national interest including kidnapping, rape, child abuse from Nigerian News headline dataset;

ii. classify the dataset using Support Vector Machine, Naïve-Bayes, Logistic Regression and Decision Tree

iii. evaluate the performance of the selected machine learning algorithms using F1 score, accuracy, precision and recall measures.

**1.4 Significance of the Study**

The significance of this study is multi-fold:

1. It can help to identify and reveal the key hidden temporal trends regarding certain phenomenal national issues, for example, rape, kidnapping, child abuse, terrorism, corruption and so on.
2. It can assist the government to identify key societal challenges and to devise more sustainable ways to address them.
3. The news insights derived from the analysis can assist the government to re-prioritize her pressing agenda.

**1.5 Scope of the Study**

This study captures dataset of Nigerian News Headlines from different newsvendors such as Punch, Vanguard, The Nation, Daily-times from 01 January 2020 to present. The selected machine learning techniques are Support Vector Machine, Naïve Bayes, Logistic Regression and Decision Tree.

**CHAPTER TWO**

**LITERATURE REVIEW**

**2.1 News**

News is an unpublished account of human activity, which seeks to interest, inform, or educate the readers. It refers to information on current events, happenings, or anything.According to Oxford Dictionary, it defines News as “New information, the report of latest incident”. Fowler (1991) also defined News as a product of journalistic process, an end result of a “systematic sorting and selecting of events and topics according to a socially constructed set of categories”. “News” may also refer to stories or information about social events “that are of interest to a sufficiently large group, or that may affect the lives of a sufficiently large group” (Reah 1998).

According to Carrel Warren, “news is usually a report which is not known to layman before its presentation. This report deals with such activities of man and is a source of interest, entertainment or information to the readers. British Journal also defines news as, “any event, idea or opinion that is timely, that interests or affects a large number of people in a community and that is capable of being understood by them”.

**2.1.1 Importance of News**

News is important for a number of reasons within a society.

* it informs our view of the world, and in response we take action and make choices based on how we perceive the world to be.
* It allows us have access to current events and up-to-date information.
* to inform the public about events that are around them and may affect them.
* for entertainment purposes.
* It allows us know what goes on in the world in many disciplines like history, everything in the economy and banking sector, politics, etc.

(Source: Samreen, 2020)

**2.1.2 Structure of News**

* **Headlines:** a news headline tells what the story is all about. It is the title of the Newspaper story or the summary of the whole news.
* **Body/content:** it tells us more information and details about the news.
* **End:** gives something to think about.

Source: (Rahul, 2015)

**2.2 Headlines**

A news headline is the title of the Newspaper story or the summary of the whole news. It is printed in large letters at the top of the story, especially on the front page. The headlines are the main points of the news which are read on radio or television. They occupy more space than the texts they refer to and where more significant words or coinages exist, they are still printed bolder than others or highlighted in italics (Collins dictionary).

The news headline is usually on the front page of a magazine or newspaper. Usually printed in bold letters, itsummarizes the main points of the story. This is in order to stress a particular word or idea. Words on the headline are usually short and crisp. Wording is determined by ideas to be expressed, technology of printing and the kind of audience associated with the paper or magazine. Technology of printing includes choices in terms of style, availability of space and accompanying photographs. Other criteria include simplicity, informality and impact (Freeborn et al, 1986).

News Headlines are vital in both capturing readers’ attention and in inﬂuencing their online reading experience of news. In fact, approximately six in 10 people limit their reading to headlines only, without clicking on a link to the full article. Eye-tracking studies have conﬁrmed this behavior empirically; many people are “entry-point readers”, who attend to headlines in order to ascertain the over view of an article, but who exhibit minimal reading activities (Holsanova, Rahm, and Holmqvist, 2006).

**2.3 News Classification**

A sample news classification system model is given as follows. It consists of the dataset acquisition stage, DataPreprocessing, Feature Extraction, Machine Learning Algorithm for Classification, Training of Classifier, Test Classifier with Trained Model and Evaluation phase.

Dataset

Dataset Pre-Processing

Feature Extraction

Machine Learning Algorithm for the Classification

Training the Classifier

Testing the Classifier

Evaluation

Figure 2.1 Architecture of a News Classification Process (Nabamita*et al.,* 2020)

**2.3.1 Dataset**

Oxford Dictionary defines a dataset as “a collection of data that is treated as a single unit by a computer”. This means that a dataset contains a lot of separate pieces of data but can be used to train an algorithm with the goal of finding predictable patterns inside the whole dataset (Iryna, 2021). The important aspect that needs to be captured when considering a dataset includes the size of the dataset, the dimension of the dataset, the number of rows and names of columns, the file extension type e.g .txt, .xlsl, and csv.

The figure below shows the sample of a dataset

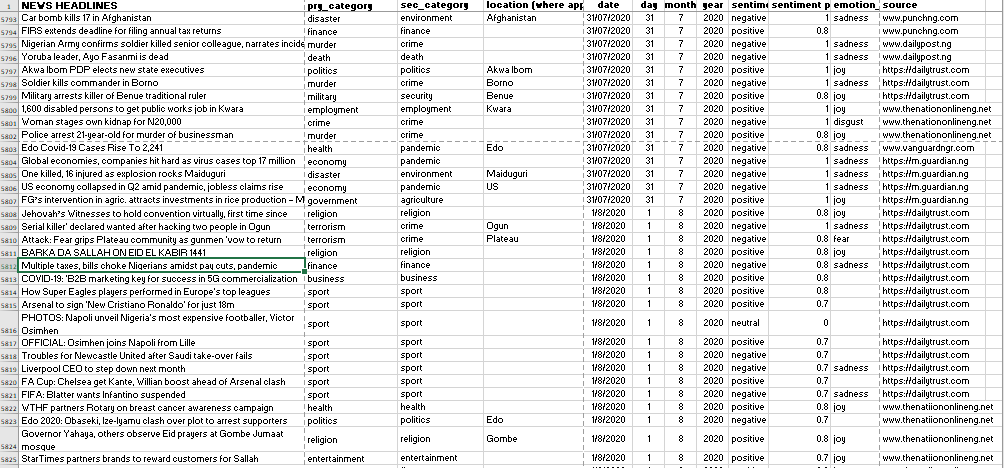


Figure 2.2 sample of the Nigerian news headline dataset

**2.3.2 Dataset Pre-Processing**

Data preprocessing is a data mining technique used to turn the raw data gathered from diverse sources into cleaner information that’s more suitable for work. In other words, it’s a preliminary step that takes all of the available information to organize it, sort it, and merge it (Ryan, 2019).

Why is Data Pre-Processing required?

Real world data is generally:

**Incomplete:** certain attributes or values or both are missing

**Noisy:** data contains errors or outliers

**Inconsistent:** Data contains differences in codes or names e.t.c

(Harshita, 2020)

The stages involved on Data Pre-Processing are listed below;

* Data Cleaning: The data cleaning process detects, removes the errors and inconsistencies present in the data and improve its quality.
* Data Integration: This task involves integrating data from multiple sources such as databases (relational and non-relational), data cubes, files, etc. The data sources can be homogeneous or heterogeneous. The data obtained from the sources can be structured, unstructured or semi-structured in format.
* Data Transformation: This involves normalization and aggregation of data according to the needs of the data set
* Data Reduction: If the data is very large, data reduction is performed. Sometimes, it is also performed to find the most suitable subset of attributes from a large number of attributes. This is known as dimensionality reduction. Data reduction also involves reducing the number of attribute values and/or the number of tuples.
* Data Discretization: It is considered as a part of data reduction. The numerical attributes are replaced with nominal ones.

(Harshita, 2020)

**2.3.3 Feature Extraction**

Feature extraction is a major dimensionality reduction to ensure the accuracy of the classification and improved time efficiency (Fagbola et al., 2012; Fagbola et al., 2017).It aims to reduce the number of features in a dataset by creating new features from the existing ones (and then discarding the original features). These new reduced set of features should then be able to summarize most of the information contained in the original set of features. In this way, a summarized version of the original features can be created from a combination of the original set (Pier,2019).

The approach to feature extraction is bag of words where each word is represented as a feature. Bag of wordsmodel is the simplest form of text representation in numbers. The approach represents the count of each word in the document as feature and does not take the order of words into account (Siwei et al., 2015).Term-frequency inverse-document-frequency (TF-IDF) is numerical statistics that reflect how important a word is to a collection of text or corpus (Purva, 2020). The TF-IDF approach of bagof words weights documents not only by their counts but also by their frequency in the whole trainingdataset(Siwei et al., 2015).

**2.3.4 Machine Learning Algorithms used for the Classification**

The next most important phase after feature selection is classification where the news headlines are classified with the aim to assign them to their respective classes. The most common news headlines classification methods are Naïve Bayes, Support Vector Machine, Neural Network, Random Forest and Decision Tree. (Nabamita*et al.,* 2020)

**2.3.5 Training of the Classifier**

In this stage, the classifiers are trained in order to understand how the given input variables are related to the class. Below is the architecture for a training data.

**Train Data**

**Test and update**

**Model (machine learning models)**

()

**Prediction**

Figure 2.3 Architecture of training Data

**2.3.6 Testing the Classifier with the Trained Model**

This is the stage where the trained dataset is tested by inputting a different dataset from testing set and then allowed to predict the members of the class based on what they’ve learned. It is used to measure the accuracy and efficiency of the algorithm used to train the machinei.eto see how well it can predict new classes based on its training. Below is the training set architecture;

**2.3.7 Evaluation**

This is the stage where the machine learning algorithms are evaluated to know how each of them performed.This step shows us how accurate our results are and how efficiently each news headline is classified into its pre-defined class. Researchers in literature have used numerous measures for this purpose i.e. accuracy, precision/recall, fallout, error, and much more. Few of these measuresare listed below;

* **Precision**: is defined as a fraction of news headlines that is relevant.
* **Recall**: is defined as fraction of relevant news headlines that is retrieved.
* **True Positive**: means that news headline is classified to its correct class.
* **False Negative**: means that news headline is classified to a wrong class.
* **True Negative**: means that news headline does not belong to that class and is misclassified.
* **Accuracy**: of a news headline is defined as the sum of true negative and true positive.

**2.4 Machine Learning Algorithms**

The Machine Learning Algorithms that will be discussed in this section are Naïve Bayes, Support Vector Machine, Logistic Regression and Decision Tree.

**2.4.1 Naïve Bayes Classifier Algorithm**

Naive Bayes classifier is a classification algorithm which is based on Bayes’ Theorem. It is a family of algorithms where all of them share a common principle, where every pair of features being classified is independent of each other. The Naïve Bayes classifier is a simple approach to the classification task that is still capable of providing reasonable accuracy (Nabamita*et al.,* 2020). It is a Probabilistic classifier based on text features. It calculates class labels and probability of classes (Gurmeet and Karan, 2016). Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles. Mathematically, it is represented as:

(1)

* P(h): the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h.
* P(D): the probability of the data (regardless of the hypothesis). This is known as the prior probability.
* P(h|D): the probability of hypothesis h given the data D. This is known as posterior probability.
* P(D|h): the probability of data d given that the hypothesis h was true. This is known as posterior probability.(source: Avinash, 2018)

**2.4.2 Support Vector Machine Algorithm**

A Support Vector Machine (SVM) is an algorithm of supervised learning which is used for fast and dependable classification that performs very well with a limited amount of data (Nabamita*et al.,* 2020). SVM has been used a lot for news text classification. It has a unique feature that it includes both negative and positive training sets which is generally not preferred by other algorithms (Gurmeet and Karan, 2016). It has helped researchers a lot for performing short text news classifications as compared to full text and have shown considerable results. Human reader emotions were classified and identified using SVM (Mazhar, Shehzad and Muhammad, 2014).

Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes (Rushikesh, 2018).SVM constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane(MMH) that best divides the dataset into classes. The figure below shows the diagram of Support Vectors (Avinash, 2019).

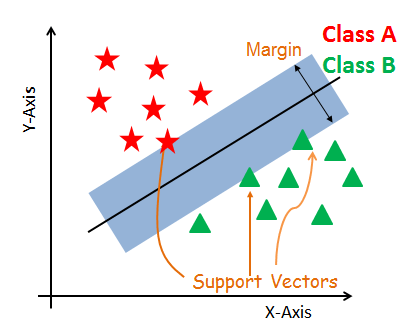


Figure 2.5diagram of support vectors (Source: Avinash, 2019)

**Support Vectors**

Support vectors are the data points, which are closest to the hyperplane. These points will define the separating line better by calculating margins. These points are more relevant to the construction of the classifier.

**Hyperplane**

A hyperplane is a decision plane which separates between a set of objects having different class memberships.

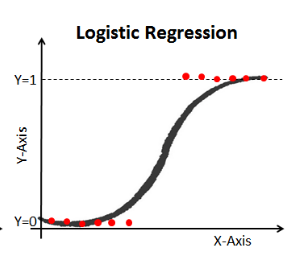
**Margin**

A margin is a gap between the two lines on the closest class points. This is calculated as the perpendicular distance from the line to support vectors or closest points. If the margin is larger in between the classes, then it is considered a good margin, a smaller margin is a bad margin.

**2.3.3 Logistic Regression**

Logistic Regression is a Machine Learning algorithm which is used for the classification problems. It is a predictive analysis algorithm and based on the concept of probability (Ayush, 2019).It is a process of modeling the probability of a discrete outcome given an input variable. The most common [logistic regression models](https://www.sciencedirect.com/topics/computer-science/logistic-regression-model) a binary outcome; something that can take two values such as true/false, yes/no, and so on (Thomas and David 2017). Logistic regression, despite its name, is a classification model rather than regression model. It is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification in industry (Abdulhamit, 2020).

The sigmoid function, also called logistic function gives an ‘S’ shaped curve that can take any real-valued number and map it into a value between 0 and 1. If the curve goes to positive infinity, y predicted will become 1, and if the curve goes to negative infinity, y predicted will become 0. If the output of the sigmoid function is more than 0.5, we can classify the outcome as 1 or YES, and if it is less than 0.5, we can classify it as 0 or NO(Avinash ,2019).



Source: (Avinash ,2019).

Figure2.6 Architecture of logistic function

**2.4.4 Decision Trees**

Decision Tree classifier is an algorithm which belongs to the family of supervised learning algorithms which can be used for solving regression and classification problems too. It is represented in a tree form of structure where the branches of tree represent weight and each leaf is a different class. The main aim of using Decision Tree is to create a training model which can be used to predict class or value of target variables by learning decision rules inferred from prior data (training data) (Nabamita*et al.,* 2020). Decision trees are easy to understand and rules can be easily generated through them. They can solve complex problems very easily (Mazhar, Shehzad and Muhammad, 2014).

A decision tree is a flowchart-like tree structure where the topmost node is known as the root node, the internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. The flowchart-like structure helps in decision making which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret. The diagram below explains the general structure of a decision tree(Avinash, 2018);

****

(Source: Avinash, 2018)

Figure 2.7 Architecture of decision tree

**2.5** **RELATED WORKS**

Pope (2007) proposed a model that determines the types of online news headlines most often selected by news websites as their Top Stories. In theresearch, headlines from four news websites were downloaded using Really Simple Syndication (RSS) feeds. Also, Supervised learning was conducted with the downloaded headlines to develop models which could automatically classify each website’s Top Story headlines, whose specific news category was unknown. The top Story headlines were also matched to headlines with known news categories from the same period to determine which news categories were most often represented as Top Stories. The results show that some news categories’ headlines, particularly those that had unique terms, were classified correctly based on the text contained in the headline. Furthermore, Headlines from World and US/UK news categories most often represented as Top Story headlines, followed by Business, Politics, and Entertainment.

Deshmukh and Kirange (2013) proposed an SVMbased news personalization system that aims at recommending news to the users as per their interests which are predefined in their profiles. The model considers news headlines which are classified in various categories using SVM. The classified news headlines are posted to the users according to their choices. They evaluated the performance of the system using 20 Newsgroup Dataset as well as using the real time dataset particularly from the Times of India news website and the experimental results show that SVM classifier performs best for classification of text.

Mazhar, Shehzad and Muhammad (2014) did a review on News Classification Based on Their Headlines. In their review, the existing classification methodswhich areNaïve Bayes, Support Vector Machine, Artificial Neural Network, Random Forest and K Nearest Neighbor are compared and contrasted based on various parameters namely criteria used for classification, algorithms adopted and classification time complexities. In their research, it is observed and clearly identified that no single classification technique can be mentioned as a general model for news headlines classification approach therefore, different classification scenarios and algorithms perform differently depending on news and data gathered.

Sandeep and Navdeep (2016) did a research on Online News Classification using Deep Learning Technique. In their work, their main aim was to increase accuracy in predicting the popularity of online news and this was implemented using Neural Network to acquire better results. The training of the Neural Networks classifier was done using newff function in MATLAB. The whole simulation of the work was done in MATLAB 2010 using various parameters like precision rate, recall rate as well as using accuracy. From their simulation result, it was concluded that Neural Networkshas the highest accuracy of 99.93 andhas provided good results with respect to traditional methodswhich are Naïve Bayes, Support Vector Machine, Artificial Neural Network, Random Forest and K Nearest Neighbor.

Mazhar *et al*., (2017) did a research work on News Headlines Classification Using Probabilistic Approach. The research work is aimed at news classification based of their headlines. After they analyzed the existing news classification methodologies, they presented a probabilistic framework which classifies each news headline to its pre-defined category by calculating its maximum probability in that category. This News headlines classification process is divided into three modules; pre-processing module, learning module, and the news headlines classification module.

Fagbola, Thakur and Olugbara (2019) did a performance comparison between two methods Artificial Neural Networks (ANN)and the Kolmogorov Complexity Distance Measure(KCDM) for addressing the news article classification problem. In their research, 2000 news articles were obtained from the publicly available BBC(British Broadcasting Corporation) News article dataset. The news article was pre-processed using Porter’s algorithm after tokenization and stop-words removal. An NTF-IDF (Normalized Time Frequency-Inverse Document Frequency) technique was used to extract and select relevant features before training and classification with the KCDM and ANN. The result of their research shows that ANN was better in terms of accuracywhile KCDM was better for developing time efficient applications.

Suleymanov and Rustamov (2019) did a research work onAutomated News Categorization using Machine Learning methods. In their research,130000 news articles were gathered along with their assigned categories. The classifier Algorithms used in their research are Naïve Bayes, Support Vector Machine and Artificial Neural networks. Count-Vectorization, TF-IDF Vectorization and removal of Stop Words was implemented for the pre processing stagewhikeChi-squared test and LASSO methods was implemented for feature selection.Hence, they tried to compare the accuracy results of classifiers with the accuracy results obtained on famous datasets. The result showed that using count vectorization yields best accuracy result for Naïve Bayes, while for SVM count vectorization yields lowest accuracy.

Nabamita*et al*., (2020) did a research on a comparative analysis of news categorization using machine learning approaches. In their work, they used a dataset from BBC(British Broadcasting Corporation) and the machine learning algorithms they used for their classification are Naïve Bayes, Support Vector Machine, Neural Network, Random Forest and Decision Tree respectively. The experimental result was analyzed for three evaluation parameters which are Accuracy, Precision and Recall. The result of their analysis show that the Naive Bayes performs better than the other four algorithms with the classification accuracy of 96.8 %, then followed by the Random Forest with accuracy 94.1 % of, Support Vector Machine (SVM) with accuracy of 96.4 %, Neural Networks with accuracy of 96.4 % and the Decision Tree with accuracy of 83.2%.

Emenike(2020), conducted a study on selected headlines of the Nigerian Vanguard newspaper. The study investigated the stylistic features of Nigerian Vanguard headlines. It adopts Halliday‘s theory of functional stylistics.The stratified sampling method was used to collect 120 headlines published between January and December, 2014. In doing this, the graphology, grammar, lexis and other notable language features were examined. The study was able to identify and explain language features that characterize the style of language use in *Nigerian Vanguard Newspaper* headlines.

Zhonglei*et al.,* (2017) proposed an efficient approach for Chinese news headline classification based on multi-representation mixed model with attention and ensemble learning. Firstly, they modeled the headline semantic both on character and word level via Bi-directional Long Short-Term Memory (BiLSTM), with the concatenation of output states from hidden layer as the semantic representation. And then, for samples with lower confidence level in the preliminary test result, they utilize ensemble learning to determine the final category of the whole test samples by sub-models voting.They then analyze the confidence level distribution of the correct and error predictive samples in development set based on the multi-representation mixed model with attention.

Their Statistical results show that, 15.34% error samples have predictive confidence above 0.80, while 8.96% correct samples below 0.80. Besides their multi-representation mixed model with attention, they selected the N-BoW and CNN as their sub-models according to the principle of "difference meets complementation" in feature extraction. The N-BoW and CNN models were trained using the same training data. After obtaining three trained sub-models, they first predict on the whole testing set using the single multi-representation mixed model with attention and the samples with lower confidence are screened as TestData-2. Then they test their three sub-models on TestData-2.The dataset was collected from several Chinese news websites and there are 18 categories in total.The macro-averaged precision, recall and F1 were used to evaluate the performance.

**CHAPTER THREE**

**METHODOLOGY**

**3.0 Research Approach**

A four-staged architecture comprising of the news headlines dataset collection, news headlines dataset pre-processing, feature extraction and classification using Naïve Bayes, Logistic Regression, SVM and Decision Tree is represented in the figure below. After the classification stage was completed, the performances of the machine learning algorithms were evaluated using F1 score, accuracy, specificity and sensitivity measures.

**Nigerian News Headlines Training Dataset**

**Nigerian News Headlines Testing Dataset**

**Dataset**

**Pre-Processing**

**Tokenization**

**Stop Words Removal**

**Word Stemming**

**Feature Extraction**

**Nigerian News Headlines Descriptor (TF-IDF)**

**Nigerian News Headlines Classification**

**Training**

* **SVM**
* **Decision Tree**
* **Logistic Regression**
* **Naïve Bayes**

**Testing**

**Classification**

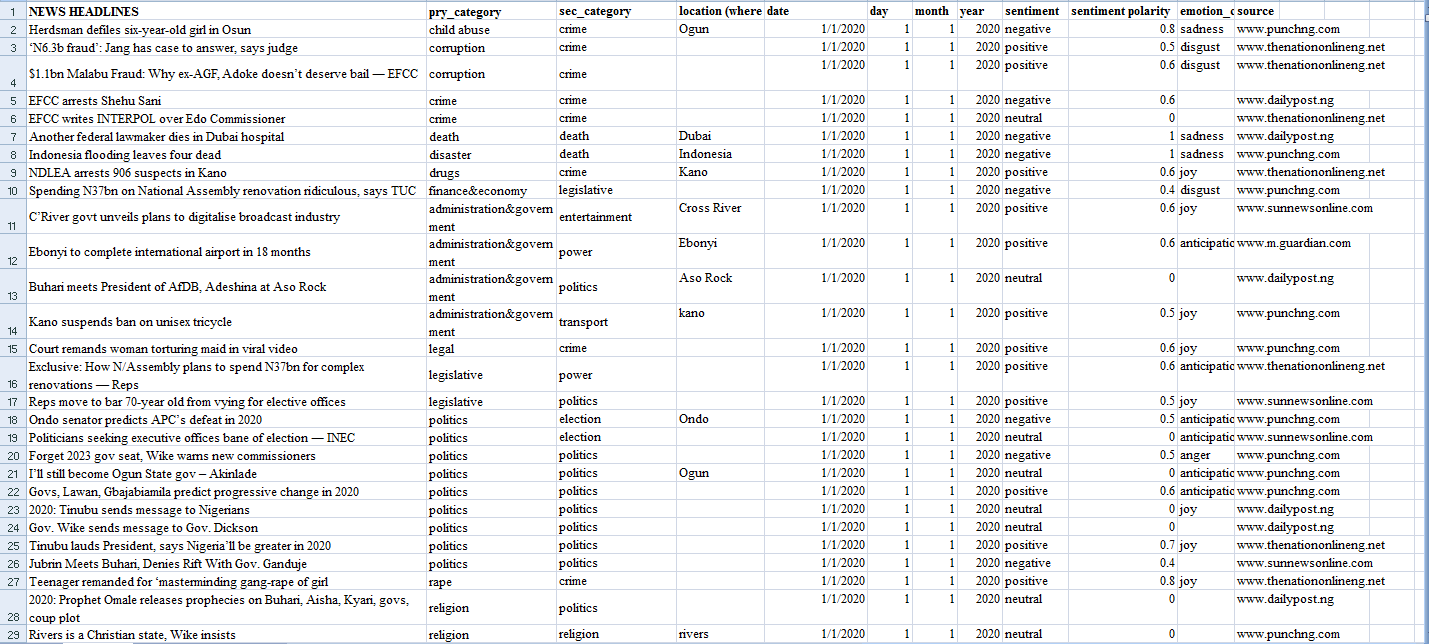
**Classification Output and performance Evaluation**

**Bag of words (BoW)**

Figure 3.1 Classification architecture used for the Nigerian News Headlines

**3.1 DATASET COLLECTION**

The news headlines were sorted from different Nigerian News Vendors which includes Vanguard, Punch, The Nations, Daily Times, Daily Trust, Sun News Online, Sports Bild, The Eagles Online and many more. The news headlines were annotated manually and were gathered from January 1st to December 31thof the year 2020. There are 8386rows in the dataset and there are 50 primary categories and 41 secondary categories which includes rape, murder, crime, politics, education, science&technology, legal, legislative, pandemic, health, business, corruption, education, death, finance&economy, drugs, employment, entertainment, food&agriculture, accident, government&administration, aviation, border control, chieftaincy, child abuse, communication, cultism, crash, democracy, disaster, disaster management, disease, drugs, social, public health, education, environment,tourism&culture, law enforcement, weather, travels, infrastructure, internally displaced persons, defense, mineral resources, oil&gas, riot&protest, religion, sports, suicide, terrorism and transport respectively. The nature of the dataset is text and its format is csv. The figure below presents the sample of the dataset.

Figure 3.2 sample of the dataset

The dataset consists of 12 columns; one has the News Headlines, followed by the Primary Category, Second Category, Location (where applicable), Date, Day, Month and Year, Sentiment, Sentiment Polarity, Emotion Class and Source.

**3.2 NEWS HEADLINES DATA PREPROCESSING**

After the collection of news headlines data preprocessing is done. The data that is obtained may be incomplete, noisy and inconsistent. Such data makes the pre-processing of data a necessity as no results could be obtained from incomplete information (Gurmeet and Karan, 2016). In this study, tokenization, removal of stop-words and stemming were performed. Tokenization involves breaking text into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes the input for further processing (Gunjit, 2018). Stop words are words which appear frequently in text. They are considered of less importance and are removed eventually because they do not change the actual meaning of the text. They are words which are frequent in many news headlines and are defined useless with respect to frequency (Mazhar, Shehzad and Muhammad, 2014). Stemming is the final and most important part of the preprocessing. It reduces inflection in words to their root forms such as mapping a group of words to the same stem. The reduced stem needs not to be a dictionary word (Hafsa, 2018). Porter Stemmer was used because of its high precision, high efficiency and simple algorithm than other stemming approaches such as S-Stemmers, Lovins Stemmer and Paice/Husk Stemmer (Gurmeet and Karan, 2016).

**3.3 FEATURE EXTRACTION**

Feature extraction is a major dimensionality reduction step to ensure the accuracy of the classification and improved time efficiency (Fagbola*et al.*, 2012; Fagbola*et al.*, 2017). The features were extracted from the terms returned from the preprocessing using TF-IDF Vectorizer. Since computers do not understand natural language,they need us to break down the text into a numerical format that’s easily readable by the machine. The news headlines (text) are converted into vectors i.e.the terms gotten from the preprocessing needs to be encoded as integers or floating-point values for use as input to a machine learning algorithm. The TF-IDF technique was adopted due to its high accurate performance when compared to other statistical approaches.

According to (Mandal and Sen, 2014), For each term i, the weight is given as:

(2)

where 𝑛𝑖 is the number of documents containing term 𝑖 and 𝑁 is the total number of documents. TF defines the measure of frequency of each term/word in a document, while IDF is the measure of how important a word is. A weight-term matrix with documents forming the rows and TF-IDF weights forming the columns was generated by calculating the TF-IDF for each feature, where 𝑤𝑖𝑗 is the weight of term 𝑖 in document 𝑗, 𝐷 is a sample document, 𝑇 represents a term and is the class of the document.

𝑇1 𝑇2 … 𝑇𝑖

𝐷1 𝑤11 𝑤12 … 𝑤1𝑖 𝑐1

𝐷2 𝑤21 𝑤22 … 𝑤2𝑖 𝑐2 (3)

: : : : : :

𝐷𝑗 𝑤𝑗1 𝑤𝑗2 … 𝑤𝑗𝑖 𝑐𝑘

The pseudocode for the feature extraction step is presented in Figure 3.3 below. The relevant-term matrix features obtained at this stage were introduced into the classifiers (Naïve Bayes, SVM, Decision Tree and Logistic Regression) for term similarity estimation and final classification.

𝑰𝒏𝒑𝒖𝒕

𝑇: 𝑡h𝑒 𝑢𝑛𝑖𝑞𝑢𝑒 𝑡𝑒𝑟𝑚𝑠 𝑖𝑛 𝑎𝑙𝑙 𝑑𝑜𝑐𝑢𝑚𝑒𝑛𝑡𝑠

𝐷: 𝑡h𝑒 𝑑𝑜𝑐𝑢𝑚𝑒𝑛𝑡𝑠 𝑜𝑓 𝑡h𝑒 𝑡𝑟𝑎𝑖𝑛𝑖𝑛𝑔 𝑠𝑒𝑡

𝑶𝒖𝒕𝒑𝒖𝒕**:** 𝑤𝑒𝑖𝑔h𝑡𝑀𝑎𝑡𝑟𝑖𝑥

𝑺𝒕𝒆𝒑:

𝑖. 𝒇𝒐𝒓 𝑒𝑎𝑐h 𝑡𝑒𝑟𝑚 𝑡𝑖∈𝑇 𝒅𝒐

𝑖𝑖. 𝒇𝒐𝒓 𝑒𝑎𝑐h 𝑑𝑜𝑐𝑢𝑚𝑒𝑛𝑡 𝑑𝑗∈𝐷 𝒅𝒐

𝑤𝑖𝑗=𝑓𝑟𝑒𝑞𝑢𝑒𝑛𝑐𝑦 𝑜𝑓 𝑡𝑒𝑟𝑚 𝑡𝑖 𝑖𝑛 𝑑𝑜𝑐𝑢𝑚𝑒𝑛𝑡 𝑑𝑗

𝑖𝑖𝑖. 𝒆𝒏𝒅𝒇𝒐𝒓

𝑖𝑣. 𝒆𝒏𝒅𝒇𝒐𝒓

Figure 3.3 the pseudocode for the normalized TF-IDF feature extraction step.

**3.4 NIGERIAN NEWS HEADLINES CLASSIFICATION**

After feature extraction, the next phase is the classification phase which is an important phase in which the aim is to classify the unseen news to their respective categories (Gurmeet and Karan, 2016). In this study, Naives Bayes classifier, Support Vector Machine, Logistic Regression and Decision Tree was used for news classification into various categories.

**Support Vector Machine (SVM) Classifier:** The Support Vector Machine classifier is one of the best machine learning techniques that outperform in most of the cases. It is a learning algorithm that works for both classification and regression problems. This classifier was selected because of its high performance even though it has a huge computational time (Khadjeh, 2015). The goal of SVM is to find the optimal separating hyperplane that gives the maximum separation margin between the hyperplane and the nearest points of both classes. The set of training data is shown in (4) below:

(4)

A hyperplane can be found to separates the two classes. The equation of a hyperplanedividing the points (for classifying) is written as (Radhika, 2020);

(5)

**Naïve Bayes Classifier:**is a probabilistic classifier which is based on text features and calculates class labels and probability of classes. It isn’t made up of a single algorithm for classification but it includes a large number of algorithms that work on a single principal for training classifiers. The principal states that the value of a particular feature is autonomous of value of any other feature specified in a class (Gurmeet and Karan, 2016).The mathematical representation is represented as follows;

(6)

P(A|B): This is known as Posterior probability. It is the Probability of hypothesis A on the observed event B.

P(B|A): This is known asLikelihood probability. It is the Probability of the evidence given that the probability of a hypothesis is true.

P(A) : This is known as Prior Probability. It is the Probability of hypothesis before observing the evidence

P(B) is Marginal Probability: Probability of Evidence.

**Logistic Regression Classifier**: it is a statistical model which is typically used to model a binary dependent variable with the help of a logistic function. It is majorly used for binary classification tasks and can also be used for multiclass classification. Another name for the logistic function is a sigmoid function and it is mathematically represented as (Khuswant, 2020);

(7)

This function assists the logistic regression model to squeeze the values from (-k,k) to (0,1).

**Decision Tree Classier:** Decision trees are constructed via an algorithmic approach which identifies ways to split a dataset based on different conditions. It is one of the most widely used and practical methods for supervised learning (prince, 2018).The main aim of using Decision Tree is to create a training model which can be used to predict class or value of target variables by learning decision rules inferred from training data (Nabita*et al.,* 2016). The mathematical representation goes thus;

(8)

The dependent variable,Y, is the target variable that we are trying to understand, classify or generalize. The vector x is composed of the features, x1, x2, x3 etc., that are used for that classification (prince, 2018).

**3.5 PERFORMANCE EVALUATION METRICS**

The F1 Score, Accuracy, Specificity and Sensitivity are the basic evaluation metrics considered in this work. These metrics assisted in determining the best-fit algorithm among SVM, Decision Tree, Logistic Regression and the Decision Tree for news headlines classification.

**3.5.1 F1 Score**

The **F1** score is the [harmonic mean](https://en.wikipedia.org/wiki/Harmonic_mean) of the precision and recall. It is calculated from the [precision](https://en.wikipedia.org/wiki/Precision_(information_retrieval)) and [recall](https://en.wikipedia.org/wiki/Recall_(information_retrieval)) of the test, where the precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive. It is represented mathematically as (Sasaki, 2007);

F1 = (9)

**3.5.2 Accuracy**

Accuracy is the proximity of measurement results to the true value;it is the closeness of the measurements to a specific value. It is the sum of true negative and true positive. It is represented mathematically as (Powers and David, 2015);

(10)

**3.5.3 Recall**

It is the fraction of relevant news that is retrieved and can also be defined with respect to either of the classes.Recall of positive class is termed sensitivity and is defined as the proportion of [truly positives cases](https://chemicalstatistician.wordpress.com/2014/05/12/applied-statistics-lesson-of-the-day-type-i-error-false-positive-and-type-2-error-false-negative/) that were classified as positive, while Recall of negative class is termed specificity and is defined as the proportion of [truly negative cases](https://chemicalstatistician.wordpress.com/2014/05/12/applied-statistics-lesson-of-the-day-type-i-error-false-positive-and-type-2-error-false-negative/) that were classified as negative (Rajkumar, 2021 ).

(11)

**3.5.4 Precision**

It is defined as a fraction of news that is relevant.The precision of negative class is the ability of the classifier not to label negative sample as positive. The precision of positive class is the ability of the classifier not to label positive sample as negative. The best value of precision is 1 and the worst value is 0(Rajkumar, 2021 ).

(12)

**CHAPTER FOUR**

**RESULTS AND DISSCUSSION**

**4.1 Results of Pre-processing**

Pre-processing simply means bringing our text into a form that is predictableand analyzable for our task. It was done by first tokenizing each news headlines, converting each text in the headlines to lower case, removing stop-words from the text, and stemming each word.

**4.1.1 Tokenization**

Each feature (news headline) was tokenized by breaking them into meaningful tokens or words. The words are then converted into lowercases and special characters were removed. The output of the output of the tokenized text is showed in the figure below

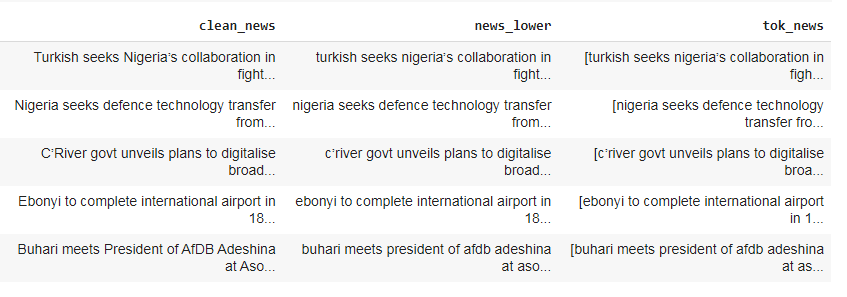
****

Figure 4.1 result of tokenization

**4.1.2 Removal of Stop-Words**

Removal of stop-words was done after the text is tokenized. This is done because they are considered to be less important and they do not change the actual meaning of the text when removed.The result of this process is displayed below;



Figure 4.2 Result of stop-words removal

**4.1.3 Stemming**

Lastly, the results gotten from the removal of stop-words were stemmed using Porter stemmer. Stemming means reducing a word to its root stem and the reduced word needs not to be in the dictionary. Porter stemmer was used because of its high efficiency and high precision. The result of the stemmed word is showed below;



Figure 4.3 result of stemmed word

**4.2 Results of Feature Extraction**

After the news headlines have been preprocessed,the text needs to be transformed to vectors so that algorithms will be able make predictions. In this case, the Term Frequency — Inverse Document Frequency (TFIDF) Vectorizer will be used to extract relevant features and to evaluate how important a word is to a document/text in a collection of documents/texts.The first column shows the row number of the trained data, the second column shows the unique integer of the number of each word in the first row, while the last column shows the score calculated by TF-IDF Vectorizer. The figure below shows the output of the extracted features

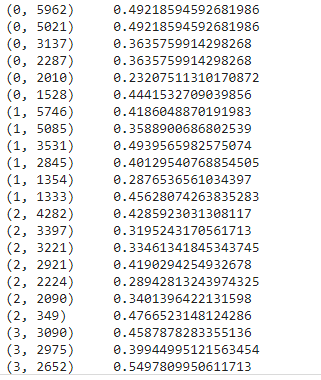


Figure 4.4 Result of the extracted features

**4.3 Result of Classification**

After tokenizing, removing punctuation, lower casing, removal of stop words and stemming was done, importance of a word is determined in terms of its frequency. The original data was divided into features (X) and label (y), which were then splitted into train (80%) and test (20%) sets. Thus, the algorithms were trained on one set of data and tested out on a completely different set of data. Four algorithms were used which are Logistic Regression, Naïve Bayes, SVM, and Decision tree. Also, test Accuracy scores, F1 scores, recall and precision were compared. For this dataset, It was revealed that Logistic Regression classifier showed the best performance compared to the other classifiers for the classification. It has the highest accuracy of 91%, followed by SVM with the accuracy of 80%, Decision tree of 73% accuracy and lastly Naïve Bayes of 67% accuracy. The figure below shows the classifiers vs their accuracy score.

Figure 4.5 Classifiers vs Accuracy score

**4.3.1 Report of the dimension of the dataset used for the Classification process**

The dataset used for this study was selected from the Nigerian news Headlines for the year 2020.The dataset comprises news headlines from ten categories. These categories include Business,Science&Technology, Entertainment, Finance&economy, Administration&government, Religion, Education, Health, Politics and Sports. Total number of the Dataset is 3,830. The Bar chart in Figure 4.6 below shows the analysis of the selected Categories used for the evaluation process.

Figure 4.6 Bar chart showing the analysis of the categories used for the Classification process

**4.3.2 Classification report for Machine learning Algorithms**

The classification report of these machine learning algorithms are discussed below. Table 4.1 shows the classification report of Naïve bayes classifier, Table 4.2 shows the classification report of SVM classifier, Table 4.3 shows the classification report of decision tree classifier and Table 4.4 for shows the classification report of logistic regression classifier. Also, the comparison between these machine learning algorithms are presented in Table 4.5 below where the average rate of precision, recall and F1 Score are stated.

**Table 4.1 Naïve Bayes Classification Report Nigerian News Headline Classification**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category of News | Accuracy  67% | Precision | Recall | F1 score |
| Health | 0.51 | 0.61 | 0.56 |
| administration&government | 0.95 | 0.46 | 0.52 |
| Sports | 0.32 | 1.00 | 0.49 |
| Politics | 0.00 | 0.00 | 0.00 |
| finance&economy | 0.38 | 0.89 | 0.54 |
| Education | 0.98 | 0.54 | 0.70 |
| Religion | 0.77 | 0.86 | 0.81 |
| Entertainment | 0.00 | 0.00 | 0.00 |
| Business | 0.00 | 0.00 | 0.00 |
| science&technology | 0.90 | 0.93 | 0.91 |

As can be seen from Table 4.1 above, the precision rate, recall rate and F1 score for the Categories of Politics, Entertainment, and Business is 0. This may be as a result of weak prediction or identification. Education, Administration&government, and Science&technology have a high precision rate of 0.98, 0.95 and 0.90 respectively.For the Recall rate, the Category Sports has the highest rate of 1.00, followed by Science&technology with 0.93, Finance&economy with of 0.89 and Religion with 0.86. For the F1 score, Science&Technologyhas the highest score of 0.91 followed by Religion with 0.81 and Education with 0.70 respectively.

**Table 4.2 SVM classification report**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category of News | Accuracy  80% | Precision | Recall | F1 score |
| Health | 0.78 | 0.67 | 0.72 |
| administration&government | 0.00 | 0.00 | 0.00 |
| Sports | 0.75 | 0.88 | 0.81 |
| Politics | 0.40 | 1.00 | 0.57 |
| finance&economy | 0.65 | 0.85 | 0.72 |
| Education | 0.88 | 0.82 | 0.85 |
| Religion | 0.89 | 0.85 | 0.87 |
| Entertainment | 0.35 | 0.75 | 0.48 |
| Business | 0.00 | 0.00 | 0.00 |
| science&technology | 0.97 | 0.87 | 0.92 |

The classification report of the table above shows thatthe precision rate, recall rate and F1 score for the Categories ofadministration&governmentand Business is 0. This may also be as a result of weak prediction or identification. Categories science&technology, Religion, finance&economy, Sports and Health has the highest precision results. Category Politics has the highest result for recall rates followed by Sports,finance&economy, education, Religionentertainment and Health respectively. Lastly, science&technology has the highest F1 Score results followed by Religion, Education, Sports, Health and finance&economy

**Table 4.3 Decision Tree Classification Report**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category of News | Accuracy  **73%** | Precision | Recall | F1 score |
| Health | 0.64 | 0.70 | 0.67 |
| administration&government | 0.25 | 0.60 | 0.35 |
| Sports | 0.69 | 0.75 | 0.72 |
| Politics | 0.30 | 0.38 | 0.33 |
| finance&economy | 0.60 | 0.68 | 0.64 |
| Education | 0.83 | 0.83 | 0.83 |
| Religion | 0.84 | 0.67 | 0.75 |
| Entertainment | 0.41 | 0.70 | 0.52 |
| Business | 0.00 | 0.00 | 0.00 |
| science&technology | 0.86 | 0.70 | 0.77 |

The above table shows that only Category Business has no report for the precision, recall rate and F1 score. Science&technology has the highest precision rate, while education has the highest recall rate. For F1 Score report, Category Education has the highest result

**Table 4.4 Logisitic Regression Classification Report**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category of News | Accuracy  91% | Precision | Recall | F1 score |
| Health | 0.93 | 0.85 | 0.89 |
| administration&government | 0.25 | 0.60 | 0.35 |
| Sports | 0.89 | 0.95 | 0.92 |
| Politics | 0.11 | 1.00 | 0.19 |
| finance&economy | 0.87 | 0.94 | 0.90 |
| education | 0.99 | 0.89 | 0.94 |
| Religion | 0.95 | 0.98 | 0.97 |
| entertainment | 0.53 | 1.00 | 0.69 |
| Business | 0.00 | 0.00 | 0.00 |
| science&technology | 0.98 | 0.91 | 0.94 |

The logistic Regression Classification Report also show that Category Business has no result for the three metrics. Category education has the highest precision rate while Category politics has the lowest precision rate. Both entertainment and Politics has the highest recall rate of 1 while Religion has the highest score for the F1 score results.

**Table 4.5 Comparison of thefour machine learning algorithm performance**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CLASSIFIERS** | **ACCURACY** | **PRECISION** | **RECALL** | **F1 SCORE** |
| **NAÏVE BAYES** | 67% | 0.48 | 0.53 | 0.43 |
| **SVM** | 80% | 0.57 | 0.70 | 0.59 |
| **DECISION TREE** | 73% | 0.54 | 0.60 | 0.56 |
| **LOGISTIC REGRESSION** | 91% | 0.65 | 0.81 | 0.68 |

The table above shows the comparison of all the evaluation metrics between the classifiers. From this table, it is revealed that Logistic regression has the highest in the accuracy score as well as in the recall rate, precision rate and F1 score. SVM has the second highest result in all the metrics, followed by the decision tree algorithm and lastly the Naïve Bayes Algorithm

**4.4 Temporal trend pattern of key terms of national interest**

This section talks about the key terms happening in the country at a particular time. The temporal trends could serve as key for pattern discovery of hidden insights regarding issues of national interest. They could also assist the government to determine and evaluate the extent to which certain policies have impacted on some issues of national interest and able to make a more informed decision. The patterns were generated using Microsoft word in which the values of each category are inputted on a monthly basis. The temporal pattern is represented in the figure below;

Figure 4. temporal trend pattern of key terms of national interest between January 2020 to December 2020

(landscape)

**CHAPTER FIVE**

**SUMMARY AND CONCLUSION**

**5.1 Summary**

In this study, collection, annotation and classification of the Nigerian News Headlines was done from January 1st 2020 to December 31st 2020. The total number 8,386 news headlines were cleaned by applying pre-processing technique such as stop words removal, tokenization, stemming and so on, then features extraction was done using the TF-IDF Vectorization. After cleaning, the selected machine learning algorithms were applied to this dataset. The results showed that ten categories have been used for the evaluation processes which areHealth, Administration&government, Sports, Politics, Finance&economy, Education, Religion, Entertainment, Business, Science&technology. Also a performance comparison among these algorithms(Logistic Regression, Naïve Bayes, SVM, and Decision tree) was conducted. From the result it has been concluded that Logistic Regression has the highest accuracy of 91%.

**5.2 Conclusion**

Summarily, this study establishes the relative importance of classifying News Headlines as a way of identifying and revealing the hidden trends of National issues in order to devise more sustainable ways to address them and to re-prioritize these pressing agenda.Furthermore, the classification process can help identify certain trade-offs associated with each algorithm and synergize decision making on what algorithm to apply to aparticular problem or a domain of interest. It is observed that no single classification technique can be mentioned as the best model for news headlines classification approach. Different classification algorithms perform differently depending on news and data gathered.

In future research, the Nigerian News Headline dataset can be used for further study such as Sentiment Analysis, Emotion Classification, Misclassification of News Headlines, Fake news Detection or Mis-Information, Classification of individual categories for example sports and exploring the different types of sports we may have, and numerous other areas that News headline Classification can be applied to.

**5.3 Limitations of the study**

As there are no existing dataset for Nigerian News Headlines, the dataset was gathered, sorted and annotated into their various categories manually. As a result, the list of categories became too much and due to this, the accuracy of the machine learning algorithms was very poor. In other to correct this, the works of Muhammad Badruddin Khan (2021) on Urdu News Classification using Application of Machine Learning Algorithms on News Headline, Nabamita*et al*, (2020) on A Comparative Analysis Of News Categorization Using Machine Learning Approaches, Gurmeet and Karan (2016)and Mark W. Pope (2007) on Automatic Classification of Online News Headlines on News Classification using Neural Networkswerereviewed.Ten most common categories that have also been used by these researchers was used for the evaluation process. These categories include; Health, Administration&government, Sports, Politics, Finance&economy, Education, Religion, Entertainment, Business, Science&technology.