Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 10, Issue 3, pp: (44-50), Month: September - December 2023, Available at: www.noveltyjournals.com

Big Data Analysis for Unstructured Data in Nigeria Court System

C. Aloy-Okwelle[1], J. Palimote[2], O.P Nweke[3]

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

DOI: https://doi.org/10.5281/zenodo.10082510

Published Date: 08-November-2023

Abstract: **Big Data Analysis for unstructured data involves analyzing and processing large amounts of unstructured data, such as text, images, and audio, to extract meaningful insights and knowledge. Techniques used in big data analysis for unstructured data include Natural Language Processing (NLP), Computer Vision, and Speech Recognition. Big data analysis can be useful in the Nigerian court system for analyzing unstructured data, such as legal documents, witness statements, and court transcripts. The goal would be to identify patterns and relationships within the data that can help make more informed decisions, improve processes, and increase the efficiency of the court system. This paper presents an improved Hybrid model for legal case document classification. The system starts by collecting legal case documents from an online domain. The collected documents are in pdf format. The collected pdf files were converted to texts using a pdf miner library in python. The converted texts were used in creating tables using the pandas library. After the creation of the dataset table, the dataset was pre-processed by removing Nan values, and non-alphanumeric values, and also performing tokenization. The tokenized data was then passed into principal component analysis for the selection of important features. The selected features were then used in training an LSTM model for the classification of the legal case documents. The result of the LSTM is outstanding, having an accuracy of 98% for training. The model was deployed to the web, for easy execution, testing, and assessment.**

Keywords: **Big Data, Legal Case, Principal Component Analysis, Long Short-Term Memory, Python flask.**

I. INTRODUCTION

Big data is a form of data with increased volume, that is difficult to analyze, process and store with traditional database technologies (Hashem *et al.* 2015). This data could be either structured, semi-structured, or unstructured. Modern organizations in the world are data dependent which results to a generation of large volume of data of sizes more than 40 zettabytes (ZB) in 2020(Qi and Tao 2018). Big Data Analysis for unstructured data involves analyzing and processing large amounts of unstructured data, such as text, images, and audio, to extract meaningful insights and knowledge. Techniques used in big data analysis for unstructured data include Natural Language Processing (NLP), Computer Vision, and Speech Recognition. Tools used in this process include Apache Hadoop, Apache Spark, and Apache Storm. The goal of big data analysis for unstructured data is to enable organizations such as judiciary to make informed decisions and gain a competitive advantage by leveraging the vast amount of data generated in today's digital world.

In Nigeria, a simple land court case can drag on for years and can move from the high court to the court of appeal and even up to the Supreme Court. The time interval between moving from one court to another can take up to several months and even years. There are several reasons why these court cases can take several years to complete. These reasons can include the absence of the defendants and/or appellants, and the absence of the judge who is to oversee the case, but most often than not, court cases are delayed because adjournments were sorted by counsels and granted by the judge. Counsels (lawyers and attorneys) usually seek adjournments in order to put their cases in order. Preparing for a court case is not an easy fit as a lot of research has to be done on the part of the attorneys in charge.

Big data analysis can be useful in the Nigerian court system for analyzing unstructured data, such as legal documents, witness statements, and court transcripts. The goal would be to identify patterns and relationships within the data that can

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 10, Issue 3, pp: (44-50), Month: September - December 2023, Available at: www.noveltyjournals.com

help make more informed decisions, improve processes, and increase the efficiency of the court system. This can involve techniques such as text mining, sentiment analysis, and network analysis. However, privacy and ethical considerations must be taken into account when dealing with sensitive legal data.

II. RELATED WORKS

[1] presented "BigBench", a proposal for an end-to-end big data benchmark. The proposal covers a data model addressing the velocity, variety and volume common in big data. Velocity is accomplished by continuous feed into the data store while variety is addressed by including structured, semi-structured and unstructured in the data model. The data model also can scale to large volumes based on as scale factor. They used PDGF as a starting point for their data generator that covers the structured part. PDGF is enhanced to produce the semi-structured and unstructured data. The unstructured component is based on a novel technique they developed that leveraged the Markov chain model. The proposal also provided a comprehensive list of workload queries and sets directions for a novel metric that focused on the different types of processing in big data. Finally, they verified the feasibility and applicability of the proposal by implementing and running it on Teradata Aster DBMS.

[10] in their work proposes a hybrid methodology that enables the integration of structured and unstructured data to support the decision-making process in public security contexts. Classifying and predicting crime in a given area is made easier by the proposed method, which enables actions to be identified based on the results to improve public security. The data was integrated in two primary steps: firstly, they importing and analyzed government-provided structured data, and secondly they ingesting, categorized, and analyzed unstructured data from digital sites like Twitter, and CityCop Based on their analysis, they conducted a series of actions intended to bring improvements to the region by the local police. They obtained an increase in the algorithms' accuracy rate of 80%, indicating that public security organizations can base their actions on the results of the proposed methodology

[2] examined the role of Big Data and Data Revolution in promoting sustainable development in Nigeria, as well the emerging opportunities for statisticians, in this regard. The paper posits that the attainment of the SDGs will be greatly hampered if statisticians do not ask the right questions; access relevant data information and crucially perform deeper analytics around data and information. Statisticians have an important role to play in promoting Nigeria's sustainable development agenda, but only if they become more entrepreneurial; and adequately master and apply the requisite technical and non-technical skills.

[3] identified the sources of assessing big data in Nigeria. They investigated how big data are generated and processed, and identified the problems of generating and processing the assessment of big data in Nigeria. Through purposive sampling technique forty-five experts in education assessment and research were selected. The instruments for data collection were interview and documents. The data collected were analysed using descriptive statistics to answer the five research questions that guided the research.

[5] in their paper presented an efficient method for storing unstructured data and presented an approach for fetching data. They made a Big Data application that gets stream of public tweets from twitter which is latter stored in the HBase using Hadoop cluster. They performed data analysis for data retrieved from HBase by REST calls is the pragmatic approach of this project.

III. DESIGN METHODOLOGY

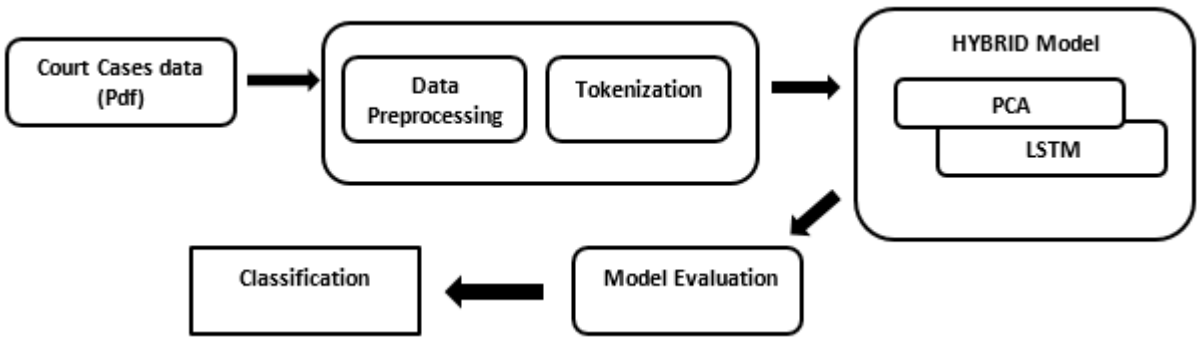

Figure 1: Architectural Design of proposed system

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 10, Issue 3, pp: (44-50), Month: September - December 2023, Available at: www.noveltyjournals.com

Court Cases (PDF): This component illustrates the document collection process, which involved gathering legal case documents from 2001 to 2022. The documents were in PDFs (Portable Document Formats) and HTML (Hypertext Markup Language) formats but were converted into text format (.txt) as required by the existing system.

Data Pre-processing: The data Pre-processing phase has to do with data cleaning and the removal of stopwords. The following are the stages of the pre-processing:

i) Removal of stop-words and non-alphanumeric words: This component illustrates the scanning of the tokens through a file containing stop words and alpha-numeric words. This stage becomes a necessity because these words occur more frequently and are of less importance in documents. Words like "an", "a", "the", etc. are stop words.

ii) Filter Tokens by length: This component illustrates the removal of words or features with a particular character length.

iii) Transform Cases: This component illustrates the conversion of the words or tokens in the text documents into a particular case, upper or lower cases to avoid duplication of the same features in different cases.

Tokenization: This component illustrates the conversion of the legal-case documents from sequences of characters to sequences of tokens (i.e. words, symbols, or phrases). Other par of the tokenization processes involves:

i. **Stemming and Lemmatization:** The aim of stemming is to inflectional words to a common base form. For grammatical reasons, text documents are going to use different forms of a word such as take, taken, and taking. Additionally, there are families of derivationally related words with similar meanings such as democracy, democratic and democratization. In many situations, it seems as if it would be useful to search for one of these words to return documents that contain another word in the set.

Output after stemming, the text becomes:

"appeal, september, company, sisters, nwugo, visit, decease, tradition, drink, order, explore, september, reconcile, wife, decease, refuse, accept, drink, ground, report, matter, church, authority, invite, appeal, sister, church, september, matter, look, deal, church, authorities"

For stemming and lemmatization, nltk.stem imp WordNetLemmatizer libraries were used to achieve those.

Hybrid Model: The hybrid model involves the use of Principal Component analysis for feature selection and the use of Long Short-Term Memory in selecting the most important features. Here is the algorithm for the PCA

Algorithm and Pseudocode for Principal Component Analysis

Step 1: Standardize the dataset.

Step 2: Calculate the covariance matrix for the input features.

Step 3: Calculate the eigenvalues and eigenvectors for the covariance matrix.

Step 4: Sort eigenvalues and their corresponding eigenvectors.

Step 5: Select k eigenvalues and form a matrix of eigenvectors.

Step 6: Perform a transformation of the original matrix.

Pseudocode for PCA

1: procedure PCA

2: Compute dot product matrix: $\mathbf{X}^T\mathbf{X} = \sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})^T(\mathbf{x}_i - \boldsymbol{\mu})$

3: Eigen analysis: $\mathbf{X}^T\mathbf{X} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$

4: Compute eigenvectors: $\mathbf{U} = \mathbf{X}\mathbf{V}\boldsymbol{\Lambda}^{-\frac{1}{2}}$

5: Keep a specific number of first components: $\mathbf{U_d} = [\mathbf{u}_1,\ldots,\mathbf{u}_d]$

6: Compute *d* features: $\mathbf{Y} = \mathbf{U_d}^T\mathbf{X}$

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 10, Issue 3, pp: (44-50), Month: September - December 2023, Available at: www.noveltyjournals.com

IV. EXPERIMENTAL RESULT

In the experiment, legal case documents were used as dataset for building a hybrid model in classifying legal cases. The legal case dataset consists of 6 categories of legal cases which comprises of criminal case, civil case, politics and governments cases, finance cases and land cases, all in pdf formats. The dataset was read into directory using os.listdir() in python. In order to have a better training data, the legal case dataset which was in pdf format was converted in text files using PDFResourceManager, PDFPageInterpreter libraries in python. For pre-processing of the text documents, tokenization which comprises of stopwords, setemming, and conversion of alphabets to lower case was used in extracting textual data from the text document dataset. The tokenized data can be seen in figure 1. A count plot of the tokenized data can be seen in figure 2. The dataset was divided into training and testing data, 80 percent of the dataset was used for training and 20% of the dataset was used for testing. In other to import features from the dataset, principal component analysis (pca) was applied to the dataset. The transformed result of principal component analysis was illustrated using a line plot which can be seen in figure 3. In other to have a better training performance, the dataset was converted to arrays. LabelEncoder was used in converting the category columns to arrays. This can be seen in figure 4. The dataset was then applied to a deep learning model for training. The Deep learning model used here is Long-Short-Term Memory (LSTM) algorithm. This was used in training a hybrid model for legal case documents classification. LSTM was chosen because of it's capability in learning long-term dependencies. The compilation and training process of the Long-Short Term memory algorithm can be seen in figure 5 and 6. The result of the Long Short-term Memory algorithm can be seen in figure 7 and 8. Figure 9 shows a classification report of the model.

	Text_Data	Category	Identifiers
0	vol lrcn adekeye v adesina prince kilani adeke...	Politics_and_Goovernment	election,pt,appeal,v,court
1	vol lrcn adekeye v adesina prince summonu ades...	Politics_and_Goovernment	election,pt,appeal,v,court
2	vol lrcn adekeye v adesinapleaded whether evid...	Politics_and_Goovernment	election,pt,appeal,v,court
3	vol lrcn adekeye v adesinaevidence pleadings w...	Politics_and_Goovernment	election,pt,appeal,v,court
4	vol lrcn adekeye v adesinapleadings whether pa...	Politics_and_Goovernment	election,pt,appeal,v,court
5	vol lrcn adekeye v adesinaissues determination...	Politics_and_Goovernment	election,pt,appeal,v,court
6	vol lrcn adekeye v adesinaalso disputed fact f...	Politics_and_Goovernment	election,pt,appeal,v,court
7	vol lrcn adekeye v adesinaon whether party app...	Politics_and_Goovernment	election,pt,appeal,v,court
8	vol lrcn adekeye v adesinaon effect validly ma...	Politics_and_Goovernment	election,pt,appeal,v,court
9	vol lrcn adekeye v adesinaon whether customary...	Politics_and_Goovernment	election,pt,appeal,v,court

Figure 1: Training data for the first ten rows

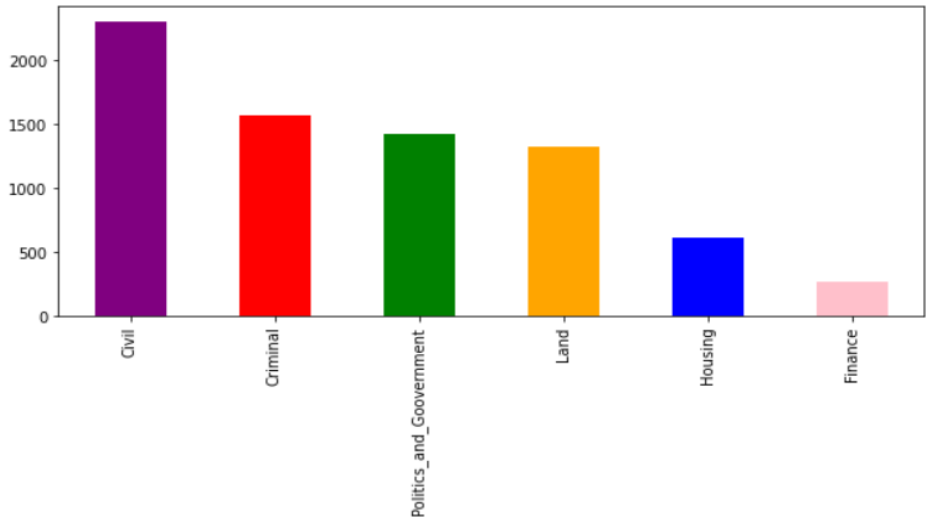

Figure 2: Histogram distribution of the training data

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 10, Issue 3, pp: (44-50), Month: September - December 2023, Available at: www.noveltyjournals.com

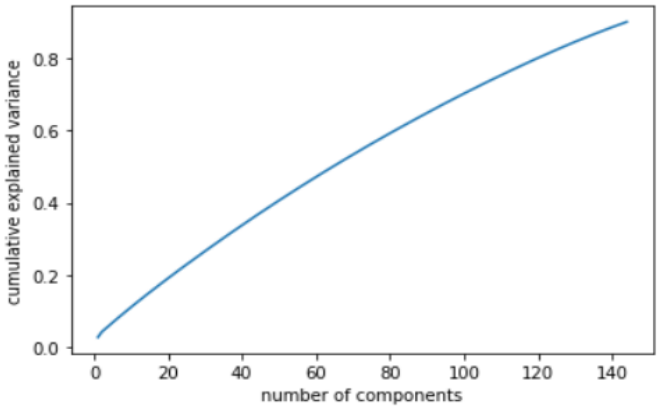

Figure 3: Transformed analysis of Principal Component Analysis

```
Shape of label tensor: (7485, 6)
[[1 0 0 0 0 0]
 [1 0 0 0 0 0]
 [1 0 0 0 0 0]
 ...
 [0 0 0 0 0 1]
 [0 0 0 0 0 1]
 [0 0 0 0 0 1]]
```

Figure 4: Transformed result of the Category column

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 250, 100)          5000000
_________________________________________________________________
spatial_dropout1d (SpatialDr (None, 250, 100)          0
_________________________________________________________________
lstm (LSTM)                  (None, 100)               80400
_________________________________________________________________
dense (Dense)                (None, 6)                 606
=================================================================
Total params: 5,081,006
Trainable params: 5,081,006
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 5: Summary of the LSTM model

```
Epoch 1/8
190/190 [==============================] - 125s 625ms/step - loss: 1.5345 - accuracy: 0.3862 - val_loss: 1.3825 - val_accuracy:
0.3576
Epoch 2/8
190/190 [==============================] - 160s 843ms/step - loss: 0.9315 - accuracy: 0.6895 - val_loss: 0.7508 - val_accuracy:
0.7552
Epoch 3/8
190/190 [==============================] - 163s 858ms/step - loss: 0.5321 - accuracy: 0.8174 - val_loss: 0.7227 - val_accuracy:
0.7730
Epoch 4/8
190/190 [==============================] - 158s 832ms/step - loss: 0.3305 - accuracy: 0.8882 - val_loss: 0.6547 - val_accuracy:
0.7893
Epoch 5/8
190/190 [==============================] - 164s 864ms/step - loss: 0.2163 - accuracy: 0.9357 - val_loss: 0.9694 - val_accuracy:
0.6454
Epoch 6/8
190/190 [==============================] - 164s 863ms/step - loss: 0.1872 - accuracy: 0.9490 - val_loss: 0.5492 - val_accuracy:
0.8398
Epoch 7/8
190/190 [==============================] - 161s 849ms/step - loss: 0.0446 - accuracy: 0.9888 - val_loss: 0.5430 - val_accuracy:
0.8501
Epoch 8/8
190/190 [==============================] - 164s 861ms/step - loss: 0.0713 - accuracy: 0.9814 - val_loss: 1.0415 - val_accuracy:
0.6662
```

Figure 6: Training steps of the LSTM model

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 10, Issue 3, pp: (44-50), Month: September - December 2023, Available at: www.noveltyjournals.com

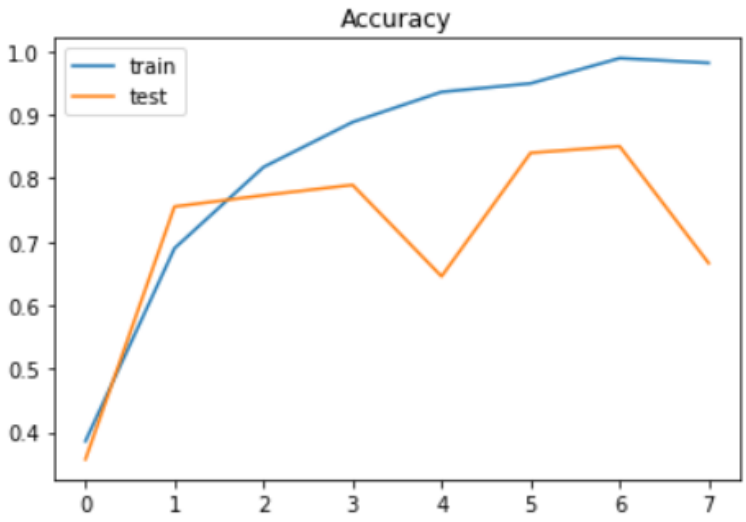

Figure 7: Accuracy of the trained Model

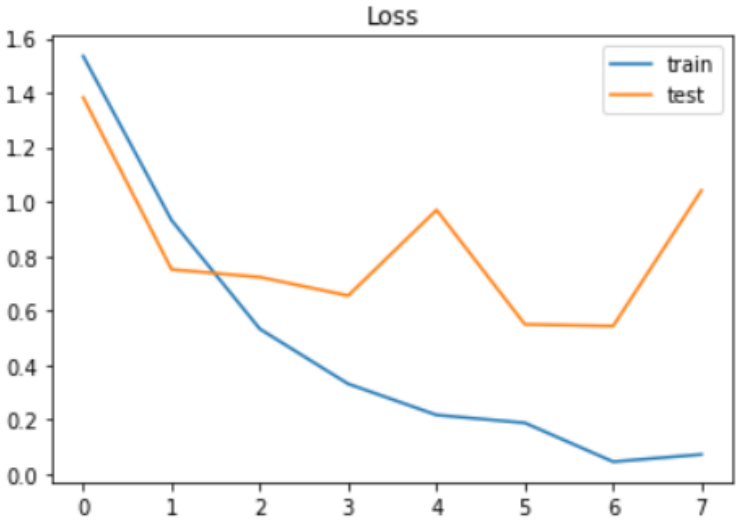

Figure 8: Loss value of the trained Model

	precision	recall	f1-score	support
0	0.94	0.73	0.82	903
1	0.91	0.97	0.94	431
2	0.29	0.85	0.44	27
3	0.48	0.99	0.64	88
4	0.84	0.84	0.84	395
5	0.87	0.92	0.89	402
accuracy			0.84	2246
macro avg	0.72	0.88	0.76	2246
weighted avg	0.88	0.84	0.85	2246

Figure 9: Classification Report

V. DISCUSSION OF RESULT

From the experiment conducted, the figure 1 shows the pre-processed data after passing through the stage of tokenization. This shows that all texts in the pdf files (legal case documents) has been extracted successfully and broken down into tokens, and it also shows that every non alphabetic characters has been removed successfully from the data. Figure 2 shows the numbers of the six categories of the text document data. Therefore, making the training data to be a total of 8600. Figure 3 shows a line graph representation of the transformed result using principal component analysis (Pca) for feature extraction. It visualizes the relationship between cumulative variance explained and number of components. It shows that the top two components can explain 90% of the original variance. Figure 5 shows a total trainable parameters of the Long Short-Term Memory algorithm. Which is about 5081006 parameters with 6 output. The 6 output indicates the six categories of legal cases (civil, criminal, finance, land, politics and governments, and housing). Figure 6 shows the performance level of the model at each training steps. The performance values comprise of accuracy and loss values for both training and testing data. Figure 7 and 8 shows a line plot, plotted against training steps and loss for both training and testing data. Figure 9 shows the classification report of the model on the test data. The evaluation performance shows that the Long Term-Short Memory algorithm had a training accuracy of 98.14%, test accuracy of 84%, precision score of 94%, f1-score of 82% and a loss value of 1.6%.

VI. CONCLUSION

This paper presents an improved Hybrid model for legal case document classification. The system begins by collecting legal case documents from an online domain. The collected documents are in pdf format. The collected pdf files were converted to texts using a pdf miner library in python. The converted texts were used in creating tables using the pandas library. After the creation of the dataset table, the dataset was pre-processed by removing Nan values, and non-alphanumeric values, and also performing tokenization. The tokenized data was then passed into principal component analysis for the selection of important features. The selected features were used in training an LSTM model for the classification of the legal case documents. The result of the LSTM is outstanding, having an accuracy of 98% for training. The model was deployed to the web, for easy execution, testing, and assessment.

REFERENCES

[1] Ahmad G., Tilmann R., Minqing H., Francois R., Meikel P., Alain C. & Hans-Arno J. (2013). “BigBench: Towards an Industry Standard Benchmark for Big Data Analytics” New York, New York, USA. SIGMOD’13, Copyright 2013 ACM 978-1-4503-2037-5/13/06.

[2] Ojijo O. & Fatima U. (2019). Harnessing Big Data for Sustainable Development in Nigeria. Published by Canadian Center of Science and Education. ISSN 1913-9063 E-ISSN 1913-9071. Vol. 12, 3.

[3] Nkechi P. E, Martins N. E., & Lydia I. E. (2020). Assessment big data in Nigeria: Identification, generation and processing in the opinion of the experts. International Journal of Evaluation and Research in Education (IJERE). Vol. 9, No. 2, June 2020, pp. 345~351 ISSN: 2252-8822, DOI: 10.11591/ijere. v9i2.20339.

[4] Hashem, I. A. T 2015). "The rise of “big data” on cloud computing: Review and open research issues." 47: 98-115.

[5] Qi, Q. & F. J. I. A. Tao (2018). "Digital twin and big data towards smart manufacturing and industry 4.0: 360-degree comparison." 6: 3585-3593.

[6] Das T.K. & Mohan K. (2013). BIG Data Analytics: A Framework for Unstructured Data Analysis. International Journal of Engineering and Technology (IJET). Vol 5 No 1. ISSN: 0975-4024.

[7] TechTarget, 'Big data'. Available at (https://searchdatamanagement.techtarget.com/definition/big-data)) Accessed 1 April 2019.

[8] The Economist, 'The world's most valuable resource is no longer oil, but data (*The Economist* , 6 May 2017) Available at(https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data) Accessed 1 April 2019.

[9] Obi, Uche Val (2020) *An Extensive Article on Data Privacy and Data Protection Law in Nigeria.* Retrieved from https://inplp.com/latest-news/article/an-extensive-article-on-data-privacy-and-data-protection-law-in-nigeria/

[10] Turet, J. G., & Costa, A. P. C. S. (2022). Hybrid methodology for analysis of structured and unstructured data to support decision-making in public security. *Data & Knowledge Engineering*, *141*, 102056.