International Journal of Innovative Information Systems & Technology Research 12(2):50-59, April-June, 2024

 www.seahipaj.org ISSN: 2467-8562

Trend Analysis In Nigerians' Tweets Using Latent Dirichlet Allocation

WAIDOR, Tamaramiebi K. & OMAMOKE, Layefa.

Department of Computer Science,
Faculty of Basic and Applied Science,
University of Africa, Toru-Orua, Bayelsa State, Nigeria
Email: zalimaxxx@gmail.com

ABSTRACT
Knowing what is trending help organisations and even government of countries to understand the sentiments of the clients or citizens in order to make decisions that would be productive and beneficial to both parties. Particularly, in Nigeria, the government can be well informed through trend analysis to deliver democratic dividends citizens. This research aims at analysing tweets by Nigerians from X (formerly Twitter) using Latent Dirichlet Allocation and train a model for trend analysis. This we hope to achieve using Python's rich Natural Language Processing libraries. The data (document) analysed were tweets made by Nigerian in for 7 days in December, 2017, and in all it has a total of 9578 tweets. From these tweets, 5 topics were selected and our findings showed that the major trending topic were Politics, Fuel Scarcity and Yuletide season.

1.0 INTRODUCTION

Knowing what is trending help organisations and even government of countries to understand the sentiments of the clients or citizens in order to make decisions that would be productive and beneficial. A government that does not listen to the governed would not be able to formulate and implement people-oriented policies. Particularly in Nigeria, the government can be well informed through trend analysis to deliver democratic dividends citizens. Trend analysis is defined as a means of projecting the appearance of future social changes based on current and historical data (Myeong-Ha Hwang et al, 2018). There are several places where trends are prevalent, depending on the context. However, Social Networking Media gives a market place of discussion of all trends. What is in vogue can be gotten through the analysis of what a particular community is talking about. Twitter is a social media platform that focuses on text messages, which number of characters is limited to a maximum of 280 characters

Trend analysis involves Topic Modelling which is the art of identifying topics in a text (Johan Risch, 2016). There are many methods of trend analysis such as Latent Semantic Analysis (LSA), Latent Dirichlet Allocator (LDA) (Siti Qomariyah, 2019) and Probabilistic Latent Semantic Analysis (PLSA) (Hidayatullah A. F. et al., 2019) etc. In this method adopted in the research is the LDA as it a better topic coherence than LSA (Siti Qomariyah et al, 2019).

The aim of this research is to do an analysis of tweets by Nigerians from twitter using LDA and train a model for trend analysis. This we hope to achieve using Python's rich Natural Language Processing libraries.

2.0 Literature Review

Trend analysis has in recent times become an object of intense research as it outcome feeds a number of eagerly awaiting stakeholders. Myeong-Ha Hwang et al, (2018) propose a trend analysis method using LDA, composed of 5 steps and performed the trend analysis by topic using the extracted result combining LDA and Top 10 keywords. The research aimed to find the trending topic from international standards documents (ITU-T), which consist 22 series from D series to Z series. categorized into a set of international standard documents related to each subject. The results of the experiments showed that technology related to cloud computing has developed recently, consequently, data transmission issues and security issues also emerged. In addition, trends for video issues to recognize the image and smart city were also in the front burner.

Research on trending issues on social media is fast becoming an emerging area of interest. In a research carried out by Annamoradnejad et al (2019), a dataset of Twitter trending topics of 2018 was collected and analyzed using criteria such as lexical analysis, time to reach, trend reoccurrence, trending time, tweets count, and language. The dataset in these six criteria according to three conditions: First rank trends, Top 10 and Top50 list. Based on their results, 77.6% of the topics that reached the Top-10 list were trended with less than hundred thousand tweet, while over 50% of the topics could not hold the position for more than an hour. English and Arabic languages comprised close to 40% and 20% of the first rank topics, respectively**.** The result shows that other languages are fast gaining on the English language in posting trending topics compared to 2011 (Lee K., et al, 2011) when more than 87% of all trending topics were in *English* language.

Even political institutions are not left out in the matter of social media trends. Any government with intention to serve properly the governed would take keen interest in the tweets of its citizens. From citizens' tweets, vital feedbacks can be gotten to enhance governance. On focusing on the tweets of Surabaya citizens, Indonesia, Siti Qomariyah et al (2019) worked on topic modeling focused on employing Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA). The aim was the evaluation LDA and LSA performance using the topic coherence. The best topic coherence for LDA was 0.1376 and the best topic coherence for LSA was 0.1264 showing that LDA gives a better result than LSA. The optimal model demonstrates that Surabaya citizen talking about 4 major topics on Twitter: "pemilu" (election), "parkir" (parking), "lalin macet" (traffic jam), and the last one is about "lalin macet karena hujan" (traffic jam because of rain). This research demonstrates the power of text mining using data from twitter.

Latent Dirichlet Allocation (Skowster, 2019):

- **Latent:** In statistics, latent variables are variables that are not directly observed but are rather inferred through a mathematical model from other variables that are observed.
- **Dirichlet:** Similar to normal distribution, but is has the useful property that all values added together sum to one, e.g. (0.1, 0.2, 0.3, 0.4) and (0.2, 0.8)
- **Allocation:** To apportion for a specific purpose or particular person or thing.

According to Skowter (2019), LDA is the use of mathematical model based on probability distributions (Dirichlet) to infer hidden (non-observable) variables and allocate words into topic.

LDA Model – Algorithm (Skowster, 2019):

Step 1: Choose the total number of topics (*K*) on desires to discover from the set of documents.

Step 2: Go through each document and randomly assign each word in the document to one of the *K* topics.

Step 3: For each document *d,*

- For each word *w* in *d*

A. For each topic t, compute 2 things

1) The proportion of words in document *d*, currently assigned to topic *t*.
P (topic *t* | document *d*)
2) The proportion of assignment to topic t over all documents that comes from this word.
P (word *w* | topic *t*)

B. Reassign w a new topic, choosing t with probability:
P (topic *t* | document *d*) * P (word *w* | topic *t*)
- this is essentially the probability that topic *t* generated word *w*.

Step 4: Repeat the above steps until topics are changing very little.

3.0 METHODOLOGY

3.1. Data Retrieval

The data (document) analysed are tweets made by Nigerian in for 7 days in December, 2017, and in all it has a total of 9578 tweets. below is a snapshot of part of the document.

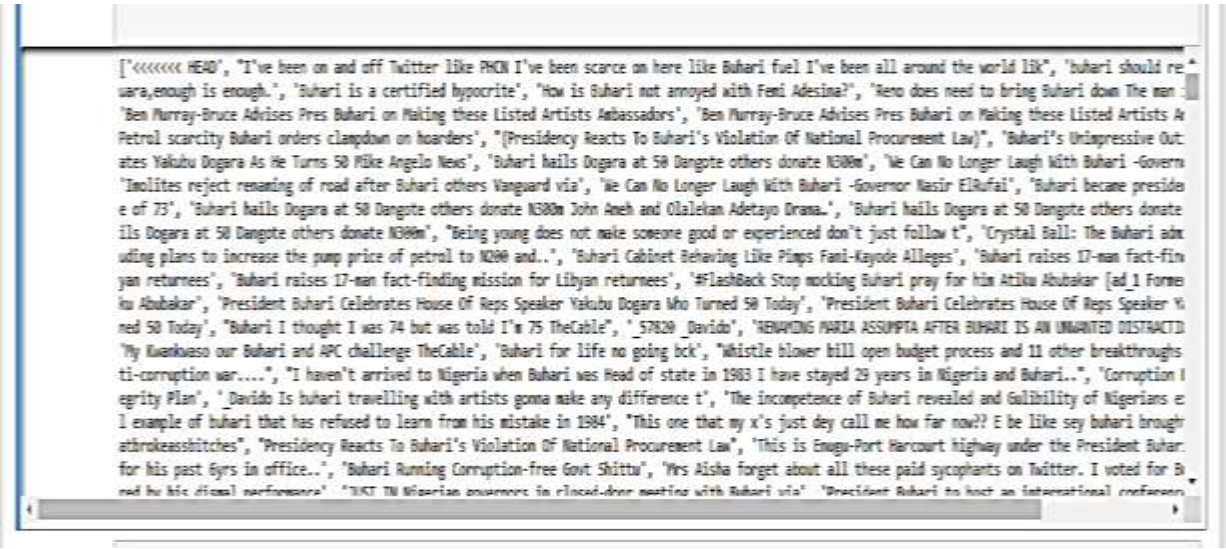

Figure 1. Snapshot of the dataset. URL: https://raw.githubusercontent.com/chibueze-oguejiofor/Project-Buhari/master/tweets.txt

Then we start by taking a look into the first 10 entries of our data. this is to ensure that the document is loaded properly into the Python editor. See figure 2 below.

Out[1]:

	<<<<<<< HEAD
0	I've been on and off Twitter like PHCN I've be...
1	buhari should resign and return to duara,enoug...
2	Buhari is a certified hypocrite
3	How is Buhari not annoyed with Femi Adesina?
4	Reno does need to bring Buhari down The man is...
5	Ben Murray-Bruce Advises Pres Buhari on Making...
6	Ben Murray-Bruce Advises Pres Buhari on Making...
7	UPDATE Petrol scarcity Buhari orders clampdown...
8	(Presidency Reacts To Buhari's Violation Of Na...
9	Buhari's Unimpressive Outing

Figure 2. First 10 entries of our data

3.2. Data Pre-Processing

Due to the unstructured nature of tweet data, some sort of pre-processing (data cleaning) is necessary as @username, URLs, Symbols, Emoticons and hashtags (#) etc. are not needed in our analysis. We start by Remove punctuation/lower casing using Python regular expression (regex) method. See a snapshot in figure 3 below after the removal of the unneeded elements and capitalization.

	HEAD
0	IVE BEEN ON AND OFF TWITTER LIKE PHCN IVE BE
1	BUHARI SHOULD RESIGN AND RETURN TO DUARAENOUG
2	BUHARI IS A CERTIFIED HYPOCRITE
3	HOW IS BUHARI NOT ANNOYED WITH FEMI ADESINA
4	RENO DOES NEED TO BRING BUHARI DOWN THE MAN IS
9573	WHEN ATIKU RERERERREDECAMPS TO APC FUTURE
9574	IN 2014 I SUPPORTED AHEAD OF GMB IN THE APC PR
9575	TOTALLY AGREE WITH YOU SIR
9576	OR THE BROOM SWEEPS YOU ALL
9577	5699E7512DB173DE1941E308F79959023598F890

9578 ROWS X 1 COLUMNS

Figure 3. dataset after removal of unwanted elements

Still in the stage of pre-processing, to help us understand the data better and ensure that the we are in the right direction, before training the model, a wordcloud of the most common words are visualized as shown in the figure 4 below.

Figure 4. Wordcloud of the most common words in the dataset.

Next step the use of Python 'gensim' and 'nltk' library to perform tokenization which is the process of converting a document to it atomic elements (Hidayatullah et al, 2019), and remove meaningless words through a process known as stop words. Examples of stop words are 'at', 'the' and 'on' etc. See a snapshot of the data in the figure 5 below are after tokenization and removal of stop words.

Figure 5. Data set after tokenization

3.3. Topic Modeling

After the pre-processing stage, the topic modeling process commences using LDA model. But first a corpus is created. This starts by constructing a document – term matrix that helps to understand how

frequently each term occurs within each document. In this regard, the corpus is a document-term matrix. A snapshot of the document – term matrix for our analysis is shown in figure 6 below.

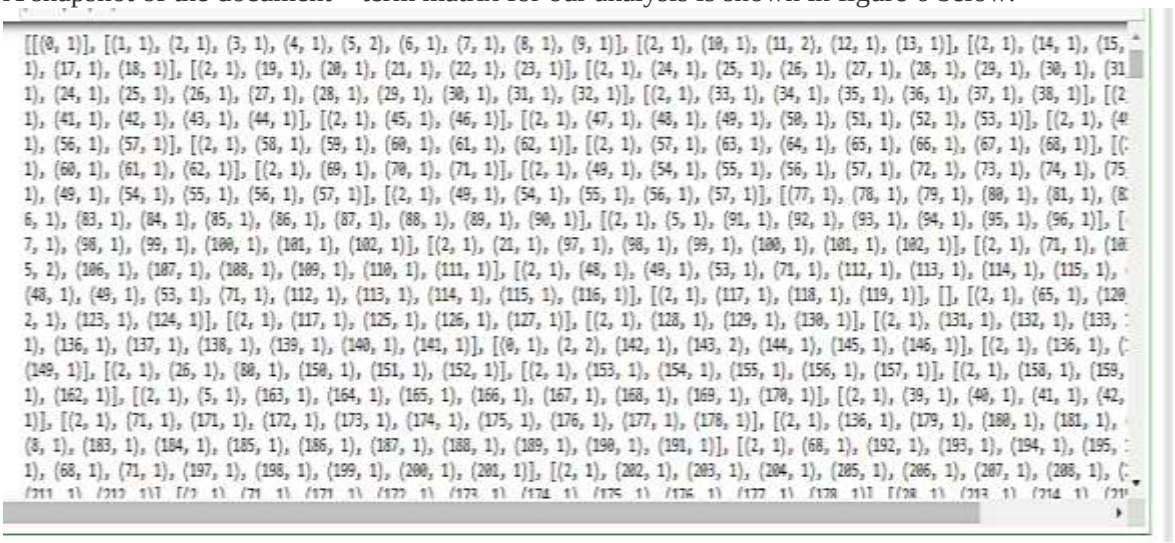

Figure 6. Snapshot of the document – term matrix

Next the LDA model object is created, visualized and analyzed.

4.0 RESULTS AND DISCUSSION

The results are visualized using the Intertopic Distance Map (See figure 7 a & b below). The words displayed in the bar charts are adjusted by Lambda, λ slider. A lambda close to 0 highlights potentially rare but more exclusive term for the selected topic. Lambda value close to 1 highlights more frequently occurring terms in the document that might not be exclusive to the topic. In this study, a lambda value close to 6 is used. The figures also show a bar chart in a descending order of the top 30 most useful terms, for interpreting a topic. The overlaid bars represent a given term's corpus-wide frequency and the topic-specific frequency (Liu et al., 2019).

From the Intertopic Distance Map, selecting topic 1, shows Buhari, president, fuel, scarcity, Nigeria in top 5 estimated term frequency within the selected topic in a descending order, while Buhari tops the chart in overall term frequency. In all 5 topics, 'Buhari' and 'atiku' tops the overall term frequency as shown in figure 7a & b.

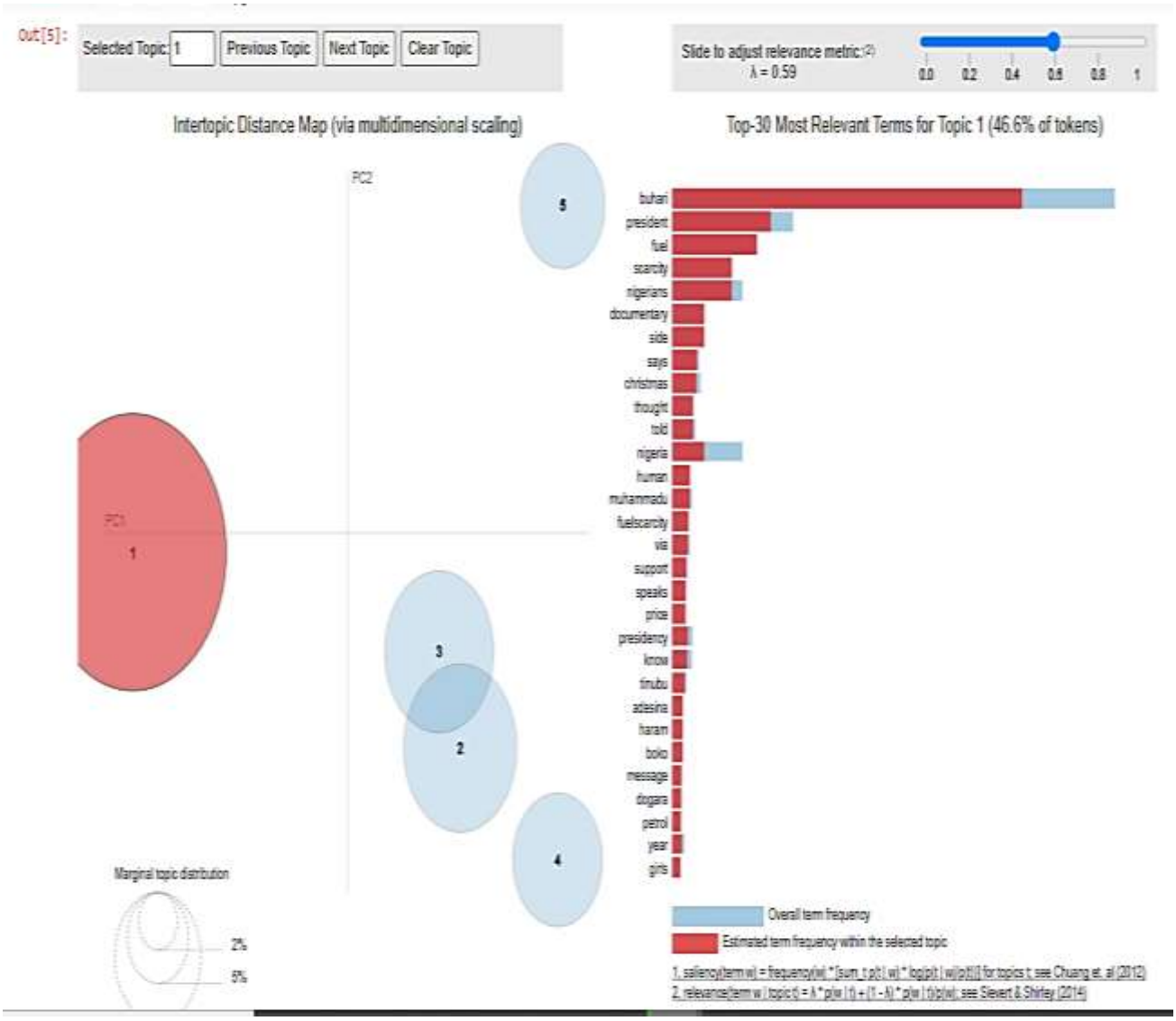

Figure 7a. Intertopic Distance Map with topic 1 selected.

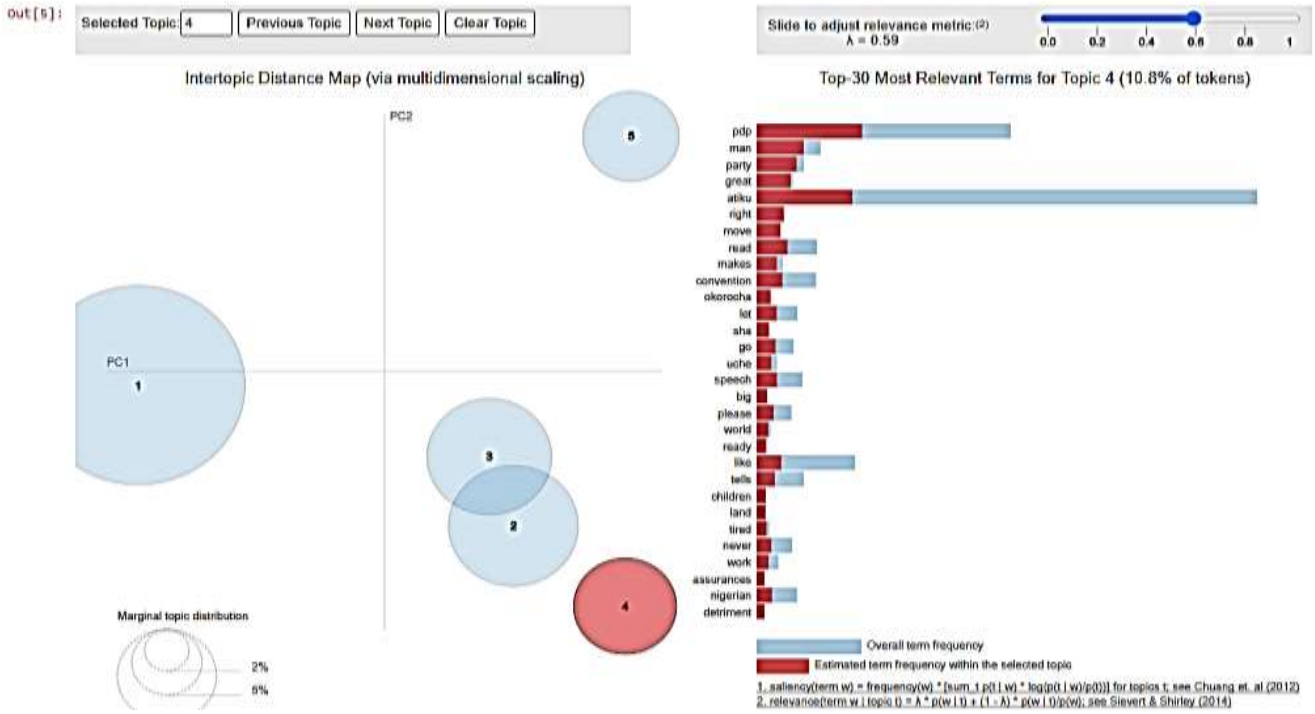

Figure 7b. Intertopic Distance Map with topic 4 selected.

```
[(0,
  '0.049*"buhari" + 0.031*"atiku" + 0.029*"president" + 0.020*"nigeria" + '
  '0.016*"good" + 0.013*"fuel" + 0.010*"scarcity" + 0.009*"reveals" + '
  '0.009*"pdp" + 0.009*"abubakar"'),
 (1,
  '0.070*"buhari" + 0.023*"atiku" + 0.019*"nigerians" + 0.019*"pdp" + '
  '0.016*"president" + 0.009*"nigeria" + 0.008*"says" + 0.007*"fuel" + '
  '0.007*"like" + 0.007*"convention"'),
 (2,
  '0.066*"buhari" + 0.025*"atiku" + 0.017*"apc" + 0.017*"president" + '
  '0.017*"nigeria" + 0.009*"fuel" + 0.008*"obasanjo" + 0.007*"becoming" + '
  '0.007*"mugabe" + 0.007*"stopped"'),
 (3,
  '0.052*"buhari" + 0.022*"fuel" + 0.019*"atiku" + 0.017*"scarcity" + '
  '0.011*"pdp" + 0.008*"president" + 0.008*"see" + 0.007*"price" + '
  '0.006*"minister" + 0.006*"nigeria"'),
 (4,
  '0.080*"buhari" + 0.009*"president" + 0.009*"pdp" + 0.009*"news" + '
  '0.008*"nigerians" + 0.008*"sir" + 0.008*"atiku" + 0.007*"christmas" + '
  '0.007*"nigeria" + 0.006*"like"')]
```

Figure 8. Output of topic generated by LDA model at one pass (iteration)

The figure 8 above shows our LDA model generated at one pass and each generated topic is separated by a comma. it can be seen that Topic One featured names such as 'buhari', 'atiku', 'president', 'nigeria', 'fuel', 'scarcity' and 'pdp' etc. This can be interpreted that the trending topic here is politics involving high ranking political figure in Nigeria. Also fuel scarcity could also be topic here with some level of proportion.

Topic two is also include terms such as 'buhari', 'atiku', 'president', 'nigerians' and 'fuel'. This is a pointer that the trending topic here is politics.

A cursory look at topic three to five also feature top political figures in Nigeria such as 'buhari', 'atiku', 'abubakar', 'obasanjo', 'fuel' speaks of politics and perhaps fuel scarcity in the country.

Topic five still of politics and Chrismas trending.

Now looking at our LDA model generated at 20 passes (iterations) tells a slightly varied story as shown in figure below. The more passes the better.

```
[(0,
  '0.040*"atiku" + 0.033*"buhari" + 0.031*"pdp" + 0.018*"apc" + '
  '0.010*"national" + 0.008*"chairman" + 0.007*"million" + 0.007*"people" + '
  '0.007*"president" + 0.007*"convention"'),
 (1,
  '0.041*"buhari" + 0.033*"good" + 0.025*"news" + 0.025*"atiku" + 0.022*"home" '
  '+ 0.017*"back" + 0.017*"jonathan" + 0.012*"luck" + 0.011*"abubakar" + '
  '0.011*"embraces"'),
 (2,
  '0.126*"buhari" + 0.035*"president" + 0.030*"fuel" + 0.021*"nigerians" + '
  '0.021*"scarcity" + 0.011*"nigeria" + 0.011*"documentary" + 0.011*"side" + '
  '0.009*"says" + 0.009*"christmas"'),
 (3,
  '0.043*"atiku" + 0.036*"nigeria" + 0.028*"reveals" + 0.026*"obasanjo" + '
  '0.024*"becoming" + 0.024*"stopped" + 0.024*"mugabe" + 0.018*"buhari" + '
  '0.015*"sir" + 0.014*"ja"'),
 (4,
  '0.036*"pdp" + 0.032*"atiku" + 0.016*"man" + 0.014*"buhari" + 0.014*"party" '
  '+ 0.012*"great" + 0.010*"read" + 0.009*"right" + 0.009*"nigerians" + '
  '0.009*"convention"')]
```

Figure 9. Output of topic generated by LDA model at 20 pass

While the topic one still points to National Politics, topic two points to perhaps the current president and former presidents embracing. Topic three points to fuel scarcity in the Yuletide season.

5.0 CONCLUSION

Topic modeling reveals the major topics in what a certain community is talking on media such as Twitter, weChat, Facebook with the intension to single out several key concerns and findings related to the topics. In the regard, data analysis and the visualization generate useful information. In the research, we able to analysis tweets from several Nigerians and the discover that the trending topics within the period covered was Politics, Fuel Scarcity and Christmas Season.

REFERENCES

Annamoradnejad, I. & Habibi, J. (2019). "A comprehensive analysis of twitter trending topics,". *5th International Conference on Web Research (ICWR), Tehran, Iran*, pp. 22-27. doi: 10.1109/ICWR.2019.8765252

Lee, D., Narayanan, R., Patwary, M. A., Agrawal, A. & Choudhary, A. (2011). Twitter trending topic classification. pp. 251–258.

Qomariyah, S., Iriawan, N. & Fithriasar, K. (2019). Topic modeling twitter data using latent dirichlet allocation and latent semantic analysis. *AIP Conference Proceedings*. https://doi.org/10.1063/1.5139825

Johan Risch (2016). Detecting twitter topics using latent dirichlet allocation. *Institutionen för informationsteknologi, Department of Information Technology, Uppsala Universitet.*

Myeong-Ha, H., Suwook H., Minkyo I & Kangchan L. (2018). A method of trend analysis using latent dirichlet allocation. International Journal of Control and Automation, *11*(5), pp.173-182. http//dx.doi.org/10.14257/ijca.2018.11.5.15

Skowster the Geek (2019, March 13). *Topic Modeling Tutorial (Latent Dirichlet Allocation) in Python* [Video]. Youtube. https://www.youtube.com/watch?v=Y79sCtzddyA

Hidayatullah, A. F., Aditya S. K., Karimah & Gardini, S. T. (2019). Topic modeling of weather and climate condition on twitter using latent dirichlet allocation (LDA). *The International Conference on Information Technology and Digital Applications IOP Conf. Series: Materials Science and Engineering.* doi:10.1088/1757-899X/482/1/012033
https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

Liu, Q, Chen, Q., Shen, J., Wu, H., Sun, Y & Ming, W.(2019). Data analysis and visualization of newspaper articles on thirdhand smoke: A topic modeling approach. .*JMIR Med Inform 7*(1)

https://raw.githubusercontent.com/chibueze-oguejiofor/Project-Buhari/master/tweets.txt

Shashank K. (2019). Topic modeling in python: latent dirichlet allocation (LDA). Https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0