Infectious Diseases in Nigeria using Topic Modelling

Abstract

In this study, we used text mining to present a list of topics and diseases that have been most researched in the last two decades in the field of infectious diseases in Nigeria, using abstracts and keywords of publications. We accessed and analyzed 15,459 publications from the Scopus database, extracting out their abstracts, keywords, years of publication and titles. We used quanteda to get n-grams and document frequencies of the keywords, which produced 39 keywords for diseases based on unigram, 55 based on bigram, and 63 for trigram. HIV/AIDS and Malaria emerged as the two primary focus in this research. Taking advantage of technological advances, we used ChatGPT to check the prevalentinfectious diseases in Nigeriaand compared output with our result, and our search turned in 14 infectious diseases in Nigeria, out of which 10 were a match to the diseases contained in the research articles' keywords. Thus, continuing measures to eradicate malaria in Nigeria, and interventions to reduce the prevalence of HIV/AIDS should be communicated. Using structural topic modelling, we were able to extract a total of 100 topics from the document abstracts in which topic 29 (sexual, HIV, partner, sex, condom, risk, men) stood out due to its higher distribution across the documents.

Keywords: Infectious disease; Systematic literature review; Topic modelling; Text mining; Nigeria.

1. Introduction

In the last 20 years, research on infectious diseases has seen unprecedented rise globally, particularly in Nigeria. All countries of the world face disease outbreaks, which often escalate to pandemics, as in the case of Ebola in 2014 and Coronavirus (COVID-19) in 2020. HIV/AIDS is another global challenge, primarily transmitted through sexual intercourse due to the human immunodeficiency virus (HIV). Malaria and yellow fever are majorly found in Africa. The 2021 report gives 95% cases of malaria in the region, as stated by WHO Africa Region.[1] These diseases outbreaks often create panic in the national and global health systems by disrupting its routine activities [3]. Nigeria, being a developing country faces numerous challenges in combating various diseases, especially in its preparedness for disease outbreaks in terms of infrastructure, human capacity, and political will [23], leading to the prevalence of infectious diseases in Nigeria. Between 2016 and 2018 alone, Nigeria recorded about 20 infectious disease outbreaks and public health emergencies [23]. The prevalence and outbreaks of infectious diseases in Nigeria have been of keen interest to numerous researchers both within and outside the nation. Some of these diseases may be endemic to Nigeria, and others may have been a pandemic that affected the whole world. Diseases,

[1] https://www.afro.who.int/

whether they are pandemics or endemics, has sparked significant research and extensive peer review in the field of medicine.

The need to know the trends of literature on infectious diseases in Nigeria is of interest. This will assist government and health facilities in their planning for possible pandemics and epidemics in the future. For example, research on COVID-19 is still being published heavily and "COVID-19". Further, we are in the world of higher uncertainty where policy makers need to prepare against future occurrences that could spark uncertainty in the global economic and financial structure, thus a thorough review of literature on disease outbreaks is necessary, taking Nigeria as a case study. The approach, detailed in this paper, will also assist researcher in the literature review management of articles.

There are various ways that reviews of literature are carried out, but most available options, like systematic literature reviews on infectious diseases in Nigeria, are not all-encompassing. They often focus on a particular disease to review, for example, [7] conducted a thorough investigation into the prevalence of chronic kidney diseases in Nigeria. [14] provided an extensive update on the monkeypox virus in Nigeria through a comprehensive review. [13] examined the occurrences and developments in cardiovascular diseases over the past two decades. Unlike machine learning methods that allow researcher to review a large corpus of literature using either supervised or unsupervised methods, to automatically review literature [25], these manual methods have limitations in terms of both the number of articles they can review and the fact that they require human intervention [2]. In contrast, machine learning techniques possess the remarkable ability to process vast amounts of articles effortlessly, eliminating the necessity for human intervention. Not only does this improve the reliability of the results, but it also significantly saves a considerable amount of time [9].

With the emergence of new approaches to conducting literature reviews, the traditional method of manual literature review is slowly becoming obsolete [2]. As a result, many individuals are adopting alternative, time- and cost-saving techniques for carrying out reviews. One such method is topic modeling in Machine Learning, which has revolutionized literature review by being able to explore large texts of literature [10]. These innovative methods, not only save valuable time and resources but also demonstrate that conducting thorough literature reviews is not limited to experts in a specific field. This advancement has revolutionized the way literature reviews are conducted in both academic and non-academic research across different disciplines. According to a study conducted by [2], it appears that the era of manual literature reviews is coming to an end. Manually examining articles is not only time-consuming but also lacks coverage. This is mainly due to high costs and extensive

time required for this demanding task. These methods do not necessitate a high level of expertise in a particular field to review them. Furthermore, these models eliminate the preferential treatment given by many journals to articles written solely by experts, considering that experts are likely to be biased in their selection articles to be reviewed, which could lead to human errors.

The integration of AI into the systematic literature review process helps to reduce human errors and obtain deeper insights from the vast amount of literature available on infectious diseases in Nigeria. ChatGPT (Chat Generative Pretrained Transformer) is one of the open AI that understands the natural language used in literature reviewing [29]. Using AI tools alone is not enough to guarantee the precision and credibility of the review's findings, it is essential to combine them with human expertise. Additionally, researchers should stay informed about the latest AI technologies and best practices in systematic literature reviews to maximize the benefits of AI in their research [26].

This work, building on the advances brought into the systematic literature review by machine learning models, intends to do a more comprehensive literature review of infectious diseases in Nigeria in the past 20 years. knowing that the keyword section contains distinct words primary to the research article, we use the document frequency (DF) to find the top diseases and the least diseases that researchers have researched most in the past 20 years. Also, using the topic representation that is produced by Structural Topic Modelling (STM) tool developed by [27], we use the abstracts of the articles to determine the topic representation of research topics from 2002 to 2022 in Nigeria. Findings in this paper will expose researchers to a comprehensive overview of the areas that have been extensively explored, and sheds light on those that have received minimal attention. Further, by using keywords, researchers identify fields of infectious diseases that have been given little attention for research. This work can be considered innovative in Nigeria, as it focuses on delving into the abstracts of research articles, as it uncovers valuable information about infectious disease research conducted over the past two decades. Also, the approach contained herein will serve as an alternative to literature review process.

2. Literature Review

Infectious disease has been a major challenge that Nigeria has been bedeviled with. Nigeria has faced various diseases over the years, such as the yellow fever disease that first surfaced in Nigeria in 1986, causing a significant impact, with over 9,800 individuals affected [8]. In 1987, yellow fever spread across Africa and affected 1,249 people, [18]. In 1996, there was

an outbreak of meningitis, affecting 10,958 people and causing 11,717 deaths [17]. The occurrence of the disease seized until 2017 when there was another outbreak of it [21]. In 1972, Nigeria recorded its first case of cholera [1], and ever since, cholera has yet to be eliminated in Nigeria. In 2022, Nigeria recorded about 23,763 cholera cases, with 592 deaths [19]. Also, in 2014 the dreaded Ebola virus was experienced in West Africa [25]. The increase in insecurities in Nigeria, such as the Boko Haram insurgence and the continuing farmers and herders' crises which remain unresolved in most parts of the country have displaced a lot of people from their ancestral homes and destroyed a lot of properties, which has led to the outbreaks of new and re-emerging diseases in Nigeria [20]. More recently, the COVID-19 pandemic shook the world, and Nigeria was no exception. Nigeria was the first country to report a case of COVID-19 in the sub-Saharan Africa on 27 February 2020 [30]. Nigeria as at 26 February 2023, recorded 266,313 confirmed cases, out of which 259,027 recovered and 3,155 died [19].

The continued and widespread transmission of infectious diseases in Nigeria has created an urgent need for the establishment and strengthening of the Nigerian Center for Disease Control (NCDC) in 2011. According to [20], this organization plays a crucial role in enabling the government to effectively mobilize resources and deliver a proactive response to disease outbreaks and other urgent public health emergencies.

Despite efforts put so far in combating infectious diseases in Nigeria, it is sad to note that certain diseases remain prevalent. One such disease is malaria. [31] in its Malaria Report 2012 obtained 213,401,238 figures as the number of malaria cases in 2012 which posed a significant malaria risk. It may sound unbelievable, but for every single person in the nation, which amounts to about 216.7 million individuals as 2022 [2] can be classified as either a high-risk or low risk in the possibility of contracting malaria. Even more astonishing is the fact that a staggering 162,911,666 individuals are deemed to be at high risk [31]. This is most worrisome because, as far back as 1948, malaria was listed among the diseases causing high death rates in Nigeria. The country is consistently recognized as the most affected by malaria globally, with the highest number of cases and fatalities [5,31].

Although Nigeria has faced numerous challenges in combating infectious diseases, it is important to highlight the significant progress the country has made in eradicating and controlling their transmission. An example is HIV/AIDs, with its first survey recorded in 1991 and having a prevalence rate of 1.8%. In the years ahead, the prevalence rate will see its

[2] https://www.statista.com/statistics/1122838/population-of-nigeria/

highest (5.8%) in the year 2001. Combating this has so far reduced HIV prevalence in Nigeria to 1.4% [22]. In 2014, Nigeria effectively managed the Ebola virus outbreak through the coordinated efforts of the Federal Ministry of Health and the NCDC, using the incident management approach. [24]. Nigeria's journey in the fight against infectious disease control has been remarkable, showcasing significant progress over the years. This research aims to highlight prominent diseases that have captured the attention of infectious disease researchers in Nigeria over the past two decades.

3. Methodology

We adopt a systematic literature review methodto examine the abstracts of articles focusing on infectious diseases in Nigeria. We utilized the extensive Scopus database to conduct the literature search for articles. Scopus stands as an unparalleled abstract database, being the largest ever constructed[6]. We used the search keywords "infect" "disease" and "nigeri". The keyword “infect,” means that words such as infectious, infection, and infecting that imply infection should be considered. When it comes to the keyword "disease," it is crucial to include in the papers the keywords that contain related terms like diseases as well, and finally, the keyword nigeri implies that papers with words such as Nigeria, Nigerian, Nigerians, or words that imply Nigeria should be considered. After choosing the keywords, we limited the articles to the years 2002 through 2022. We further limited our search to 14 fields and limited it to articles written in English language. Finally, we limited the articles to those focusing on Nigeria only. After the search keywords and limitations were done, we obtained a search result of 15,459 documents, out of which 15,458 were successfully exported to Excel. The exported Excel document contains a wealth of valuable information which includes variables such as Authors, Authors' full names, Authors' ID, Title, Year, Source title, Volume, Issue, Art No., Page Start, Page End, Page count, Cited by, DOI, Link, Abstract, Author Keywords, Index Keywords, Document type, Publication stage, Open access, Source, and EID. It provides a comprehensive overview of the research data, making it an essential resource for analysis and referencing. Out of these variables, only “Year, Title, Abstracts, and Keyword” were used for this research. This search was carried out on 25 May 2023, and it is expected that anyone who keys in the query in Appendix 1 in the Scopus database should get the same result.

After collating this dataset, the next step is to prepare the data and carry out the analysis. For this, we used the R software. Two packages were instrumental in this work; the first is the quanteda package developed and maintained by [4]. This package is very

resourceful in the preprocessing of text data. With this package, we convert all text to lowercase, and remove the following: stop words (“Smart”, “Stopwords.iso”, “stopwords(), custom medical stop words contain in the R codes in appendix III), special characters, numbers, lemmatized words, punctuations, and symbols. We created a document term matrix (DTM) and then built a document features matrix (DFM). The DFM helps us to generate the document frequency for each word, allowing us to know how many documents a word wascaptured. Furthermore, for our keyword analysis, this package was used to produce Bi-grams and Tri-grams, and the term frequency for the Uni-gram, Bi-gram, and Tri-gram. After using our keywords to determine the uni-gram, bi-gram and tri-gram, we will use ChatGTP to conduct a comprehensive search using the specific query "Comprehensive List of Infectious Diseases Reported in Nigeria in the Past 20 Years" in order to gather information on infectious diseases in Nigeria over the past 20 years. This will enable us compare our results with those gotten from ChatGPT and see how reliable the ChatGPT search engine is, when providing information on infectious diseases in Nigeria.

The second package used is STM (Structural Topic Modeling), developed in[27]. This package was used to read the text and develop topics from the abstracts of the articles collected. It was further used to match the topics generated with the article topics to find the most closely related articles to the topics generated. The uniqueness of STM lies in its ability to incorporate metadata into its model. This integration allows for the visualization of shared topics among various documents in the corpus, each employing distinct grammatical constructs. This further allows us to estimate topics with documents-level covariates [27] to examine the relationship between the generated topics and other variables. Due to its remarkable flexibility, STM can easily accommodate a wide variety of datasets. Not only can it effortlessly handle large volumes of data, but it also effectively manages missing data within the datasets and these are its uniqueness over other text modelling algorithms.

The operations of STM are illustrated in Figure 1.

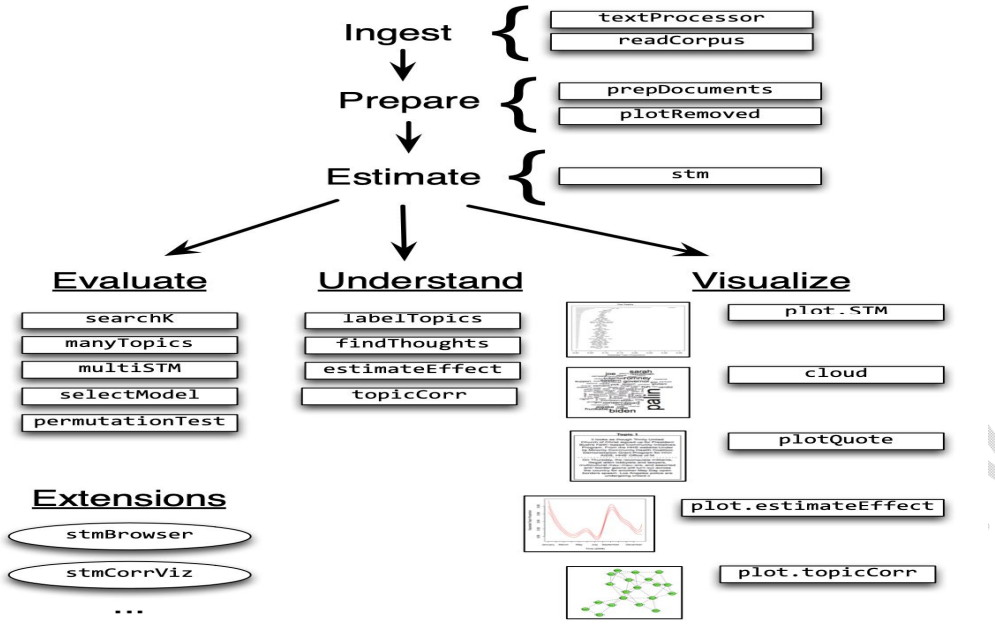

Figure 1: Heuristic description of selected STM package features. Source: [27]

The STM is a generative model of word counts as described in Figure 2. This implies that the model generates the data for each document and uses the data to find values for the parameters in the model. The process begins at the top with document topics and topic word distributions generating documents that have meta-data associated with them. Given this, [30] defines a topic as a mixture of words, where each word has a probability of belonging to a topic. while defining a document as a mixture of topics.The generative process for each document d with vocabulary of size V for an STM model.

1. Draw document-level attention to each topic from a logistic-normal generalized linear model based on a vector of document covariates X_d, i.e.,

   $$\vec{\theta}_d | X_{d\gamma}, \Sigma \sim LogisticNormal(\mu = X_{d\gamma}, \Sigma) \quad (1)$$

   where X_d is a $1 - by - p$ vector γ is a $p - by - (k-1)$ matrix coefficients, and $\sum$ is a $(k-1) - by - (k-1)$ covariance matrix.

1. Given a document-level content covariate y_d from the document-specific distribution over words representing each topic (k) using the baseline word distribution (m), the topic specific deviation$k_k^{(t)}$, the covariate group deviation $k_{yd}^{(c)}$ and the interaction between the two $k_{yd,k}^{(i)}$.

   $$\beta_{d,k} \propto \exp\left(m + k_k^{(t)} + k_{yd}^{(c)} + k_{yd,k}^{(i)}\right) \quad (2)$$

where m, $k_k^{(t)}$, $k_{yd,k}^{(i)}$ and $k_{yd,k}^{(i)}$ are V-length vectors containing one entry per word in the vocabulary. When no convent covariate is present, β can be formed as $\beta_{d,k} \propto \exp(m + k_k^{(t)})$ or simply point estimated (the latter behavior is the default).

2. For each word in the document, $(n \in \{1, \ldots, N_d\})$, draw word's topic assignment based on the document-specific distribution over topics:

$$z_{d,n}|\vec{\theta}_d \sim Multinomial(\vec{\theta}_d) \quad (3)$$

Conditional on the topic chosen, draw an observed word from that topic based on the distribution:

$$\omega_{d,n}|z_{d,n}\beta_{d,k=z_{d,n}} \sim Multinomial(\beta_{d,k=z_{d,n}}) \quad (4)$$

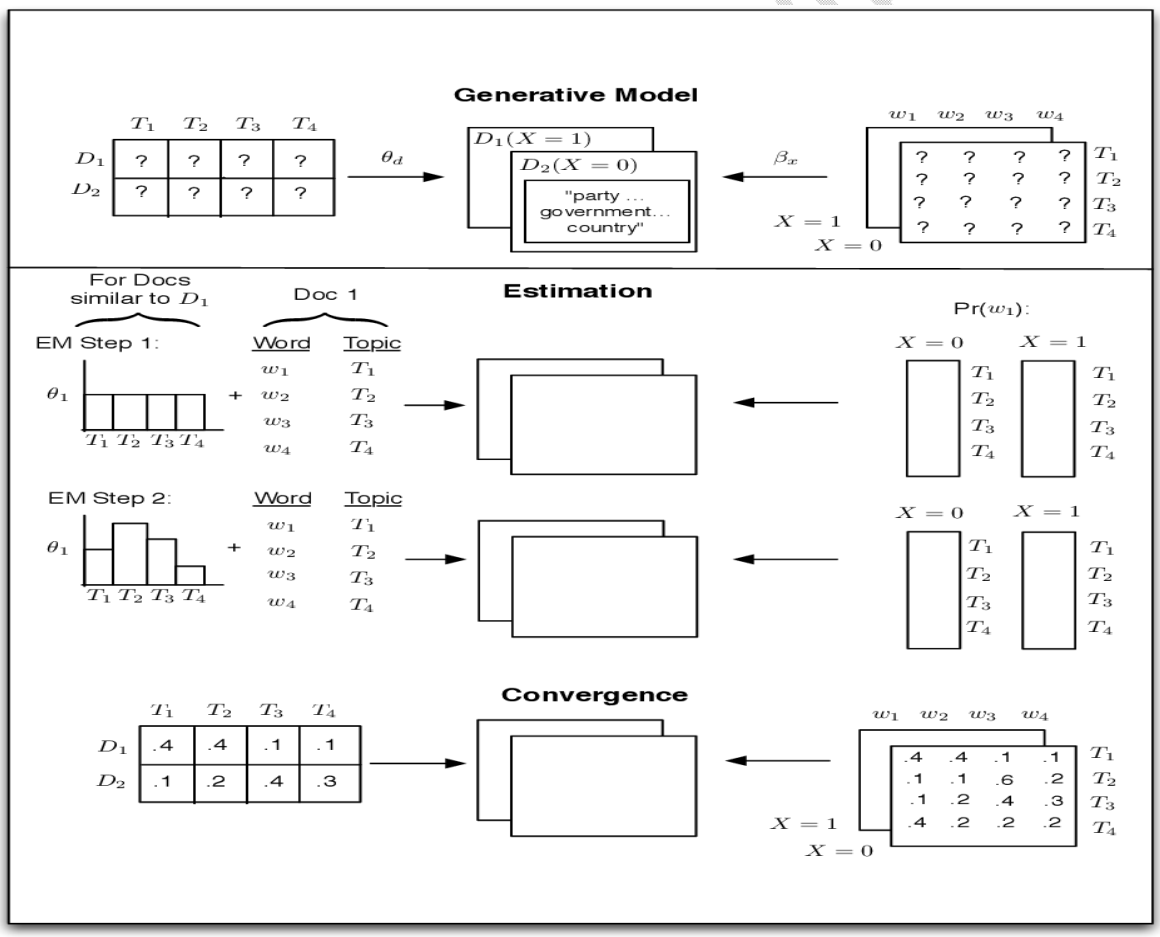

Figure 2: Heuristic description of generative process and estimation of the STM. Source: [28]

3.2 Estimation Procedure

In this section, we shall give a detailed procedure on how the estimation will be carried out. The following are the procedures for the estimation process:

I. We shall be importing all the packages needed for our analysis into R.

II. We shall import the datasets into R.

III. We shall then proceed to carry out Exploratory Data Analysis (EDA), to prepare our data for descriptive statistics. This involves the following steps:
   - Reduce the data set to the variables in consideration.
   - Create additional list of stop words to be used.
   - Lower case, remove stop words, remove punctuation, stem the words, set word length to at least 3, limit to only English words, reduce to only characters, remove customized stop words.

IV. After carrying out EDA, we shall proceed to carry out a Descriptive Statistics Analysis (DSA). The DSA done are:
   - Use the column for year to visualize the years and their publications.
   - Use the Key words to determine the term frequencies to get diseases most researched about.
   - Use ChatGPT to determine the most prevalent diseases in Nigeria and compare it to those most researched in Nigeria to determine those the available data set reveal they have been researched on.

V. After DSA, we shall proceed to build STM discussed in the methodology above.

VI. From our models, we shall determine the number of topics to be generated and generate the topics accordingly.

4. Results and Discussion

In this section, we present the results from the analysis carried out using R software. These results are in two parts; the first is the term frequencies for uni-grams, bi-grams, and tri-grams for the article keywords. The second part is the STM topic modelling results for the abstracts of the articles.

First, we want to look at the years of publication and the number of articles published for each year. From Figure 3, the number of articles reporting infectious diseases has increased astronomically, from 266 in 2002 to 1205 in 2022, while in 2011 to 2013, there was a stall in the progress having reached 851 in 2011. This further dropped to 763 in 2014 and continued to progress astronomically until 2022.

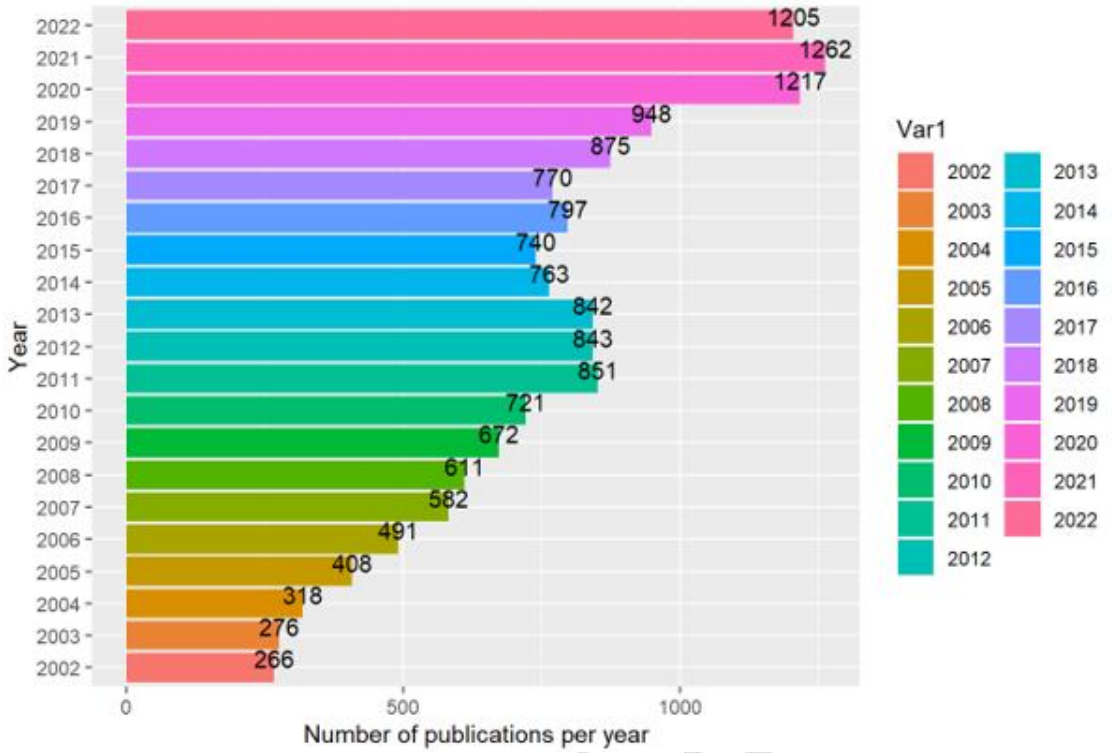

Figure 3: Number of articles published by year

Following [11], the N-gram helps us to find the repeated word pattern in a given text. Table 1 displays the output for words with their corresponding document frequency(df)and how many documents (articles) captured the diseases using the fact that keywords of an article defined the content of that article. Table 1 therefore informs us the number of documents and corresponding keywords as discussed as part of the subject. Our interest is in identifying diseases, and how many documents they appeared in, to give us an understanding of how frequently such diseases were discussed in the past two decades. Table 1 below shows a summarized collection of the diseases from the output of our results. Appendix I contains the diseases as they appeared in different forms, and how they were summed up to give us the Table 1 below.

From the results of Unigrams displaced in Table 1, HIV, Malaria, COVID-19, Diabetes and Cancer, were the top 5 most researched in Nigeria. Then, we had diseases such as Ebola, Measles, Schizophrenia, Zoonosis, and Ulcer as the least 5 researched. For the Bigram, HIV AIDs, Sickle Cell, Cancer Nigeria, Malaria Nigeria, and Diabetes Miletus were the top 5 most researched diseases. Diseases such as Tinea Capitis, Renal Failure, Nigeria, Swine Fever, genital Mutilation, and Fever Nigeria were the least researched. For the Trigram, Sickle Cell Diseases, Human Immunodeficiency virus, Chronic Kidney diseases,

Hepatitis Virus Hepatitis, and Nigeria Plasmodium Falciparum are the top 5 most researched diseases. The least researched diseases for the Trigram were Systemic Lupus Erythematosus, Viral Hemorrhagic fever, Virus Hepatocellular Carcinoma, Nephrotic Syndrome Nigeria, and Swine Fever Virus.

Table 1: The Document Frequency representation for Uni-, Bi-, and Tri-Grams.

S/N	Unigram	DF	Bigram	DF	Trigram	DF
1	HIV	1599	HIV aids	1252	Human immunodeficiency virus	545
2	Malaria	686	Malaria Nigeria	486	Sickle cell disease	339
3	Covid-19	438	Sickle cell	374	Hepatitis virus hepatitis	118
4	Diabetes	412	Cancer Niger	302	Chronic kidney disease	110
5	Cancer	394	Diabetes Miletus	216	Nigeria plasmodium falciparum	101
6	Pregnancy	380	Disease Nigeria	210	Nigeria pregnant women	72
7	Tuberculosis	300	Infection Nigeria	201	Urinary tract infection	63
8	Anaemia	238	Pregnant Women	185	Type diabetes mellitus	56
9	Hypertension	224	Covid-19 Nigeria	152	Cervical cancer screening	52
10	Hepatitis	203	Hepatitis virus	146	Sexually transmitted infections	43
11	Fever	137	Cell disease	145	Covid-19 Nigeria Pandemic	39
12	Injury	125	Kidney disease	129	Neglected tropical disease	37
13	Stroke	108	Mental health	124	Coronary heart disease	35
14	Stress	105	Lassa fever	71	Lassa fever Nigeria	35
15	Schistosomiasis	104	Oxidate stress	71	Hypertension left ventricular	32
16	Blindness	100	Mycobacterium tuberculosis	67	Ebola virus disease	28
17	Obesity	95	Virus Nigeria	66	Respiratory tract infection	28
18	Illness	80	Hypertension Nigeria	63	Mycobacterium tuberculosis Nigeria	26
19	Epilepsy	77	Human papillomavirus	62	Suppurative otitis media	26
20	Lassa	67	Cardiovascular disease	61	Cell disease stroke	25
21	Influenza	66	Staphylococcus aureus	61	Human papilloma virus	25
22	Staphylococcus	65	Heart disease	54	Pathogenic avian influenza	25
23	HBV	62	Blood pressure	52	Blood pressure control	23
24	Infertility	58	Mosaic Virus	41	Nigeria prostate cancer	17
25	Infertility	58	Heart failure	39	Newcastle disease Nigeria	16
26	Asthma	56	Avian Influenza	36	Mental illness Nigeria	15
27	Sars-cov-2	54	Maternal mortality	34	Left ventricular hypertrophy	14
28	Malnutrition	52	Newcastle disease	32	Acute flaccid paralysis	13
29	Pneumonia	52	Ebola virus	31	Cassava mosaic disease	13
30	Trauma	52	Lymphatic filariasis	29	Infection prevention control	13
31	Polio	51	Nigeria polio	28	Adverse drug reactions	12
32	Diarrhoea	50	Flaccid Paralysis	27	Benign prostatic hyperplasia	12
33	HCV	48	Urinary schistosomiasis	27	Chronic renal failure	12
34	HPV	47	Anxiety depression	25	Oral polio vaccine	12
35	Ebola	46	Cardiovascular disease	25	Female genital mutilation	11
36	Zoonosis	46	Hearing loss	22	Infectious bursal disease	11
37	Schizophrenia	45	Toxoplasma gondii	19	Intimate partner violence	11
38	Ulcer	42	Nigeria Obesity	18	Maternal mortality Nigeria	11
39	Measles	41	Renal disease	16	Rift valley fever	11
40			Yellow Fever	16	Lymphatic filariasis Nigeria	10
41			Epilepsy Nigeria	15	Nigeria non-communicable diseases	10
42			Erectile dysfunction	15	Cardiovascular disease risk	9

43			HBV HCV	15	Foot mouth disease	9
44			Nigeria sars-cov-2	15	Squamous cell carcinoma	9
45			Liver disease	14	Breast cancer Nigeria	8
46			Ventricular hypertrophy	14	Nigeria oxidated stress	8
47			Cerebral Palsy	13	Herpes simplex virus	7
48			Musculoskeletal disorder	13	West Nile virus	7
49			Nigeria Onchocerciasis	13	Chronic obstructive Pulmonary	6
50			Parkinson disease	13	Spinal cord injury	6
51			Renal failure	13	Body mass obesity	5
52			Tinea capitis	13	Chronic liver disease	5
53			Fever Nigeria	12	Co-infection hbvhcv	5
54			Genital mutilation	12	Cucumber mosaic virus	5
55			Swine fever	12	Methicillin-resistant staphylococcus aureus	5
56					Nigeria pandemic sars-cov-2	5
57					Nigeria Parkinson disease	5
58					Systemic Lupus Erythematosus	5
59					Viral hemorrhagic fever	5
60					Virus hepatocellular carcinoma	5
61					Anxiety depression Nigeria	4
62					Nephrotic syndrome Nigeria	4
63					Swine fever virus	4

We used ChatGPT to conduct a comprehensive search using the specific query "Comprehensive List of Infectious Diseases Reported in Nigeria in the Past 20 Years" to gather information on infectious diseases in Nigeria over the past 20 years. We compared the outcome of a publication search with the use of AI tools. The importance of AI tools has been firmly established due to their ability to automate tasks, improve efficiency, and effectively handle large volumes of data [15]. The outcome of the ChatGPT search was 14 infectious diseases of which 4 (Cholera, Meningitis, Typhoid Fever, Leprosy) were not captured in the list of identified infectious diseases from the publication search while nine of the diseases were captured, as shown in Table 2. Despite the criticism on the usage of ChatGPT for information and literature search, the AI tool has been used effectively in the field of healthcare and medicine to assist in diagnosing, summarizing medical research, providing patients with medical information in a comprehensible way and enabling collaboration between healthcare experts.

Table 2: A comparison between ChatGPT identified disease and those in the publication.

S/N	ChatGPT generated list of Infectious diseases in Nigeria	Publication search comparison
1	Malaria	Yes
2	HIV/AID's	Yes
3	Lassa Fever	Yes
4	Tuberculosis	Yes
5	Cholera	No
6	Yellow Fever	Yes

7	Polio	Yes
8	Measles	Yes
9	Hepatitis	Yes
10	Ebola	Yes
11	Meningitis	No
12	Typhoid Fever	No
13	Leprosy	No
14	COVID-19	Yes

In the second part of the analysis, we utilize the STM package to preprocess the text. This enables us to ascertain the optimal number of topics to be generated. In the preprocessing, after building corpus, converting to lower case, removing punctuation, removing stop words and custom stop words, removing numbers, and stemming, we had 15458 documents, and 40707-word dictionary. We used the prepDocuments command to further clean the data, and we had 31108 removed from the dictionary of 40707 words, and 50211 tokens were removed out of 1275935 tokens. In the end, we had 15458 documents, 9599 words and 1225724 tokens in our corpus or analysis. We then proceeded to generate topics from the abstracts of the compiled articles. When it comes to choosing the topics to be generated [12], there is no one-size-fits-all method that can accurately determine the ideal number of topics to be selected. In this work, we use the function searchK to determine the optimal number of topics. We first use a selection method recommended by [16], where we set the initialization type to Spectral, and specified $K = 0$. This suggested topics numbers 105, 106, and 116 at different runs. The main obstacle that we faced was how to maintain consistency. Despite setting seed to ensure that the same result is reproduced, there was a different output for every run of the code. Seeing that the results produced were above 100, K was set to generate topics in the range of 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, and 140. Based on the plot in Figure 4, we have determined that the optimal number of topics falls within the range of 80 to 100. Since topic selection is an intuitive matter, we have set these boundaries to guide our decision-making process, thus we decided to generate 100 topics.

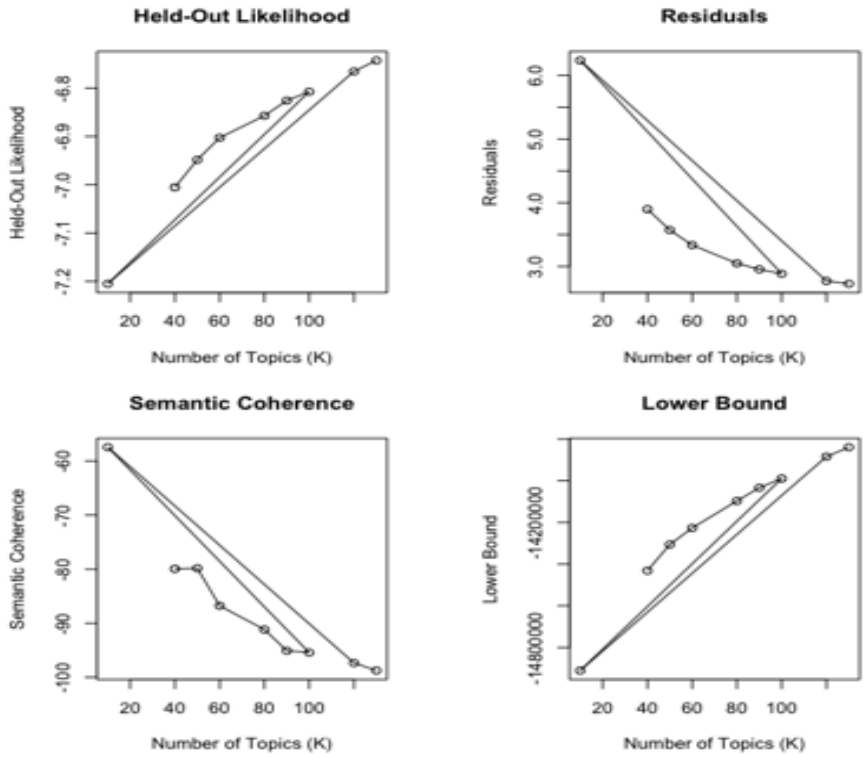

Figure 4: Diagnostics tests for selecting optimal number of topics, K

Table 2 displays the results for the 100 generated topics and the number of documents relating to each topic. Table 3 provides an insightful overview of the results, revealing a fascinating trend: each topic appears across various documents, demonstrating a remarkable degree of similarity in the research conducted. We have observed that topic 29 has a significantly broad presence, encompassing approximately 415 documents. Furthermore, our code was unable to accurately capture the number of documents associated with topic 2, as a result, we had to impute "NA," which stands for "not available."

Table 3: Estimated Topics with corresponding numbers of documents

S/N	Generated Topics (High Probability)	No. of Document
1	medicin, privat, afford, tropic, societi, abstract, press	178
2	avail, free, limit, despit, freeli, nigeria, make	NA
3	distribut, attribut, term, open, common, access, articl	35
4	medic, adher, self, hospit, reason, non, poor	87
5	school, adolesc, parent, children, girl, pupil, age	77
6	isol, strain, sequenc, nigeria, genet, genom, analysi	155
7	oil, activ, compound, acid, delta, antioxid, niger	58
8	cancer, cervic, screen, hpv, women, risk, nigeria	177

9	wound, bone, fractur, heal, limb, amput, head	87
10	mutat, subtyp, gene, sequenc, resist, associ, genotyp	92
11	ckd, kidney, renal, diseas, creatinin, iron, chronic	157
12	seed, fungal, growth, isol, pathogen, fungi, infect	120
13	present, old, rare, clinic, featur, syndrom, diseas	307
14	season, preval, dri, infect, rate, wet, month	96
15	genotyp, cassava, diseas, resist, genet, popul, allel	121
16	injuri, trauma, emerg, death, caus, hospit, accid	139
17	extract, rat, induc, mgkg, effect, activ, dose	251
18	mortal, sever, score, morbid, death, admiss, associ	136
19	hiv, infect, posit, count, hivaid, immunodefici, haart	261
20	year, femal, male, age, diseas, mean, hospit	121
21	virus, influenza, infect, antibodi, nigeria, detect, sampl	145
22	practic, hand, wast, worker, hospit, safeti, infect	130
23	lant, tradit, medicin, cell, diseas, herbal, activ	137
24	food, consumpt, exposur, consum, intak, household, dietari	84
25	asthma, respiratori, lung, symptom, worker, function, control	115
26	ill, famili, mental, caregiv, care, social, health	167
27	level, serum, control, day, vitamin, plasma, compar	145
28	hepat, infect, hbv, hcv, virus, hbsag, preval	262
29	sexual, hiv, partner, sex, condom, risk, men	415
30	heart, diseas, failur, congenit, cardiac, defect, chd	118
31	oral, clinic, day, devic, infant, hospit, care	37
32	children, age, childhood, month, year, five, diarrhoea	97
33	salmonella, poultri, antimicrobi, farm, chicken, isol, resist	116
34	infect, parasit, intestin, helminth, preval, sampl, stool	192
35	tuberculosi, pulmonari, rate, sputum, smear, nigeria, ptb	141
36	week, exercis, ultrasound, mean, method, assess, pain	59
37	depress, score, life, qualiti, assess, scale, symptom	371
38	cryptosporidium, infect, msp, divers, sampl, nigeria, human	69
39	stroke, nigeria, year, noma, brain, manag, prevent	87
40	dog, count, cell, blood, pcv, volum, haematolog	118
41	research, implement, health, develop, nigeria, countri, polici	316
42	obes, bodi, weight, bmi, mass, index, age	149
43	cost, drug, per, prescrib, prescript, cent, nigeria	153
44	breast, cancer, stage, year, present, diseas, surviv	196
45	communiti, rural, urban, area, household, nigeria, peopl	181
46	risk, blood, hypertens, factor, pressur, cardiovascular, transfuse	153
47	content, protein, diet, composit, acid, rang, sampl	74
48	extract, infect, mice, activ, dose, mgkg, day	127
49	resist, isol, gene, esbl, coli, strain, antibiot	231
50	resist, root, nematod, infect, popul, plant, soil	79
51	pcr, sampl, reaction, detect, chain, dna, polymeras	77
52	eye, visual, blind, ocular, glaucoma, cataract, examin	213
53	women, pregnant, pregnanc, antenat, preval, age, vagin	148
54	health, care, servic, facil, provid, healthcar, access	269
55	incid, yield, diseas, varieti, plant, crop, nigeria	112
56	cholesterol, metabol, lipid, densiti, hdl, level, high	151
57	surgic, surgeri, oper, complic, procedur, day, hospit	319
58	hypertens, left, ventricular, control, abnorm, systol, dysfunct	114
59	disord, epilepsi, neurolog, seizur, psychiatr, clinic, mental	128
60	men, prostat, urinari, urethr, tract, psa, age	94
61	respond, knowledg, data, statist, health, practic, educ	306
62	africa, african, nigeria, west, countri, sub, outbreak	133
63	antibodi, donor, infect, blood, igg, screen, seropreval	215
64	diabet, type, mellitus, control, glucos, tdm, diseas	151
65	fever, lassa, outbreak, diseas, evd, virus, nigeria	122
66	malaria, children, net, day, drug, antimalari, act	198
67	tick, infect, anim, nigeria, infest, cattl, speci	125
68	test, posit, valu, sensit, diagnosi, negat, specif	114
69	data, nigeria, model, estim, countri, diseas, burden	196

70	water, mosquito, snail, speci, collect, sampl, area	157
71	associ, factor, preval, aor, risk, odd, age	208
72	knowledg, student, respond, attitud, questionnair, practic, awar	316
73	cell, sickl, scd, anaemia, sca, diseas, haemoglobin	185
74	farmer, product, farm, diseas, state, market, practic	172
75	lesion, malign, year, histolog, tumour, seen, carcinoma	203
76	vaccin, immun, coverag, measl, state, polio, nigeria	203
77	isol, antibiot, infect, bacteri, resist, cultur, bacteria	366
78	renal, kidney, fistula, dialysi, diseas, transplant, aki	127
79	infertil, hormon, thyroid, tubal, male, fertil, abnorm	99
80	present, clinic, hospit, manag, year, period, age	183
81	liver, level, control, serum, diseas, healthi, signific	106
82	parasit, speci, nigeria, host, preval, infect, veget	105
83	activ, chicken, bird, stress, oxid, antioxid, diseas	120
84	chronic, pylori, ulcer, symptom, gastric, gastrointestin, diseas	125
85	covid, pandem, nigeria, coronavirus, sar, cov, measur	206
86	extract, activ, plant, concentr, antimicrobi, inhibit, mgml	165
87	matern, neonat, birth, deliveri, babi, mortal, death	296
88	cattl, anim, pig, goat, preval, nigeria, slaughter	224
89	skin, drug, effect, onchocerciasi, efficaci, ivermectin, herbal	78
90	tooth, periodont, teeth, extract, dental, loss, molar	80
91	malaria, infect, blood, parasit, falciparum, preval, plasmodium	183
92	concentr, sampl, risk, metal, health, smoke, level	190
93	oral, dental, status, health, children, malnutrit, socioeconom	113
94	loss, hear, ear, media, otiti, impair, cleft	86
95	arteri, surgeri, vascular, heart, year, nigeria, procedur	53
96	art, month, therapi, hiv, antiretrovir, initi, clinic	223
97	mother, hiv, infant, child, transmiss, pmtct, feed	121
98	control, post, compar, effect, baselin, random, applic	36
99	infect, preval, schistosomiasi, urin, area, age, state	219
100	nigeria, human, infect, cov, mer, camel, east	6

4.1 Topic Interpretation

After generating our topics, another task we have to carry out is to make meaning out of the words that have been given to us for each topic. To do this, we used the findthoughts function to match the 100 generated topics to the titles of the articles, so we that could use the topics of the articles to infer what the combined words are saying. We chose not to match the topics to the abstracts because we wanted to see if the topics generated actually aligned with the topics of the articles we analyzed. However, where we were unable to use the matched topics to interpret the generated topics, we matched the generated topics to the abstracts to give a more comprehensive understanding of the generated topic. To do this, we programmed the findthoughts function to match our topics to 5 article topics. From these topics matched, we discerned to find topic interpretation. When we were unable to come to a consensus on a topic interpretation, we use the findthoughts function to match the generated topics to abstracts. Before applying these methods, we confirmed that documents matched by topic titles and abstracts are the same. To ensure this was true, we matched the topics generated and searched them in the Excel document to view their abstracts. After multiple trials, we discovered that the documents of the topic titles generated were the same document generated

when the generated topics are matched with abstracts, and in cases where we do not get clarity from the topics, we read the abstracts manually to help us interpret the generated topics.

Topic 1 covers research works that are surveys on different diseases and their available medicines. Topic 2 covers research works on health care services delivery to communities to combat varies disease; and assessment of performance. Topic 3 covers research works on the causes of certain infectious disease spread. Topic 4 covers research on patients consent and adherence to treatment. Topic 5 covers research on the behavioral patterns of children living with different diseases. Topic 6 covers research on genome analysis in Nigeria. Topic 7 covers research works on medicine extracts from plants and leaves. Topic 8 covers research on female genital infections such as cervical cancer, human papilloma virus. Topic 9 covers research on various treatment of bone injuries. Topic 10 covers research on animal genes and HIV strains in Nigeria.

Topic 11 covers research on chronic kidney disease in Nigeria. Topic 12 covers research on spread of infectious diseases in vegetables. Topic 13 covers research on different diseases such as juvenile dertatomyositis in girls, hypohlycaemic hemiparesis in alcoholics, and subacute cerebral infarcts in males. Topic 14 covers research on transmission and spread of diseases such as simuliumdamnosun, onchocerca infections in rural communities. Topic 15 covers research on cassava mosaic related diseases. Topic 16 covers research on different injuries experienced by Nigerian Northerners. Topic 17 covers research on Anti-inflammatory and Anti-nociceptive effects. Topic 18 covers research on stroke mortality in Nigeria. Topic 19 covers research on HIV patients suffering from other infectious diseases in Nigeria. Topic 20 covers research on issues that occur in the medical wards in hospitals, mostly in rural communities.

Topic 21 covers research on virus gotten from farm produces. Topic 22 covers research on compliance of health workers to use of medical tools. Topic 23 covers research on cytotoxic activities in Nigerian plants. Topic 24 covers research on food safety consumptions on rural areas in Nigeria. Topic 25 covers research on respiratory system diseases in Nigeria. Topic 26 covers research on care givers of chronically sick and mentally unstable people. Topic 27 covers research on vitamin effectiveness in children. Topic 28 covers research on Hepatitis infection and spread in Nigeria. Topic 29 covers research on Sexual risk of being infected with HIV. Topic 30 covers research on different patterns of heart diseases in Nigeria.

Topic 31 covers research that compares which treatment method is more effective for children: drugs or injections. Topic 32 covers research on rotavirus infection among under-5 in Nigeria. Topic 33 covers research on salmonella disease. Topic 34 covers research on parasites common in rivers Benue and Niger, and the Lagos Lagoon. Topic 35 covers research on trend of tuberculosis detection in different states. Topic 36 covers research on Knee and hip disease healing. Topic 37 covers research on depression and Dementia symptoms and how they can be measured. Topic 38 covers research on Molecular detection and characterization of Cryptosporidium. Topic 39 covers research on management and characterization of Ischemic stroke. Topic 40 covers research on assessment of certain diseases in animals such as dogs, squirrel found in Nigeria.

Topic 41 covers research on health research principles and ethics and their development. Topic 42 covers research on simian and Sydney crease and relationships between physical activities and blood pressure. Topic 43 covers research on patterns of prescription of medications to patients by doctors and pharmacist. Topic 44 covers research on different breast diseases in Nigeria. Topic 45 covers research on education of rural communities on diseases such as leprosy and other endemic diseases. Topic 46 covers research on impact of different diseases such as sickle cell, cardiovascular risk factors e.t.c. Topic 47 covers research on nutritional benefits of certain plans such as mushroom, cassava and sorghum in Nigeria. Topic 48 covers research on possible combination of antimalaria plants. Topic 49 covers research on carbapenem resistance from bacterium in hospitals in Nigeria. Topic 50 covers research on different meloidegyne infestation in Nigeria.

Topic 51 covers research on disease prevalence in rodents, dogs and other animals. Topic 52 covers research on causes of visual impairment in Nigeria. Topic 53 covers research on vaginal candidiasis and Vaginosis in pregnant women. Topic 54 covers research on easing accessibility to healthcare services of treatment of diseases such as HIV, DR-TB and others in rural communities. Topic 55 covers research on genetic yield in grains and effect of sesame diseases. Topic 56 covers research on analysis of samples of serum collected in males, and different samples in female. Topic 57 covers research on laparoscopy in Nigeria. Topic 58 covers research Ventricular in patients in Nigeria. Topic 59 covers research on treatments of mental illness. Topic 60 covers research on Prostrate and urinary tract in Nigerian Men.

Topic 61 covers research on awareness creation of infectious diseases such as diabetes mellitus, ebola virus, etc. Topic 62 covers research on disease outbreaks such as foot-and-mouth, pathogenic avian influenza, e.t.c. Topic 63 covers research on prevalent diseases

among blood donors. Topic 64 covers research on diagnosis and complication of diabetes. Topic 65 covers research on measure to control diseases caused by rodents such as Lassa fever etc. Topic 66 covered research on effectiveness of different malarial drugs. Topic 67 covers research on diseases found in cattle's and vegetables in Nigeria. Topic 68 covers research on laboratory reagents used in carrying out test for different diseases. Topic 69 covers research on reviews on mortality, HIV and other diseases in Nigeria. Topic 70 covers research on river water contamination in different parts of the country.

Topic 71 covers research on prevalence of diseases such low back pain, psychiatric morbidity, work related musculoskeletal discomfort, etc. Topic 72 covers research on Health professional knowledge and experience, and Nigerian secondary and tertiary institution students' awareness on infectious diseases. Topic 73 covers research on patients with sickle cell and diseases they suffer from. Topic 74 covers research on crop, aquaculture, bird and animal farmers challenges and management of infectious diseases. Topic 75 covers research on different kind of tumors. Topic 76 covers research on vaccination coverages in rural and urban communities. Topic 77 covers research on bacterial isolation in wounds, urine, and other specimens. Topic 78 covers research on complications and challenges faced in dialysis centers in Nigeria. Topic 79 covers research in different causes of infertility in women. Topic 80 covers research on reviews of management of paralysis related diseases in Nigerian hospitals.

Topic 81 covers research on the assessment of liver enzymes in Nigeria. Topic 82 covers research on Parasitic infections emanating from different places and damage it causes on different parts of the body. Topic 83 covers research on velogenic Newcastle disease. Topic 84 covers research on upper gastrointestinal endoscopy, nasal diseases and chronic simple rhinosinusitis in Nigeria. Topic 85 covers research on the impact of the media of coronavirus education. Topic 86 covers research on antimicrobial activities and properties from plants leaves and stems. Topic 87 covers research on life threatening pregnancies. Topic 88 covers research on prevalence of bovine brucellosis, serological and sero-prevalence in abattoirs and milks from sheep's and goats. Topic 89 covers research on clinical evaluation of different ointments and lotions for the treatment of skin diseases such as scabies etc. Topic 90 covers research on causes an treatments of tooth decay.

Topic 91 covers research on malaria prevalence in difference communities in Nigeria, both urban and rural. Topic 92 covers research on natural minerals found in soils and he health impact of this minerals. Topic 93 covers research on child hygiene observation in schools. Topic 94 covers research on causes of hearing loss in adult and children. Topic 95

covers research on heart snuggeries in Nigeria. Topic 96 covers research on virologic failure and impact treatment. Topic 97 covers research on effectiveness of different HIV treatments. Topic 98 covers research on control of parasite and disease infestation in plants. Topic 99 covers research on urinary-tract infection among different communities. Topic 100 covers research on coronavirus infect and spread according to categories of people and what the nature of their jobs.

Conclusion

The paper considered a systematic literature review of infectious diseases in Nigeria using topic modelling which is an aspect of machine learning. Specifically, it considered the period 2002 to 2022. Having done the initial preprocessing, a structural topic modelling approach was used. Up to 15,459 publications were accessed from the Scopus database, teasing out their corresponding abstracts, keywords , year and title. HIV/AIDS and Malaria were diseases most researched in Nigeria. Up to 100 structural topic models were obtained and discussed in the paper given the number of documents they reflect in, like topic 29 which has a connection to about 415 documents, and topic 100 which has just 6 documents it is linked to. The second objective of this work was to identify diseases that have been most researched in the past two decades. This work has identified diseases that have been most discussed in the past two decades by researchers and has also presented diseases that have not been discussed in large volume by researchers. To achieve this, the work relied on the keywords to determine how often a disease was considered as part of the primary discussion. By meeting the second objective, this work met its third objective, which is to inform researchers of diseases that have been least researched in the past two decades. This was achieved by identifying the diseases with the least appearance in the large volume of articles analyzed. This work has also made further reviews easy for researchers by identifying the documents associated with each topic.

In terms of limitation encountered, this work is unable to tell us how much each document connected to a topic contributes to that topic. This means that we are unable to measure how relevant each associated document is to our generated topic. Second, while we can determine the number of documents that are linked to each topic, we are unable to tell if the topics with larger documents connected to them are better captured than those with smaller documents associated with them, and the reverse applies to both. Finally, the topics given that the corpus is very large gives us only general overview, without little depth to the contents of the topics.

Despite having its limitations, this work presents some findings that researchers would build on to guide them into a more indebt analysis of the generated topics. We recommend that in the future, some researchers may choose to pick one of the 100 topics generated, sift out the documents associated with it, and carry out an in-depth review. This also presents an avenue for researchers with information on diseases that have not been explored in much depth that they should look into [26]. Further work could harness the balance between AI-assisted innovation and human capability by including more keywords derived from AI.

Reference

1. Adagbada, A. O., Adesida, S. A., Nwaokorie, F. O., Niemogha, M. T., & Coker, A. O. (2012). Cholera epidemiology in Nigeria: an overview. Pan African Medical Journal, 12(1).
2. Asmussen, C. B. and Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. Journal of Big Data, 6(1), 1-18.
3. Binagwaho, A. and Mathewos, K. (2022). Infectious disease outbreaks highlight gender inequity. Nature Microbiology, 7(3), 361-362.
4. Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. Journal of Open Source Software, 3(30), 774-774.
5. Bruce-Chwat, L. J. (1951). Malaria in Nigeria. Bulletin of the World Health Organisation 1951, 4: 301-327.
6. Burnham, J. F. (2006). Scopus database: a review. Biomedical digital libraries, 3(1), 1-8.
7. Chukwuonye, I. I., Ogah, O. S., Anyabolu, E. N., Ohagwu, K. A., Nwabuko, O. C., Onwuchekwa, U., ... &Oviasu, E. (2018). Prevalence of chronic kidney disease in Nigeria: systematic review of population-based studies. International journal of nephrology and renovascular disease, 165-172.
8. De Cock, K. M., Nasidi, A., Enriquez, J., Craven, R. B., Okafor, B. C., Monath, T. P., ... &Sorungbe, A. (1988). Epidemic yellow fever in eastern Nigeria, 1986. The Lancet, 331(8586), 630-633.
9. DiMaggio, P., Nag, M. and Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. Poetics, 41(6), 570-606.
10. Elgesem, D., Steskal, L. and Diakopoulos, N. (2019). Structure and content of the discourse on climate change in the blogosphere: The big picture. In Climate Change Communication and the Internet (pp. 21-40). Routledge.
11. García, M., Maldonado, S. and Vairetti, C. (2021). Efficient n-gram construction for text categorization using feature selection techniques. Intelligent Data Analysis, 25(3), 509-525.
12. Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political analysis, 21(3), 267-297.
13. Ike, S. O., & Onyema, C. T. (2020). Cardiovascular diseases in Nigeria: What has happened in the past 20 years? Nigerian Journal of Cardiology, 17(1), 21-26.

14. Kabuga, A. I. and El Zowalaty, M. E. (2019). A review of the monkeypox virus and a recent outbreak of skin rash disease in Nigeria. Journal of Medical Virology, 91: 533-40.
15. Madakam, S., Holmukhe, R. M., & Jaiswal, D. K. (2019). The future digital work force: robotic process automation (RPA). JISTEM-Journal of Information Systems and Technology Management, 16.
16. Mimno, D., and Lee, M. (2014). Low-dimensional embeddings for interpretable anchor-based topic inference. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1319-1328).
17. Mohammed, I., Nasidi, A., Alkali, A. S., Garbati, M. A., Ajayi-Obe, E. K., Audu, K. A., ... & Abdullahi, S. (2000). A severe epidemic of meningococcal meningitis in Nigeria, 1996. Transactions of the Royal Society of Tropical Medicine and Hygiene, 94(3), 265-270.
18. Nasidi, A., Monath, T. P., DeCock, K., Tomori, O., Cordellier, R., Olaleye, O. D., ... &Oyediran, A. B. O. (1989). Urban yellow fever epidemic in western Nigeria, 1987. Transactions of the Royal Society of Tropical Medicine and Hygiene, 83(3), 401-406.
19. NCDC (2023). Nigerian Centre for Disease Control Cholera Situation Report. Monthly Epidemiological Report, 11, Epidemiological week 48 – 52 (28 November to 1 January 2023) Available from: file:///Users/Adeda/Downloads/An%20update%20of%20Cholera%20outbreak%20in%20Nigeria_221222_52.pdf [Accessed 7, Sept. 2023]
20. Njidda, A. M., Oyebanji, O., Obasanya, J., Ojo, O., Adedeji, A., Mba, N., ... &Ihekweazu, C. (2018). The Nigeria centre for disease control. BMJ global health, 3(2), e000712.
21. Nnadi, C., Oladejo, J., Yennan, S., Ogunleye, A., Agbai, C., Bakare, L., ... &Ihekweazu, C. (2017). Large outbreak of Neisseria meningitidis serogroup C—Nigeria, December 2016–June 2017. Morbidity and mortality weekly report, 66(49), 1352.
22. Okoroiwu, H. U., Umoh, E. A., Asanga, E. E., Edet, U. O., Atim-Ebim, M. R., Tangban, E. A., ... &Povedano-Montero, F. J. (2022). Thirty-five years (1986–2021) of HIV/AIDS in Nigeria: bibliometric and scoping analysis. AIDS Research and Therapy, 19(1), 1-15.
23. Olumade, T. J., Adesanya, O. A., Fred-Akintunwa, I. J., Babalola, D. O., Oguzie, J. U., Ogunsanya, O. A., ... &Osasona, D. G. (2020). Infectious disease outbreak preparedness and response in Nigeria: history, limitations and recommendations for global health policy and practice. AIMS Public Health, 7(4), 736.
24. Otu, A., Ameh, S., Osifo-Dawodu, E., Alade, E., Ekuri, S. and Idris, J. (2018). An account of the Ebola virus disease outbreak in Nigeria: implications and lessons learnt. BMC public health, 18(1), 1-8.
25. Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H. and Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. American Journal of Political Science, 54(1), 209-228.
26. Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems.
27. Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder□Luis, J., Gadarian, S. K., ... and Rand, D. G. (2014). Structural topic models for open□ended survey responses. American journal of political science, 58(4), 1064-1082.

28. Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. Journal of Statistical Software, 91, 1-40.
29. Salvagno, M., Taccone, F. S., and Gerli, A. G. (2023). Can artificial intelligence help for scientific writing?. Critical care, 27(1), 1-5.
30. Siwatu, G. O., Palacios-Lopez, A., Mcgee, K. R., Amankwah, A., Vishwanath, T., & Azad, M. (2020). Impact of covid-19 on Nigerian households: baseline results.
31. WHO (2022). World Malaria Report 2012. Available from: https://www.who.int/publications/i/item/9789241564533.[Accessed 12, Oct. 2023].

Statement of Ethics:

The study being a systematic review of literature did not need the approval of a formal ethical committee because the data used was sourced from the Scopus database.

Conflict of Interest Statement:

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Sources:

This work did not receive any funding.

Author contribution:

AYANTSE, Cornelius and YAYA, S. Olaoluwa, conceptualized this idea, sourced, and collated the data, determined the methodology, carried out the analysis, and wrote the first draft of the paper. Together all the authors contributed to reviewing and finalizing the article.

Data availability statement:

This data is a public data accessing on Scopus data base.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

Option 2:

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc. have been used during the writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1.

2.

3.

Appendix 1

Table 1: Work that produced table 1

From Table 1 above, looking at the unigram column, the word anemia appeared in the form's anaemia (150) and anemia (88), making a total of 238. The word disease appeared in the form's disease (951) and diseases (314), making a total of 1265. The word infection takes the form's infection (448), infections (185), co-infection (106), and infectious (81), making it a total of 820. The word pregnancy appeared in the form's pregnancy (248) and pregnant (132), making it a total of 380. The word malaria appeared in the form's malaria (540) and falciparum (146), making it a total of 686. The word diabetes appeared as diabetes (274) and mellitus (138), making it a total of 412 The word HIV appeared in the form's HIV (1207) and aids (392), making it a total of 1599. The word cancer appeared as cancer (346) and carcinoma (48), making it a total of 394. The word blindness appeared as blindness (51) and onchocerciasis (49), making it a total of 100. The word Covid-19 appeared in the form's Covid-19 (375) and coronavirus (63), making it a total of 438. The word tuberculosis appeared in the form's tuberculosis (250) and tb (50), making it a total of 300. The word injury appeared in the form injury (78) and wound (47), making it a total of 125.

From the Table 2, looking at the bigram column, the sentence HIV aids appeared in the form's HIV aids (304), HIV Nigeria (141), human immunodeficiency (125), immunodeficiency virus (119), aids Nigeria (71), HIV infection (50), heart HIV (29), acquired immunodeficiency (27), immunodeficiency syndrome (26), HIV HIV (24), living HIV (24) virus HIV (24), therapy HIV (24) HIV malaria (24) HIV testing (22), HCV HIV (21), children HIV (18), hepatitis HIV (17) HIV knowledge (16), HIV human (16), aids HIV (15), co-infection HIV (14), HIV positive (14), Africa HIV (13), HIV status (13), HIV prevention (13), epidemiology HIV (13), condom HIV (13), cancer HIV (12), and HBV HIV (12), making it a total of 1252. The sentence mental health appeared in the form's mental health (81), mental illness (34) and mental disorder (14), making it a total of 129. The sentence Infection Nigeria appeared in the form's infection Nigeria (46), infection control (28), infectious disease (26) tract infection (49), tract infections (20), infection prevention (16) and infection prevalence (16), making it a total of 201. The sentence malaria Nigeria appeared in the form's Nigeria malaria (115), plasmodium falciparum (112) Nigeria plasmodium (37), plasmodium berghei (34), Malaria plasmodium (27), falciparum malaria (25), HIV malaria (24), malaria pregnancy (24), asymptotic malaria (21), malaria parasitemia (19), severe malaria (18), malaria prevalence (16) and knowledge malaria (14), making it a total of 486. The sentence diabetes mellitus appeared in the form's diabetes mellitus (135), type diabetes (70) and diabetes foot (11), making it a total of 216 documents. The sentence sickle cell appeared in the form's sickle cell (278), cell anaemia (56) an cell anemia (40), making it a total of 374. The sentence pregnant women appeared in the form's pregnant woman (123), Nigeria pregnant (40) and pregnancy prevalence (22), making it a total of 185. The sentence coronavirus covid-19 appeared in form's coronavirus covid-19 (29), covid-19 Nigeria (34), covid-19 health (25), covid-19 pandemic (17), covid-19 healthcare (17), coronavirus disease (16), and covid-19 covid-19 (14), making it a total of 152. The sentence disease Nigeria appeared in the form's disease Nigeria (59) non-communicable diseases (27), infectious disease (26), disease virus (20), tropical diseases (15), disease severity (14), diseases Nigeria (12), disease control (13) chronic disease (12) and disease burden (12), making it a total of 210. The sentence virus Nigeria appeared in the form's virus Nigeria (53) and virus humas (13) making a total of 66. The sentence mycobacterium tuberculosis appeared in the forms mycobacterium tuberculosis (25), Nigeria tuberculosis (20) and pulmonary tuberculosis (22), making it a total of 67. The sentence anxiety depression appeared in the form's anxiety depression (13) and depression Nigeria (12), making it a total of 25 documents. The sentence Hepatitis virus appeared in the form's hepatitis virus (89),

hepatitis hepatitis (21), virus hepatitis (18), and hepatitis surface (18), making it a total of 146. The sentence Nigeria cancer appeared in the form's cancer screening (23), cancer risk (18), cancer Nigeria (17), prostrate cancer (32), breast cancer (91), cervical cancer (81), hepatocellular carcinoma (15), cancer cervical (13), cancer HIV (12), making it a total of 301. The sentence kidney disease appeared in the form's kidney disease (96), kidney injury (17) and acute kidney (16), making it a total of 129 documents. The sentence human papillomavirus (28), human papilloma (17) and papilloma virus (17), making it a total of 62. The sentence Nigeria polio appeared in the form's Nigeria polio (16) and polio eradication (12), making it a total of 28. The sentence Lassa virus appeared in the form's Lassa virus (16) and Lassa fever (55) making it a total of 71. The sentence hypertension Nigeria appeared in the form's hypertension Nigeria (48) and hypertension left (15), making it a total of 63. The sentence mosaic virus appeared in the form's mosaic virus (14), cassava mosaic (14) and mosaic disease (13), making it a total of 41. The sentence flaccid paralysis appeared in the form's flaccid paralysis (14) and acute flaccid (13) making it a total of 27.

From Table 3, in the trigram columns, the sentence human immunodeficiency virus appeared in the form's, human immunodeficiency virus (119), HIV aids Nigeria (70), acquired immunodeficiency syndrome (26), antiretroviral therapy HIV (22), HIV aids knowledge (21), people living HIV (19), immunodeficiency virus Nigeria (17), immunodeficiency virus HIV (14), HBV HCV HIV (12), living HIV aids (12), virus human immunodeficiency (10), quality life HIV (9), HIV infection Nigeria (9), acquired immunodeficiency syndrome (8), immune deficiency syndrome (8), cervical cancer HIV (8), therapy HIV aids (8), heart HIV aids (8), HIV knowledge Nigeria (8), hepatitis virus HIV (8), HIV testing Nigeria (7), aids knowledge Nigeria (7), immunodeficiency virus acquired (6), HIV human immunodeficiency (6), HIV status Nigeria (6), virus acquired immunodeficiency (6), mother-to-child transmission HIV (6), HIV Nigeria pmtct (6), HIV care continuum (5), HIV drug resistance (5), HIV counselling testing (5), health system HIV (5), HIV Nigeria pregnancy (5), cell count HIV (5), HBV HIV Nigeria (5), HIV Nigeria tuberculosis (5), hepatitis hepatitis HIV (5), condom HIV aids (5)m immunodeficiency virus infection (5), attitude HIV aids (5),

attitudes HIV aids (5), HIV aids prevalence (5), HIV aids sexual (5), hepatitis human immunodeficiency (4), making it a total of 547. The sentence sickle cell disease appeared in the form's sickle cell disease (143), sickle cell anaemia (56), sickle cell anemia (40), Nigeria sickle cell (38), screening sickle cell (11), sickle cell trait (10), disease sickle cell (7), anemia sickle cell (5), sickle cell stroke (5), prevent sickle cell (5), cell disease sickle (5), anaemia sickle cell (5), iron deficiency anaemia (5) and care sickle cell (4), making it a total of 339. The sentence malaria Nigeria plasmodium appeared in the form's Nigeria plasmodium falciparum (31), malaria plasmodium falciparum (13), malaria Nigeria plasmodium (10), plasmodium falciparum Nigeria (7), berghei plasmodium falciparum (7) birth weight malaria (6), plasmodium berghei plasmodium (6), malaria Nigeria parasitemia (6), malaria pregnant women (5), prevalence treatment malaria (5), malaria Nigeria pregnancy (5), making it a total of 101. The sentence chronic kidney disease appeared in the form's chronic kidney disease (77), acute kidney injury (16), kidney disease Nigeria (8), HIV kidney disease (5), kidney injury children (4), making it a total of 110. The sentence urinary tract infection appeared in the forms, urinary tract infection (42), urinary tract infections (11), urinary tract symptoms (6), bacteriuria urinary tract (4), making it a total of 63. The sentence Nigeria pregnant women appeared in the form's Nigeria pregnant women (38), pregnant women prevalence (16), Nigerian pregnancy prevalence (9), pregnant women seroprevalence (9), making it a total of 72. The sentence Hepatitis virus Nigeria appeared in the form's hepatitis virus hepatitis (17), virus hepatitis virus (15), hepatitis surface antigen (15), hepatitis virus Nigeria (12), hepatitis virus human (11), hepatitis virus HIV (8), hepatitis virus infection (5), antigen hepatitis virus (5), surface antigen hepatitis (5), hepatitis hepatitis HIV (5), genotype hepatitis virus (4), hepatitis human immunodeficiency (4), making it a total of 118. The sentence type diabetes mellitus (27), diabetes mellitus Nigeria (10), diabetes mellitus hypertension (7), Nigeria type diabetes (6), diabetes mellitus type (6), making it a total of 56. The sentence

blood pressure control appeared in the forms, blood pressure control (10), blood pressure hypertension (9), ambulatory blood pressure (4), making it a total of 23. The sentence hypertension Nigeria prevalence appeared in the form's hypertension left ventricular (12), blood pressure hypertension (9), hypertension Nigeria prevalence (6), pressure control hypertension (5), making it a total of 32. The sentence covid-19 Nigeria pandemic appeared in the form's covid-19 knowledge Nigeria (8), covid-19 healthcare worker (8), covid-19 Nigeria pandemic (7), coronavirus disease covid-19 (6), anxiety covid-19 depression (5), attitude covid-19 knowledge (5), making it a total of 39. The sentence sexually transmitted infections appeared in the forms sexually transmitted infections (24), sexually transmitted infection (8), sexually transmitted diseases (6), sexually transmitted disease (5), making it a total of 43. The sentence cervical cancer screening appeared in the form's cervical cancer screening (13), cervical cancer cervical (8), cervical cancer human (7), cervical cancer knowledge (7), awareness cervical cancer (7), cancer cervical cancer (6), invasive cervical cancer (4), making it a total of 52. The sentence suppurative otitis media appeared in the form's suppurative otitis media (10), sensorineural hearing loss (8), chronic suppurative otitis (8), making it a total of 26. The sentence neglected tropical disease appeared in the form's neglected tropical diseases (12), neglected tropical disease (7), tropical disease Nigeria (6), disease virus Nigeria (4), peadiatric infectious disease (4), infectious diseases immunization (4), making it a total of 37. The sentence Lassa fever Nigeria appeared in the form's Lassa fever Nigeria (10), fever Lassa virus (7), knowledge Lassa fever (5), control Lassa fever (5), Lassa virus Nigeria (4), Lassa fever mastomys (4), making it a total of 35. The sentence avian influenza Nigeria appeared in the form's avian influenza Nigeria (7), pathogenic avian influenza (9), avian influenza h5nl (5), and influenza virus Nigeria (4), making it a total of 25 documents. The sentence mental illness Nigeria appeared in the form's mental illness Nigeria (9), and countries mental health (6), making it a total of 15. The sentence Newcastle disease

Nigeria appeared in the form's Newcastle disease Nigeria (9) and New castle diseases (7), making it a total of 16. The sentence respiratory tract infections appeared in the form's acute respiratory infections (6), respiratory tract infection (8), respiratory syncytial virus (5), respiratory tract infection (5) and acute respiratory infection (4), making it a total of 28 documents. The sentence chronic renal failure appeared as chronic renal failure (6), and stage renal disease (6), making it a total of 12. The sentence oral polio virus appeared in the form's oral polio virus (7) and polio eradication initiative (5), making it a total of 12. The sentence cardiovascular disease risk appeared in the form's cardiovascular disease risk (5) and cardiovascular disease chronic (4), making it a total of 9. The sentence Nigeria prostrate cancer appeared in the form's Nigeria prostrate cancer (7), prostrate specific antigen (5), prostrate cancer prostrate (5), making it a total of 17 documents. The sentence rift valley fever appeared in the form's rift valley fever (6), valley fever virus (5), making it a total of 11. The sentence stroke risk factor appeared in form's cell disease stroke (6), stroke risk factor (5), sickle cell stroke (9), quality life stroke (5), making it a total 25. The sentence human papilloma virus appeared in the form's human papilloma virus (16), human papillomavirus Nigeria (5), human papillomavirus vaccine (4), making it a total of 25. The sentence coronary heart disease appeared in the form's coronary heart disease (8), and heart disease Nigeria (5) and congenital heart disease (22), making it a total of 35. The sentence foot mouth disease appeared in the form's foot mouth disease (5) and foot-and-mouth disease virus (4), making it a total of 9. The sentence mycobacterium tuberculosis Nigeria appeared in the form's mycobacterium tuberculosis Nigeria (8), HIV Nigeria tuberculosis (5), Nigeria pulmonary tuberculosis (5), treatment outcome tuberculosis (4) and rifampicin resistance tuberculosis (4), making it a total of 26 The sentence lymphatic filariasis Nigeria appeared in the form's lymphatic filariasis Nigeria (5), and lymphatic filariasis lymphoedema (5), making it a total of 10.

Appendix II

Scopus Search Query:

TITLE-ABS-KEY ((infect* OR disease) AND (nigeri*)) AND (LIMIT-TO (PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015) OR LIMIT-TO (PUBYEAR , 2014) OR LIMIT-TO (PUBYEAR , 2013) OR LIMIT-TO (PUBYEAR , 2012) OR LIMIT-TO (PUBYEAR , 2011) OR LIMIT-TO (PUBYEAR , 2010) OR LIMIT-TO (PUBYEAR , 2009) OR LIMIT-TO (PUBYEAR , 2008) OR LIMIT-TO (PUBYEAR , 2007) OR LIMIT-TO (PUBYEAR , 2006) OR LIMIT-TO (PUBYEAR , 2005) OR LIMIT-TO (PUBYEAR , 2004) OR LIMIT-TO (PUBYEAR , 2003) OR LIMIT-TO (PUBYEAR , 2002)) AND (LIMIT-TO (SUBJAREA , "MEDI") OR LIMIT-TO (SUBJAREA , "IMMU") OR LIMIT-TO (SUBJAREA , "BIOC") OR LIMIT-TO (SUBJAREA , "AGRI") OR LIMIT-TO (SUBJAREA , "PHAR") OR LIMIT-TO (SUBJAREA , "VETE") OR LIMIT-TO (SUBJAREA , "ENVI") OR LIMIT-TO (SUBJAREA , "NURS") OR LIMIT-TO (SUBJAREA , "MULT") OR LIMIT-TO (SUBJAREA , "HEAL") OR LIMIT-TO (SUBJAREA , "NEUR") OR LIMIT-TO (SUBJAREA , "PSYC") OR LIMIT-TO (SUBJAREA , "CHEM") OR LIMIT-TO (SUBJAREA , "DENT")) AND (LIMIT-TO (PUBSTAGE , "final")) AND (LIMIT-TO (AFFILCOUNTRY , "Nigeria")) AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English"))

Appendix III

R code:

```
setwd("/Users/Adeda/Documents")
library(tm)
library(tidyverse)
library(topicmodels)
library(tidytext)
library('ggraph')
library("stm") # Used to build structural topic Models.
library(NLP) #used for analyzing text data
library("quanteda") # used for preparing data for LDA
library("stopwords") # generates a set of stopwords to be removed in the text
library("tm")   # uses the corpus as its main structure
library("quanteda.textstats") # used to get the stats of text data
library("quanteda.textplots") # used to plot graphs.
library("slam")
library("geometry")
library("Rtsne")
library("rsvd")


data =  read.csv("Infeceous_Diseases_Data3.csv", header = T)  #importing the data to be used

colnames(data)
```

```
datam = data[, c("Year", "Title", "Abstract", "Author.Keywords")] # extracting the variables to be used

summary(datam)


# Part One of the Work
# select the author keywords variable and create a corpus
a.Korpus = corpus(datam$Author.Keywords)

# create tokens and clean the data
a.token_corpus = token_corpus = tokens(a.Korpus, what="word",
remove_numbers = TRUE,
remove_punct = TRUE,
remove_symbols = TRUE,
remove_separators = T,
remove_hyphines=T)
# lower letters
a.token_corpus = tokens_tolower(a.token_corpus)

SW = stopwords("en", source = "stopwords-iso")

# remove stopwords
a.token_corpus = tokens_select(a.token_corpus, stopwords(source = "smart"),
                        selection= "remove")
a.token_corpus = tokens_select(a.token_corpus, SW,
                        selection= "remove")
a.token_corpus = tokens_select(a.token_corpus, stopwords(),
                        selection= "remove")
#create dfm
a.corpus_dfm = dfm(token_corpus)

# check for frequent words
textfreq = textstat_frequency(a.corpus_dfm, n=300)
View(textfreq[,c("feature","docfreq")])

# set seed and plot cloud
set.seed(123)
textplot_wordcloud(a.corpus_dfm, min.freq = 100)

# create bigrams to repeate the same action for authro keywords avriable
bigram = tokens_ngrams(a.token_corpus, n=2L, concatenator = " ")

#dfm for bigram
bdfm = dfm(bigram)
btextfreq = textstat_frequency(bdfm, n=200)

view(btextfreq[,c("feature","docfreq")])

plot(btextfreq["feature"], btextfreq["docfreq"])
```

```
btextfreq("docfreq")
#create trigram to repaete the same action for author keywords
trigram = tokens_ngrams(a.token_corpus, n=3L, concatenator = " ")

#dfm for trigram
tdfm = dfm(trigram)
ttextfreq = textstat_frequency(tdfm, n=200)

view(ttextfreq[,c("feature","docfreq")])

k = as.data.frame(table(datam$Year))

ggplot(k, aes(x = Var1, y=Freq, fill = Var1)) +
geom_bar(stat = "identity") +
coord_flip()+
geom_text(aes(label = Freq), vjust = 0)+
labs(title = "Annual Publications Bargraph ", y="Number of publications per year", x="Year")

# Part Two of the work

library(tm)
library(tidyverse)
library(topicmodels)
library(tidytext)
library('ggraph')
library("stm") # Used to build structural topic Models.
library(NLP) #used for analyzing text data
library("quanteda") # used for preparing data for LDA
library("stopwords") # generates a set of stopwords to be removed in the text
library("tm")   # uses the corpus as its main structure
library("quanteda.textstats") # used to get the stats of text data
library("quanteda.textplots") # used to plot graphs.
library("slam")
library("geometry")
library("Rtsne")
library("rsvd")

getwd()
setwd("/Users/Adeda/Documents")
data =  read.csv("Infeceous_Diseases_Data3.csv", header = T)  #importing the data to be used

colnames(data)
datam = data[, c("Year", "Title", "Abstract", "Author.Keywords")] # extracting the variables to be used

summary(datam)

corpus = Corpus(VectorSource(datam$Abstract)) # create a corpus
corpus
```

```
corpus <- tm_map(corpus, content_transformer(tolower))
dtm<- DocumentTermMatrix(corpus)

#create custom stopwords to further clean the data set
words = c("administered", "achieved","adverse","accordingly","better",
        "both","between","cannot","case","cases","cohort","cohorts",
        "common","confirmed","depending","diagnosed","did",
        "difference","different","does","each","either",
        "eligible","even","events","every", "excluded","experienced",
        "group","groups","having","higher","included","ineligible",
        "intervention","interventions","least","less","lower","many",
        "may","minimum","more","most","must","none", "number",
        "observed","occurred","one","outcome","over","part",
        "participant","participate", "participants","participating",
        "participation","patient","patients","primary","prior",
        "receive","received","receiving","recent","recently","record",
        "recorded","report","reported","require","required","requires",
        "response","results","same","secondary","several","still",
        "study","subjects","suffered","take","treated","treatment",
        "trial","types","undergo","unit","use","used","uses","who",
        "will","would","another","afterwards","against","among","amongst",
        "analyze","anything","anyhow","anywhere","anyone","applicable",
        "apply","arise","around","assume","become","because","become",
        "becomes","becoming","before","beforehand","beside","besides",
        "between","beyond","both","but","came","compare","could","dealing",
        "department","depend","discover","does","done","during","effected",
        "either","elsewhere","enough","especially","ever","every","everyone",
        "everything","everywhere","except","find","found","further","gave",
        "give","having","hence","here","hereafter","hereby","herein",
        "hereupon","moreover","must","much","mostly","moreover","might",
        "meanwhile","mainly","latterly","latter","investigate","indeed",
        "important","importance","immediately","however","himself",
        "herself","myself","namely","necessarily","neither","nevertheless",
        "next","nobody","normally","noted","nothing","obtain","often",
        "others","other","ought","ourselves","overall","owing",
        "particularly","perhaps","predominantly","precede","present",
        "previously","promptly","quickly","quite","rather","readily",
        "really","recent","thereupon","therein","therefore","thereby",
        "thereafter","thence","themselves","their","theres","take",
        "sufficiently","such","substantially","studied","strongly",
        "somewhere","somewhat","sometimes","something","someone",
        "somehow","slightly","since","significantly","shows",
        "showed","should","seriously","seems","relate","regarding",
        "though","through","throughout","together","toward","towards",
        "under","unless","until","useful","usefulness","using","using",
        "various","whatever","whence","whenever","whereafter","whereas",
        "whereby","wherein","whither","whoever","whom","within","without",
        "yourselves","yourself")

# we use the textProcessor function to clean and prepare our text data for analysis
```

```
poliblog5k.proc <- textProcessor(documents=datam$Abstract,
                        metadata = data,
                        lowercase = TRUE, #*
removestopwords = TRUE, #*
removenumbers = TRUE, #*
removepunctuation = TRUE, #*
                        stem = T, #*
wordLengths = c(3,Inf), #*
sparselevel = 1, #*
                        language = "en", #*
                        verbose = TRUE, #*
onlycharacter = TRUE,
customstopwords = words, #*
                        v1 = FALSE) #*


# the prepdocument command converts the cleaned data into a data structure that can be manioulated using the STM package
out <- prepDocuments(poliblog5k.proc$documents, poliblog5k.proc$vocab, poliblog5k.proc$meta, lower.thresh=5)



# This command helps us to find the number of topics to be generated from the data
storage <- searchK(out$documents, out$vocab, K = 0,data = out$meta, max.em.its = 10, init.type = "Spectral", seed=250)

# we compare different topic numbers to see which is better and suitable for our data
#searchk<- searchK(out$documents, out$vocab, K = c(40, 50,60,80,90,100,10,120,130),data = out$meta, max.em.its =1, init.type = "Spectral", seed=250)
#plot(searchk, main = "Diagnostic Value by Number of Topics")

# We build our model using 100 topics as intuited from the searchK function
poliblogPrevFit<- stm(documents = out$documents, vocab = out$vocab,
              K = 100, max.em.its = 10,
              data = out$meta, init.type = "Spectral", seed=120)

# we view the top 30 words in the 100 topics generated
as.data.frame(t(labelTopics(poliblogPrevFit, n = 10)$prob))

# this helps us get the summary of the fitted model
summary(poliblogPrevFit)


# we use this to find the titles of articles that are close in relation to each 100 ropics
thoughts3<-findThoughts(poliblogPrevFit,texts=datam$Title, n=5,topics=1:100)
thoughts3
```

```
# we find the documents that match each of the 100 topics
theta_100k = make.dt(poliblogPrevFit)

data$Rank = NA

for (i in 1:nrow(datam)){
  column = theta_100k[i,-1]
maintopic = colnames(column) [which(column== max(column))]
data$Rank[i] = maintopic
}

table(data$Rank)
```