An Exploratory Data Analytics of Multivariate Observational Metrics on Generative AI

Wilson Ahiara[1], Temitope Abioye[2], Tochukwu Chiagunye[3], Taiwo Olaleye[4]

[1.3] *Department of Computer Engineering, Micahel Okpara University of Agriculture Umudike, Nigeria*
[2] *Computer Science and Information Technology Department, Bells University of Technology, Ota, Nigeria*
[4] *Department of Computer Science, Federal University of Agriculture. Abeokuta*

Abstract

The level of traction recently gained by the family of generative artificial intelligence tools has amplified calls for an inclusive training set of the language models in order to ensure the tools serves the purpose of less-popular languages as the English language. It is therefore important to ascertain the level of interest in the intelligent tools by societies across the globe. The purpose of this study is to explore the interest of Nigerians in the generative AI tools, as well as the potential socioeconomic factors that may influence their awareness of its use cases. A multivariate metrics containing both socioeconomic and demographic data and as well as Nigeria's web analytics metrics from Google Trends are used for the study. An exploratory data analysis is implemented on the data attributes using python programming to infer actionable insights. Experimental result showed that there is no positive correlation between the literacy level, poverty index, population distribution of Nigerians and their awareness and interest in generative AI tools. The study also revealed that the most popular keywords related to generative AI tools in Nigeria were "Generative AI" and "ChatGPT" even in the northern region with lower literacy level, just as only Lagos returned inquiries on language models.

Keywords 1

Artificial Intelligence, Generative AI, Language Model, ChatGPT, Nigeria, Data Science

1. Introduction

People all across the world have been fascinated by and interested in artificial intelligence (AI). However, different cultures and geographical areas may have varying level of interest in and involvement with the technology. As a branch of artificial intelligence that deals with the production of original content, such as literature, images, and music, generative AI (gAI) has attracted a lot of attention in recent years. Investigating the degree of interest and involvement in this technology over time across cultures and countries is crucial given its potential effects on creative sectors and the larger society. In recent past, both Language Models (LM) and gAI have gained traction with varying use case perceptions. Indeed, an AI algorithm known as a LM is trained to forecast the likelihood of the subsequent word in a string of words [1]. A LM's training data often comprise of text from sources like books, papers, and websites. In order to produce convincing and cohesive text, a LM learns the patterns and structures of language from the training data. However, any AI system that can produce new content, such as text, graphics, or music, depending on a collection of input data is referred to as a gAI [2]. One use case of gAI is the LM, but there are other kinds of gAI systems that work with other kinds of input data. A gAI is therefore a broader context of the family of human language-based AI use cases. These AI use cases, especially gAI, have a complex and multifaceted impact on digital literacy across

MoMLeT+DS 2023: 5th International Workshop on Modern Machine Learning Technologies and Data Science, June 3, 2023, Lviv, Ukraine
EMAIL: ahiara.wilson@mouau.edu.ng (W. Ahiara); elizatope_2005@yahoo.com (T. E. Ogunbiyi); tchiagunye@yahoo.com (T. T. Chiagunye); agsobaolaleyetaiwo@gmail.com (O. T. Olaleye)
ORCID: 0000-0002-7220-1835 (W. Ahiara); 0000-0002-3373-396X (T. E. Ogunbiyi); 0000-0002-0622-3822 (T. T. Chiagunye); 0000-0001-6222-7575 (O. T. Olaleye)

cultures and regions. These technologies provides potent tools for enhancing communication, education, and access to information [3]. Functionalities of chatbots and virtual assistants that can interact with users in natural language, answering questions and offering content-generation assistants, are some of the important utilitarian values. There are also concerns that these technologies may worsen existing inequalities and biases in the global society [4]. A LM trained on a textual data written in English or other languages spoken by more privileged populations may not be as effective at generating text in other languages or dialects that are less well-represented in the training data. Whatever improvement that can scale the representations of less-represented languages in its training set should stem from the popularity of the AI tools in affected world regions. The aim of this study therefore is to analyze and compare the level of awareness and interest in gAI, over time, in different cultures and regions that makes up Nigeria. The adoption of Nigeria serves the tripod purpose of being the most populated country in Africa [5], the largest economy in Africa [6], and the world's largest black nation [7]. Nigeria's socioeconomic and demographic data are acquired for this study, alongside its web analytics metrics on gAI. Thereby, the problem statement is to investigate the relationship between demographic, socioeconomic, and digital indicators of Nigeria ethnicities as it relates to their level of gAI awareness and interest. By analyzing these data sets, the observational study would infer insights into how factors such as population, poverty index, literacy level, and digital connectivity metrics relates to Nigerian gAI utility. The rest of the paper is structured in the following ways; section II discusses existing literature while section III explained the EDA methodology. Experimental result is discussed in section IV and the study is concluded with recommendations in section VI.

2. Review of Existing Studies

Artificial intelligence is not new to the global information technology lexicon, as well as its use cases of machine learning, expert system, natural language processing, fuzzy logic, etc. The sudden interest in data science is not also completely lost on the less-developed societies across in this generation of Internet of Things, and an improved discussions on the importance of smart cities. These tools and their unprecedented functionalities continue to shape global discuss in the academia and industries. Researchers believe their perceived ease of use and usefulness has helped gain traction in academia [8]. The entry of ChatGPT by OpenAI in November 2022 is believed to be a watershed in the developmental history of generative artificial intelligence [9]. Interchangeably described as either a Generative Artificial Intelligence (gAI) [10], Language Model (LM) [11], or Generic Artificial Intelligence (GAI) [12], ChatGPT and its likes has triggered research interest in recent past. The gAI models are trained to create new data that is equivalent in structure and content to existing data [13], in contrast to standard AI models trained to classify or recognize current data [14]. For software engineering, the tool has been used to create test cases in the testing phase of the software development life cycle, and as well as producing artificial data for training machine learning models [11]. Other studies on ChatGPT include a study that investigated the sentiment about the gAI tool. The study identified concerns on plagiarism, referencing, citation, and literature reviews as expressed by reviewers of gAI [1]. The appropriateness of Google Trend data for nowcasting the growth of a new concept was established by Kohnsand Bhattachrjee [15]. The study revealed that a high dimensional collection of search keywords on Google Trend could return a reliable perspective on the future of a search term at its formative stage like the ChatGPT. The Google trend data was likewise employed in [16] to investigate the awareness and interest of inbound tourists. The web analytics data was employed to find terms associated with foreign travels into China thereby generating monthly search frequency for each of the keywords.

3. Methods and Materials

This study involves two phases of data acquisition and the exploratory data analytics. Observational metrics on Nigeria, referring to quantitative indicators including population, poverty index, literacy level, and internet subscription, are used in this study. The data attributes of each state of the country include their Multidimensional Poverty Index (MDI), the Population Size (POP), the Internet Subscription rate (INT_SUB), their Male gender literacy level (M_LIT_L), the Female gender literacy level (F_LIT_L), their actual literacy level (LIT_RATE); their search rate for keywords 'chatgpt' (CHATGPT), 'generative AI' (gAI), 'artificial intelligence' (AI), and 'language model' (LM). These data are acquired from the National Bureau of Statistics [17] and the Nigeria Communication Commission [18], and represent observations within the last one year. The observational data is used with web analytics data about Nigeria from GT, within the last 12 months (April 2022 – April 2023). Search words including 'generative AI', 'chatgpt', 'artificial intelligence', and 'language model' are used to query GT real time. The EDA, implemented using Python programming, helps to condense the observational metrics towards gaining insights on the awareness and interest of Nigerians and as well as the factors that influences their gAI awareness or interest. The EDA is an efficient data science tool for data mining functionalities [19] and the following tools will be implemented in this study.

3.1 Interquartile Range (IQR)

A dataset's IQR is a measure of variability that sheds light on the distribution of the middle 50% of the data [20]. The IQR is specifically the variation between a dataset's third and first quartiles (Q3 and Q1, respectively). Compared to the range or standard deviation, it is a reliable measure of dispersion that is less susceptible to outliers. The amount of variability inside the middle 50% of the data is one insight that may be obtained from the IQR. A lower IQR indicates that the data are closely grouped around the median, whereas a higher IQR indicates that the data are spread out more widely. The IQR can also be used to spot probable outliers in a dataset, which are often identified as observations that are more than 1.5 times the IQR below Q1 or above Q3 in the dataset.

$$Q1 = \{\frac{n+1}{4}\}^{th} \quad (1)$$

depicts the most centered value in the 1st half of the rank-organized dataset;

$$Q3 = \{3\frac{n+1}{4}\}^{th} \quad (2)$$

is the most centered value in the 2nd half of the rank- organized dataset, while Q2 is the median and computed as:

$$Q2 = Q3 - Q1 \quad (3)$$

3.2 Correlation Coefficient

The correlation coefficient is a statistical indicator that displays the strength and direction of the linear relationship between two variables [21]. It is a number between -1 and 1, where a value of -1 implies an idealized absence of correlation between the variables, a value of 1 denotes an idealized presence of correlation, and a value of 0 denotes an idealistic negative correlation. When a variable

rises, the other variable also tends to climb if the correlation coefficient is positive. A low correlation value indicates that the other variable tends to decline when the first variable increases. The magnitude of the correlation coefficient, with larger absolute values indicating stronger correlations, determines the strength of the association. In this study, the correlation coefficient will be computed to investigate the linear relationship between the multivariate metrics. The correlation is computed thus:

$$\text{Correlation} = P = \frac{COV(x,y)}{\sigma X \sigma Y} \quad (4)$$

where x and y are the attributes being investigated for likely positive or negative correlative relationship. The *COV* is the covariance and σ is the standard deviation of the two attributes. The heat plot will be employed to display the correlation coefficient matrix of the multivariate data set of this study. Standard deviation and other measures of dispersion including the mean and median will be computed.

3.3 Standard Deviation

The rate of deviation of variables contained in the multivariate data from the mean is captured by the value of the standard deviation. Evaluating the variability or dispersion of each variable (such as population, poverty index, literacy level, and internet penetration indices) throughout the Nigerian states with respect to their interest in gAI would be instructive for this study.

$$\sigma = \sqrt{\frac{\sum(Xi-\mu)^2}{N}} \quad (5)$$

where σ is the population standard deviation, N is the population size, x_i is each value from the population, and μ is the the population mean

4. Result and Discussion

The eleven (11) attributes represented in the study data, as described earlier in section 3, are for the 36 states and the Federal Capital Territory of Nigeria. The instances describes the socioeconomic and demographic metrics of each of the thirty-seven (37) federating units of the country. The data is further analyzed along the six geo-political zones of the country in order to fully domesticate the findings across Nigeria's multi-ethnic regions. The statistical summary of the data is presented in Table 1 and Table 2 for the entire multivariate metrics with respect to their measures of dispersion as well as the IQR. Figure 1 shows the rate at which Nigerians enquire about the keywords 'chatgpt', 'generative AI'. 'language model', and 'artificial intelligence' within the last 12 months on Google search engine. As could be observed, the line plot summed all search per month for the various keywords. The degree of the keyword search per thirty seven federating units is plotted in Figure 2, while Figure 3 shows the plot of Nigeria's population, multidimensional poverty index, and the literacy level in terms of the geo-political zones. The geo political zone-based plot is likewise presented in Figure 4 for searches on the keywords. The correlation coefficient matrix of the attributes is presented in the heat plot of Figure 5 where the correlation value of each attribute pair is presented across the rows and columns.

Table 1
Summary statistics of Nigeria's multivariate observational metrics on gAI

S/N	MPI	M_LIT_L	F_LIT_L	POP	LIT_RATE
mean	0.263973	78.36757	69.43889	5.36E+06	50.33784
std	0.096425	19.69605	27.88117	2.54E+06	19.54937
min	0.095	39.8	20.1	2.39E+06	14.5
Q1 (25%)	0.185	62.6	43.975	3.84E+06	33.9
Q2 (50%)	0.289	88.3	82.35	4.78E+06	49.3
Q3 (75%)	0.328	95	94.625	5.86E+06	65.7
max	0.409	99	98.2	1.43E+07	92.1

Table 2
Summary statistics of Nigeria's multivariate observational metrics on gAI

S/N	INT_SUB	CHATGPT	gAI	AI	LM
mean	3.79E+06	81.45946	0.027027	10.56757	0.027027
std	3.01E+06	25.94181	0.164399	9.833448	0.164399
min	1.02E+06	1	0	0	0
Q1 (25%)	1.95E+06	80	0	0	0
Q2 (50%)	2.82E+06	85	0	14	0
Q3 (75%)	4.28E+06	100	0	18	0
max	1.75E+07	100	1	28	1

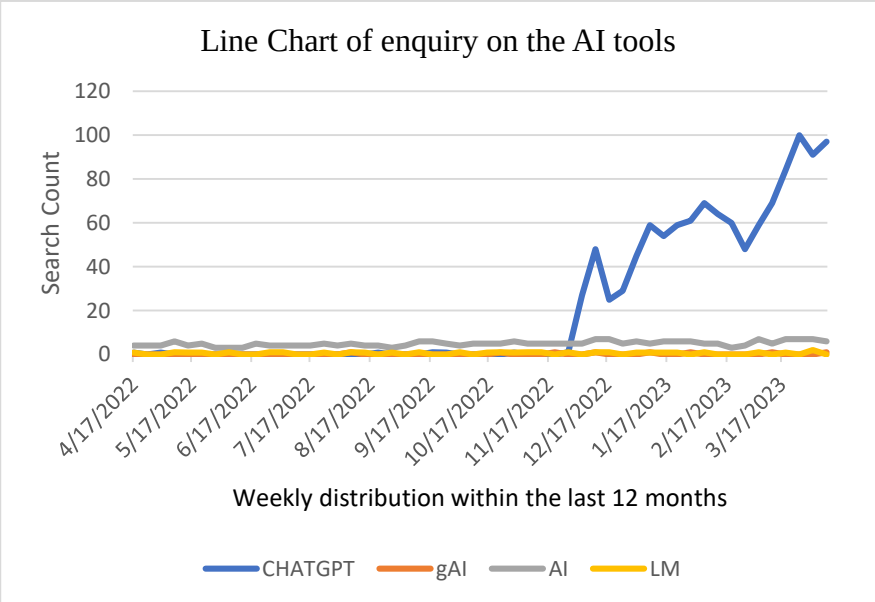

Figure 1: Line chart of Nigeria searches on Google Trend for 'generative AI', 'artificial intelligence', 'lamguage model', and 'chatgpt' per week within the last 12 months.

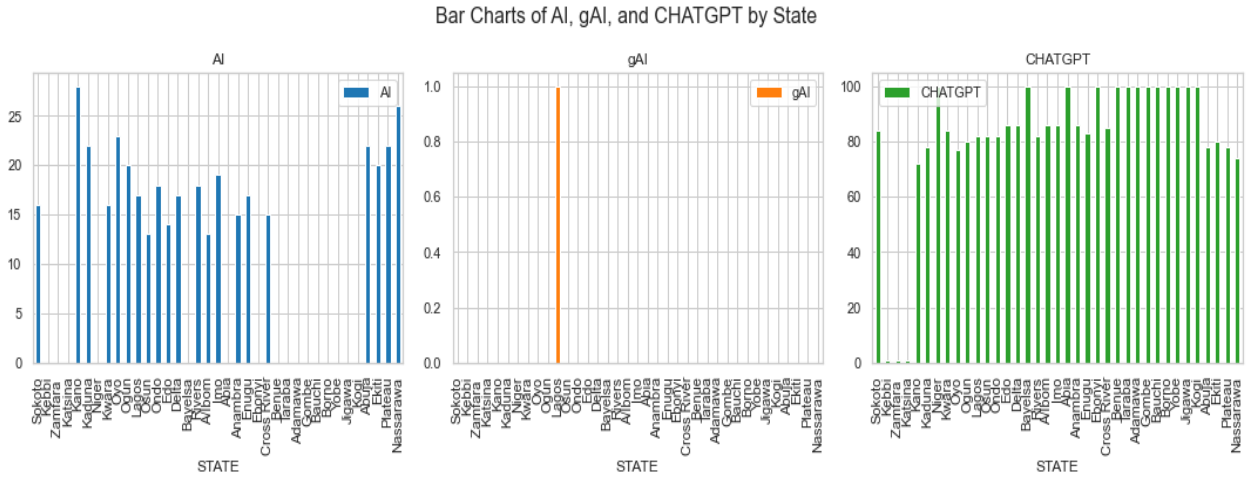

Figure 2: Stacked bar plots of Nigeria searches on Google Trend for 'generative AI', 'artificial intelligence', and 'chatgpt' keywords by state.

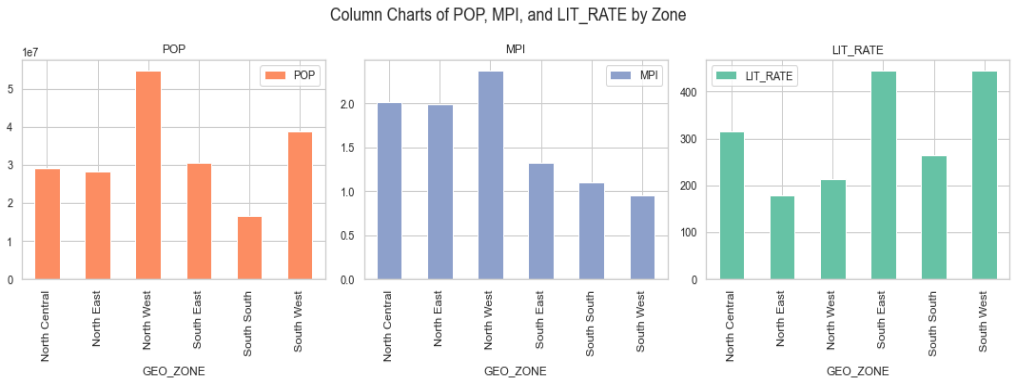

Figure 3: Stacked bar plots of Nigeria's population, multidimensional poverty index, and the literacy level by geo-political zones.

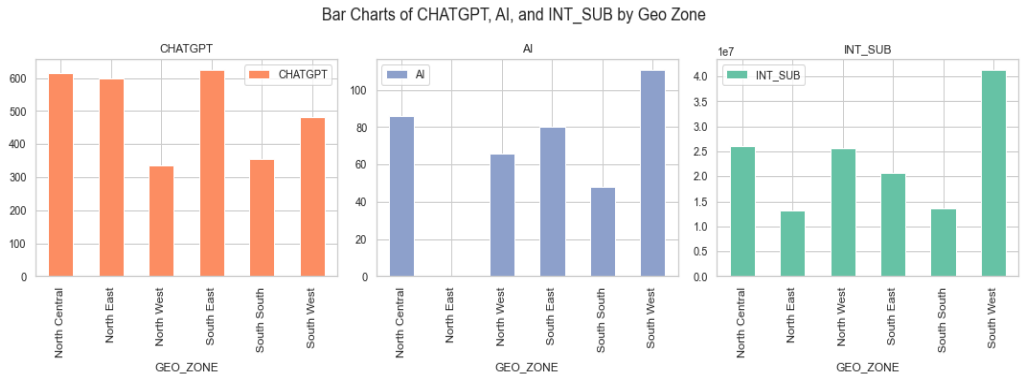

Figure 4: Stacked bar plots of Nigeria searches on 'chatgpt' and 'artificial intelligence', including Nigeria's internet subscription rate by geo-political zones.

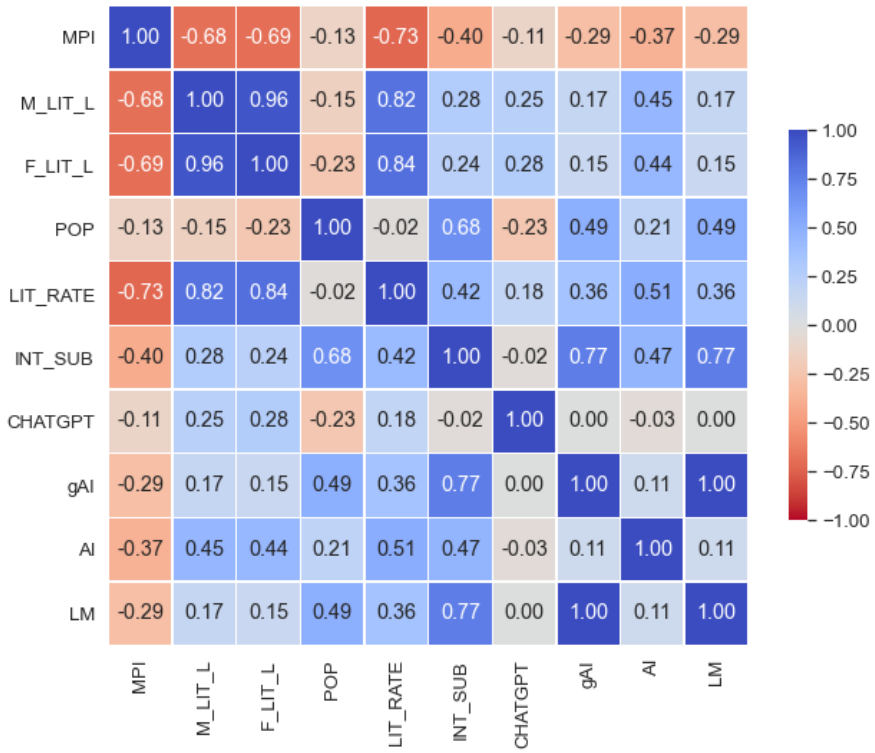

Figure 5: Heat plot showing the correlation coefficient matrix of the multivariate observational data on Nigeria with relation to gAI

The statistical summary table is revealing of the spread of data points across critical points. The multidimensional poverty index of the Nigerian states returns a mean value of 0.2634 for a population with an average of 5 million people per state. The literacy level of the country within the last one year is put an average of 50 and their internet subscription average 3.7. With the somewhat high literacy and internet subscription rate, an average of 81%, 10%, 0.027% and 0.027% enquiries were made on 'chatgpt', 'generative AI', 'language model' and 'artificial intelligence' itself respectively. ChatGPT keyword is the most popular across the country returning a maximum search percentage of 100% per time with artificial intelligence having a maximum of 28% search enquiry per time. The 25% or less of total search per time is 80% about chatgpt only. Considering 50% of search history per time, 85% are about chatgpt while 14% will be about artificial intelligence. Not more than 75% of the entire search entries are 100% of chatgpt and some 18% of them focused on artificial intelligence keyword. The data indicates that until the Q2 (median) of the IQR, there are no traces of generative AI, artificial intelligence, and language model keywords in the enquiry efforts of Nigerian within the last year. As observed from the statistical summary table, the mean poverty index per a state in Nigeria is 0.264 with an average population of 5 million people. The mean literacy rate is 50.338, with an average internet subscription rate of 3 million people per state. Given this, a substantial number of Nigerians has had internet access in the last one year across each state with over 85% of their search enquiries on 'chatgpt', 10% on 'artificial intelligence' and an insignificant part of the population on 'language model' and 'generative AI' keywords. As observed from Figure 1, there was no record of 'chatgpt' in the search history of Nigerians until November 2023, with 'artificial intelligence' having a relative enquiry check by Nigerians. This is followed by 'language model' search. The unprecedented surge of interest in 'chatgpt' peaked in March 2023 after initial plunge in December 2023 and late February 2023. Trying to establish the linear relationship between Nigerians' internet savviness and their socioeconomic indices can be inferred from the heat map of Figure 5. As can be observed, there exist a positive correlation between the internet subscription (INT_SUB) figure of the states with their respective population (POP) [0.68], and with rate of search enquiry on gAI (0.77) and LM (0.77). This implies

that increase in population increases the rate of internet subscription and enquiry on gAI and LM expectedly. Whereas, enquiries on gAI during the period under review came from Lagos only (Figure 2b). Notwithstanding the surge on 'chatgpt' enquiry across the states except for Kebbi, Zamfara, and Katsina (Figure 2c), there exist almost an insignificant relationship between the variable and MPI (-0.11), POP (-0.23), and the INT_SUB rate (-0.02) as seen on Figure 5. The population of each states shows a somewhat positive correlation with the gAI and LM (0.49) enquiry respectively. This could mean that states with higher population ordinarily will experience higher research enquiry into the AI tools, though the matrix shows a weak association. The literacy level of the states shows a similar pattern of insights as observed on the heat plot with some upsets. There is a negative correlation of -0.68 and -0.69 between the poverty index of states and their respective male and female literacy levels (M_LIT_L & F_LIT_L) which does not necessarily affect the level of enquiries into 'chatgpt' (0.25 and 0.28); 'gAI' (0.17 and 0.15); 'AI' (0.45 and 0.44), and the 'LM' (0.17 and 0,15) search. Observation of non-existence of negative correlation between the literacy level of the states and their level of enquiry into 'chatgpt' (0.18); 'gAI' (0.36); 'AI' (0.51) and 'LM' (0.36) indicates that literacy level may not necessarily determine the level of awareness or interest in the AI technologies. This is further observed in Figure 2a and 2c where supposed less literate state researched more into 'chatgpt' and 'AI'. Majority of the less literate states are located in the North East and North West of the country plotted in Figure 3c., with the highest rate of multidimensional poverty index as plotted in Figure 3b. With the high population of the North Central, North East and North West, and its towering poverty index and low literacy rate, the search for 'chatgpt' (Figure 4a) and 'gAI' (figure 4b) is high in the regions (except for the North East region with no enquiry on AI), which is indicative of their awareness level and interest in the AI tools. Their internet subscription rate is likewise relatively encouraging when compared with other literate states as plotted in Figure 4c. On the different search keywords, there is a strong positive relationship (1.00) between 'gAI' and 'LM' search words respectively. This shows aside the 'chatgpt' search keyword, the two words are used more interchangeably by Nigerians on the search engines. On the basis of Nigeria's geo-political zoning.

5. Conclusion and Recommendation

The study employed Nigeria's socioeconomic and demographic data together with web analytics metrics from Google Trend for an exploratory data analysis. The aim is to discover the relationship between the awareness and interest levels on Nigerians on generative artificial intelligence tools and the various factors that could influence the predisposition of different cultures towards the AI tools. Experimental result reveals the popularity of ChatGPT over other keywords like generative AI and language models and as well as the awareness of the intelligent tools even at regions with low literacy and high multidimensional poverty index.

6. Acknowledgements

Authors profoundly appreciate the painstaking efforts of the reviewers.

7. References

[1] U. Bukar, M. S. Sayeed, S. F. A. Razak, S. Yogarayan and O. A. Amodu, "Text Analysis of Chatgpt as a Tool for Academic Progress or Exploitation," *SSRN*, p. 4381394, 2023.

[2] M. Mijwil and M. Aljanabi, "Towards Artificial Intelligence-Based Cybersecurity: The Practices and ChatGPT Generated Ways to Combat Cybercrime," *Iraqi Journal For Computer Science and Mathematics*, vol. 4.1, no. 2023, pp. 65-70, 2023.

[3] J. Qadir, "Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education," 2022.

[4] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu and A. Radford, "Release strategies and the social impacts of language models," *arXiv preprint*, p. 1908.09203, 2019.

[5] O. J. Akintande, O. E. Olubosoye, A. F. Adenikinju and B. T. Olanrewaju, "Modelling the determinant of renewable energy consumption: Evidence from the five most populous nations in Africa," *Energy*, vol. 1, no. 206, p. 117992, 2020.

[6] A. Lateef, M. A. Azeez, O. B. Suaibu and G. O. Adigun, "A decade of nanotechnology research in Nigeria (2010-2020): a scientometric analysis," *Journal of Nanoparticle Research*, vol. 23, no. 2021, pp. 1-27, 2021.

[7] B. O. Ajibade, "A critical analysis of Nigeria's educational system," *Global Journal of Human-Scoial Science, Linguistics and Education* , vol. 19, no. 8, 2019.

[8] N. Curtis, "To ChatGPT or not to ChatGPT? The impact of artificial intelligence on academic publishing," *The Pediatric Infectious Disease Journal*, vol. 42, no. 4, p. 275, 2023.

[9] M. Shidiq, "The use of artificial intelligence-based ChatGPT and its challenges for the world of education; from the viewpoint of creative writing skills," in *Proceeding of International Conference on Education, Society and Humanity*, 2023.

[10] A. M. Alkalbani, A. M. Ghamry, F. K. Hussain and O. K. Hussain, "Sentiment Analysis and Classification for Software as a Service Reviews," in *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, 2019.

[11] S. M. Lakew, M. Cettolo and M. Federico, "A comparison of transformer and recurrent neural networks on multilingual neural machine translation," *arXiv preprint arXiv*, p. 1806.06957, 2018.

[12] G. Cooper, "Cooper, G. (2023). Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence," *Journal of Science Education and Technology*, pp. 1-9, 2023.

[13] R. V. Yampolskiy, "Predicting future AI failures from historic examples," *Foresignt* , vol. 21, no. 1, pp. 138-152, 2019.

[14] T. Olaleye, O. T. Arogundade, S. Misra, A. Abayomi-Alli and U. Kose, "Predictive Analytics and Software Defect Severity: A Systematic Review and Future Directions," *Scientific Programming*, pp. 1-18, 2023.

[15] D. Kohns and A. Bhattacharjee, "Nowcasting growth using Google Trends data: A Bayesian Structural Time Series model," *International Journal of Forecasting*, 2022.

[16] Y. Fenga, G. Lia, X. Suna and J. Li, "Forecasting the number of inbound tourists with Google Trends," in *7th International Conference on Information Technology and Quantitative Management (ITQM 2019)*, 2019.

[17] NBS, "Nigeria Population by State," National Bureau of Statistics, Abuja, 2023.

[18] NCC, "Subscriber Statistics," Nigeria Communications Commission , Abuja, 2023.

[19] T. Olaleye, A. Abayomi-Alli, K. Adesemowo, O. T. Arogundade, S. Misra and U. Kose, "SCLAVOEM: hyper parameter optimization approach to predictive modelling of COVID-19 infodemic tweets using smote and classifier vote ensemble," *Soft Computing,* vol. 27, pp. 1-20, 2022.

[20] K. Rashid, M. A. Islam, R. A. Tanzin, M. L. Labib and M. Khan, "Heart Disease Prediction Using Interquartile Range Preprocessing and Hypertuned Machine Learning," in *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2022.

[21] M. Sajjad, W. Sałabun, S. Faizi, M. Ismail and J. Wątróbski, "Statistical and analytical approach of multi-criteria group decision-making based on the correlation coefficient under intuitionistic 2-tuple fuzzy linguistic environment," *Expert Systems with Applications,* vol. 193, p. 116341, 2022.