MALARIA PREVENTION USING SOCIAL MEDIA AND TEXT MINING

Ibrahim Umar

(40805)

A Thesis submitted to the Faculty of Computer Science at the African University of Science and Technology

In Partial Fulfilment of the Requirements for the degree of Master of Science in the

Computer Science Department

JULY 2021

African University of Science and Technology [AUST]

Knowledge is Freedom

APPROVAL BY

Supervisor

Surname: Prasad

First name: Rajesh

Signature 

The Head of Department

Surname: Prasad

First name: Rajesh

Signature 

ABSTRACT

The battle with malaria especially in the African continent still exists and has been taking the lives of many in the area, so there is a need to keep fighting the battle, monitor progress and challenges. One way is the usage of social media particularly twitter as a tool to fight malaria. Research has been done on malaria twitter data to classify tweets as malaria and non-malaria cases using support vector machine (SVM) which is used in the predictions of future tweets to avoid outbreaks. Malaria twitter data has also been used to find trends and patterns on public opinions regarding malaria topics which is used by health sectors in managing funds allocation and making informed decisions. The objective of this study is to tap into Nigerian Malaria twitter data to understand public opinions of tweets relating to malaria, gain insight into the data to find trends and patterns and compare results with WHO battle against malaria. We describe a combine approach of sentiment analysis, word cloud and topic modelling using LDA. The sentiment analysis is for assessing public opinion about malaria in Nigeria. Word cloud for data visualization and LDA to find hidden topics which is compared to WHO fight against malaria. Despite the small size of the data set, the word cloud visualized topics with the highest frequency and this could be labelled as topics creating awareness on malaria, malaria treatment, testing before treating malaria and the goal of having a malaria free Nigeria. The LDA result correlated well with WHO's battle against malaria and issues the battle is still facing like adverse effect of malaria on pregnant women and young children under age 5. The sentiment analysis provided us sentiment and public opinion of tweets with 42.6% positive, 15.6% negative and 41.8% neutral.

DEDICATION

This work is dedicated to the whole world, particularly Africa dealing with the problems of malaria.

ACKNOWLEDGEMENT

I acknowledge the presence of Almighty Allah who has made this project a success. A big thanks to my supervisor who has assisted me with all the necessary guidance, time and patience. I also wish to thank African University of Science and Technology (AUST) for the full scholarship that had made my M.Sc. dream come to realization.

TABLE OF CONTENTS

ABSTRACT ii

DEDICATION iii

ACKNOWLEDGEMENT iv

TABLE OF CONTENTS v

LIST OF FIGURES viii

CHAPTER ONE 1

1.0 Introduction 1

1.1 Problem Statement 1

1.2 Aim and Objectives 1

1.3 Research questions 2

1.4 Significance of the study 2

1.5 Definition of terms 2

1.6 The scope 3

CHAPTER TWO 4

Literature review 4

2.0 Introduction 4

2.1 Malaria 4

2.2 Twitter 5

2.3 Data mining 6

2.3.1 Text mining 7

2.4 Word Cloud 7

2.5 Sentiment analysis 8

2.6 Topic Modelling 8

2.7 Latent Dirichlet Allocation (LDA) 9

2.7.1 Notations used in describing LDA 10
2.8 Related Works 11
2.9 Research Gap 12
2.10 Significance of the Study 12
CHAPTER THREE 13
3.0 Introduction 13
3.1 Proposed model 13
3.1.1 Data collection 14
3.1.2 Data Pre-processing 14
3.1.3 Data Cleaning 14
3.2 Word cloud 15
3.3 Sentiment Analysis 15
3.3.1 Subjectivity, polarity, and Analysis 16
3.4 Latent Dirichlet Allocation (LDA) implementation 16
3.4.1 Lemmatization 16
3.4.2 Tokenization 16
3.4.3 Dictionary generation 17
3.4.4. Document term matrix (DTM) 17
3.4.5 LDA Topics 17
3.4.6 Visualization of the LDA 17
CHAPTER FOUR 18
Result and discussion 18
4.0 Introduction 18
4.1 Word Cloud 18
4.2 Sentiment Analysis 19

4.3 Visualization of sentiment analysis result 19
4.4 Percentage of tweets (positive, negative and neutral) 20
4.5 Bar Graph of sentiment analysis 21
4.6 Latent Dirichlet Allocation 21
4.7 Discussions 27
CHAPTER FIVE 30
Conclusion, recommendation, and future work 30
5.0 Conclusion 30
5.1 Recommendation 30
5.2 Future work 30
REFERENCES 31

LIST OF FIGURES

Figure 2.1 Source: Jiawei Han and Micheline Kamber (2011), Data Mining: Concepts and Techniques, Third Edition, Elsevier 22 6
Figure 3.1 Proposed model for the research work 13
Figure 3.2 Sample of the data set 14
Figure 3.3 Sample screenshot of the cleaned and processed tweets 15
Figure 3.4 Tweets with their subjectivity, polarity and analysis 16
Figure 3.5 Sample tweet before tokenization 16
Figure 3.6 Sample Tweet after Tokenization 17
Figure 4.1 Word cloud generated from the cleaned and processed data 18
Figure 4.2 Tweets with their subjectivity, polarity and analysis 19
Figure 4.3 Scatter Plot showing the Polarity and Subjectivity of Tweets 19
Figure 4.4 Percentage of Positive Tweets 20
Figure 4.5 Percentage of Negative Tweets 20
Figure 4.6 Percentage of Neutral Tweets 20
Figure 4.7 Bar Graph for the Positive, Neutral and Negative Tweets 21
Figure 4.8 Topic generated by the LDA 21
Figure 4.9 Diagrammatic view of the 5 generated topics 22
Figure 4.10 Topic 1 23
Figure 4.11 Topic 2 24
Figure 4.12 Topic 3 25
Figure 4.13 Topic 4 26
Figure 4.14 Topic 5 27

CHAPTER ONE

1.0 Introduction

With the rise of the internet and so many internet users worldwide, a large bank of data exists and keeps increasing exponentially, this has given humans the power to convert this huge amount of data into information that can be used to make decisions. Data itself does not make any meaning until it has been converted to information. The idea of data mining helps collect data, preprocess that data by cleaning it and gaining insights into such data to find information. One particularly good source of data is Twitter data which comes in the form of tweets. The twitter API can be used to gather tweets using various forms like searching using keywords, accounts or topics. This has made gathering electronic data easier. Twitter has been used as a good source of data for much research ranging from public health, elections, feedback and so much more. machine learning algorithms like supervised and unsupervised learning algorithms have been used for prediction, clustering and many more on this data. This shows the usefulness of twitter as a good source of data. Malaria is a disease that has killed and is still killing people, especially in Africa. Nigeria has a remarkably high cases of malaria which is caused by mosquito. This brought our attention to gathering data related to malaria in Nigeria, performing sentiment analysis, LDA and word cloud on this data to find insights, trends and patterns and compare the result with WHO fight against malaria. This will help enlighten the public on how to care and prevent this deadly disease disturbing the African continent

1.1 Problem Statement

The availability of huge amounts of data has given us the opportunity to extract information and gain insight into such huge data. Nigeria and other African continents suffer from malaria. Although effective malaria drugs are available and are effective in working. There is a need to derive measures of preventing malaria to reduce the mortality rate caused by malaria.

1.2 Aim and Objectives

The aim of this work is to prevent malaria using twitter as a social media platform. This work focuses on Nigeria which still battles with malaria. The following objectives shall be achieved.

- Performing sentiment analysis
- Performing a Latent Dirichlet Allocation (LDA)

- Interpreting the results from word cloud, sentiment analysis and LDA and comparing them with reports on malaria by WHO.

1.3 Research questions

1. What insight can be obtained from twitter data regarding malaria in Nigeria?
2. What topics or trends can be obtained from the twitter data set?
3. What correlation can be obtained from twitter data and reports by WHO.

1.4 Significance of the study

This research work will benefit the world especially Nigeria's health sector in the fight against malaria. The support offered by the WHO and other bodies in terms of campaign awareness, cure and prevention would be measured to see how social media is being used to assist in this fight. Furthermore, the power of twitter in disseminating information shall be explored.

1.5 Definition of terms

Twitter API: API stands for application programming interface. The Twitter API provides the tools you need to contribute to, engage with, and analyze the conversation happening on Twitter.

Data: is information that is stored which can be text, images, videos, audio clips and much more.

WHO: WHO is the United Nations agency that connects nations, partners and people to promote health, keep the world safe and serve the vulnerable – so everyone, everywhere can attain the highest level of health.

LDA: LDA stands for Latent Dirichlet Allocation is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

CDC: Stands for the Centers for Disease Control and Prevention. The United States Centers for Disease Control and Prevention is the national public health agency of the United States. It is a United States federal agency, under the Department of Health and Human Services, and is headquartered in Atlanta, Georgia.

SVM: stands for support vector machine. SVM is a supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.

HIV: Human immunodeficiency virus (HIV) is a sexually transmitted infection (STI). It can also be spread by contact with infected blood or from mother to child during pregnancy, childbirth, or breast-feeding. Without medication, it may take years before HIV weakens your immune system to the point that you have AIDS.

AIDS: Acquired immunodeficiency syndrome (AIDS) is a chronic, potentially life-threatening condition caused by the human immunodeficiency virus (HIV). By damaging your immune system, HIV interferes with your body's ability to fight infection and disease.

COVID-19: Coronavirus disease 2019 is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first known case was identified in Wuhan, china in December 2019. in December 2019. The disease has since spread worldwide, leading to an ongoing pandemic.

NLTK: Natural Language Toolkit is an open-source collection of libraries, programs, and education resources for building NLP programs**.**

Gensim: This is a free open-source Python library for representing documents as semantic vectors, as efficiently (computer-wise) and painlessly (human-wise) as possible.

1.6 The scope

This research focuses solely on leveraging tweets from twitter related to malaria in Nigeria. It takes advantage of Nigeria's tweets to find insights to help in the fight against malaria. This work can be used to checkmate the support offered by international bodies in combating malaria and its deadly cause.

CHAPTER TWO

Literature review

2.0 Introduction

This chapter presents the literature review which is important in finding what has been done and what needs to be done. According to Parajuli (2020), reviewing the literature is one of the most important steps of a research. The author further listed identifying the problem of choosing subject, formulating the objectives, formulating the hypothesis, generalizing and writing report as other steps of a research. Hart (2018), also mentioned the importance of literature review in which he stressed its significance in understanding a topic, knowing what has been done and how it was done. He also mentioned identifying issues that need to be addressed as an importance.

2.1 Malaria

Malaria is a common and popular disease known worldwide especially in Africa where it is still existing despite measures put in place to tackle it. The Centers for Disease Control and Prevention (CDC, 2021) defines:

> "Malaria is a serious and sometimes fatal disease caused by a parasite that commonly infects a certain type of mosquito which feeds on humans. People who get malaria are typically very sick with high fevers, shaking chills, and flu-like illness."

According to the World health organization (WHO), 229 million malaria cases were estimated in 2019 globally. Amongst this, 215 million estimated cases were from the World Health Organization (WHO) African region. Additionally, Nigeria holds the highest percentage with 27% out of the 95% in twenty-nine countries as accounted.

Kwenti (2018), describes Malaria and HIV as two of the world's most deadly diseases which are widespread, he added that they are prevalent in sub-Saharan Africa. Although the covid-19 pandemic has affected the world globally, Nghochuzie et al., (2020), advised on a collaborative efforts to monitor both covid-19 and malaria. The author further advised performing malaria diagnosis and covid-19 screening and testing to avoid misdiagnosis and achieve ease of management. He also stressed favoring covid-19 at the expense of malaria could be detrimental for global health.

Talapko et al. (2019) did a research on malaria: the past and the present and described malaria as a disease transmitted to humans through the bit of a female mosquito, the disease is serious and is known to be a leading cause of death around the world. The author added malaria is the most prevalent disease in Africa and some Asian countries. This shows malaria has caused a lot of death and Africa has a major share of mortality rate as a result of malaria disease. Furthermore, the author listed the global malaria control program by the WHO as consisting of focus on primary health care, early diagnosis of disease, timely treatment and prevention of disease. This shows the fight against malaria by the WHO is indeed a big battle.

2.2 Twitter

Murthy (2013, pp. 1–3) explained twitter as a microblogging technology particularly created to publicize short messages to a large audience. This audience can extend beyond a user's direct social network. The author stressed the usage of twitter in improving awareness and its power of broadcasting to a large audience. The beauty of this is that both friends and non-friends can see it.

Dewing (2012) describe the term "social media" as services used for online exchanges which are mobile, and internet based. The author further mentioned twitter as both a social network site and a service for status update. Additionally, reaching out to a large audience by companies, organization and governments is possible through social media.

Jordan et al. (2019) in a paper titled "Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response "stated, due to its wide reach, social media platform, Twitter, has proven to be a key source of health-related information shared by citizens and organizations, thereby providing researchers with a real-time source of public health data on a global scale and becoming valuable in public health research. While there are challenges such as lack of standard methods of determining accuracy of extracted data and levels of expert research in the field, there are strong potentialities of social media as an untapped source of data and many great opportunities for improvement of different aspects of public health. The author concluded that the research provided a basis for improving machine learning algorithms especially in exploiting the big data of social media to help improve public health.

2.3 Data mining

With the increase in number of internet users and internet usage, so much data is being generated, such data when put into good use can help a lot in decision making. There is need to mine such data to find insights into what information can be obtained. Han et al. (2011, pp. 13–14) views data mining as a process in which interesting patterns and knowledge can be obtained or discovered from large amounts of data. He added sources of this data can be from data warehouses, the web, other information repositories or dynamically streamed data.

Data mining cuts across many fields as the Figure 2.1 shows.

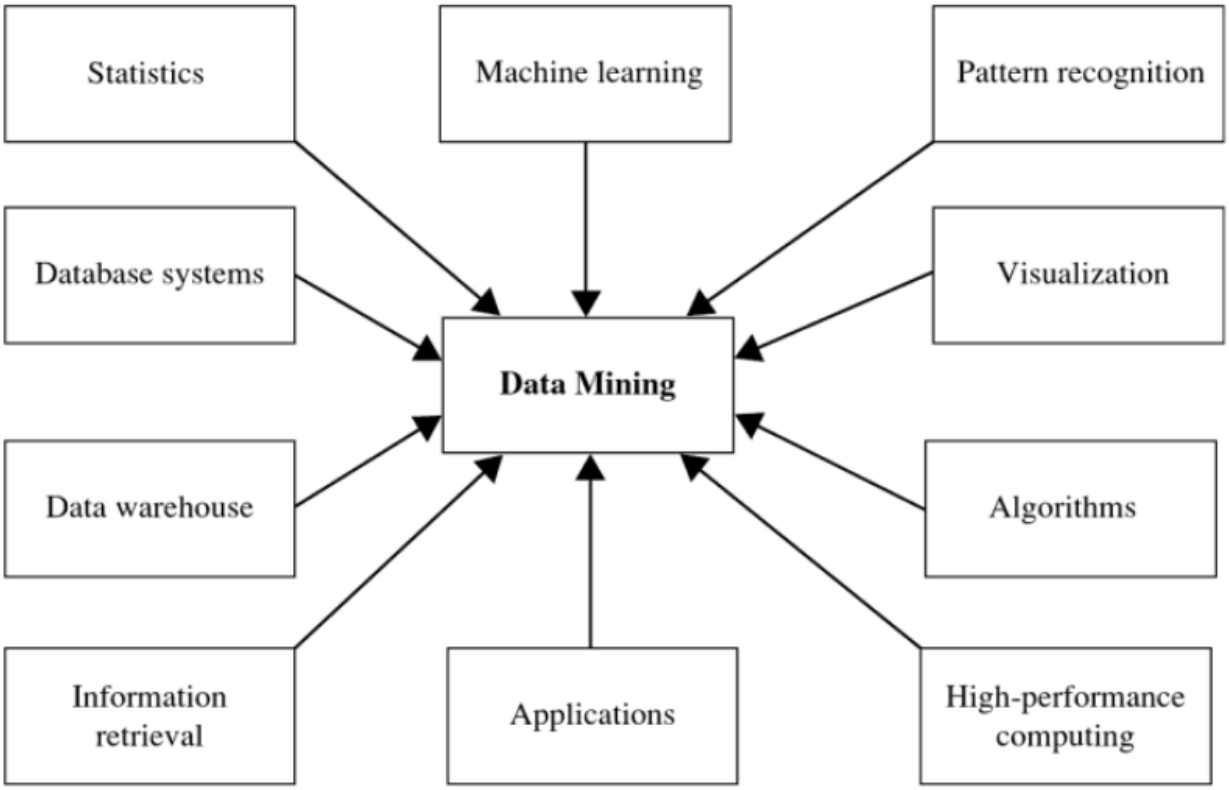


Figure 2.1 Source: Jiawei Han and Micheline Kamber (2011), Data Mining: Concepts and Techniques, Third Edition, Elsevier 22

Han et al. (2000) agrees with Figure 2.1 as he pointed out that data mining is an interdisciplinary subject that draws ideas from statistics, machine learning, pattern recognition, database technology and other disciplines. The author also introduced data mining as the process of looking for fascinating information from a large database.

2.3.1 Text mining

Kwartler (2017, pp, 1-3), while acknowledging many technical definitions of text mining defines "text mining is the process of distilling actionable insight from text", this definition was based on the primary goal of text mining in the extraction of useful output such as visualization or structured table of outputs. The author identified some benefits of text mining in which based on all relevant information, it provides novel insights or strengthens existing perception. Additionally, extraction of information requires little to no sampling.

Aggarwal and Zhai (2012, pp. 1–3), describes the goal of information access as "connecting the right information with the right users at the right time with less emphasis on processing or transformation of text information.", the author further described text mining as an extension of information access to assist users perform analysis to understand and help in decision making.

Han et al. (2011, pp. 13–14) provided an example of text mining usage in which hot topics in literature on data mining from the past ten years can be identified. He also provided another example of mining user comment on products to determine how customers feel about the products. Sentiment analysis is used for this purpose.

2.4 Word Cloud

Heimerl et al. (2014), describe a word cloud as a visually appealing text method for visualization. The author further mentioned its usage in displaying words with the highest frequency in many contexts. The author also referred word cloud as tag cloud and explained its use in showcasing words that occur most often within a text.

Yuping (2019), stated in a paper titled "development of a word cloud software based on python" that a word cloud is a type of list that is weighted to visualize data in form of text and has gained noticeable attention and application on big data. The author notified that current word clouds were available online for users in English but do not support non-English characters. So, the author developed a word cloud that supported multi language environments so it can be used for wider applications especially in solving big text data problems.

2.5 Sentiment analysis

Taboada et al. (2011), describes sentiment analysis as a technique that can be used in drawing out subjectivity and polarity from text. Sentiment analysis is performed to find opinions. The technique can be applied to many text data to find opinions of people.

Hasan et al. (2018) did a research on machine learning based sentiment analysis for twitter accounts in which he used a hybrid approach involving a sentiment analyzer that included machine learning. The author mentioned the progress made in opinion mining and sentiment analysis as rapid. The hybrid approach described by the author was due to a dire need for a more advanced approach to performing sentiment analysis or opinion mining. Additionally, a comparison between Naïve Bayes and support vector machine (SVM) supervised machine learning algorithms was performed.

Nirmala et al.(2015), performed sentiment analysis using tweeter data related to unemployment. To reduce the challenge associated with manually labeling or annotating tweets, the author automated the labels by scoring tweets based on dictionaries of positive and negative terms. For this method, score less than 0 indicated an overall positive opinion for the tweets, score greater than 0 indicated an overall negative opinion of the tweet and neutral was a score equal to 0.

2.6 Topic Modelling

A topic modelling is a process of discovering abstract topics that occur in a collection of documents. It is used in machine learning and natural language processing. It helps us extract or uncover hidden/latent topics in a dataset. Some of the topic models in existence are:

i. Latent Semantic Analysis (LSA)
ii. Probabilistic Latent Semantic Analysis (PLSA)
iii. Latent Dirichlet Allocation (LDA)
iv. Correlated Topic Model (CTM)
v. Explicit semantic analysis
vi. Hierarchical Dirichlet process

This study makes use of latent Dirichlet Allocation (LDA) which is one of the most used topic models.

Silge and Robinson (2017), explained topic modelling as a technique that can be used for classification of documents which follows the principle of unsupervised learning. Like clustering of numeric data, it helps us find natural groups of items even when we are not sure of what to expect. It can be used to divide blog post or news articles into natural groups so we can understand them separately.

Martin and Johnson (2015) described topic modelling as one of the ways in which themes from large documents collection can be observed. The author added topic modelling sees documents as a collection of multiple latent topics. The study made a comparison between three topic models in which the first was obtained using raw news corpus, the second using lemmatized version of the news corpus and the third from the lemmatized news corpus reduced to nouns only. The study found elimination of all words except nouns improved topics' semantic coherence.

2.7 Latent Dirichlet Allocation (LDA)

LDA means Latent Dirichlet Allocation and was developed by David Blei, Andrew Ng, and Michael Jordan in 2003. LDA is a type of topic model and Blei et al. (2003) described LDA as

> "a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document."

Jelodar et al. (2019), described LDA as a method for topic modelling and one of the most popular in that field. The author mentioned that various models based on LDA exist and further made a research on LDA to discover recent developments, trends, and intellectual structure of topic modelling. The investigation covered articles between 2003 to 2016.

Silge and Robinson (2017), describes LDA as one of the most common and popular methods for topic modelling which is based on unsupervised learning and is similar to clustering of numeric data. The author further described it as guided by two principles, namely:

1. Every document is a mixture of topics

   The author explains each document may consist of words from several topics having a specific proportion. Using a two-topic model as an example, document 1 is 90% topic A and 10% topic B, while document 2 is 30% topic A and 70% topic B.

2. Every topic is a mixture of words

   The author explains this principle by giving an example of a two-topic model of American news, one topic for "politics" and the other for "entertainment". Politic may have most common words to be "president". "congress" and "government" while entertainment may have most common topics to be "movies", "television" and "actors". The author further stated essentially that the possibility of words to be shared between topics is possible.

2.7.1 Notations used in describing LDA

To have proper understanding of LDA, certain concept of words needs to be defined. Blei et al. (2003) provided definitions on some important terms used in LDA.

Word: considered as a basic unit of discrete data, a word is an item from a vocabulary indexed by {1,…,V}. its representation utilizes unit-basis vectors that have a single component equal to one and all other components equal to zero.

Document: A document represents a chain of N words defined by $w=(w_1,w_2,\ldots,w_N)$, where w_n is the nth word in the chain.

Corpus: A corpus (plural: corpora) is defined as a collection of M documents denoted by $D=\{w_1,w_2,\ldots,w_M\}$.

How LDA Works

Blei et al. (2003) described LDA as a generative probabilistic model of a corpus where documents are random mixture over hidden topics. These hidden topics are also referred to as latent topics and each topic is presented by a distribution over words. An assumption based on the idea of generative process for each document w in a corpus D is given:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   a. Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   b. Choose a word w_n from $p(w_n \mid z_n, \beta)$ a multinomial probability conditioned on the topic z_n.

Hence, Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

2.8 Related Works

Maurice et al. (2019, p. 591), used malaria twitter data and precipitation as a method for monitoring and reporting malaria instances. The author made use of Support vector machine (SVM) classifier to classify Nigeria twitter messages into malaria case related and non-malaria case related tweets. A high correlation between the malaria related case and average precipitation in Nigeria was obtained. The author highlighted the use of twitter in monitoring and reporting malaria instances directly and predicting malaria outbreak in Nigeria and places where malaria is endemic. This shows the importance of using twitter data in the health sector.

Boit and El-Gayar (2020), used a text mining approach on twitter social media to mine malaria topics. This study was done using the crimson social media analytics software to examine public discourse, trends and emergent themes related to malaria discussion. The author highlighted the importance of the study in understanding patters and trends of public opinion regarding malaria. The author also provided examples of how the insight can assist in making informed decisions such as health cost reduction by effective drug development, acting in advance on allocation of resources and well-timed distribution of kits in affected areas and improving public health management overall.

Dewi et al. (2020), used sentiment analysis to predict the success of social distancing made by the government of Indonesia to its people. This was in relation to the COVID-19 pandemic. The hashtag used was (#dirumahaja) which means “stay at home”. This study made use of Naïve Bayes and Random forest models and concluded that highest accuracy of classification was obtained using Random Forest algorithm compared to Naïve Bayes with a yield of 95.98%. The study further found out that positive sentiments were greater than negative which means the people of Indonesia agree to social distancing program made by the Indonesian government.

Nkiruka et al. (2021), did a study on malaria epidemic prediction systems built to reduce the increase in disease outbreak in some African countries and discovered the need to have better models. Such models should have improved prediction capability based on non –seasonal variations in climatic conditions. Factors that contributed to malaria outbreak such as precipitation,

temperature and surface radiation was considered for this study. The author concluded that the improved system outperformed other classification modes and further stated that the model serves as an early detection mechanism in monitoring the spread of malaria.

Hasan et al. (2018), did a study on tweets relating to Indonesian presidential elections in which the author used a combined approach of sentiment analysis and machine learning. The author made use of two supervised machine learning algorithms namely, Naïve Bayes and support vector machine (SVM) and made a comparison between the two. This shows tweets can be used for many research purposes including election predictions and more.

2.9 Research Gap

Research has been done in topic modelling and sentiment analysis using twitter malaria data in the past. Techniques such as support vector machine, naïve Bayes classifiers has been used and compared on such data. Also, topic modelling has been used to find topical themes on such data, but none has used twitter malaria data to find topical themes and correlated it with WHO mission to fight malaria.

2.10 Significance of the Study

This study will help to find ways of preventing malaria via social media by finding public opinions of tweets and using those tweets on a Latent Dirichlet Allocation (LDA) to find topics or themes and correlating it with WHO battle against malaria.

CHAPTER THREE

3.0 Introduction

This chapter introduces the methods and methodology used in carrying out the main work of the project. The main source of data comes from twitter which is collected and saved as a. JSON extension file. This data must undergo some data pre-processing using some data mining techniques. Python programming language is used for the implementation of this work. The proposed model of the study is depicted in Figure 3.1

A sentiment analysis is carried out on the data and visualizations using scatter plots and bar graphs is used. Also, Latent Dirichlet Allocation (LDA) is used to cluster the topics. A word cloud is generated to visualize the topics from the LDA.

3.1 Proposed model

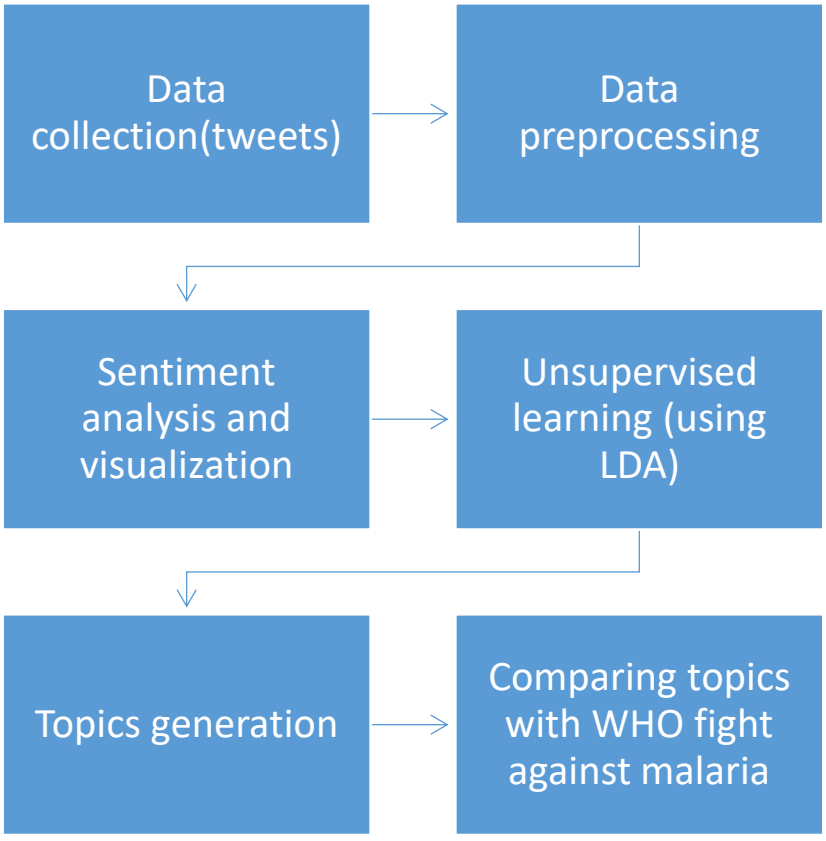
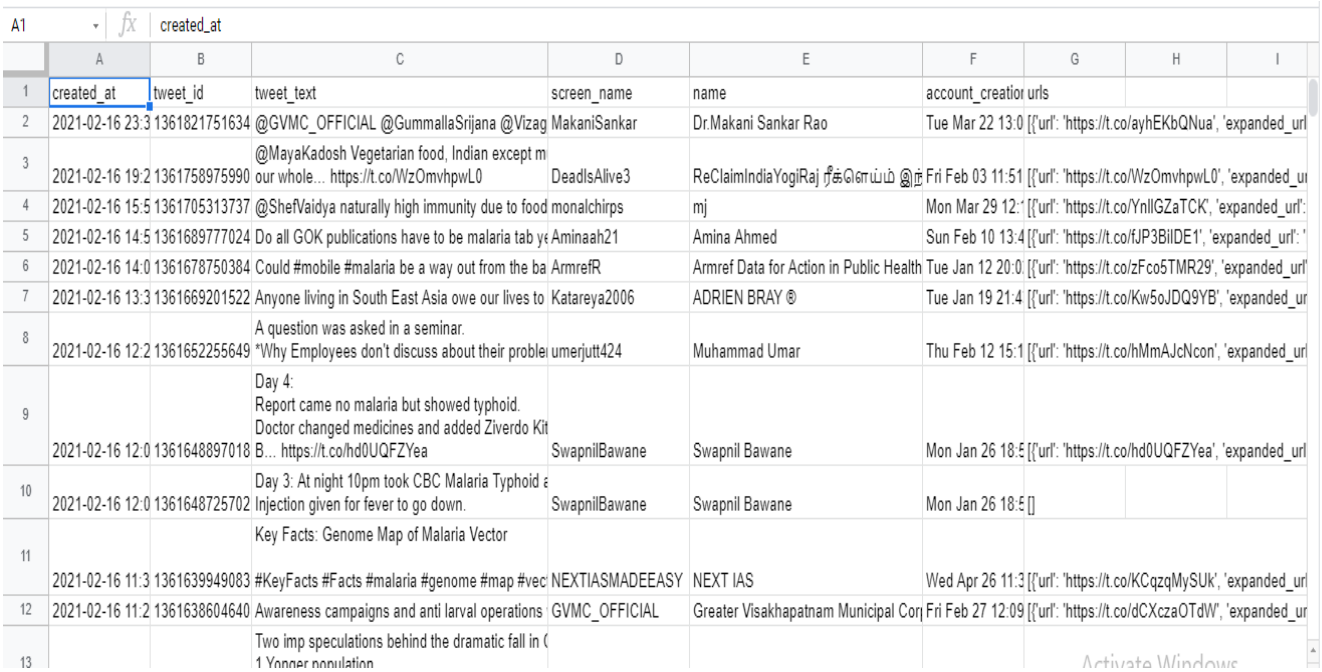


Figure 3.1 Proposed model for the research work

3.1.1 Data collection

The data to be used for this experiment is sourced from twitter using the twitter API tools. This data is searched using the keyword “malaria Nigeria” from 2017 to 2021. This data is saved in a .JSON file and has seven columns and the number of rows is equal to the number of tweets searched during data collection. Tweet_text column is being considered for the purpose of this research.



A1 | created_at

	A	B	C	D	E	F	G	H	I
1	created_at	tweet_id	tweet_text	screen_name	name	account_creatio	urls		
2	2021-02-16 23:3	1361821751634	@GVMC_OFFICIAL @GummallaSrijana @Vizag	MakaniSankar	Dr.Makani Sankar Rao	Tue Mar 22 13:0	[{'url': 'https://t.co/ayhEKbQNua', 'expanded_url		
3	2021-02-16 19:2	1361758975990	@MayaKadosh Vegetarian food, Indian except m our whole... https://t.co/WzOmvhpwL0	DeadIsAlive3	ReClaimIndiaYogiRaj ரீக்ளெய்ம் இ	Fri Feb 03 11:51	[{'url': 'https://t.co/WzOmvhpwL0', 'expanded_u		
4	2021-02-16 15:5	1361705313737	@ShefVaidya naturally high immunity due to food	monalchirps	mj	Mon Mar 29 12:	[{'url': 'https://t.co/YnllGZaTCK', 'expanded_url':		
5	2021-02-16 14:5	1361689777024	Do all GOK publications have to be malaria tab y	Aminaah21	Amina Ahmed	Sun Feb 10 13:4	[{'url': 'https://t.co/fJP3BilDE1', 'expanded_url': '		
6	2021-02-16 14:0	1361678750384	Could #mobile #malaria be a way out from the ba	ArmrefR	Armref Data for Action in Public Health	Tue Jan 12 20:0	[{'url': 'https://t.co/zFco5TMR29', 'expanded_url'		
7	2021-02-16 13:3	1361669201522	Anyone living in South East Asia owe our lives to	Katareya2006	ADRIEN BRAY ®	Tue Jan 19 21:4	[{'url': 'https://t.co/Kw5oJDQ9YB', 'expanded_ur		
8	2021-02-16 12:2	1361652255649	A question was asked in a seminar. *Why Employees don't discuss about their proble	umerjutt424	Muhammad Umar	Thu Feb 12 15:1	[{'url': 'https://t.co/hMmAJcNcon', 'expanded_url		
9	2021-02-16 12:0	1361648897018	Day 4: Report came no malaria but showed typhoid. Doctor changed medicines and added Ziverdo Kit B... https://t.co/hd0UQFZYea	SwapnilBawane	Swapnil Bawane	Mon Jan 26 18:5	[{'url': 'https://t.co/hd0UQFZYea', 'expanded_url		
10	2021-02-16 12:0	1361648725702	Day 3: At night 10pm took CBC Malaria Typhoid a Injection given for fever to go down.	SwapnilBawane	Swapnil Bawane	Mon Jan 26 18:5	[]		
11	2021-02-16 11:3	1361639949083	Key Facts: Genome Map of Malaria Vector #KeyFacts #Facts #malaria #genome #map #vec	NEXTIASMADEEASY	NEXT IAS	Wed Apr 26 11:3	[{'url': 'https://t.co/KCqzqMySUk', 'expanded_url		
12	2021-02-16 11:2	1361638604640	Awareness campaigns and anti larval operations	GVMC_OFFICIAL	Greater Visakhapatnam Municipal Cor	Fri Feb 27 12:09	[{'url': 'https://t.co/dCXczaOTdW', 'expanded_ur		
13			Two imp speculations behind the dramatic fall in (1 Yonger population						

Figure 3.2 Sample of the data set

3.1.2 Data Pre-processing

The data collected needs to be cleaned in order to process it and get accurate results. Data cleaning is especially important in text mining. Before data cleaning is performed, the tweet column is selected for the purpose of the analysis to be performed.

3.1.3 Data Cleaning

The following processes are used in cleaning the tweet data.

- Tweeter data (in a .csv file) is loaded into a data frame using pandas.
- A specific column “tweet_text” from the loaded data frame is loaded into a new data frame.

- A function is created in python to clean the text. It performs the function of removing retweets (RT), hashtags (#), "@" symbol and hyperlinks (https).
- The data frame is converted into lower case.
- Special characters are also removed.
- Stop words such as "this" "at" "your" are also removed.

```
0      channelstv this manifestation acute malaria do...
1      must malaria drugs bitter wickedness like smel...
2      s3ns3iowen dfkm your time goat bossed child wa...
3                              malaria still killing sigh
4      funmiscute mosquitos weakness small bite like ...
                              ...
495    ericgarland with only young children pregnant ...
496    ericgarland with only young children pregnant ...
497    editoroferic with only young children pregnant...
498    isimenancy with only young children pregnant m...
499    bellahtyrah have typhoid test maybe should tre...
Name: tweet_text, Length: 500, dtype: object
```

Figure 3.3 Sample screenshot of the cleaned and processed tweets

3.2 Word cloud

A word cloud is generated using the cleaned and processed data set. The word cloud for five years ranging from 2017 to 2021 was generated for the purpose of gaining insights into the data. The word cloud is displayed in Figure 4.1.

3.3 Sentiment Analysis

Sentiment analysis is carried out on the clean and processed data set. The result is displayed using the bar graph and scatter plot. The percentage of positive, neutral, and negative tweet's result is also obtained. The sentiment analysis model implemented followed the method used by Nirmala et al.(2015), which used automated labels by scoring sentiment of tweets based on dictionaries of positive, negative and neutral terms. This idea was to reduce challenges of manual label classification which is time consuming. For this method,

score < 0 indicated an overall positive opinion for the tweets

score > 0 indicated an overall negative opinion of the tweet

score = 0 indicated an overall neutral opinion of tweet.

3.3.1 Subjectivity, polarity, and Analysis

Sentiment analysis shows opinion of users via their tweets. the polarity ranges from [-1,1] with positive tweet being 1 and negative tweet being -1. The subjectivity refers to opinion and it lies between [0,1].

	content	subjectivity	polarity	Analysis
29259	when will nigeria be malaria free?	0.8	0.4	Positive
29260	for those who think nigeria can never get rid of malaria	0.0	0.0	Neutral
29261	pls china, any and all effort to improve our collective wellbeing is welcome. mosquitoes and malaria remains a problem in nigeria	0.9	0.8	Positive
29262	increasing resistance to chloroquine in falciparum malaria in sokoto, north western nigeria.	0.0	0.0	Neutral
29263	this is cheery news.\nseeing what malaria costs us yearly, do nigeria have any plan to achieve this strategic feat? \n	1.0	0.7	Positive

Figure 3.4 Tweets with their subjectivity, polarity and analysis

3.4 Latent Dirichlet Allocation (LDA) implementation

A detailed description of LDA is given in Section 2.7. LDA is one of the most common topic models and is used for this work. Gensim in Python programming language is used to implement the LDA algorithm. The following processes are used to generate the topics.

3.4.1 Lemmatization

Lemmatization is the process of deriving the root of words by removing prefixes and suffixes. We used the lemmatization module in Python's NLTK library to perform lemmatization on our data.

3.4.2 Tokenization

This is the process of breaking down our data into single words. " " is used as the delimiter for tokenization.

```
must malaria drugs bitter wickedness like smell bitter it's going looking tablets 🤦
```

Figure 3.5 Sample tweet before tokenization

```
['drug', 'bitter', 'wickedness', 'tablet']
```

Figure 3.6 Sample Tweet after Tokenization

3.4.3 Dictionary generation

A dictionary is generated which helps us find the list of unique words in our corpus. This is needed to be fed into our topic model after lemmatization and tokenization of our processed data. The python Gensim library has a dictionary module that generate the dictionary of our corpus.

3.4.4. Document term matrix (DTM)

Using our dictionary, we create a document term matrix which will be passed to the LDA model to build the model. The document term matrix is mathematical matrix used to describe the frequency of terms that occur in a collection of documents. In a DTM, columns correspond to terms (words) while rows correspond to documents. This is generated based on the dimension of the document. DTM is passed to the LDA to create the LDA model. Parameters passed to the model include, the corpus, which is the same as our DTM, the dictionary, number of topics, random state, chunk size, number of passes, and number of iterations respectively.

3.4.5 LDA Topics

LDA was performed using 5 topics. This enabled generation of 5 topics. Some of the topics overlapped while some had no intersection at all. The result of all five topics was diagrammatically presented in Figure 4.9

3.4.6 Visualization of the LDA

The LDA generated five topics and inter-topic distance map (multidimensional scaling) was used to represent each topic with a circle. An intersection of two or more circles shows topics having something in common. The five topics are presented in Figure 4.10 to Figure 4.14.

CHAPTER FOUR

Result and discussion

4.0 Introduction

This chapter discusses the results obtained from tweets retrieved from 2017 to 2021. Result of the word cloud from the 5 years will be discussed. The results from the sentiment analysis from the five years is also discussed. The topics generated using LDA shall also be discussed, and topics obtained analyzed. This will be used to make a comparison with WHO works.

4.1 Word Cloud

The word cloud shows the frequency of occurrence of words or topics. The result of the word cloud generated on malaria tweets gives us some information. The word cloud as depicted in fig 4.1 shows the most frequently mentioned topics as “Malaria Drugs”,” Malaria Free”,” Free Malaria”, “anti malaria”, “Malaria Prevention”, “Signs”, “Test before treating Malaria”, “World Malaria Day”, “Malaria typhoid”,” visitors high”, “Malaria care”, and more.

Figure 4.1 Word cloud generated from the cleaned and processed data

4.2 Sentiment Analysis

	content	subjectivity	polarity	Analysis
29259	when will nigeria be malaria free?	0.8	0.4	Positive
29260	for those who think nigeria can never get rid of malaria	0.0	0.0	Neutral
29261	pls china, any and all effort to improve our collective wellbeing is welcome. mosquitoes and malaria remains a problem in nigeria	0.9	0.8	Positive
29262	increasing resistance to chloroquine in falciparum malaria in sokoto, north western nigeria.	0.0	0.0	Neutral
29263	this is cheery news.\nseeing what malaria costs us yearly, do nigeria have any plan to achieve this strategic feat? \n	1.0	0.7	Positive

Figure 4.2 Tweets with their subjectivity, polarity and analysis

4.3 Visualization of sentiment analysis result

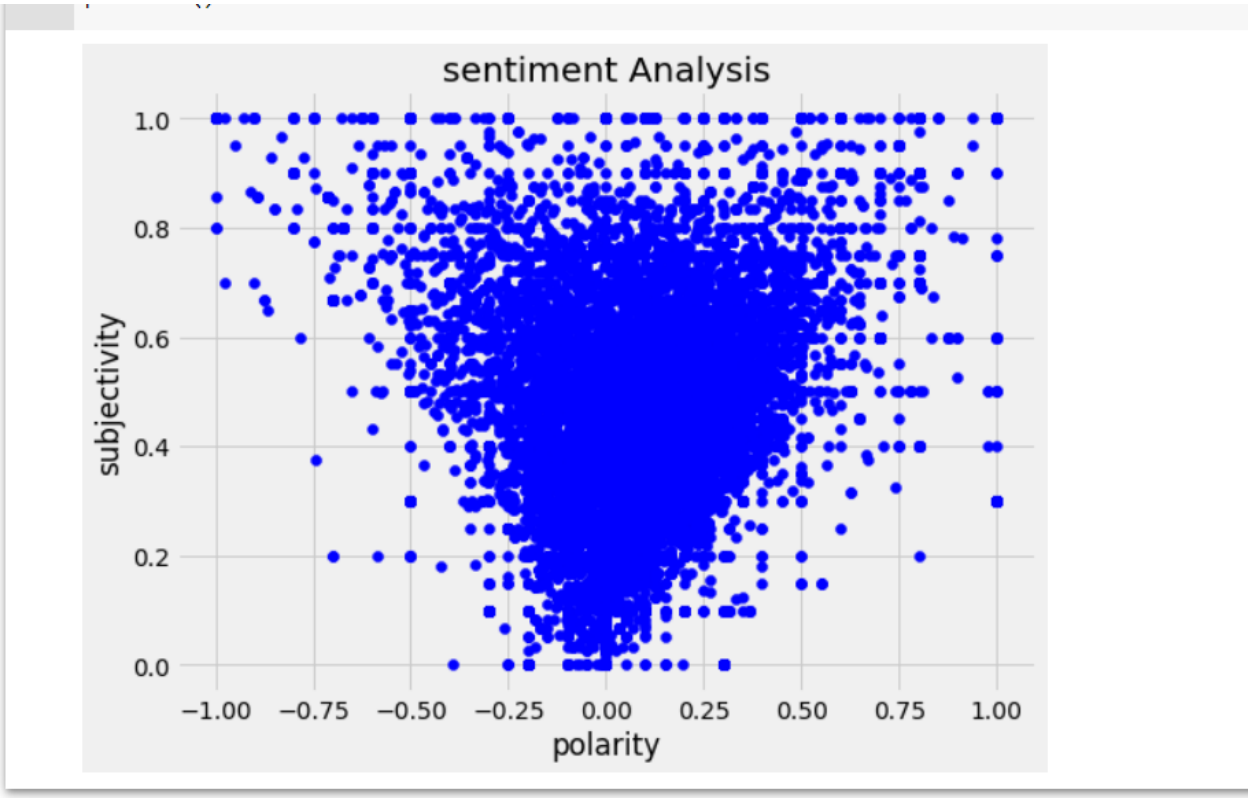


Figure 4.3 Scatter Plot showing the Polarity and Subjectivity of Tweets

4.4 Percentage of tweets (positive, negative and neutral)

```
[109] #get percentage of Positive tweets
      ptweets = dfC[dfC.Analysis == 'Positive']
      ptweets = ptweets['content']

      round( (ptweets.shape[0] / dfC.shape[0]*100),1)

      42.6
```

Figure 4.4Percentage of Positive Tweets

```
[106] #get percentage of negative tweets
      ntweets = dfC[dfC.Analysis == 'Negative']
      ntweets = ntweets['content']

      round( (ntweets.shape[0] / dfC.shape[0]*100),1)

      15.6
```

Figure 4.5 Percentage of Negative Tweets

```
#get percentage of Neutral tweets
neutweets = dfC[dfC.Analysis == 'Neutral']
neutweets = neutweets['content']

round( (neutweets.shape[0] / dfC.shape[0]*100),1)

41.8
```

Figure 4.6 Percentage of Neutral Tweets

4.5 Bar Graph of sentiment analysis

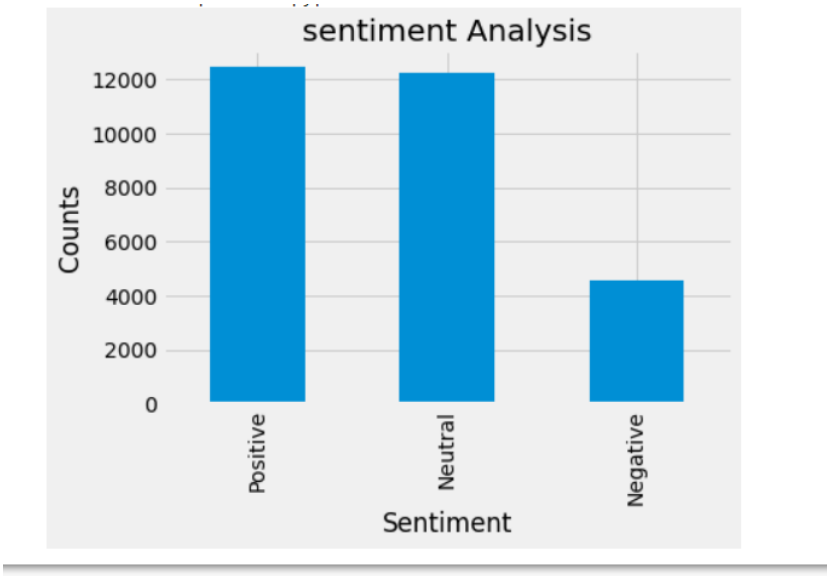


Figure 4.7 Bar Graph for the Positive, Neutral and Negative Tweets

4.6 Latent Dirichlet Allocation

The LDA generated five topics as indicated by the circles in fig 3.4. The topics are labelled 1-5.

```
[(0,
  '0.073*"support" + 0.067*"child" + 0.064*"young" + 0.063*"pregnant" + 0.061*"mum" + 0.041*"net" + 0.041*"love" + 0.037*"month" + 0.016*"drug" + 0.014*"time"'),
 (1,
  '0.033*"typhoid" + 0.015*"much" + 0.012*"time" + 0.011*"thing" + 0.011*"doctor" + 0.011*"betterhalf" + 0.011*"ment" + 0.008*"money" + 0.008*"blood" + 0.008*"medical"'),
 (2,
  '0.051*"people" + 0.049*"friend" + 0.044*"donation" + 0.032*"chip" + 0.032*"american" + 0.028*"life" + 0.026*"brother" + 0.025*"money" + 0.024*"democracy" + 0.021*"treatment"'),
 (3,
  '0.046*"today" + 0.035*"message" + 0.030*"leader" + 0.028*"poor" + 0.028*"continent" + 0.024*"health" + 0.024*"life" + 0.024*"youth" + 0.023*"world" + 0.017*"sure"'),
 (4,
  '0.078*"child" + 0.076*"young" + 0.069*"mum" + 0.069*"pregnant" + 0.061*"able" + 0.049*"free" + 0.027*"malaria" + 0.012*"people" + 0.011*"mosquito" + 0.011*"system"')]
```

Figure 4.8 Topic generated by the LDA

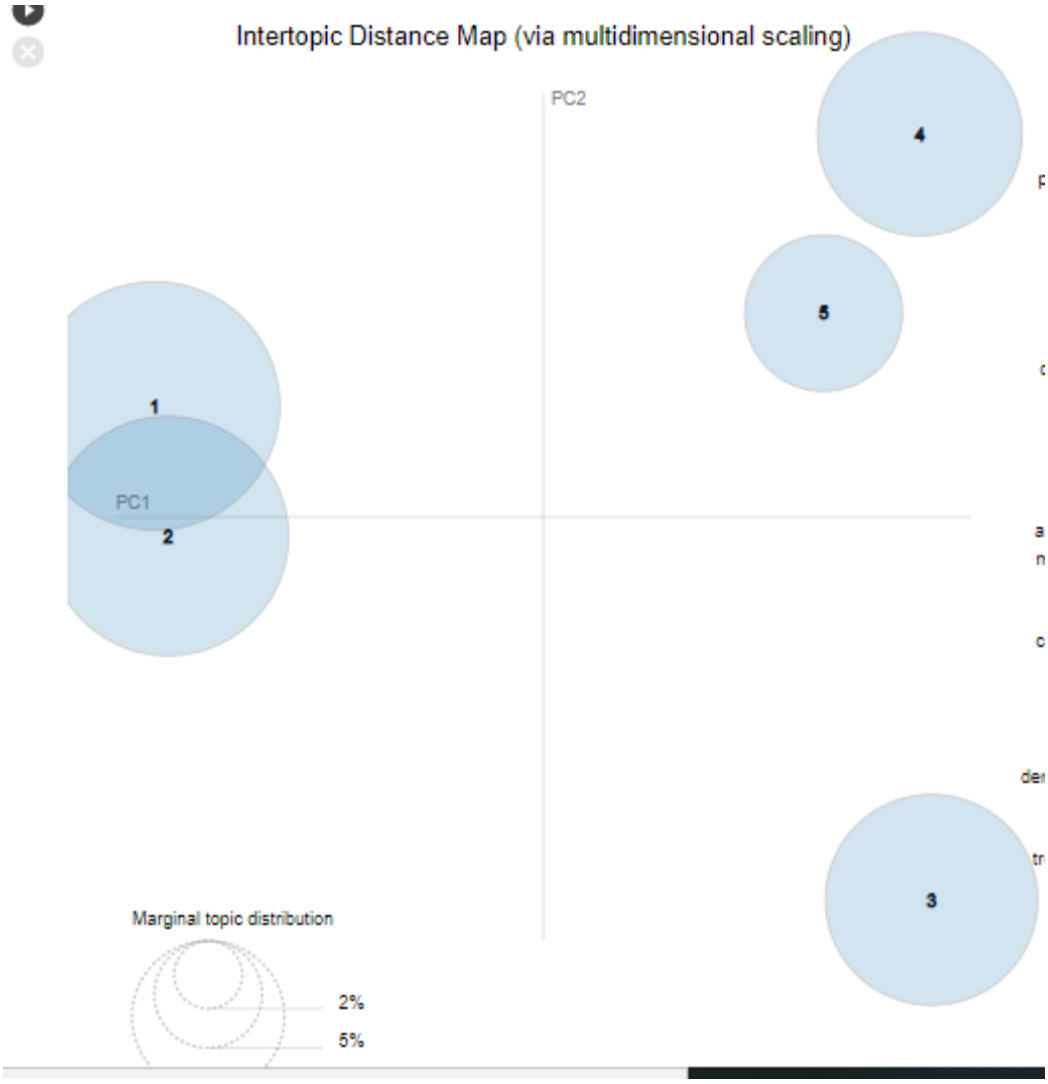


Figure 4.9 Diagrammatic view of the 5 generated topics

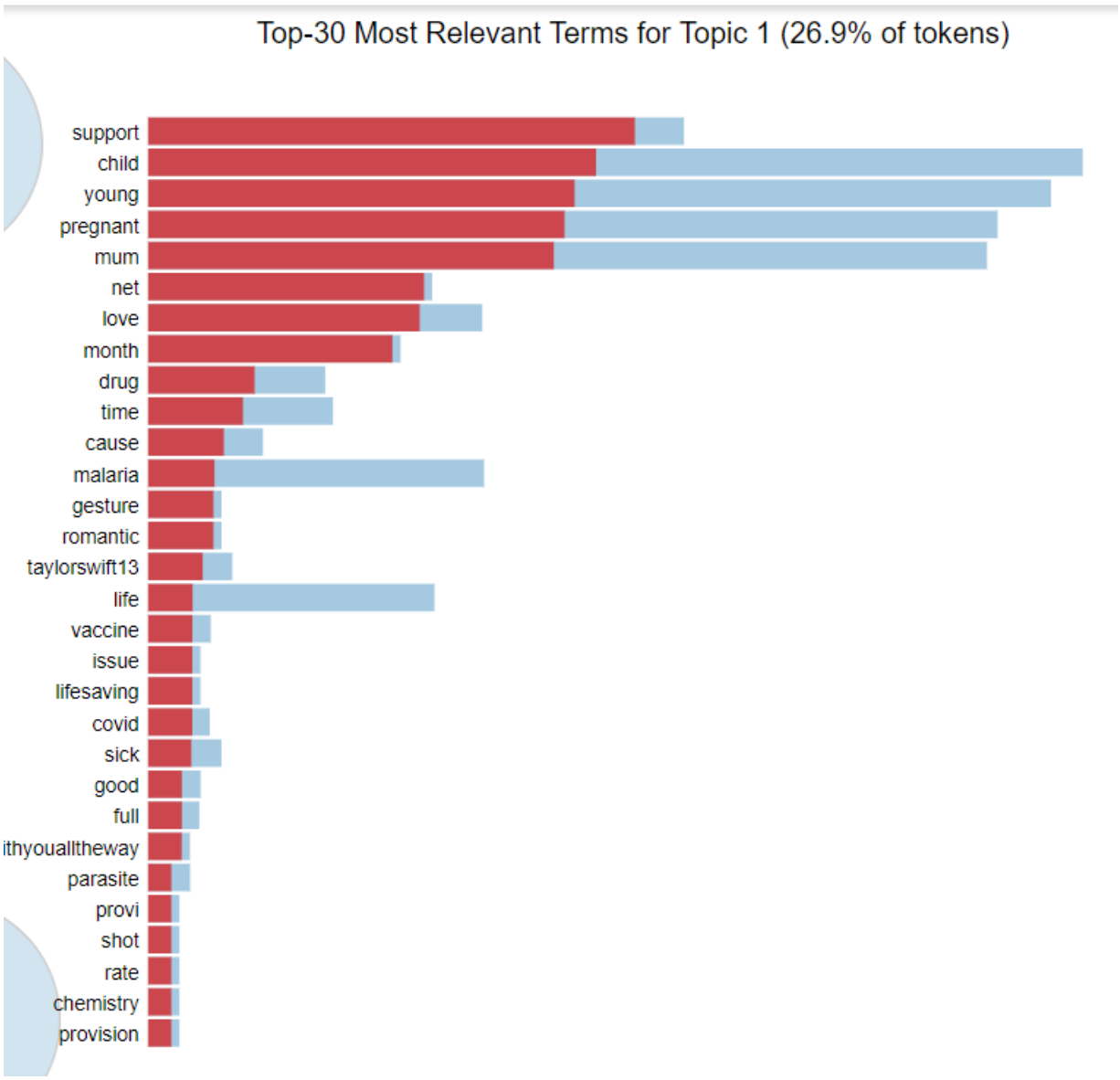


Figure 4.10 Topic 1

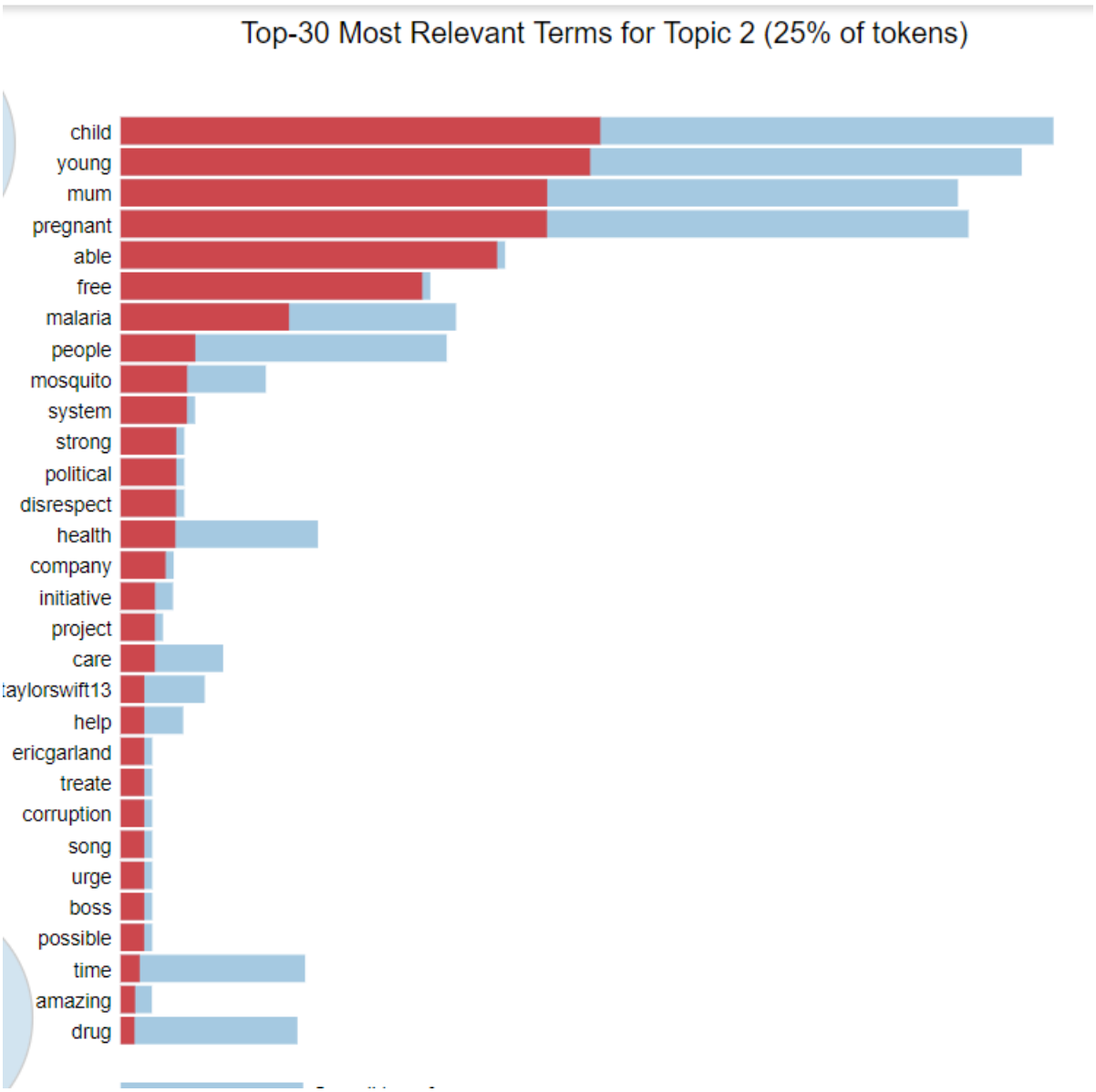


Figure 4.11 Topic 2

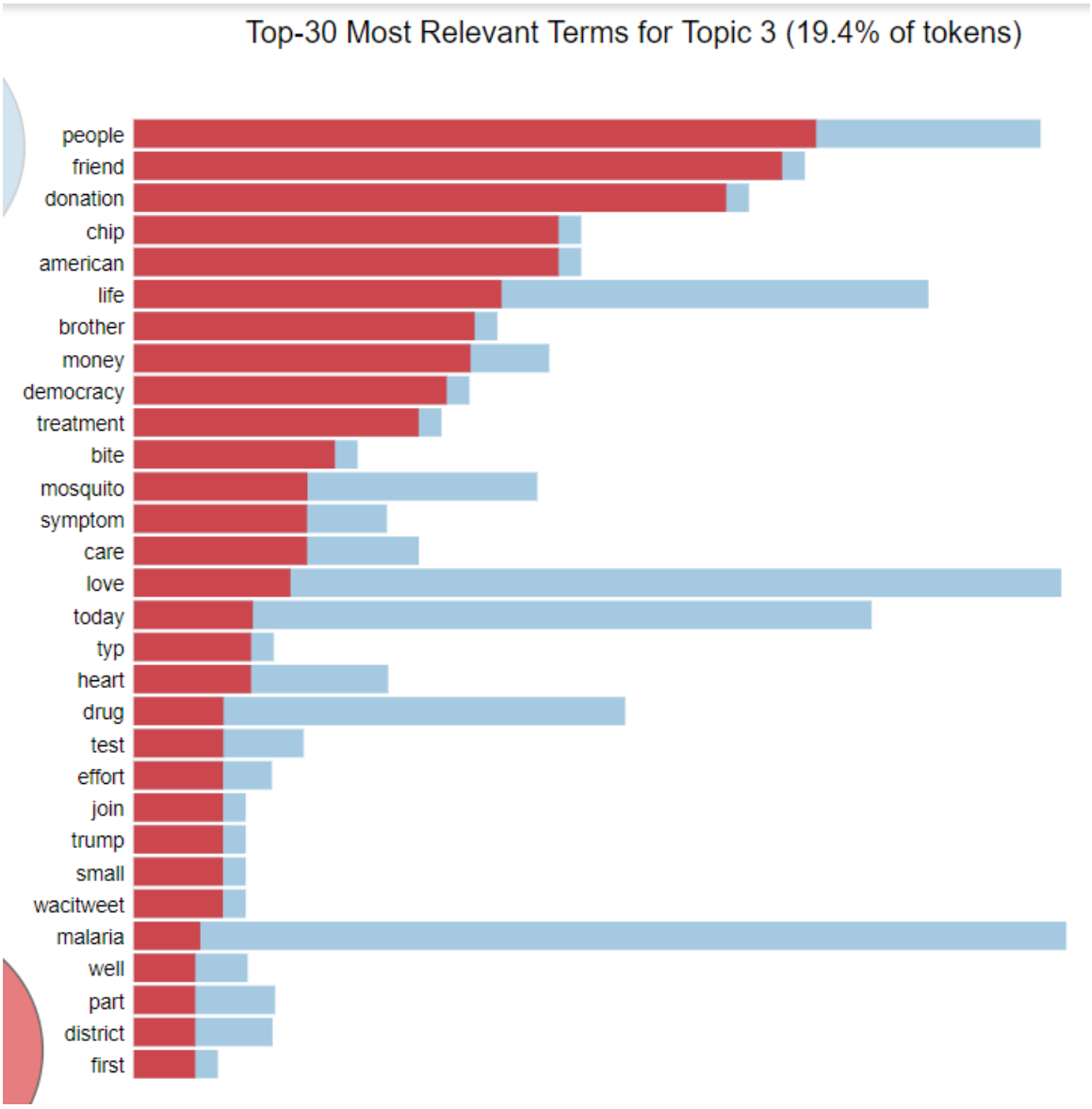


Figure 4.12 Topic 3

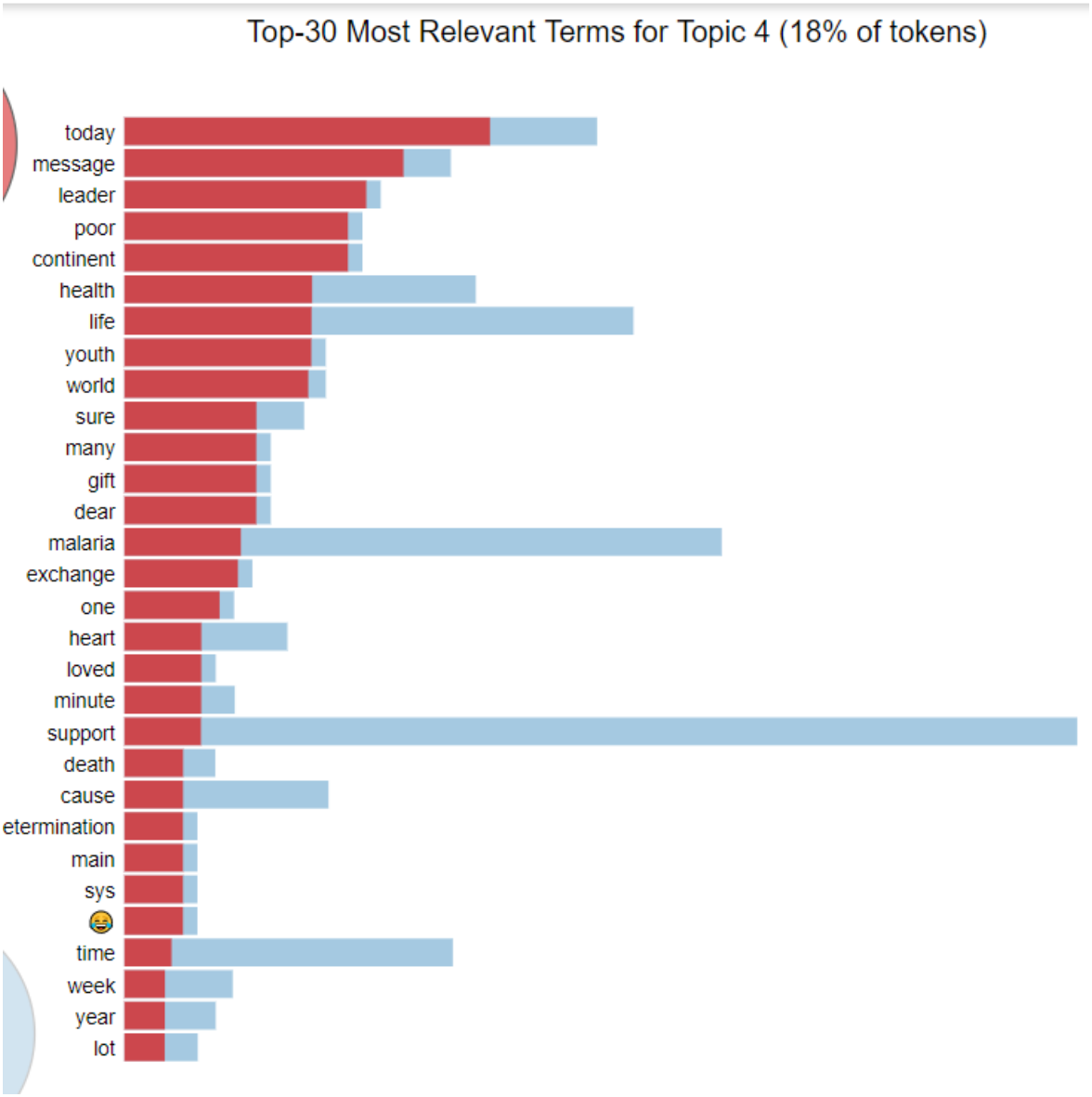


Figure 4.13 Topic 4

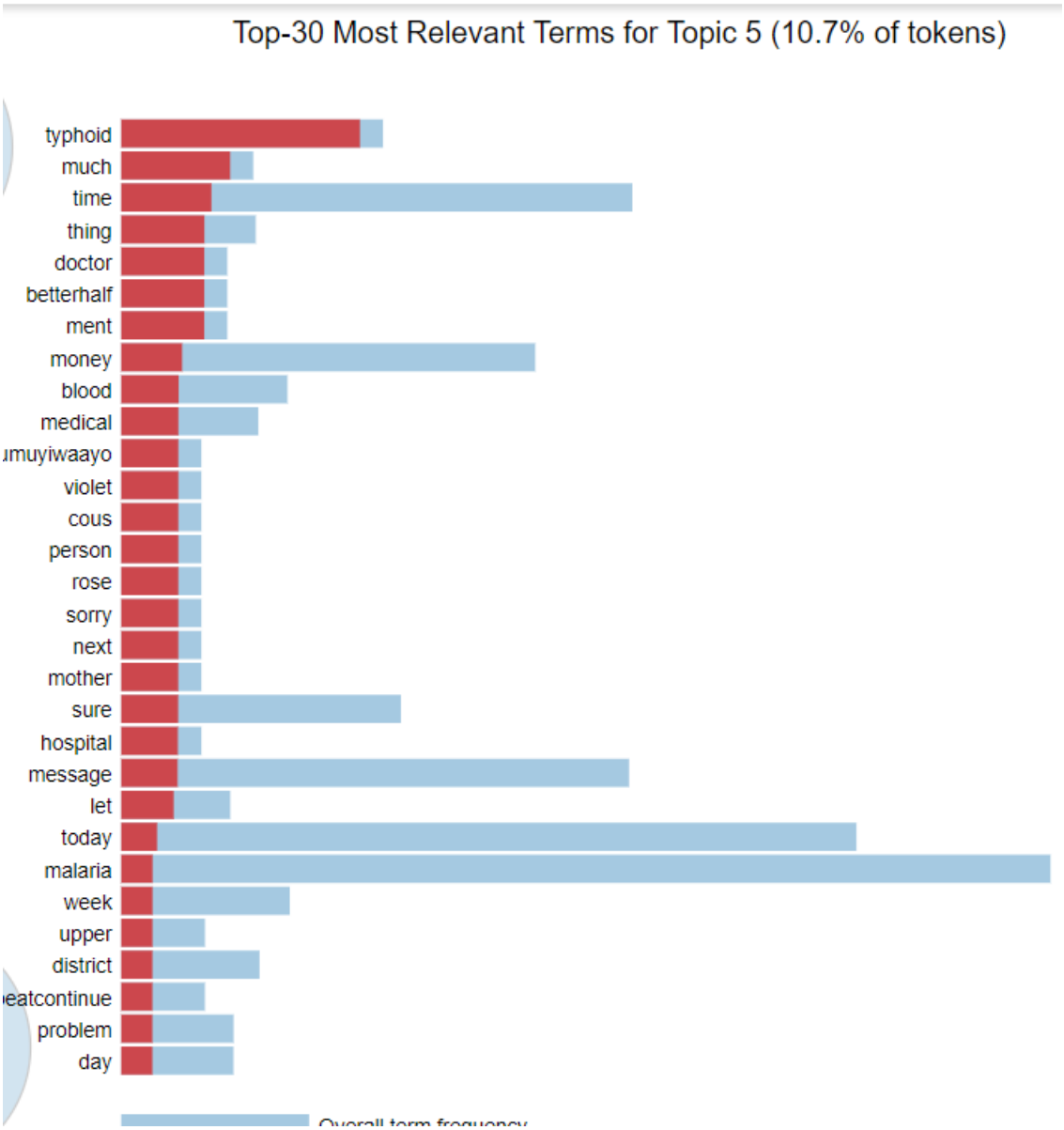


Figure 4.14 Topic 5

4.7 Discussions

The result from word cloud shows topics with the most frequency. Topics like world malaria day, malaria care, signs, and malaria free serve as a way of enlightening people on malaria, creating awareness and serve as campaigns on informing the public on better ways to prevent, manage and combat malaria. “Malaria free” and “free malaria “conforms to the vision of the current National Malaria Strategic Plan 2014-2020 of achieving a malaria-free Nigeria (Nigeria Ministry of Health, 2014). Topics like malaria drugs and anti-malaria gives insight into treatment of malaria. Topics

like test before treating malaria, malaria typhoid shows getting proper test to see if really malaria is the cause of sickness or typhoid to avoid blind treatment.

The result from sentiment analysis shows that 42.6% of the tweets are positive, 15.6% are negative and 41.8% are Neutral. This shows positive sentiment is highest among the tweets. This is further visualized using the scatter plot and bar graph as depicted in Figure 4.3 and Figure 4.7, respectively. The positive tweets indicate the fight against malaria is getting better grounds.it also indicates that efforts to combat malaria and malaria drugs are being effective. Awareness regarding malaria is also effective. The neutral indicates the tweets are neither positive nor negative. This shows neutrality of the tweets such as campaigns on world malaria day, and other awareness programs. The negative percentage shows death as regard malaria, sickness and disease mentioned in such tweets. overall, we can deduce fight against malaria contributed to the positive tweets which means it is being effective.

The result from LDA generated 5 topics. Topics 1 in Figure 4.10 and topic 2 in Figure 4.11 intersect which implies they have similar topics in common. Topics 4 in Figure 4.13 and topic 5 in Figure 4.14 did not intercept but have very related topics in their cluster. Topic 3 in Figure 4.11 looks like an outlier as it is far from all the 4 topics. Topics 1 and 2 have topic relating to child, support, young, pregnant, and mum. This shows malaria is still affecting young children and pregnant mothers are also at risk. Both topics conform to the key message of World Malaria report,2019 which reported hard strike of malaria against pregnant women and children. Furthermore, the report advised countries to prioritize on pregnant women and children. (World Malaria report, 2020)

"Net" and "drug" are also mentioned which can be attributed to prevention and cure of malaria. This correlates to the recommendation of WHO that all children living in malaria-affected areas to sleep under ITN. (WHO, 2019)

Topic 3 generally mentions topics like people, donation, money, treatment, bite, mosquito, symptom, drugs, efforts, test and more. This correlates to campaigns on malaria, malaria awareness, malaria symptoms, and mosquito as a cause/source. This topic can be labelled as "Malaria awareness/campaign.

Topic 4 mentions today, message, leaders, poor, continent, health, life, youth, world and more. These topics can be labelled as “factors contributing to malaria”. Nigeria health system is perturbed by poor leadership and governance. This is more government related.

Topic 5 mentions typhoid, doctor, money, blood, medical, hospital and more. This can be attributed to “treatment or cure for malaria.”

CHAPTER FIVE

Conclusion, recommendation, and future work

5.0 Conclusion

with the availability of data comes information which can help in decision making. This available data can be used to find insights into what people are saying about a particular topic like malaria. This gives people, governments and international bodies fighting or supporting the fight against malaria good information on the effectiveness of their campaign awareness, on effect of anti-malaria drugs used, on the outcomes obtained from using insecticide treated nets and so much more. The WHO has a report titled “world malaria report” each year. This research shows a lot of correlation with what this international organization is doing on the war against malaria.

5.1 Recommendation

Based on the insights found using twitter data, we recommend leveraging twitter for more research on other topics such as health, security, sports, agriculture, technology and more. this will help in finding insights that can be used in other areas to enable concerned bodies in making better decisions.

5.2 Future work

Due to time limitation, we were not able to manually label our twitter data and as such our sentiment analysis might be prone to errors. Future research can manually annotate the tweets to use support vector machine, Naïve Bayes or Random forest classifiers to train and make predictions of new malaria tweets. This will give more accurate opinions of the tweets as positive, negative, or neutral. Also, LDA can be used to find topical themes for other years and compare with future work of WHO is fighting malaria

REFERENCES

Adrover, C., Bodnar, T., Huang, Z., Telenti, A., & Salathé, M. (2015). Identifying adverse effects of HIV drug treatment and associated sentiments using twitter. *JMIR Public Health and Surveillance*, *1*(2).e7 https://doi.org/10.2196/publichealth.4488

Aggarwal, C. C., & Zhai, C. (2012). *Mining Text Data* (2012th ed.). Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Boit J, El-Gayar O (2020) Topical mining of malaria using social media. A text mining approach. In: Proceedings of the 53rd Hawaii International Conference on System Sciences 7-01-2020, 3810- 3820 https://doi.org/10.24251/HICSS.2020.466.

Dewing, M (2012). Social media: an introduction. In: Library of Parliament Publication no. 2010-03-E, pp. 1–5

Dewi, T. B. T., Indrawan, N. A., Budi, I., Santoso, A. B., & Putra, P. K. (2020). Community Understanding of the Importance of Social Distancing Using Sentiment Analysis in Twitter. *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*, *Yogyakarta*, *Indonesia*, 336–341. https://doi.org/10.1109/IC2IE50715.2020.9274589

Hart, C. (2018). Doing a literature review: Releasing the research imagination (2nd ed.). Sage.

Centers for Disease Control and Prevention. (2021). Malaria. SAGE Publications Ltd; Second edition. https://www.cdc.gov/malaria/about/faqs.html

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)* (3rd ed.). Morgan Kaufmann.

Hand, D. J., Blunt, G., Kelly, M. G. & Adams, N. M. (2000). Data mining for fun and profit. *Stat. Sci.* **15**, 111 –131

Heimerl, F., Lohmann, S., Lange, S. and Ertl, T. (2014), "Word cloud explorer: Text analytics based on word clouds", *System Sciences (HICSS) 2014 47th Hawaii International Conference*, 6-9 Jan, Waikoloa, HI, USA, pp. 1833-1842.

Hasan, A.; Moin, S.; Karim, A.; Shamshirband, S. 2018. Machine Learning-Based Sentiment Analysis for Twitter Accounts. *Math. Comput. Appl., 23(1),* 11. https://doi.org/10.3390/mca23010011

Jelodar, H., Wang, Y., Yuan, C., Feng,X., Jiang, X., Li,Y., Zhao,L. (2019) *et al.* Latent Dirichlet Allocation (LDA) and topic modeling: models, applications, a survey. *Multimed Tools Appl* **78,** 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Jordan, S.E.; Hovet, S.E.; Fung, I.C.-H.; Liang, H.; Fu, K.-W.; Tse, Z.T.H. (2019). Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response. *Data*, *4*, 6. https://doi.org/10.3390/data4010006

Kwartler, T. (2017). *Text Mining in Practice with R* (1st ed.). Wiley.

Kwenti T.E. (2018). Malaria and HIV coinfection in sub-Saharan Africa: Prevalence, impact, and treatment strategies. *Res. Rep. Trop. Med.* 2018; 9:123–136. doi: 10.2147/RRTM.S154501.

Martin, F and Johnson, M. (2015) More efficient topic modelling through a noun only approach Proceedings of the Australasian Language Technology Association Workshop, pp. 111-115. U15-1013.pdf (aclanthology.org)

Maurice, N., Aicha, S., Young, H. S., Eon, K. J., Hoon, K., Junseok, P., & Won-Joo, H. (2019). Malaria Epidemic Prediction Model by Using Twitter Data and Precipitation Volume in Nigeria. *Journal of Korea Multimedia Society*, *22*(5), 588–600. https://doi.org/10.9717/kmms.2019.22.5.588

Murthy, D. (2013). *Twitter: Social Communication in the Twitter Age (Digital Media and Society)* (1st ed.). Polity Press.

Nigeria Federal Ministry of Health, (2014) National Malaria elimination program: Guidelines for Malaria Advocacy, Communication and Social Mobilisation Programmes. Nigeria Malaria ACSM Guide.pdf (thecompassforsbc.org)

Nirmala, C. R., Roopa, G. M., & Kumar, K. R. N. (2015). Twitter data analysis for unemployment crisis. Proceedings of the 2015 International Conference on Applied and Theoretical

Computing and Communication Technology, ICATccT, Davangere, India, 29-31 Oct. 420–423. https://doi.org/10.1109/ICATCCT.2015.7456920

Nghochuzie, N. N., Olwal, C. O., Udoakang, A. J., Amenga-Etego, L. N., & Amambua-Ngwa, A. (2020). Pausing the Fight Against Malaria to Combat the COVID-19 Pandemic in Africa: Is the Future of Malaria Bleak?. *Frontiers in microbiology*, June *11*, 1476. https://doi.org/10.3389/fmicb.2020.01476

Nkiruka, O., Prasad, R., & Clement, O. (2021). Prediction of malaria incidence using climate variability and machine learning. *Informatics in Medicine Unlocked*, *22*, 100508. https://doi.org/10.1016/j.imu.2020.100508

Parajuli, J. (2020). Significance of Literature Review in the Research of Social Sciences. *Journal of Population and Development*, *1*(1), 96-102. https://doi.org/10.3126/jpd.v1i1.33108.

Silge, J. and Robinson, D. (2017). Text Mining with R: A Tidy Approach. O'Reilly Media. isbn 9781491981627. https://books.google.com.ng/books?id=qNcnDwAAQBAJ

Taboada, M., Brooke,J. , Tofiloski, M. , Voll, K., Stede,M. (2011); Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*; 37 (2): 267–307. doi: https://doi.org/10.1162/COLI_a_00049

Talapko, J.; Škrlec, I.; Alebić, T.; Jukić, M.; Včev, A. (**2019)** Malaria: The Past and the Present. *Microorganisms*, *7*, 179. https://doi.org/10.3390/microorganisms7060179

World Health Organization. (2019). World malaria report 2019. World Health Organization. https://apps.who.int/iris/handle/10665/330011. License: CC BY-NC-SA 3.0 IGO

World malaria report 2020: 20 years of global progress and challenges. Geneva: World Health Organization; 2020. License: CC BY-NC-SA 3.0 IGO.

Yuping, J. 2017. Development of Word Cloud Generator Software Based on Python. Procedia Engineering. Volume 174, Pages 788-792. https://doi.org/10.1016/j.proeng.2017.01.