Exploration of Big Data in Nigerian Digital News Media Using Machine Learning Approach (Naïve Bayes Algorithm)

Bisallah HI*, Owolabi O*, Aminat A*

*Department of Computer Science, University of Abuja, Nigeria
Hashim.bisallah@uniabuja.edu.ng

Abstract: This research focuses on exploration of big data in Nigerian digital news media using Naïve Bayes Algorithm. Social media data set was collected from three major twitter handles of some selected newspapers and analysed with intend to revealing velocity, variety and volume of the data. The research shows that most of the news posted, politics seems to be the most discussed topic followed by business and sports. The result also shows Nigerian online readership followers of all the newspapers online are the most interactive followed by African countries. The research also reveals that majority of the topics mentioned in the tweets of all the newspapers contains word that are political in nature. On the topic trends, the followers exclusively focused on sports and politics, revealing behavioral patterns in the network, discourse focus and network connectivity, demographics as determinant of interactions and connections, patterns and messages speed at the expense of business areas. The big data that has been analysed will help online content news editors, online marketers, individuals, governments and businesses understand the essence of trends of reader's perspective in other to boost readership reach, appropriate online news content to be posted by contend editors, effective decisions making having identified many patterns within it.

[Bisallah HI, Owolabi O, Aminat A. **Exploration of Big Data in Nigerian Digital News Media Using Machine Learning Approach (Naïve Bayes Algorithm).** *N Y Sci J* 2019;12(6):66-77]. ISSN 1554-0200 (print); ISSN 2375-723X (online). http://www.sciencepub.net/newyork. 9. doi:10.7537/marsnys120619.09.

Key words: Big Data, New Media, Social Media, Topic modelling, Latent Semantic Analysis, Multinomial Naïve Bayes, Latent Dirichlet Allocation, Visualization, Exploration

1.0 Introduction

As the new newspapers are springing up every day, Internet is also changing how stories are being published and patterns of people's reactions to them (stories). No doubt, heavy Internet penetration and deployment has given the citizens the opportunity to demand for accountability and transparency from their leaders through various platforms and groups offered which the conventional media cannot. Internet has increased the visibility and online presence of Nigerian newspapers to the whole world beyond the traditional circulation of hardcopy. Their uptake of the world-wide platform offered by the internet has given a wider reach than what the conventional circulation could offer. This has invariably widened the audience base of the newspapers, including the youths whose attachment to anything online is huge.

Media practitioners have been empowered by the Internet to deepen their investigative journalism practice and emboldened those to publish what ordinarily shouldn't have been easily published in the conventional media. This powerful tool has brought pool of information and data hitherto difficult to obtain at their fingertips; with a few clicks on the internet through various search engines, millions of information are aggregated and sorted for journalists to build and enrich their story and discourse. According to McCaughey and Ayers cited by Dare (2011), the immediacy of the Internet and its centrality to the work of the media has been established in Nigeria through their robust investigative reports. Also, the potential it holds as a tool for social change empowering the individual became quite obvious to millions of Nigerians. Dare (2011) succinctly espouses that the Internet is immediate, even more immediate than a daily newspaper and it could be more interactive than television. The world has never seen a communication development that can give power to individuals like the internet.

The advent of website and social networks, facilitated by the Internet, for news generation and dissemination has dramatically altered the traditional media landscape in Nigeria. Dare (2010:18-19) observes that "through the instrumentality of the Internet, emerging sites that aggregate views, information, news, comments and diverse opinions have opened a new frontier of possibilities in the way news is produced, distributed and consumed. More importantly, it has brightened the hope for a participatory and interactive process. With the emergence of the Internet and its various social media and social networking tools (Facebook, Twitter, My Space, You Tube, Blogosphere and so on), newspaper owners and publishers and media practitioners cannot afford to be traditional in their professional practice of news gathering and news dissemination.

Readers' contribution via comments, replies, share and retweets has increased volume of messages being disseminated by various newspapers on their websites. This has resulted into the alignment of data journalism with the big data era. As observed while referring to a number of scholars, availability of readers' contribution is changing the way researchers are thinking about news media generation of big data (Sudhahar, Lansdall-Welfare, Flaounas and Cristianini, 2012). This is hinged on the fact that the speed of creating data is increasing every second, the types and volumes are also growing exponentially. Evidently, big data has reached a point which majority of businesses especially the Fortune 1000 firms cannot do without it. Businesses are now viewing big data as very important or critical to their operational effectiveness and attainment of corporate goals (New Vantage Partners, 2016).

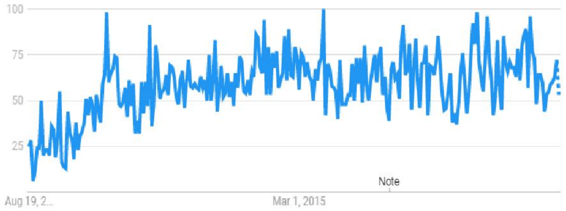


Figure 1: Big Data Trends in Books and Literature 2012-2017
Source: Google Trends, 2017

Based on the foregoing, the study seeks to investigate news dissemination Internet-enabled channels that generate big data for Nigerian newspapers and news patterns (politics, business and sports) they had the most. The study also intends to determine time it took readers to reply to tweets or posts of the newspapers and the likelihood of the previous users to be replied again. The geographical locations of the people who replied to the news published on Twitter is part of the investigations.

1.1 Statement of the Problem

Available studies reveal that scholars have directly or indirectly focused on investigation of 4 Vs-velocity, veracity, variety and volume of big data. Salehi, Moradi and Pour (2010) exclusively investigated the veracity involved in financial reporting through the examination of quality and manner of financial reporting of corporations listed in Tehran Stock Exchange on the internet. The scholars did computational analysis of the narratives of the US Elections 2012 on the Internet.

None of these studies has explored volume, variety and velocity of news stories published by the Nigerian newspapers on their websites and social networking sites. This is the gap the current study seeks to fill.

1.2 Purpose of the Study

The study will exclusively focus on the exploration of data abound on twitter platforms of the Nigerian newspapers with additional intent of reveling velocity, variety and volume of the data. Factors such as message age, previous interactions and sending rate, affecting users or readers of the newspapers on the two Internet-enabled information dissemination will be explored. In specific terms, number of stories or information created and disseminated posted on Twitter will be investigated. The outcomes of the study are expected to enable the researcher develop ***Bisallah's News Network Model*** **(BNNM)** for the newspapers. The model is expected to be **predictive** and useful in determination of users' engagement patterns and interconnectivity with the news stories published by the newspapers.

1.3 Scope of the Study

This study will be limited to exploration of big data available on twitter sites of the Nigerian newspapers. The study does not cover electronic media and at the same time exclude magazine aspects of the Nigeria's print media industry.

1.4 Significance of the Study

Outcomes of the study will be useful to a number of stakeholders in print media industry and academic researchers. Media practitioners, especially online news editors and Information Technology Specialists would know likely existence of interconnectivity between the news stories and readers on their Internet-enabled news dissemination platforms. Further academic research would be able to pinpoint areas that need for investigation in journalism and other facets of the Nigeria's society where varied data are being churned out every second from different sources.

1.5 Research questions

1. Which of the news dissemination Internet-enabled channels generate big data for Nigerian newspapers?

The main thrust of this question is to unearth the how twitter that generate large data for the selected newspapers. This becomes imperative on the view of knowing appropriateness of the media in disseminating the news and public's reactions to them.

2. What is the relationship between data on the twitter networks of the newspapers?

This research question aims at revealing correlation that exists between the various data available on the newspapers twitter networking sites created by the newspapers and their followers on the sites. Specifically, the intention is to find out significant relationship between existing data on online platforms of the newspapers and readers' reactions to the posted news.

3. Which of the news patterns (sports, business and political) the newspapers had the most?

Since newspapers gather and disseminate different happenings daily, the focus of this question is to determine categories of news that dominate the twitter sites of the newspapers mostly considering the various socio-economic and political incidents in the country.

4. What are the geographical locations of the people who replied to the news published on the posted on Twitter?

Readers are expected to express their views in the comment section that accompanies each story after reading it. At the same time, they are also likely to react to the story by commenting, replying, sharing and rewetting the story on the twitter media sites. Thus, it is necessary to know how the readers are connected in terms of geographical closeness and connectivity.

5. How long does a tweet wait in the timeline to be replied?

With this question, I intend to know how many second, minute and hour a story posted on social media spent before readers reacted to it. This will go in a long way of knowing the readers' eagerness of being abreast of happenings around them within the context of emerging new technologies, especially social-inclined ones.

6. Are previously replied users more likely to be replied again?

This question will help me in understanding the extent to which social connection or network is being formed among the readers. In other words, are there passive or active commenters to the stories posted on the twitter sites of the newspapers?

Table 1: Derivation of Study's Research Questions from Network Theory

S/N	Research Question	Network Theory's Proposition and Assumption
1.	Which of the news dissemination Internet-enabled channels generate big data for Nigerian newspapers?	The amount of information sent depends on the importance of the vertex itself and the strength of the relation to the receiving vertex.
2.	What is the relationship between data twitter network of the newspapers?	Vertex and edge are the two constructs of the theory. They help in understanding connection which exists between or among discrete objects.
3.	Which of the news patterns (sports, business and political) the newspapers had the most?	Vertices or edges appearing or disappearing
4.	What are the geographical locations of the people who replied to the news published on the Twitter?	Eigenvector centrality is applicable to weighted, unweighted, directed and undirected graphs. This is an iterative process, where in each step a vertex sends information to its neighbours.
5.	How long does a tweet and post wait in the timeline to be replied?	The average distance from the source vertex to any other vertex within the graph is usually measured.
6.	Are previously replied users more likely to be replied again?	Betweenness centrality measures to what degree actors have to go through a specific actor in order to reach other actors

2.0 Methodology/Theoretical Framework

Newspaper, new media, social media, social network and big data are the basic concepts that is discussed and understood that enable us develop our thoughts in relation with the dependent (Newspapers) and independent variables (news categories) of this research. The current study is better understood through propositions and assumptions of Network Theory. The theory hinges on graphs as a depiction of either symmetric relations or of asymmetric relations between discrete objects (Lee, 2009).

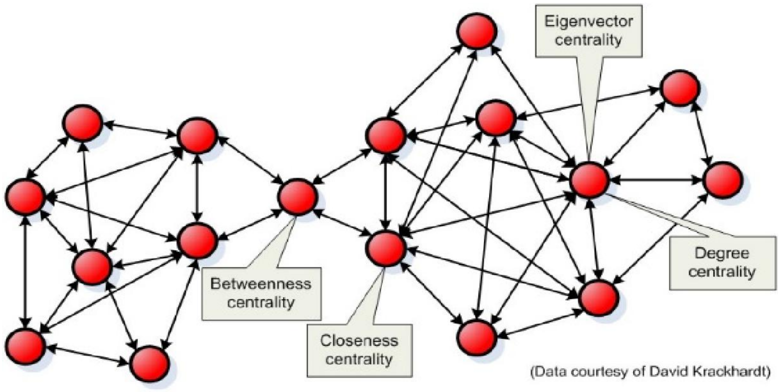


Fig 1: A Simple Network

The approaches to this study are presented thus; the diagrammatic representation of the research approach is described in figure 2:

i. Requirement Specification
ii. Data Collection
iii. Pre-processing of tweets collected (cleaning and removing of duplicate tweets)
iv. Running of Sentiment Analysis on each newspaper
v. Social Network Analysis/Visualization

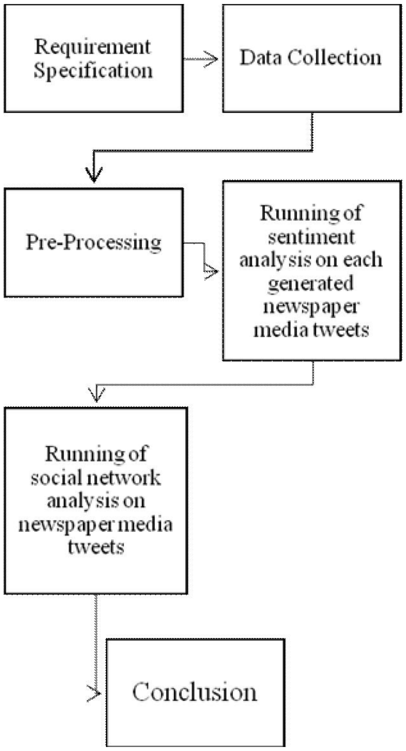

Requirement Specification
Data Collection
Pre-Processing
Running of sentiment analysis on each generated newspaper media tweets
Running of social network analysis on newspaper media tweets
Conclusion

Fig 2 Diagrammatic representation of the research approach

2.1 Requirement Specification

Tweets are extracted and filtered automatically. The tweet shall not contain any images since only English texts are considered for the research study.

2.2 Data Collection

Data collection with respect to this research is the process of getting tweets from Twitter that is related to the selected newspapers.

The method used to collect the tweets from twitter was searching of tweets matching to the hash-tags. This method was adopted in this study because the need to cover the newspaper dissemination context by using hash-tags, therefore reducing the number of unused tweets. The tweets collected were saved in excel spreadsheet format.

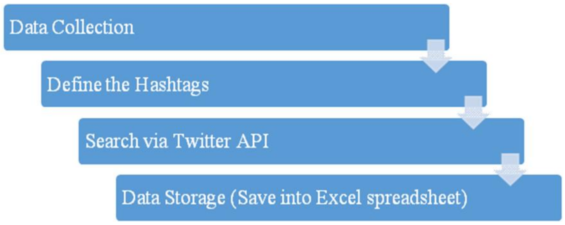


Fig 3: Methods for Data Collection for Analyzing Tweets using Python

Table 1. Newspapers' Hashtags

Newspaper selected	Search Keywords (Hashtags)
Vanguard newspaper	#Vanguardngr
Punch newspaper	#Punchngr
Guardian newspaper	#Guardianngr

The data generated from Twitter API is stored in an Excel spreadsheet. The format used was. CSV (comma separated values) for the collected files because the data consists of several fields. The fields are separated with a comma; therefore, it is very convenient to access the particular field which consists of text.

Separate directories were used to store tweets of different newspaper files. The stored data was imported to the snippet and further proceed for analysis. Once the tweets are stored, there is need to pre-process the data before applying and running network analysis classification because the data collected from twitter API is not fit for data mining.

2.3 Data Pre-processing

Data Pre-processing plays an important role to reduce the noise in the dataset. A significant amount of techniques is applied to data in order to reduce the noise of text, reduce dimensionality, and the cleaning of data by removing html references, punctuation symbols, stop words, numbers and candidate names and addresses, also the removal of retweets to remove duplicate tweets.

The techniques include:

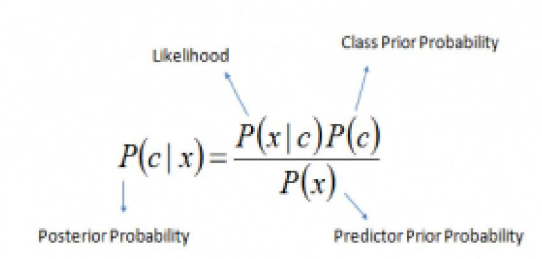
i. Removal of URLs
ii. Removal of duplications
iii. Removal of punctuation Remove stop words
iv. Converting all Uppercase to Lowercase

2.4 Model Development

2.4.1 Machine Learning Approach

Multinomial Naive Bayes is a supervised learning technique which is used for the structuring and the categorization of data. The model adopted for this research is based on it.

A set of bag of words (S) from a corpus is given as an input to the Naive Bayes classifier.


Likelihood
Class Prior Probability
$P(c\,|\,x)=\frac{P(x\,|\,c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability

$$P(c\,|\,X) = P(x_1\,|\,c)\times P(x_2\,|\,c)\times\cdots\times P(x_n\,|\,c)\times P(c)$$

- ***P*(*c*|*x*) is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).**
- ***P*(*c*) is the prior probability of *class*.**
- ***P*(*x*|*c*) is the likelihood which is the probability of *predictor* given *class*.**
- ***P*(*x*) is the prior probability of *predictor*.**

The dataset contains the topics sports, politics, and business as class labels, hence an attempt to classify the corpus based on each category. The corpus is characterized into bag of words and it follows the Bayes theorem rule for further classification (F. Semastiana, 2002).

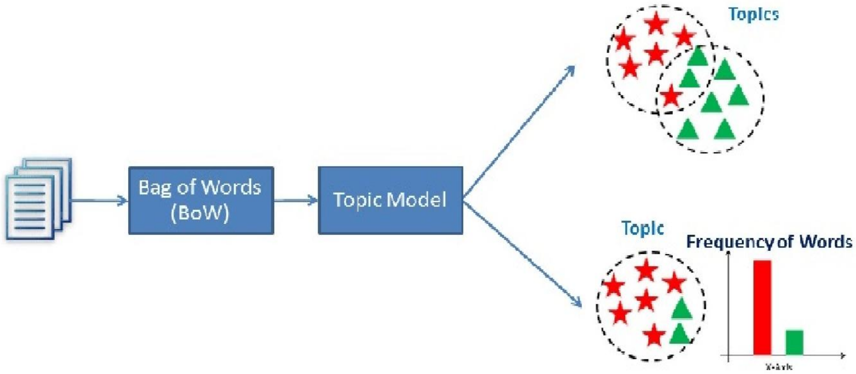


Fig 4: A set of bag of words (S) from a corpus is given as an input to the Naive Bayes classifier.

2.4.2 Naive Bayes pseudocode

```
Procedure:
Input: Set of bag of words {Term Document Matrix}
Output: $P(c_i/w_d)$ – conditional probability of class $c_i$ given word $w_d$
    1. For each $c_i$ belongs to C,
           • find prior probability of $P(c_i)$
    2. Count the words ($w_d$)
    3. Find conditional probability of word w for given class $c_i$- $P(w_d/c_i)$
    4. Compute $P(c_i/w_d) = P(w_d/c_i)\ P(c_i)\ /\ P(w_d)$
```

The Pros and Cons of Naive Bayes

Pros:

It is easy and fast to predict class of test data set. It also perform well in multi class prediction. When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.

It perform well in case of categorical input variables compared to numerical variable (s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

Cons:

If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as "Zero Frequency".

Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

2.5 Topic Modelling Approaches

2.5.1 Latent Semantic Analysis (LSA)

LSA is a technique used to extract and infer the relationship between the related words in a given context. Building a LSA does not involve any human-computer interaction or user-defined dictionaries, grammars, semantic networks etc (Landauer, T. Foltz, D. Laham, 1998). LSA receives the unprocessed data as input, parses them to get terms (unique character strings). These processed strings are separated into meaningful groups such as sentences or paragraphs.

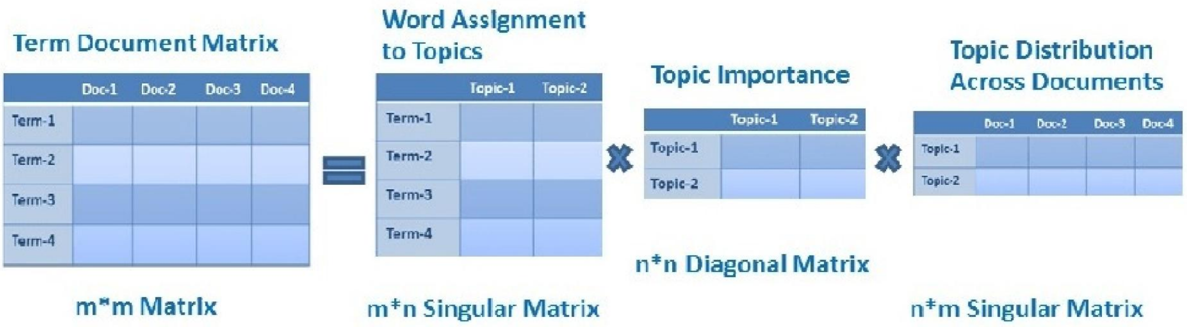


Fig 5 Term Document Matrix Steps involved in LSA Algorithm

2.5.2 Text Matrix

Corpus has been represented in matrix format which specifies the frequency of the word for each term present in the document. Column [1..n] specifies the document and row 1 specifies the terms. Corpus is represented as bag of words using which TF-IDF matrix is created.

2.5.3 Generating TF-ID (Term Frequency-Inverse Document Frequency)

TF-ID which describes the number of occurrence of word in a document. It is used for removing stop words.

It is mathematically represented as follows:

$$\text{Tf-id} = (N_{i,j} / N_{*j}) * \log (D/D_j)$$

Ni, j – No. of times word i appear in document in j

N*j – No. of total words in document j

D - No. of document (Represented in number of Column)

Dj- No. of document in which words i appears.

2.6 Implementation using Python

2.6.1 Dataset Description

The dataset being used here was obtained from the different twitter handles of the newspaper platforms using the hashtags belonging to this newspapers.

2.6.1 Preprocessing

In text mining techniques, pre-processing plays a significant role. Preprocessing is the initial step in text mining (Vikasthada, Dr. Vivekjaglan, 2013). Here, the first step is to convert unformatted data to plain text files, then remove all numbers, signs, symbols, non-English letters, stop words, convert all English letters to lowercase and perform stemming (K. R. Bindu, Latha Parameswaran, K. V. Soumya, 2015).

2.6.2 Python-Packages

With this, a new latent semantic space can be constructed over a given document-term matrix []. To ease comparisons of terms and documents with common correlation measures, the space can be converted into a text matrix of the same format as text matrix (Deerwester, S. Dumais, S. Furnas, G. Landauer, R. Harshman, 1990).

The library functions used for this research are: -

Sno	Python Libraries	Uses or Functions
1	Numpy (np)	Is a multidimensional array used to store values of datatype
2	Pandas (pd)	It provides high-performance, easy to use structures and data analysis tools.
3	Matplotlib (plt)	Used to create 2D graphs and plots by using python scripts.
4	Sciklearn (skl)	Transforming input data such as text for **use** with machine learning algorithms
5	Networkx (nx)	package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks
6	Pylab (pl)	Used with Matplotlib to compute graphs

3.0 Discussion of results results

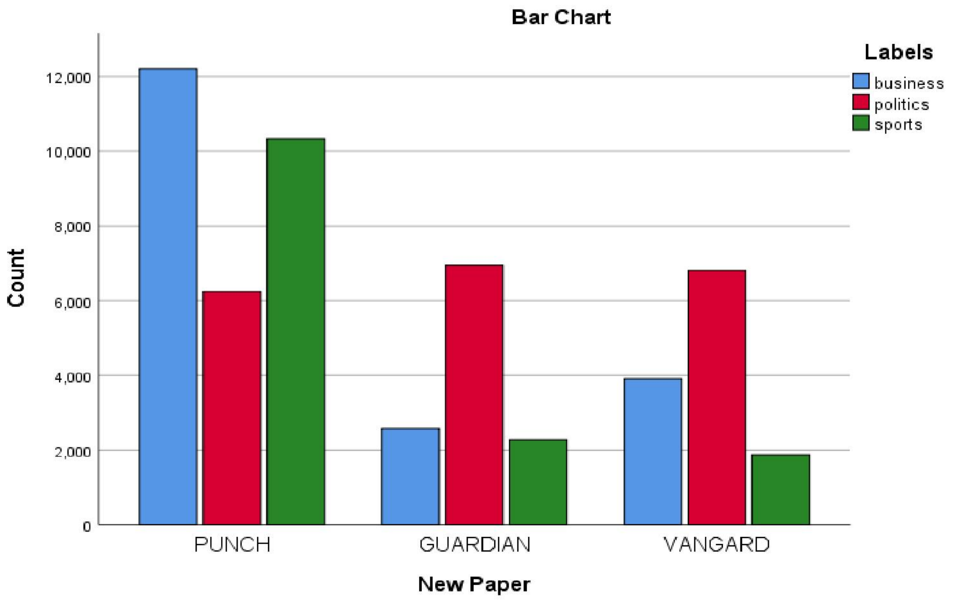
3.1 Machine Learning Approach Results

Here, the machine learning classification and clustering algorithms for predicting random text gives a near perfect prediction. On experimented with a giving input using different forms of random tweet texts, the results shows the prediction was done by the model as either business, politics or sports.

3.2 Exploratory Approach Results

Cross tabulation (Newspaper vs Headline Type)

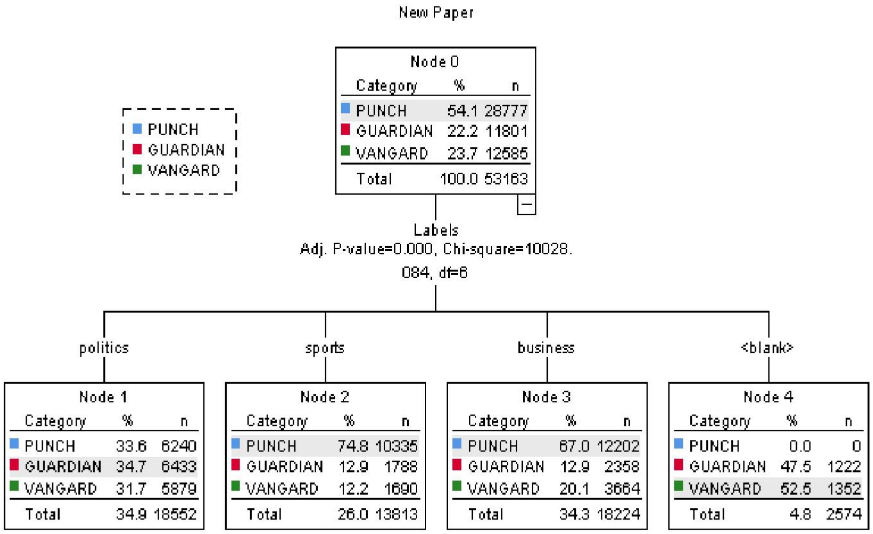
			Labels			Total
			Business	politics	sports	
Newspaper	PUNCH	Count	12202	6240	10335	28777
		% within Labels	65.3%	31.2%	71.4%	54.1%
	GUARDIAN	Count	2578	6951	2272	11801
		% within Labels	13.8%	34.8%	15.7%	22.2%
	VANGARD	Count	3906	6804	1875	12585
		% within Labels	20.9%	34.0%	12.9%	23.7%
Total		Count	18686	19995	14482	53163
		% within Labels	100.0%	100.0%	100.0%	100.0%



3.3 Model Summary

	Independent Variables	Labels
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	Labels
	Number of Nodes	5
	Number of Terminal Nodes	4
	Depth	1

This is a decision tree table by the CHAID algorithm. The acronym CHAID stands for *Chi*-squared Automatic Interaction Detector. The name is derived from the basic algorithm that is used to construct (non-binary) trees, used for classification problems (when the dependent variable is categorical in nature) relies on the *Chi*-square test to determine the best next split at each step.



It was observed that at node 0, the punch newspaper represents most the news categories we have (Politics, Sports and Business) with 54.1% to 22.2% of Guardian and 23.7% of Vanguard. Furthermore, for the politics news category is mostly reported by Guardian newspaper with 34.7% to Punch newspaper of 33.6% and Vanguard with 31.7% politics reporting index. For the sport news category, punch newspaper is observed to have most sport reporting index of 74.8%, while the Guardian and Vanguard newspaper reports sport at a little above 12% reporting index. In the node 3, it was also observed that the punch newspaper reports most of the business news with 67.0% reporting index. Interesting the punch newspaper under the node 4 has 0% for other news, this implies that all the news published by the Punch newspaper is basic of the three categories (politics, business and sports).

3.3.1 Target Category: PUNCH

Gains for Nodes						
Node	Node		Gain		Response	Index
	N	Percent	N	Percent		
2	13813	26.0%	10335	35.9%	74.8%	138.2%
3	18224	34.3%	12202	42.4%	67.0%	123.7%
1	18552	34.9%	6240	21.7%	33.6%	62.1%
4	2574	4.8%	0	0.0%	0.0%	0.0%
Growing Method: CHAID Dependent Variable: Newspaper						

The table showing the distributions of Punch news headlines across the 4 nodes (Politics, business, sport and others).

3.3.2 Target Category: GUARDIAN

Gains for Nodes						
Node	Node		Gain		Response	Index
	N	Percent	N	Percent		
4	2574	4.8%	1222	10.4%	47.5%	213.9%
1	18552	34.9%	6433	54.5%	34.7%	156.2%
2	13813	26.0%	1788	15.2%	12.9%	58.3%
3	18224	34.3%	2358	20.0%	12.9%	58.3%
Growing Method: CHAID Dependent Variable: Newspaper						

The table showing the distributions of Guardian news headlines across the 4 nodes (Politics, business, sport and others).

3.3.3 Target Category: VANGARD

Gains for Nodes						
Node	Node		Gain		Response	Index
	N	Percent	N	Percent		
4	2574	4.8%	1352	10.7%	52.5%	221.9%
1	18552	34.9%	5879	46.7%	31.7%	133.9%
3	18224	34.3%	3664	29.1%	20.1%	84.9%
2	13813	26.0%	1690	13.4%	12.2%	51.7%
Growing Method: CHAID Dependent Variable: Newspaper						

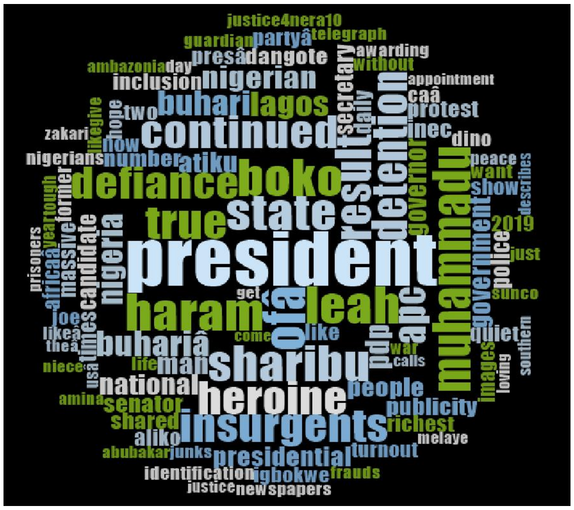
The table showing the distributions of Guardian news headlines across the 4 nodes (Politics, business, sport and others).

3.4 Word frequency Guardian

Word	Length	Count	Weighted Percentage (%)
President	9	1653	1.34
Ofâ	3	1120	0.91
State	5	1071	0.87
Book	4	1067	0.87
Haram	5	1061	0.86
Leah	4	1061	0.86
Sharibu	7	1030	0.84
Muhammadu	9	1000	0.81
True	4	965	0.78
Result	6	948	0.77

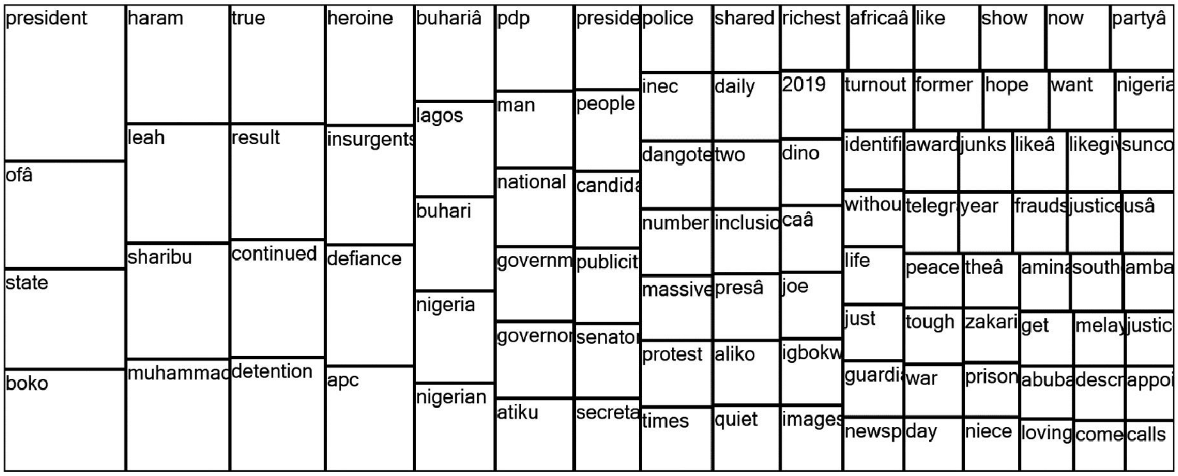
Word frequency showing the top most frequently mentioned words in the headlines for Guardian newspaper. The choice of words observed suggests strongly that the newspaper twitter post more of political news as seen in the register of words above.

3.4.1 Word cloud Guardian



The word cloud presents the first 100 most frequent words in all the headline tweets of the newspaper. The words font size is proportional to its frequency of occurrence.

Tree map Guardian



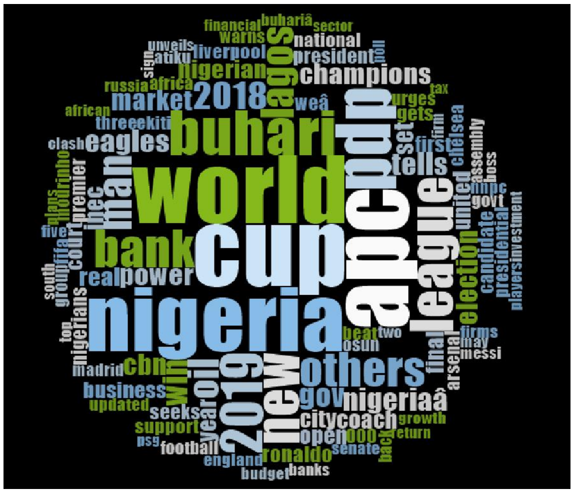
The tree map is a chart that further explains the links within words in the headline tweets of the newspaper company. Supporting the findings of the word frequency.

Word frequency Punch

Word	Length	Count	Weighted Percentage (%)
Cup	3	1440	0.72
Apc	3	1390	0.70
World	5	1306	0.65
Nigeria	7	1233	0.62
Pdp	3	986	0.49
Buhari	6	866	0.43
League	6	803	0.40
New	3	793	0.40
2019	4	758	0.38
Bank	4	691	0.35

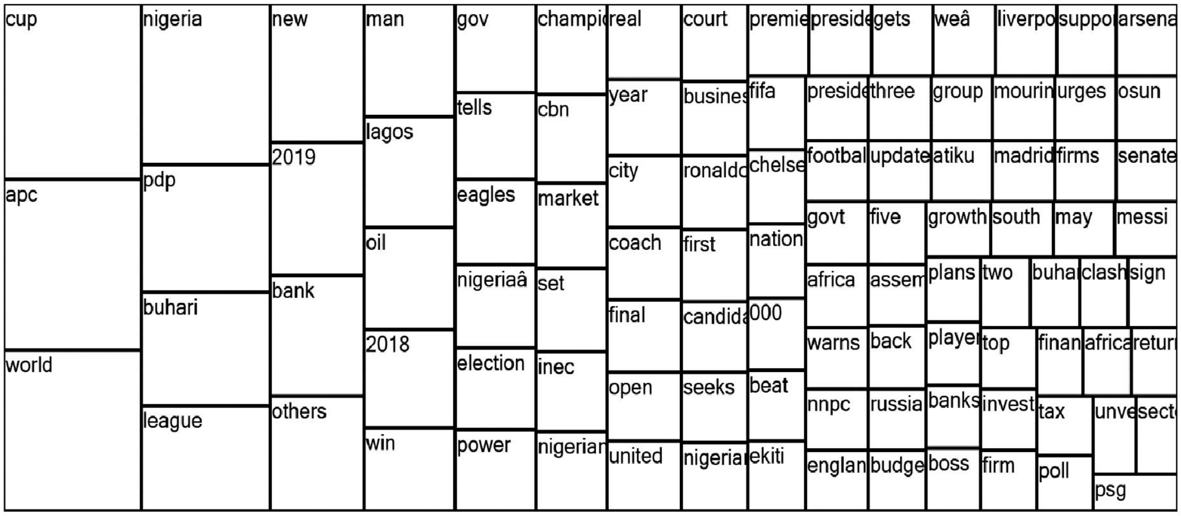
Word frequency showing the top then most frequently mention words in the headlines for Punch newspaper. The choice of words observed suggest strongly that the newspaper twitter post more of sports and political news as seen in the register of words above.

Word cloud Punch



The word cloud presents the first 100 most frequent words in all the headline tweets of the newspaper. The words font size is proportional to its frequency of occurrence.

Tree map Punch



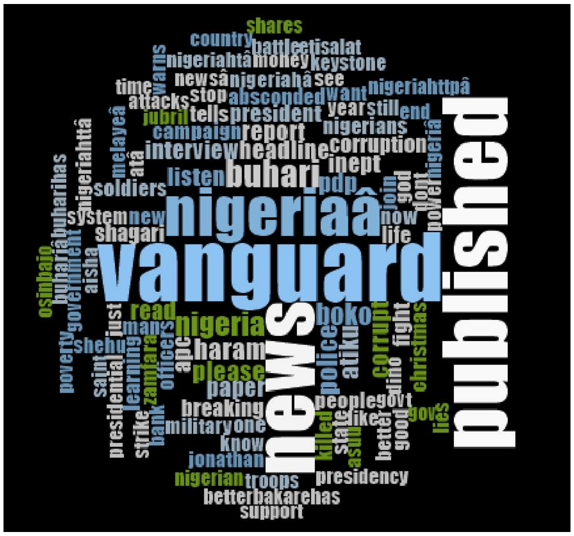
The tree map is a chart that further explains the links within words in the headline tweets of the newspaper company. Supporting the findings of the word frequency.

Word frequency Vanguard

Word frequency showing the top then most frequently mention words in the headlines for Vanguard newspaper. The choice of words observed suggest strongly that the newspaper twitter post more of sports and political news as seen in the register of words above.

Word cloud Vanguard

Word	Length	Count	Weighted Percentage (%)
Vanguard	8	4679	4.49
Published	9	4416	4.24
News	4	4059	3.89
Nigeriaâ	8	2941	2.82
Buhari	6	1120	1.07
Nigeria	7	808	0.78
Book	4	646	0.62
Haram	5	615	0.59
Police	6	573	0.55
Please	6	532	0.51



The word cloud presents the first 100 most frequent words in all the headline tweets of the newspaper. The words font size is proportional to its frequency of occurrence.

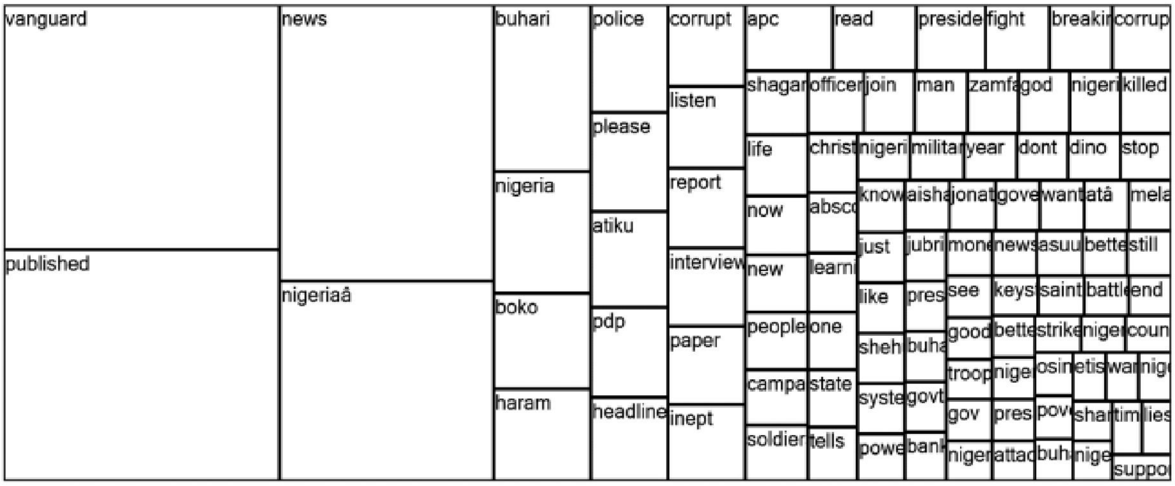
Tree map Vanguard

The tree map is a chart that further explains the links within words in the headline tweets of the newspaper company. Supporting the findings of the word frequency.

Summary and Conclusions

The research focuses on exploration of big data in Nigerian digital news media. Data set was collected from three major twitter handles of some selected newspaper and analysed with intend to revealing velocity, variety and volume of the data.

The big data that has been analysed will help online content news editors, online marketers, individuals, governments and businesses understand the essence of trends of reader's perspective in other to boost readership reach, appropriate online news content to be posted by contend editors, effective decisions making having identified many patterns within it.



The research also shows that most of the news posted, politics seems to be the most discussed topic followed by sports and business.

The result also shows that Nigerian online readership of all the newspapers online followers are the most interactive followed by other African countries.

The research also reveals that majority of the topics mentioned in the tweets of all the newspapers contains word that are political in nature.

On the topic trends, the followers exclusively focused on sports and politics, revealing behavioral patterns in the network, discourse focus and network connectivity, demographics as determinant of interactions and connections, patterns and messages speed at the expense of business areas.

In conclusion, having applied machine learning approaches (Naive Bayes, Topic Modeling (Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), the analysis made and the resultant prediction, the model can now be referred to as **Bisallah's News Network Model (BNNM)** that can be used by the Nigerian Newspapers online editors, online marketers and academic researchers.

References

1. Archenaa1, J., and Anita, E.A.M., (2015) "A Survey Of Big Data Analytics in Healthcare and Government" *Procedia Computer Science* Vol.50 Pp408–413 Available online at www.sciencedirect.com.
2. Bollen, J., Mao, H. and Zeng, X. (2010) 'Twitter mood predicts the stock market', *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8.
3. Dare, S., (2011) "The Rise of Citizen Journalism in Nigeria – A Case Study of Sahara Reporters" *Reuters Institute Fellowship Paper, University of Oxford.*
4. Gomaa, W.H., and Fahmy, A.A., (2013) "A Survey of Text Similarity Approaches" *International Journal of Computer Applications* Volume *68*– No.13 Pp 13-18.
5. Hassan, I., Latiff, M. N., and Atek, E. S., (2015) "Readers' Motivations towards Online Newspaper Reading in North Western Nigeria" *International Journal of Academic Research in Business and Social Sciences* Vol. 5, No. 8 Pp 197-209 Available on DOI: 10.6007/IJARBSS/v5-i8/1776 URL: http://dx.doi.org/10.6007/IJARBSS/v5-i8/1776.
6. Hemsley, J., and Eckert, J., (2014) "Occupied with Place: Exploring Twitter Resistance Networks" b In*i Conference 2014 Proceedings* pp372–387 doi:10.9776/14114.
7. Henrich, N., and Holmes, B., (2013) "Web news readers' comments: Towards developing a Methodology for using on-line comments in social inquiry" *Journal of Media and Communication Studies* Vol. 5(1), pp. 1-4 Available online at http://www.academicjournals.org/JMCS.

6/19/2019