International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V69I2P104

Volume 69 Issue 2, 22-28, February 2021

Original Article

Giving Structure to Unstructured Text Data by Employing Classification

Ngor Gogo[1], Matthias Daniel[2], Alabo Gift[3]

[1,2,3] *Department of Computer Science, Rivers State University, Port Harcourt, Nigeria*.

Received Date: 15 December 2020
Revised Date: 27 January 2021
Accepted Date: 30 January 2021

Abstract *- As relevant as the need to have information readily available and well manage; quite a volume of information are inaccessible and locked up in a huge volume of text documents (unstructured data) that could be applied in the economy by the government, individuals, and corporate organization to ameliorate on the state of life and develop better working system; this cannot be overemphasized, therefore the need to extract this information and give a structure that will expedite adequate management, storage, and access when required because of their importance. The aim of this research is to implement a Classification Algorithm as a technique for giving Structure to Unstructured Data (Text document). The Multinomial Naïve Bayes classifier Algorithm was deployed for the purpose of classifying these unstructured data to give structure to it. There are two major phases involved in this: first is the pre-processing phase (Tokenization, Stemming, and Stop Word Removal), and second the Classification phase. The system built performed better, as shown from the result, that it can be used to classify text documents for proper and easy management, storage, and accessibility.*

Keywords *- Structure, Unstructured data, Classification, Multinomial Naïve Bayes classifier, Algorithm, Pre-processing*

I. INTRODUCTION

The challenges inherent in the management, accessibility, and storage of the huge amount of data produced and disseminated by various nations, industries, organizations, businesses, etc. brings to the fore the need for business organizations to pay closer attention to the relevance of unstructured data in today's intensely competitive world. These unstructured data occupy about 70% of the entire digital space. Goutam, in this publication (Analysis of unstructured data), showed that a significant volume of the digital world is dominated by unstructured data [1]. On an hourly basis, there has been a continuous increase in the volume of unstructured data (text documents). On a daily basis, there has been a continuous rise in the volume of unstructured data (text documents) and the amount of information closed-in and inaccessible. Consequently, the need to consider an efficient Classification technique that will enhance the fast and appropriate categorization of these text documents that will result in facilitating easy access and storage of this essential information.

Document classification is the method of searching for commonalities in documents in a collection and grouping them into an earlier defined label (supervised learning) on the grounds of a unifying idea that is a recurrent element displayed by the documents. Like the Incident Response (IR) processes, the classification of documents has a number of applications and is an essential part of text analytics. The following are some areas document classification applicable. News article classification, spam detection, call center routing, e-mail forwarding. Statistical Analysis System (SAS) is an advanced analytic software suite for business intelligence, predictive analytic, and data management. It has the capability of handling the categorization of documents that in reality display multiple themes and are not having a mutually exclusive class; limiting such documents to only one category will pose a challenge to large documents. In a situation where documents require being restricted to only one class, text clustering comes in handy rather than the extraction of text topic, which is common to document classification. The clustering algorithm helps to expose that the collection is really made up of categories like real estate, car, sales, etc. [2].

Document classification is a basic learning challenge confronting information retrieval and management. In application, document classification plays a very significant role in searching, organizing, representing, and classifying huge volumes of data; in a move to better the accuracy of document classification, a Feature algorithm was employed to achieve context of development centric topic. This was accomplished from two points of consideration. (a) Topics that are much focused (b) Regional specific features which are the result of the influence of local language and cultural undertone that may compromise the topic. The algorithm put together popularity and rarity as different feature extraction metrics that explain a topic [3].

An algorithm is a machine-like, simple, exact, brief, and unambiguous executable sequence of elementary instruction. It is a finite set of instructions used in solving the problem [2]. Fore gave an informal description of the algorithm as a procedure with the capability such that: if (the procedure has been spelled out) we pick any question of the class (category), the procedure will give us how to carry out a step by step performance, which after a limited number of steps, will produce the answer to the question we picked. In carrying out these steps, it is of utmost importance to follow the set down rules mechanically, like robots. No invention or pre-knowledge is needed of us. At the end of each step, if we do not arrive at the result yet, the instructions alongside the prevailing situation will inform us what we have to do next. The set down rules to be followed will give us the ability to recognize when we come to the terminal point in the steps and to obtain the answer to the question (yes or No) based on the result of the situation. In specific terms, no human operation can make use of an infinite volume of information. The set of rulers that describes the procedure must be finite [4].

II. RELATED LITERATURES

An Efficient Algorithm for Document Classification is a necessity that intends to unravel problems and challenges connected with proper organization, storage, and retrieval of unstructured data.

Unstructured data, according to [5] is information, whose forms differ in many ways and doesn't fit into the conventional data models; and obviously isn't suitable for a mainstream relational database.

Appreciably is the egress of other platforms for storing and managing such data; it is progressively prevalent in IT systems and is used by organizations in diverse operational intelligence and analytics applications.

In dealing with unstructured data, there is a need to look at the subject of Big Data, which is a broad umbrella under which unstructured data is a type. Big data is not a recent concept. History is furnished with volumes of data/information overflow characterizing as a result of the advent of social transformation and the introduction of new technologies. Within this period, individuals, various administrative powers, organizations, firms, and industries have come-up with essential data sets, organized at the time in a sequence that is logical and coherent. To be translated into relevant information about the past, making it as useful as necessary in present times [6].

Document classification is the job of organizing documents into categories based on their content. Document classification is an essential learning problem that is at the center of many information management and retrieval jobs. Document classification plays an essential role in various applications that deal with organizing, classifying, searching, and concisely representing an important amount of data.

Document classification has been an existing problem in information retrieval, which has been well undertaken [7]. Automatic document classification can be extensively categorized into three:

These are Supervised document classification, Unsupervised document classification, and Semi-supervised document classification.

A. Supervised Document Classification

In Supervised document classification, some processes occur externally to the classification model (generally human), providing relevant information related to the appropriate document classification. Therefore, in the case of Supervised document classification, it is much easier to prove the accuracy of a document classification model.

B. Unsupervised Document Classification

In Unsupervised document classification, there is no external information provision process at all [8].

C. Semi-Supervised Document Classification

In the case of Semi-supervised document classification, segments of the documents are tagged by an external mechanism [9].

There are two main factors that influence document classification task:

(a) feature extraction; (b) topic ambiguity. Firstly, Feature extraction is concerned with filtering out the right set of features that correctly depicts the document and helps in the formation of a good classification model. Secondly, many broad topic documents are themselves so ambiguous that it becomes challenging to allocate them into any particular class/category. For instance, a document whose dealing/ content is on theocracy. In such a document, it would be difficult to specify whether the document should be allocated under the class of politics or religion. In addition, expanded topic documents may carry terms that have multiple meanings due to different contexts and may appear a number of times within a document in different contexts [7].

D. Techniques Employed in Document Classification

Some of the methods that are adopted for document classification includes:

Expedition maximization, Naïve Bayes classifier, Support Vector Machine, Decision Trees, Neural Network, etc.

Some of the applications that make use of the above techniques for document classification is listed below:

- E-mail routing: Directing and redirecting of an e-mail to a general address, to a specific address or mailbox relying on the topic of the e-mail.
- Language identification: Automatically ascertain the language of a text. It can be handy in many use cases, one of which could be the direction in which the language should be processed. For instance, Hebrew and Arabic are processed from right to left. And the contrary, when processing the English language. This

concept can then be applied along with language identification in accurately processing text in any language.

- Readability assessment: Automated determination of the state of readability of any document is for a group of persons of a particular age bracket.
- Sentiment analysis: Ascertain the biasness of a speaker due to the content of the document [9].

E. Classification Algorithms

a) Naïve Bayes

The Naïve Bayes Classifier technique [10] is relied on the Bayesian theorem and is particularly suited for inputs which have high dimension. In spite of its simplicity, it can often outperform more sophisticated classification methods. The Naive Bayes continuous works similarly with the continuous variable as input [11].

b) Multilayer Perceptron

It is the most known network architecture in the world today. The units each perform a biased weighted sum of their inputs and pass this activation level through a transfer function to produce their output. The units are arranged in a layered feed-forward topology. The network has a simple input-output model with weights and thresholds. Such networks can model functions of almost arbitrary complexity, with the number of layers, and the number of units in each layer, determining the function complexity. The important issues in Multilayer Perceptron are the design specification of the number of hidden layers and the number of units in these layers [11].

c) Random Forest Tree (RND Tree)

A Random Tree consists of a set of integrated, simple tree forecaster, each having the ability to produce a response when given a set of forecasted values. For classification problems, this response takes the form of a class membership, which associates or classifies a set of autonomous forecaster values with one of the categories available in the dependent variable. For regression problems, the tree feedback is estimated for the dependent variable given by the forecasters [11]. A recent study estimated the neonatal levels in Nigeria using Random Forest Regression Model [12].

III. METHODOLOGY

Constructive Research Methodology

The constructive research methodology was adopted in this research because it targets resolving practical issues as well as generating theoretical contributions that are academically acceptable. This term construct points to an important plus being created, which includes: a framework, algorithm, model, theory, and software. Acquiring expanse knowledge of the problem domain and associated theories could form a strong foundation that may enhance the construction of the desired solution.

IV. USE CASE DIAGRAM FOR THE PROPOSED SYSTEM

The Use Case Diagram for the proposed system is a representation of the relationship and interactions between the actor (user) and the system when it is fully developed.

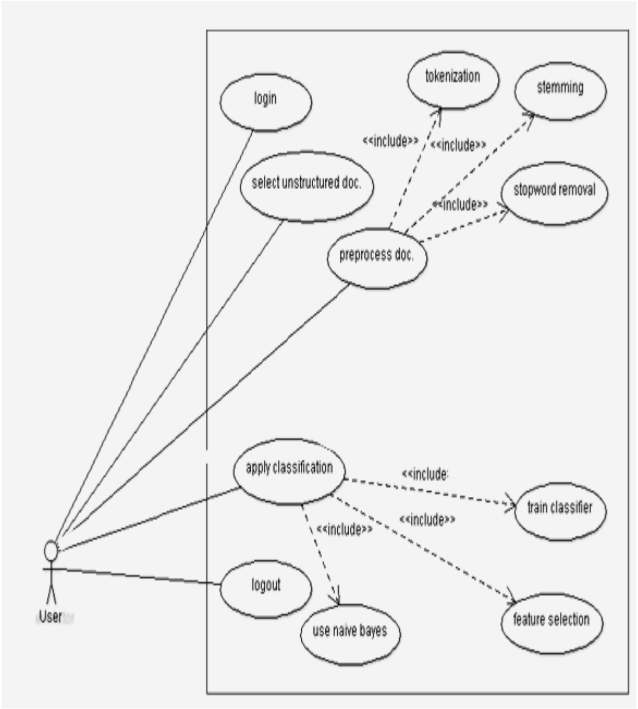


Fig. 1 Use Case Diagram for the proposed system

V. SYSTEM ARCHITECTURE

The architectural design of the proposed system is the very initial step in the modeling of the software development process. It functions as the major connection between the design and requirements engineering, as it always indicates the main structural parts in the system and the relationships between them. It is a high-level representation of the proposed system.

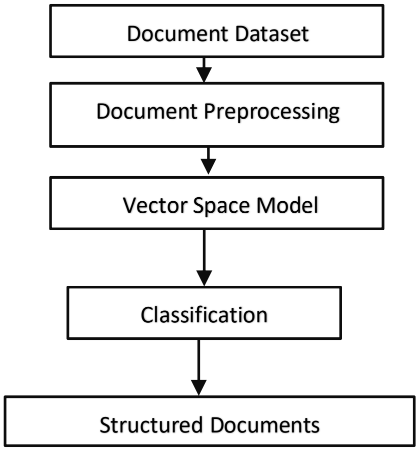


Fig. 2 Proposed System Architecture

The proposed system accepts text documents (unstructured data) as its input dataset. The dataset undergoes a Pre-processing Phase:

- Tokenization involves the process of breaking down words/characters into smaller parts or crunches while the security and integrity of the word are maintained.
- Stemming involves returning words to their root or morpheme state (by either removing their prefix or suffix)
- Stop word removal involves removing words that would appear to be less important in selecting a document that would be compatible with meeting a user's prerequisite; these words are completely expelled from the vocabulary. This technique increments the effectiveness and efficiency of the result of classification documents. Examples of the stop words are a, is, that, those, then, the, when, etc.

The output of the pre-processing phase is a Vector Space Model (Weighted Matrix). The Vector Space Model of text data can be regarded as a word-by-document matrix, whose rows are the words and columns are document vectors, where each entry W_j depicts the weight of word ***i*** in the document j. The weight W_j can be determined in many ways. The frequency computes the number of occurrences of a term t_i in the document j.

$$t_{ij}=f_{ij}=\textit{frequency of the word in document j}$$

This phase is concluded by creating a weight matrix, and this becomes the input to the Clustering Component (Phase)

$$\begin{pmatrix} T1 & & T1 \cdots & & Ti \\ D1 & w11 & w12 \ldots w1i & & c1 \\ D2 & w21 & w22 \cdots w2i & & c2 \\ : & : & : & : & : \\ Dj & wj1 & wj2 \ldots wji & & ck \end{pmatrix}$$

Fig. 3 Vector Space Model (Weighted Matrix)

Classification Component

When pre-processing has been concluded, the system creates the vector space model, which serves as an input for the classification module, the details of the classification component are shown in Figure 3.

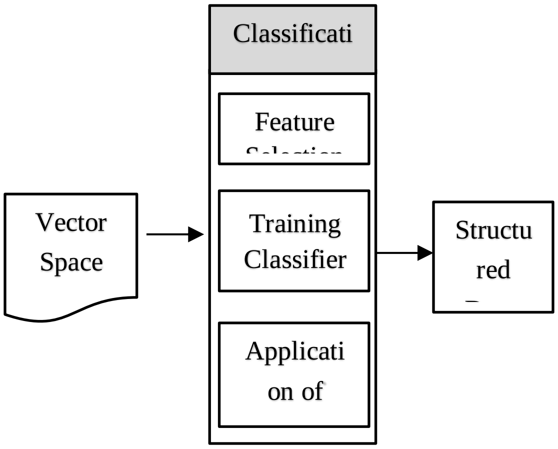


Fig. 4 Details of the Classification Component

a) Feature Selection

Apart from eliminating the stop words and substituting each term by its stem in the pre-processing phase, the occurrences of words in a weight matrix made in the vector space model are still very large. Thus, the feature selection phase is used for the reduction of the dimensionality of the feature set by eliminating the unimportant features. The aim of this phase is to enhance the efficiency of classification correctness and to minimize the computational needs. Feature selection is carried out by retaining the features with the highest score agreeing to the features' relevance [13]. The most popularly applied technics for features' rating include Information Gain, Chi-square statistic, Mutual Information, and latent semantic analysis.

b) Training Classifier

Training classifier is the major part of the text classification process. The function of this phase is the construction of a classifier or bring forth a model by training it, and Training is done by applying an earlier defined document (sets of documents already labeled) that will be employed to classify unlabeled documents. The Multinomial naive Bayes classifier is chosen to construct the classifier in our model.

Given document d fixed set of classes C = {c1, c2, …, cn}

A training set of m documents that we have pre-determined to belong to a specific class, we train our classifier using the training set and result in a learned classifier. We can then use this learned classifier to classify new documents.

Notation: we use Y(d) = C to represent our classifier, where Y () is the classifier, d is the document, and c is the class we assigned to the document.

c) Application of Naïve Bayes

This is the part of the classification component responsible for assigning an unlabeled document to the correct class of that document using Naïve Bayes. For determining which class that document d is assigned to, it's required to calculate the probability of assigning that document to each class and selects the class with the highest probability. Below are the steps are taken to apply the Naïve Bayes:
Calculate prior probabilities. These are the probability of a document being in a specific category from the given set of documents.

P(Category) = (No. of documents classified into the category) divided by (Total number of documents).

After that, Calculate the Likelihood. The likelihood is the conditional probability of a word occurring in a document, given that the document belongs to a particular category.

P(Word/Category) = (Number of occurrences of the word in all the documents from a category+1) divided by (All the words in every document from a category + Total number of unique words in all the documents)

Then, Calculate P(Category/Document) = P(Category) * P (Word 1 /Category) * P (Word 2 /Category) * P (Word 3 /Category) …* P (Word n /Category).

Therefore, the most probable category for a document to fall into is the one with the highest probability among its peers.

VI. SYSTEM REQUIREMENT / SET UP

The software (EADCC) is a web-based application and was built using HTML 5, CSS 3, and JavaScript (ECMA 6) as its front end, while PHP (7.2.3) was used to handle the backend.

Google Chrome browser was used during the testing phase of the software development.

The system was tested with a dataset of 737 BBC sports text documents having 5 natural classes from http://mlg.ucd.ie/datasets/bbc.html.

BBC Sports Dataset:

Documents: 737, Terms: 4613, Natural Classes: 5 (athletics, cricket, football, rugby, tennis).

The details of the dataset are given in Table 1 below.

Table 1. Dataset of BBC Sport

Natural Classes	Number of Documents	Total Words	Total Unique Words
Athletics	101	1018	458
Cricket	124	1032	287
Football	265	997	389
Rugby	147	977	531
Tennis	100	589	162
Total	737	4613	1827

Classifier Training

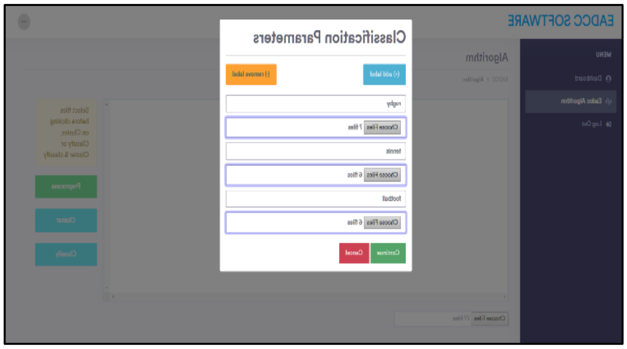


Fig. 5 Classifier Training page

Classification Output

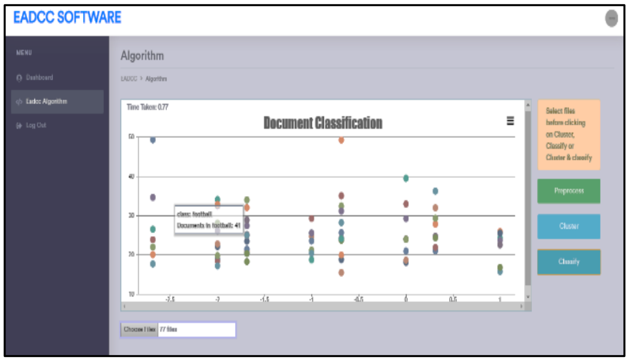
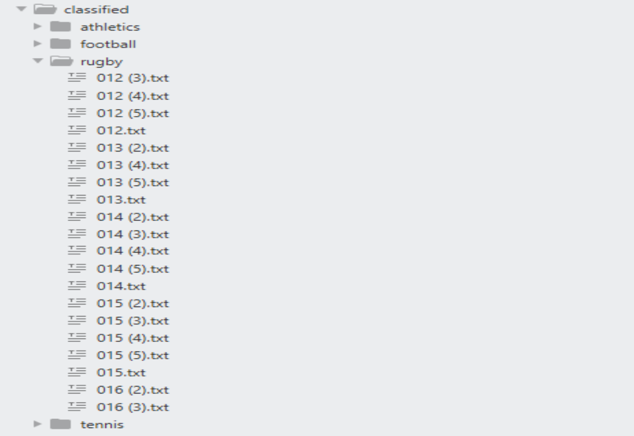


Fig. 6 Classification Sample output page (Scattered Plot)

Documents Classified into Folders



```
▾ classified
  ▸ athletics
  ▸ football
  ▾ rugby
      012 (3).txt
      012 (4).txt
      012 (5).txt
      012.txt
      013 (2).txt
      013 (4).txt
      013 (5).txt
      013.txt
      014 (2).txt
      014 (3).txt
      014 (4).txt
      014 (5).txt
      014.txt
      015 (2).txt
      015 (3).txt
      015 (4).txt
      015 (5).txt
      015.txt
      016 (2).txt
      016 (3).txt
  ▸ tennis
```

Fig. 7 Documents Classified into Folders

VII. RESULT AND DISCUSSION

The intended system was built with the appropriation of the Multinomial Naïve Bayes Classifier method, which was tested with a dataset from the BBC sports repository. The Dataset were absolutely text documents that were upload into the system. It was pre-processed and finally subjected to classification. The number of classes expected is derived from the number of classifiers derived during the training phase. Figure 6 presents a scattered plot which is the output of the developed classification system showing the various classes of documents, and each dot represents a document with some overlapping others. The overlap is a result of their degree of similarity. The system also displays the various class folders, their labels, and text documents in that particular class labeled folder, as shown in Figure 7.

Table 2 displays the various classes and the number of documents in them; this is demonstrated with the aid of a column chart in figure 8. Reading from the chart, the class with the highest number of documents was "Cricket," and "Rugby" had the least number of a text document in its class.

Table 2. Number of Documents Classified in each Class.

Class label	number of documents
Cricket	101
Football	76
Rugby	72
Tennis	93

Finally, Table 3 shows the time performance and efficiency of our developed system when compared to three (3) other existing systems. Our developed system classified a thousand text documents at the least time of 197 seconds; this shows a significant performance and efficiency in the developed system with reference to classification time. This was also graphically displayed on a line graph in Figure 8.

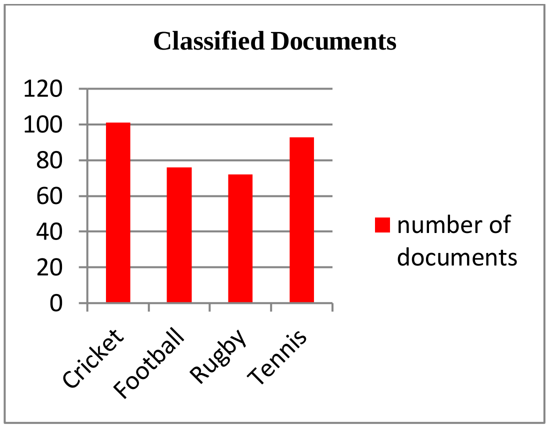


Fig. 8 Column Chart Displays the number of Documents Classified in each Class label in the Developed Systems in table 2

Table 3. Classification Time Comparison of Systems

Systems	Time Taken To Classify 1000 Documents (sec.)
System 1 (Jyotismita Goswami, 2015)	660
System 2 (Anna Huang,2017)	720
System 3 (Santra & Josephine, 2012)	960
New System	602

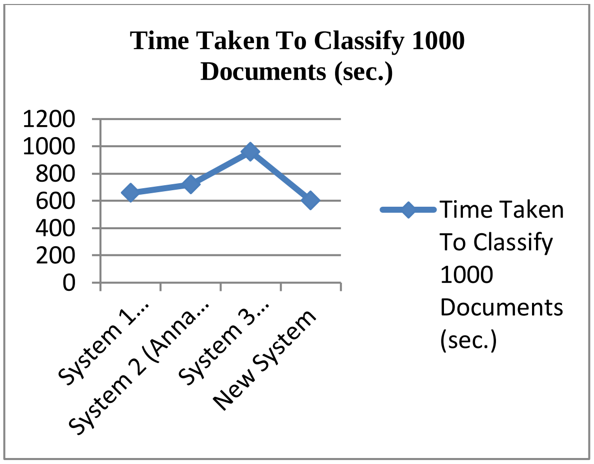


Fig. 9 Line Graph Displays Classification Time Comparison of Systems in Table 3

VIII. CONCLUSION

In view of the fact that unstructured data account for about 70% of relevant information, the need to classify this information for a more effective and efficient means of extraction, storage, and accessibility to users. In Nigeria, for instance, like other nations of the world, a huge volume of this information that could support learning, research, and the development of the country are locked up in piles of text documents that are unstructured. Information in this unstructured form is difficult to truly store and retrieve when required for use. The application of Multinomial Naïve Bayes Algorithms to this research work has made possible the production of a system with an efficient text document classification potential. This makes it easier for unstructured text documents to be classified based on their similarities and stored for easy access to relevant information contained in them.

REFERENCES

[1] Chakraborty, G., Pagolu, M., Text Mining and Analysis, Practical Methods, Examples and Case studies using SAS. SAS Institute Inc. Press, Cary, North Carolina, USA., (2014).

[2] Goutam Chakraborty., Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining. Department of Marketing, Spears School of Business, Oklahoma State University., (2015).

[3] Praveen, P. & Rama, B., A k-means Clustering Algorithm on Numeric Data. International Journal of Pure and Applied Mathematics 117(7)(2017)157-164. ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version).

[4] Fore, N. K., A Contrast Pattern-Based Clustering Algorithm for Categorical Data, Wright State University Core Scholar., (2010).

[5] Jyotismita, G., A comparative Study on clustering and classification Algorithms, International Journal of Scientific and Applied Science (IJSEAS) 1(3)(2015) 70-177

[6] Fredrick, J. & Leonardo S., Data Clustering, its application, and benefits, Semantic Scholar, (2017).

[7] Russell, P., Jay, C., Trishank, K., & Lakshminarayanan S., Document Classification for Focused Topics. International Journal of Computer Applications, 31(5)(2010).

[8] Anna, J. K., Data Clustering: 50 Years Beyond K-Means, King-Sun Fu Prize lecture delivered at the 19th International Conference on Pattern Recognition (ICPR), Tampa, FL, (2008).

[9] Quan, Y., Gao, C., & Nadia, M., Enhancing Naive Bayes with Various Smoothing Methods for Short Text Classification, (2013).

[10] Phimphaka, T., & Sudsanguan, N., Incremental Adaptive Spam Mail Filtering Using Naïve Bayesian Classification, 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking, and Parallel/Distributed Computing. Materials Research, (2011) 171-172,543-546.

[11] Shomona, G. J., & Geetha, R., Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multiclass Categorization of Breast Tissue Data, International Journal of Computer Applications,32(7)(2011) 201-213.

[12] Managwu, C., Matthias, D. and Nwiabu N., Random Forest Regression Model for Estimation of Neonatal Levels in Nigeria. SSRG International Journal of Computer Science and Engineering, 8(20)(2020) 1-4.

[13] Thaoroijam, K., A Study on Document Classification using Machine Learning Techniques. International Journal of Computer Science 11(1)(2014) 165-172.