Predicting Nigerian Stock Market Returns based on Daily Business News Headlines using Support Vector Machine

Preprint · January 2020

DOI: 10.13140/RG.2.2.27060.19849

4 authors, including:

OlaOluwa Simon Yaya
University of Ibadan
111 PUBLICATIONS **240** CITATIONS

SEE PROFILE

Olanrewaju Ismail Shittu
University of Ibadan
53 PUBLICATIONS **191** CITATIONS

SEE PROFILE

Ngozi Victor Atoi
Central Bank of Nigeria
11 PUBLICATIONS **49** CITATIONS

SEE PROFILE

Predicting Nigerian Stock Market Returns based on Daily Business News Headlines using Support Vector Machine

OlaOluwa S. Yaya
Computational Statistics Unit, Department of Statistics, University of Ibadan, Ibadan, Nigeria
Email address: os.yaya@ui.edu.ng; o.s.olaoluwa@gmail.com

Cornelius Ayantse
Computational Statistics Unit, Department of Statistics, University of Ibadan, Ibadan, Nigeria
Email address: adeda2222@gmail.com

Olanrewaju I. Shittu
Mathematical Statistics Unit, Department of Statistics, University of Ibadan, Ibadan, Nigeria
Email address: oi.shittu@hotmail.com

Ngozi V. Atoi
Department of Statistics, Central Bank of Nigeria
Email address: nvatoi@cbn.gov.ng

Abstract

The impact of the Nigerian daily news headline on the daily returns of the Nigerian Stock Exchange (NSE) is investigated in this paper. Vanguard newspaper is used as source of information, to gather texts in the corpus, and text mining technique such as Term Frequency-Inverse Document Frequency (TF-IDF) is applied to pre-determine important words combination and their influences on the daily market returns, while Support Vector Machine (SVM), with five kernel types are used to predict the direction of stock returns based on words' combination. A combination of two, three and four words are considered, to see how they influence the stock market. Using sensitivity, specificity, negative-predictive value, positive predictive value and accuracy as judging criteria from the SVM, the results show high predictive evidence of news headlines for NSE indices, but our findings are limited to available small sample size, meanwhile, it is hoped that the sensitivity will increase with large sample, which is left for future research.

Keywords: Nigerian Stock Market; News Headlines; Support Vector Machines; Term Frequency-Inverse Document Frequency (TF-IDF)

JEL Classification: C00, C63, H54, R53

Corresponding Author:
OlaOluwa S. Yaya
Computational Statistics Unit
Department of Statistics
Faculty of Science
University of Ibadan
Nigeria

Email: os.yaya@ui.edu.ng
Phone: +2347052185573; +2348094841881

1. Introduction

The Nigerian Capital Market is a channel of mobilizing long-term funds by providing a mechanism for private and public savings through financial instruments (equities, debentures, bonds, and stocks) with major components consisting of the Security and Exchange Commission (SEC) and the Nigerian Stock Exchange (NSE). Founded in 1960, the NSE is the second-largest market in sub-Saharan Africa with a fully automated exchange that provides the listing and trading services as well as electronic Clearing, Settlement and Delivery (CSD) services through Central Securities Clearing System (CSCS). The exchange keeps on evolving as a competitive market and meeting the needs of investors. It operates fair, orderly and transparent markets with over 200 listed equities and 258 listed securities, and had attracted the best of African enterprises as well as the local and global investors (Dallah and Adeleke, 2018).

The NSE has series of financial news, published in the newspapers over the years, and coupled with the fact that political and other structural influences hinder NSE operations and pricing, therefore, there is the need to embrace alternative computational strategy to predict activities in the stock market other than usual econometrics and time series models. Data mining has been a growing area for the analysis of stock market prices. This has given rise to the analysis of all forms and kind of data, which before the arrival of data mining techniques, were not possible. One such method is the analysis of text data, which has been on an increase and text data has been in relation to different spheres of life. One of such sphere is in the area of stock market.

Text mining is the process of distilling actionable insights from text. (Ted, 2017). The user interacts with a document collection over time and filtration of text is done by certain machine learning tools. During the process, useful information is extracted from data sources by identifying and exploring patterns. Data sources for Text mining are document collections,

and interesting patterns are in the unstructured textual data in the document collections (Feldman and Sanger, 2007). Text mining technique is mainly used for extracting a pattern from unstructured data (Yogapreethi and Maheswari, 2016). With the high level of textual data on daily basis, and with much interest in the stock market and the capacity to predict with much accuracy, whether there will be increase or decrease in a stock market, we would like to see the impact of Nigerian daily headline textual data on the Nigerian stock market index.

The term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistic, commonly used in information retrieval and Text mining, which serves as a weighting factor to reflect how important a term (word) is to document relative to the whole document collection or corpus. The basic intuition captured by the TF-IDF is that, the more often a term occurs in a document, the more it is a representative of its content while the more documents a term occurs in, the less discriminating the term becomes. Thus, the computed TF-IDF value increases proportionally to the number of occurrences of a word in the document.[1] Variations in TF-IDF values are often used in search engines, stop-words filtering and text summarization. The TF-IDF gives mere descriptive measures, whereas Support Vector Machine (SVM), as a supervised learning model, based on algorithms makes classification using a regression approach. Apart from linear classification, the SVM effectively computes the nonlinear classification using kernel tricks. Both TF-IDF and SVM are used in this work. TF-IDF is used for description of the dataset while SVM is used for prediction.

Machine learning approach is therefore used in this paper to predict stock price movement in the NSE using daily data from May 27th, 2014 to November 1st, 2019. Vanguard online newspaper was used for web scraping as this was the only well-known Nigerian

[1] 83 percent of research paper-based recommender systems in digital libraries use TF-IDF.

newspaper with a long historical record of news headlines we could access at the time of the analysis, for business news.

The remainder part of the paper is structured as follows: Section 2 presents relevant literature review. Section 3 presents the machine learning methods used, section 4 presents the data and results obtained while section 5 concludes the paper.

2. Literature Review

The need to understand the movement of stocks has been an interesting area of study. Having a model with a higher degree of precision has been of utmost importance and as such, researchers have developed various models with the intent of increasing the accuracy at which one will be able to determine the future outcome of stock movement. Most of the models that have been on the ground for years make use of numerical data but with the arrival of machine learnings methods, models have also been developed to use text mining in machine learning to see if the information gathered from text will help to improve the prediction of stock market.

Newspaper headlines serve two functions which makes it easier to filter information. One, the headline summarizes the article, and the second, it serves as a pragmatic one in which the heading is coined in such a way to attract attention of readers and provoke them in reading the article (Kuiken et al, 2017). Kuiken et al. (2017) further found that readers preferred headlines that are creative, more confusing or less informative at times, that is readers valued headlines for what they picture rather than creating an impression that attracts them to read a particular article which eventually they find less interesting or challenging. The era of reading hardcopy newspapers is phasing out little by little, rather online newspapers are subscribed to and trending news/topic pops up on our androids and are read. The readers depend on more

news articles read online using their mobile phones and laptops as long as they are connected to the internet as they did when newspapers were distributed to individual offices or homes.

Olley and Chile (2015), in studying readers' perception of Nigerian newspapers on the internet found that the internet is a relevant medium of communication in our modern society especially in the media environment vis-à-vis the newspaper owing to the various potentials which the internet displays. This development has also given Nigerian readers a medium of replying to publications in the Newspaper, as the various Newspapers online have established a feedback mechanism on the internet for readers to post their views or reactions. Even with this, many Nigerians still do not fancy the reading of Newspapers online because they believe that the internet is a place where any faceless individual can post whatever he likes for public consumption. Again, there is no maximum utility in the feedback avenues offered by such newspapers for online newspaper readers. However, the reverse is gradually becoming the case, as could be observed in their study, where there is an improvement in the reader's beliefs, patronage and the needs for feedbacks. Again, there is no maximum utility in the feedback avenues offered by such newspapers for online newspaper readers. Additionally, they observed that there is a gross and less awareness of Nigerian online newspapers among many Nigerians. For this reason, utility in readership of both Nigerian online Newspapers and foreign ones is far less than expected. From the above findings, it can still be upheld that audience perception and utilization of the Nigerian online Newspaper is fast-growing and taking a positive approach.

3. Methodology

3.1 The TF-IDF

The TF-IDF is the product of two statistics: the term frequency (TF) and the Inverse Document Frequency (IDF). The TF uses the raw count of a term in a document, that is, the number of times term occurs in document d. The IDF measures how much information the word provides, that is, if it is common or rare across all documents. It is the logarithmic scaled inverse function of the documents that contain the word.

The mathematical expression of the TF-IDF is given as follows:

$$TF - IDF_{(t_k, d_j)} = TF_{(t_k, d_j)} . \log \frac{D}{DF_{(t_k)}} \quad (1)$$

where $TF_{(t_k, d_j)}$ implies the Term-Frequency – the number of times term t_k appear in document d_j, while $DF_{(t_k)}$ implies Document Frequency – the number of documents in which terms t_k appears and D denotes the total number of document in the collection.

The linear Kernel SVM with n points training data set is defined as,

$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), ..., (\vec{x}_n, y_n) \quad (2)$$

where y_1 are either 1 or -1, each indicating the class of $\vec{x}_i$ point, and each $\vec{x}_i$ is a p-dimensional real vector. The idea is to obtain a hyperplane which divides points $\vec{x}_i$ for which $y_i = 1$. A hyperplane is written as the set of points $\vec{x}_i$ satisfying,

$$\vec{w}.\vec{x} - b = 0 \quad (3)$$

where $\vec{w}$ is the vector of the hyperplane, and the parameter $\frac{b}{\|\vec{w}\|}$ determines the offset of the hyperplane from the region along the normal vector $\vec{w}$.

In a case where the training data are linearly separable, we can identify two parallel hyperplane that separate the two data classes, that is,

$$\vec{w}.\vec{x} - b = 1 \tag{4}$$

and,

$$\vec{w}.\vec{x} - b = -1 \tag{5}$$

Such that the distance between the two hyperplanes is $\frac{2}{\|\vec{w}\|}$. In order not to allow data points from falling into the margin, we apply the constraints,

$$\vec{w}.\vec{x}_i - b \geq 1 \text{ if } y_i = 1 \tag{6}$$

and,

$$\vec{w}.\vec{x}_i - b \leq -1 \text{ if } y_i = -1 \tag{7}$$

communicating that each data points lies correctly on the margin. We can then easily write (6) and (7) together in compact form as,

$$y_i(\vec{w}.\vec{x}_i - b) \geq 1, \text{ for all } 1 \leq i \leq n \tag{8}$$

and $\vec{w}$ and b are solved by minimizing $\|\vec{w}\|$ subject to the function for $i = 1, 2, ..., n$. Our classifier is then determined by

$$\vec{x} \rightarrow \text{sgn}(\vec{w}.\vec{x} - b) \tag{9}$$

Thus, the max-margin hyperplane is determined by those $\vec{x}_i$ that lie nearest to it, and these $\vec{x}_i$ are the support vectors. This linear case describes the case of Hard-Margin SVM.

The case of Soft-margin SVM is when the data are not linearly separated, and thus, hinge loss function,

$$\max[0,1-y_i(\vec{w}.\vec{x}_i-b)] \tag{10}$$

is used. If $\vec{x}_i$ lies correctly on the margin, (10) becomes 0, and if $\vec{x}_i$ lies wrongly on the margin, the function is equated to the proportional distance from the margin. The interest now is to minimize,

$$\left\{\frac{1}{n}\sum_{i=1}^{n}\max[0,1-y_i(\vec{w}.\vec{x}_i-b)]\right\}+\lambda\|\vec{w}\|^2 \tag{11}$$

where λ is the trade-off between increasing the margin size and ensuring that $\vec{x}_i$ lies correctly on the margin.

Different nonlinear kernels are often used in SVM modelling. These kernels help in further fitting the maximum-margin hyperplane. Also, working in a hyper dimensional space reduces the error in SVM, and the algorithm performs better as samples become larger.

Common kernels are:

1. Linear: $\vec{x}_i.\vec{x}_j$
2. Polynomial: $\left(\vec{x}_i.\vec{x}_j\right)^d$ or $\left(\vec{x}_i.\vec{x}_j+1\right)^d$
3. Radial basis function: $\exp\left(-\gamma\left\|\vec{x}_i-\vec{x}_j\right\|^2\right)$, for $\gamma>0$
4. Sigmoid: $\tanh\left(\gamma.\vec{x}_i.\vec{x}_j+1\right)$
5. Hyperbolic tangent: $\tanh\left(\kappa.\vec{x}_i.\vec{x}_j+c\right)$, for some $k>0$ and $c<0$.

Generally, the computation of nonlinear SVM classifiers amounts to minimizing the corresponding function to (10) (based on the kernels above), and parameters and statistical properties of the SVM model are obtained. We obtain the accuracy with its confidence limit, specificity, negative-predictive value, positive-predictive value and its sensitivity criteria.

4. Data and Empirical Results[2]

4.1 Data Description

We used text data to carry out our analysis. The dataset was obtained from Vanguard online newspaper using web scrapping. The data are Vanguards news headlines from May 27th, 2014 to November 1st, 2019 and we have chosen Vanguard newspaper because, it was the only well-known Nigerian newspaper with a long historical record of its news headlines, that we could access. The news headlines were scrapped from the website of the newspaper. https://www.vanguardngr.com/category/business and we limited our headline news to business headlines. To determine the returns in the stock market, we collected ASI of th NSE data from https://ng.investing.com/indices/nse-all-share-historical-data, and we compted the daily price change, with positive change implying stocks increase and negative change implying stocks decrease. Thus, stock increases are labelled 1 while stock decreases are labelled 0. R Software was used to carry out all analyses, including the web scraping which was used to scrap out news headlines from the newspaper. The appendix of this work is attached to the R script that was used in collecting the data from the Vanguard newspapers website.

Our dataset has 3 columns: the first column contains the Headlines, the second column contains the Dates and the third column contains the Label of the daily stock. The label is a record from the change in the movement of daily stock, where we have 1 as an increase in NSE all-share index, and 0 as a decrease in all-share index, as mentioned earlier.

In assigning for the stock data,we had missing data on the label column as a result of this, leading us to remove all the dates that fall on a weekend as this could not have a stock trade records. We also had some lapses from the records on the part of the records on the

[2] R codes for web scraping and for text analysis are found in Appendix 1 and 2, and details of other results that are not included in the paper are available on request.

newspaper as we discovered, that there where dates or days which news records were found missing, but such dates were very few.

In using R to carry out our analysis, we followed these steps to prepare our data for analysis, and exploration. After cleaning our data, we explored the data by checking for most common words that appear most when there is an increase in the NSE all-share index and the words that appear most when there is a decrease. We also went further to check for two, three, and four pairs of words that appear most when there is an increase in NSE all share Index and those that appear most when there is a decrease in the ASI,We worked with the TF-IDF and it assisted in picking out the most important words in the text, rather than just selecting words that occur more frequently. We also plotted the TF-IDF using network diagrams, which enabled vitualize how words were connected to each of the labels, with the increased label which is 1 and with the decrease label which in 0.

4.2 Results

Below is a sample of the data that were used to carry out the analysis. In the table 1, we see that the data had three columns The ‘headline’ contained our text data, which was a collection of our daily business news headlines from the vanguard newspaper.

Our text data were collected by dates, while we were scraping out our data, we ensured that each headline came with a date, to enable us know which date the news headline was broadcasted. This was also meant to guide us in creating our dummy variables. This means, when we downloaded our Nigerian stock exchange all share index, we created our dummy variables for the data. This means, For when there was an increase and decrease, and this was append as a third column, but this was done according to the dates that had already reflected for each of the business headlines.The last column is the Label,containing large values 1 or 0 for ASI.

Table 1: Sample of the data collected from Vanguard.

S/N	Headline[3]	Date	Label
1	cadbury nigeria revenue hits n28.9bn in 9 months border closure: if maintained for 2 yrs, security challenges will be eradicated emefiele cbn unconventional policy to grow economy paying off emefiele patronise locally made products, buhari tells nigerians naicom bemoans poor corporate governance in insurance sector capital market key to sustainable economic growth finance minister cac records 300,000 new business registrations for smes	1-Nov-19	0
2	navy seizes 3,378 bags of smuggled rice in akwa ibom ghana govt plans to ban rice, poultry imports in 3 years minister nigeria, vietnam sign visa waiver agreement foodco entry boosts organized retail market space oando's profit up by 26 to n13.1 bn in q3	31-Oct-19	1
3	39 businesses to exhibit at waccse 2019 flour mills finance cost drops by 21 to n8.8 bn union bank profit up by 5 to n15.6 bn in q3 renewed interest: nse crucial market indicators up by 0.25 why afriexim bank is investing n18 billion in imo ihedioha nestle declares n25 interim dividend transcorp total assets hit n313.07 bn in q3 bua obu cement, ccnn set to merge as group consolidates business contravention of import duty: nagaff seeks equal treatment for all goods more controversy trails kaduna dry port n100bn needed to revive cotton, textile, garment subsector emefiele uighur criticism not beneficial for trade talks, china warns u.s. report of nigerian ship registry review committee not shrouded in secrecy ilori n226bn chinese loan for Lekki Port construction controversy over portal service at ports auto dealers split over plans to sue customs afreximbank, shippers council float sealink shipping project	30-Oct-19	1
4	cbn injects billions into textiles, cotton industries cbn votes n500m for recycling of old naira notes, others bailout: governors reject fg n162m monthly repayment plan maritime operators task fg, world bank over ease of doing business ranking world bank pledges to assist nigeria to increase igr just in commodities trading: lcfe, cotecna in talk over certification lagos power oil walkhearton 4.0 holds just in: seplat declares 29m interim dividend to shareholders in q3 sterling bank net interest income up by 19 in q3 fintech will increase capital market participation sec updated: 28 million bank customers have no credit history cbn cbn votes n500m for recycling of old naira notes, others breaking: 28 million bank customers have no credit history cbn breaking: ecobank cancels charges for ussd transactions breaking: cbn will license more firms for currency processing, cashintransit operations okoroafor dappman faults nnpc sole importation of petroleum products gridlock: fcta orders banks, others to revert to original land use in 72 hrs access bank completes second phase postmerger integration access bank better positioned to serve customers nnpc sets deadline for accurate crude oil, fuel consumption data legal issues, global energy transition, others may impede on energy legal issues, global energy transition, others may impede on energy	29-Oct-19	1
5	updated: amcon lobbies national assembly on emergence of arik as national carrier breaking: gtbank removes all bank charges for young undergraduates breaking: amcon lobbies national assembly on emergence of arik as national carrier breaking: gtbank removes all bank charges for young undergraduates updated: cashintransit firms must have n1bn to operate nationwide cbn updated: currency processing firms must have n3bn capital to get national license cbn cbn denies directing merchants to charge customers n50 on pos transactions breaking: currency processing firms must have n3bn capital for national license cbn breaking: cash in transit firms must have n1bn capital to get national license cbn primary market listing on nse rises to n2.7 trillion in 6 months rise in bonds prices to persist as n413bn inflow boosts interbank ndic ready to tackle emerging threats to banks ibrahim	28-Oct-19	1
8	border closure: mixed reactions greet closure in southwest mtn clarifies controversy over ussd charge how my agric business became billion naira venture exmilitant border closure: we need nigeria more than they need us ghana minister	25-Oct-19	0

[3] Source: Nigerian Vanguard newspapers.

9	ministry of finance will capture remaining batch on ippis minister nnpc signs mou with russian oil company, lukeoil mou presidency how to succeed in the business of advertising yinka onigbinde world bank ease of doing business ranking: our strategy working buhari breaking: nigeria ranks 131 in world bank ease of doing business presidency boasts southern nigerians more willing to emigrate than northerners survey revenue: nis remits over n11.68bn, 22,972 to fg in 8 months eedc improves power supply in south east oil slips below 61 as weak demand outlook weighs fgn bond auction over subscribed in october, says dmo nigeria moves 15 places in world bank ease of doing business ranking	24-Oct-19	0
10	fgn bonds attract oversubscription gas flaring: fec approves conversion of natural gas to methanol border closure: we are crippled by nigeria restriction niger traders nse: market capitalization rebounds by n15bn revealed: fg spends n383bn on fuel subsidy in 7 months nigeria loses about 15bn to tax evasion annually, says fowler afreximbank intervention in african ports hits n152bn in 3yrs softbank clinches deal to take over wework sources pound falls against dollar, euro after british lawmakers reject brexit timetable	23-Oct-19	1

We carried out tokenization, which enabled us to make the words stand as individual words, and no longer as a sentence. This helped us to group each word according to the numbers of times they appeared in our news headline by grouping them according to the label. Since our interest is in determining the words that affect the change in the Nigerian stock exchange ASI and from tasce 2 we would see the words in our data, and the number of times they reflected according to the positive and negative change in the Nigerian stock exchange all share index. The data were cleaned by lowers all capital letters to small letters, removing all number from the text as numbers do not have any significant impact on the text data, we removed all punctuations, we removed all stop words or qualifier words as we can call them as they have not significant impact on the text and we also went on to remove all spaces in between words.

Table 2: Common words, and their number of occurrence n.

S/N	**word**	**Label**	**n**
1	nigeria	0	782
2	bn	0	758
3	nigeria	1	595
4	bn	1	588
5	bank	0	447
6	fg	0	415
7	market	0	415
8	oil	0	403
9	fg	1	353

10	bank	1	325
11	cbn	0	306
12	market	1	297
13	oil	1	264
14	cbn	1	234
15	investors	0	230
16	sector	0	201
17	insurance	0	200
18	banks	0	194
19	naira	0	194
20	business	0	186

Table 3 presents the term frequency results,and we see that we find the total number of occurrences of all words, and divide it by the total number of occurrence of each word, which gives us the term frequency, and the rank is based on the words by their groupings, with the highest occurrence to the lowest. When the word with the highest occurrence is ranked as 1 and the ranking continues to the last word, which occurs less in the data.

Table 3: Term frequency table for words and their occurrence.

S/N	word	Label	N	total	rank	term frequency
1	nigeria	0	782	44146	1	0.0177139
2	bn	0	758	44146	2	0.0171703
3	nigeria	1	595	33175	3	0.0179352
4	bn	1	588	33175	4	0.0177242
5	bank	0	447	44146	5	0.0101255
6	fg	0	415	44146	6	0.0094006
7	market	0	415	44146	7	0.0094006
8	oil	0	403	44146	8	0.0091288
9	fg	1	353	33175	9	0.0106405
10	bank	1	325	33175	10	0.0097965
11	cbn	0	306	44146	11	0.0069315
12	market	1	297	33175	12	0.0089525
13	oil	1	264	33175	13	0.0079578
14	cbn	1	234	33175	14	0.0070535
15	investors	0	230	44146	15	0.0052100
16	sector	0	201	44146	16	0.0045531
17	insurance	0	200	44146	17	0.0045304
18	banks	0	194	44146	18	0.0043945
19	naira	0	194	44146	19	0.0043945
20	business	0	186	44146	20	0.0042133

The result from our term frequency-inverse document frequency in table 4 seems to reflect different words from our term frequency since, unlike our term frequency that bases its

words on the highest number of occurrences, the term frequency-inverse document frequency picks words and arranges them based on how important they seem to reflect in the text, not minding if it occurred highest or not.The result table shows us the first twenty TF-IDF words, That is, the first twenty words with high importance in our news headline.

Table 4: Result of TF-IDF

S/N	word	Label	N	tf	idf	tf_idf
1	vessels	0	19	0.0004304	0.6931472	0.0002983
2	kenya	0	12	0.0002718	0.6931472	0.0001884
3	japaul	1	8	0.0002411	0.6931472	0.0001671
4	remits	0	9	0.0002039	0.6931472	0.0001413
5	resume	0	9	0.0002039	0.6931472	0.0001413
6	sponsors	0	9	0.0002039	0.6931472	0.0001413
7	hallmark	0	8	0.0001812	0.6931472	0.0001256
8	maersk	0	8	0.0001812	0.6931472	0.0001256
9	tvc	1	6	0.0001809	0.6931472	0.0001254
10	adoption	0	7	0.0001586	0.6931472	0.0001099
11	arabia	0	7	0.0001586	0.6931472	0.0001099
12	drag	0	7	0.0001586	0.6931472	0.0001099
13	dump	0	7	0.0001586	0.6931472	0.0001099
14	exclusion	0	7	0.0001586	0.6931472	0.0001099
15	panel	0	7	0.0001586	0.6931472	0.0001099
16	track	0	7	0.0001586	0.6931472	0.0001099
17	alisonmadueke	1	5	0.0001507	0.6931472	0.0001045
18	die	1	5	0.0001507	0.6931472	0.0001045
19	employ	1	5	0.0001507	0.6931472	0.0001045
20	mmm	1	5	0.0001507	0.6931472	0.0001045

From our wordcloud, we are able to see the words that reflected most in our text data. We may have noticed that one word may have appeared twice in our wordclouds, e.g bn, Nigeria, naira, etc. this is because our data has been group by labels, and the wordcloud reflected the words according to the labels and highest number of occurrences. This implies that if there are two words that both reflected in the label for increase and the label for decrease, and the occurrence of such words both cases were high, since our plot reflects the first 50 words, if this word happens to appear twice as a result of the labeling, then it will be reflected twice on the plot. From our plot, we see that words such as Nigeria, bn, market, bank, fg, cbn, investor, were among the very frequent occurring words in the news headline.

Figure 6: The wordcloud display of most frequent words

The bar graphs below reflect the TF-IDF results of the text. The plot is done by the label, where each word is plotted according to the label they belonged to In figure 7, we see two bar graphs, one is for case of increase in stocks, which is 1 and the plot there shows us by TF-IDF if the words that are more important to the text, that contributed greatly the increase in the Nigerian stock market. Likewise, the other graph which is labelled 0,which contains by TF-IDF, all the important words that contribute greatly, to the decrease of the Nigerian stock market.

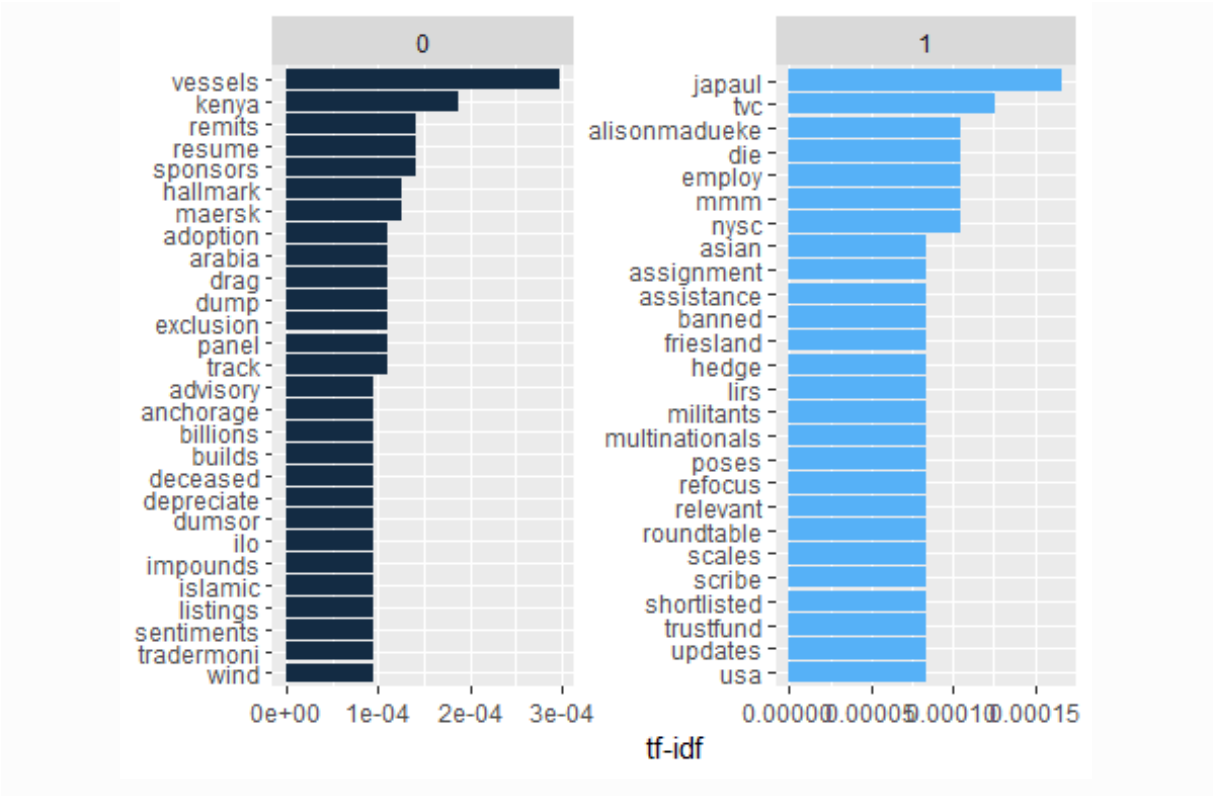

Figure 7: Bar graph by TF-IDF for most important words

In the n-grams results in Table 5, our interest is not to pick a number of words together and see how the combination of these words impacts the text which we are analyzing, instead we will be combining a number of words to see how they impact the increase and decrease in prices in the Nigerian stock market. In this case, we had a combination of two words to see how these two words have impact and our text. The number of occurrences for the first 10 paired words are presented in Table 5.

Table 5: Results of the n-gram for when n=2

S/N	word1	word2	n
1	Stock	market	104
2	Capital	market	89
3	World	bank	80
4	Naira	appreciates	73
5	Access	bank	72
6	Naira	depreciates	72
7	Stanbic	ibtc	68
8	Oil	price	67
9	Oil	gas	65
10	Parallel	market	62

We went further to look for the TF-IDF of this combined words (see Table 6) and we see below in the table, the first six words by TF-IDF that are of great importance to the daily headline news, and by the label of increase and decrease in the stock market.

Table 6: TF-IDF for N-gram when n=3

S/N	Label	Bigram	n	tf	idf	tf_idf
1	0	insurance penetration	10	0.000319	0.693	0.000221
2	1	bank plc	7	0.000298	0.693	0.000207
3	0	naira depreciation	9	0.000287	0.693	0.000199
4	0	equities pricing	8	0.000255	0.693	0.000177
5	0	trade volume	8	0.000255	0.693	0.000177
6	1	empower smes	6	0.000255	0.693	0.000177

We went on to plot the bar graph for this combined word too by their labeling, to determine which words have great importance in the increase and decrease of the Nigerian stock market.

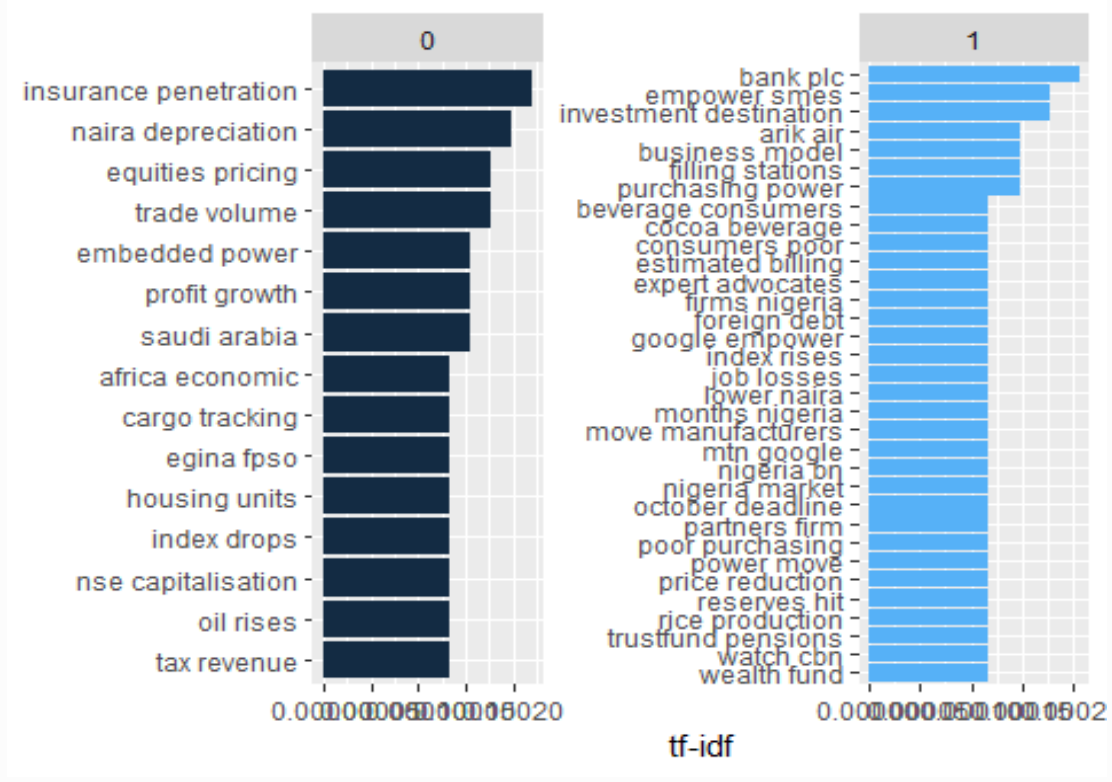

Figure 8: The bar graph for n-gram, when n=2

The network diagrams for the results in Tables 5 and 6 are also plotted in Figures 9 and 10, respectively, to determine the connection it words have with each other. From the network analysis, we are able to see that in combining the words, there happens to be a connection between the words in the sense that, the same pair of words that could cause an increase, could also cause a decrease. But from our bar graph, we are able to determine to what extent this word has impacted the increase or decrease in the Nigerian stock market performance.

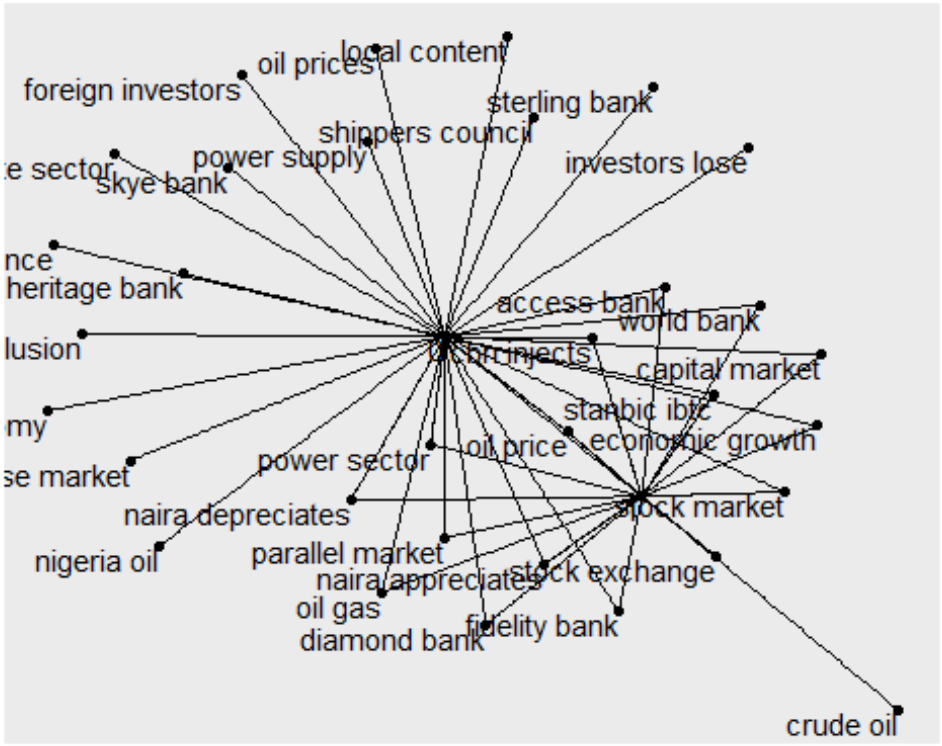

Figure 9: The Network Diagram for n=2

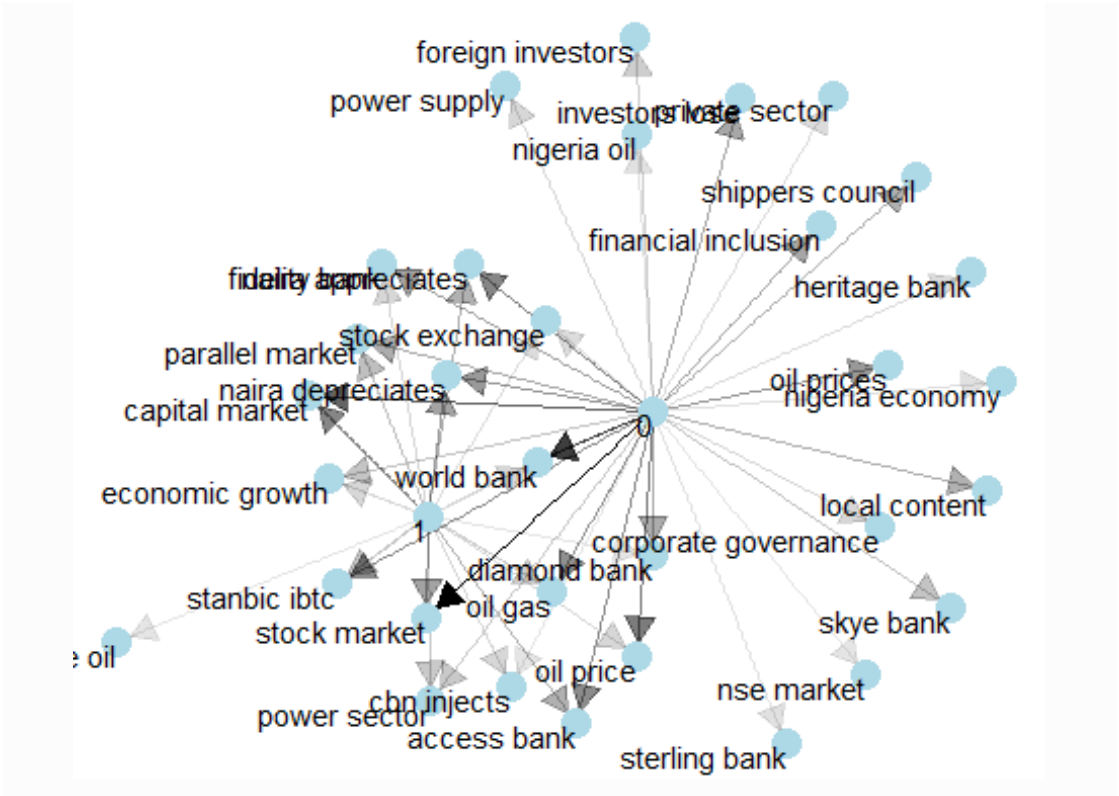

Figure 10: The Network Diagrams for n-gram when n=2

We also made combinations of three words to see how these three words have an impact and our text, and results are presented in Table 7 which gives us the number of occurrences for the first 10 words we have combined together.

Table 7: Results of the n-gram for when n=3.

S/N	word1	word2	word3	n
1	top	performing	stocks	35
2	nigerian	Stock	exchange	26
3	nse	market	capitalisation	17
4	nse	market	indices	13
5	banks	Wema	bank	12
6	market	investors	lose	12
7	Stock	market	investors	12
8	Improving	Nigeria	development	11
9	Tony	Elumelu	foundation	11
10	Banks	Skye	bank	9

We went ahead to look for the TF-IDF of these combined words and we see below in Table 8 the first six words by TF-IDF, that of great importance to the daily headline news, and by the label of increase and decrese in the daily stock market.

Table 8: TF-IDF for N-gram when n=3

S/N	Label	bigram2	n	tf	idf	tf_idf
1	1	oil gas sector	5	0.000304	0.693	0.000211
2	1	beverage consumers poor	4	0.000243	0.693	0.000168
3	1	cocoa beverage consumers	4	0.000243	0.693	0.000168
4	1	consumers poor purchasing	4	0.000243	0.693	0.000168
5	1	google empower smes	4	0.000243	0.693	0.000168
6	1	mtn google empower	4	0.000243	0.693	0.000168

We went on to plot the bar graph for this combined words too by their labeling, to determine which words have great importance in the increase and decrease of the Nigerian stock market. The barchart is given in Figure 11.

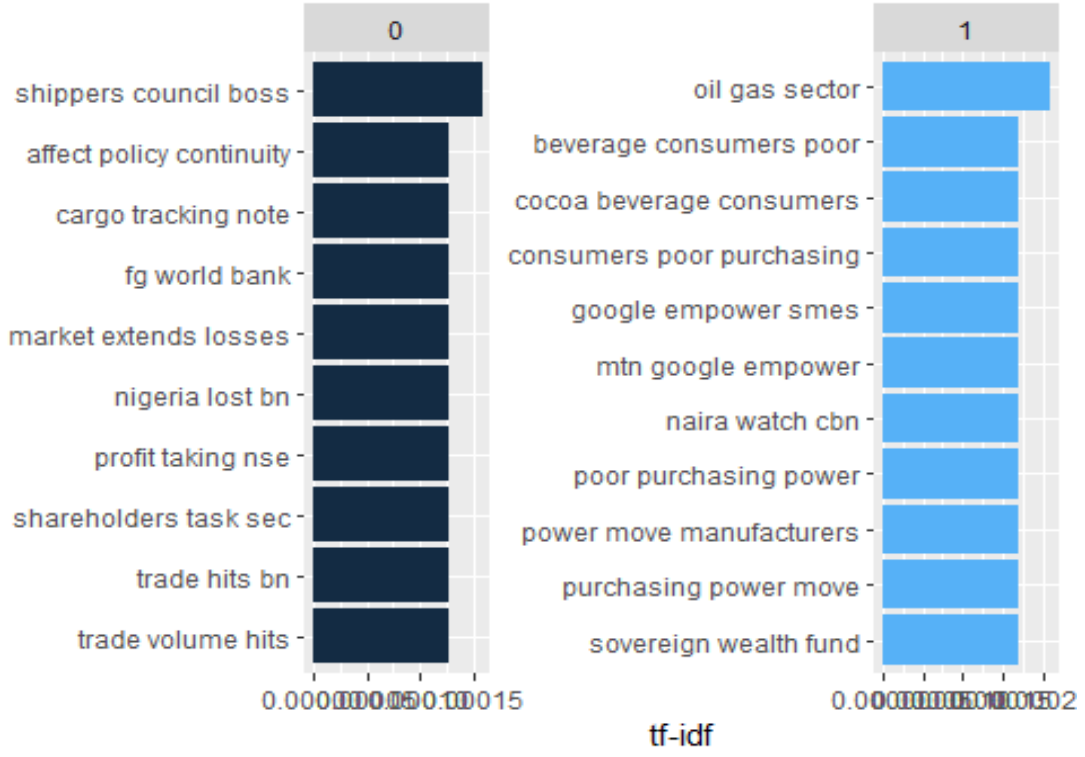

Figure 11: The Bar Graph for n-gram, when n=3

In this case, we did a combination of four words to see how these four words have an impact and our text, and Table 9 gives us the number of occurrences for the first 10 words we combined together.

Table 9: Results of the n-gram for when n=4

S/N	word1	word2	word3	word4	n
1	Stock	market	investors	lose	12
2	Wema	bank	partners	dbn	6
3	Fees	ceo	magnartis	finance	5
4	Gap	united	capital	md	5
5	Infrastructure	gap	united	capital	5
6	Listing	fees	ceo	magnartis	5
7	Review	listing	fees	ceo	5
8	Ships	arrive	lagos	ports	5
9	Streamline	share	transmission	process	5
10	Trn	infrastructure	gap	united	5

We went ahead to look for the TF-IDF of this combined words and we see below in the table, the first six words by TF-IDF, that of great importance to the daily headline news, and by the label of increase in the daily stock market and decrease in the daily stock market.

Table 11: TF-IDF for N-gram when n=4

S/N	Label	bigram	n	tf	idf	tf_idf
1	1	beverage consumers poor purchasing	4	0.000353	0.693	0.000245
2	1	cocoa beverage consumers poor	4	0.000353	0.693	0.000245
3	1	consumers poor purchasing power	4	0.000353	0.693	0.000245
4	1	mtn google empower smes	4	0.000353	0.693	0.000245
5	1	poor purchasing power move	4	0.000353	0.693	0.000245
6	1	purchasing power move manufacturers	4	0.000353	0.693	0.000245

We went on to plot the bar graph for this combined words too by their labeling, to determine which words have great importance in the increase and decrease of the Nigerian stock market (see Figure 12).

Figure 12: Bar Plot for n-gram for when n=4

In Table 13, the results of SVM using varius kernels are presented. Here SVM model was used for categorization and prediction of the text data. In doing this, we carried out our analysis using four support vector kernels. For the first one, we did not use any kernel, for the second we used the linear kernel, for the third we used the radial kernel, for the third we used the sigmoid kernel and the fourth, we used the polynomial kernel. From the results, we are able to see that the results for the sigmoid kernel, the polynomial kernel, and the radial have the same accuracy, the same sensitivity. The results for these three seem to be the same, which implies the same performance from these three kernels on our data.

Table 13: Results for all kernel types

no kernel function	Accuracy	0.5648
	95% CI	(0.5089, 0.6195)
	Sensitivity	0.04762
	Specificity	0.99435
	Pos Pred Value	0.87500
	Neg Pred Value	0.55696
`linear` kernel function	Accuracy	0.5278
	95% CI	(0.4718, 0.5832)
	Sensitivity	0.5876
	Specificity	0.4558
	Pos Pred Value	0.5652
	Neg Pred Value	0.4786
`polynomial` kernel function	Accuracy	0.5463
	95% CI	(0.4903, 0.6014)
	Sensitivity	1.0000
	Specificity	0
	Pos Pred Value	0.5463
	Neg Pred Value	NaN
`radial` kernel function	Accuracy	0.5463
	95% CI	(0.4903, 0.6014)
	Sensitivity	1.0000
	Specificity	0
	Pos Pred Value	0.5463
	Neg Pred Value	NaN
`sigmoid` kernel function	Accuracy	0.5463
	95% CI	(0.4903, 0.6014)
	Sensitivity	1.0000
	Specificity	0
	Pos Pred Value	0.5463
	Neg Pred Value	NaN

5. Concluding remarks

From the results, we were able to note words such as Nigeria, bn, market, bank, fg, cbn, investor, to have a direct impact or the increase and decrease in prices of the Nigerian stock. This, of course, is based on the frequency of words of occurrence. But on carrying out the TF-IDF, we see the this words do not reflect that much, but we words vessels, Kenya, remits, resumes, sponsors, hallmark, Maersk, adoption, arabia, drag, dump, exclusion, panel, track, advisory, acnchorage, billios, build, deceased, depreciate, dumsor, ilo, impounds, Islamic, listing, dentiments, tradermoni, wind where important to the decrease of the Nigerian stock market,

while japaul, tvc, alisonmadueke, die, employ, mmm, nysc, Asian, assignment, assistance, banned, friesland, hedge, lirs, militants, multinational, poses, refocus, relevant, rountable, scales, scribe, shortlisted, trustfund, updates and usa were important to the increase of the Nigerian stock market.

In carrying out the n-grams, we found out that for when we combined two words, the following pairs of words, insurance penetration, naira depreciation, equities pricing, trade volume, embedded power, profit growth, Saudi arabia, Africa economic, cargo tracking, egina fpso, housing units, index drops, nse capitalization, oil rise, tax revenue, were important to the decrease of the Nigerian stock market while words such as bank plc, empower smes, investment destination, arik air, business model, filling model, purchasing power, beverage consumers, cocoa beverage, consumers poor, estimated billing, expert advocate, firms Nigeria, foreign debt, google empower, were important to the increase of the Nigeria stock market.

We went ahead to also get the tri-gram, which has to do with combining three words together and we found words such as, shippers council boss, affect policy continuity, cargo tracking note, fg world bank, market extends losses, Nigeria lost bn, profit-taking nse, shareholders task sec, trade hits bn, trade volume hits, was important to the decrease of stock, while words such as, oil gas sector, beverage consumer poor, cocoa beverage consumer, consumer poor purchasing, google empowers smes, naira watch cbn, poor purchasing power, power move manufacturers, purchasing power move, sovereign wealth fund.

We also carried out the n-gram for when n=4 and words such as, agric transportation labs, bn amind sell pressure, firms owe Nigerian banks, energy firms owe Nigeria, ideal national investor protection, Nigeria fiscal policy fbn, Nigeria offshore rig count, nse capitalisation drops, were important to the decrease of the Nigerian stock market, while words such as beverage consumers poor purchase, cocoa beverage consumer poor, mtn google

empower, purchasing power move manufacturer, word bank loan, nse market capitalisation, were important to the increase in the Nigerian stock market.

In carrying out svm, because of the limited amount of data, the accuracy was not very high but we had an accuracy that was between 0.5278 to 0.5648. If we had a larger amount of data, we would be able to achieve a higher accuracy as the model will have enough sufficient data to study and predict better the data. I light of that, we would recommend that a rather limiting the text data to just business headline news, one could extend this work by not limiting the news to just business headlines.

Reference

A study of newspaper and net paper reading. In The Mind's Eye (pp. 657-670). North-Holland. Afego, P. N. (2017). Effects of changes in stock index compositions: A literature survey. International Review of Financial Analysis, 52, 228-239.

Ajao, M. G., & Osayuwu, R. (2012). Testing the weak form of efficient market hypothesis in Nigerian capital market. Accounting and Finance Research, 1(1), 169-179.

Awad, M., & Khanna, R. (2015). Efficient learning machines: theories, concepts, and applications for engineers and system designers. Apress.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Bola, A. A., Adesola, A. G., Olusayo, O. E., & Adebisi, A. A. (2013). Forecasting movement of the Nigerian stock exchange all share index using artificial neural and Bayesian networks. Journal of Finance and Investment Analysis, 2(1), 41-59.

Breitinger, C., Bela, G. and Stefan, L. (2015). Research-paper recommender systems: a literature survey. International Journal on Digital Libraries. 17 (4): 305–338.

Corporate Finance Institute. Stock Market, Public Markets for Issuing, Buying and Selling Stocks. Retrieved from https://blog.apastyle.org/apastyle/2010/11/how-to-cite-something-you-found-on-a-website-in-apa-style.html.

Dallah, H., & Adeleke, I. (2018). Macorvian characteristics of the Nigerian stock market. AFRREV STECH: An International Journal of Science and Technology, 7(2), 67-77.
Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press.

Frey, L., Botan, C., & Kreps, G. (1999). Investigating communication: An introduction to research methods. (2nd ed.) Boston: Allyn & Bacon.

Holmqvist, K., Holsanova, J., Barthelson, M., & Lundqvist, D. (2003). Reading or scanning?
Huang, M. Y., Rojas, R. R., & Convery, P. D. (2018). News Sentiment as Leading Indicators for Recessions. arXiv preprint arXiv:1805.04160.
Kalyani, J., Bharathi, P., & Jyothi, P. (2016). Stock trend prediction using news sentiment analysis. arXiv preprint arXiv:1607.01958.

Kirange, D. K., & Deshmukh, R. R. (2016). Sentiment Analysis of News Headlines for Stock Price Prediction. Compusoft, An international journal of advanced computer technology, 5(3), 2080-2084.

Kuiken, J., Schuth, A., Spitters, M., & Marx, M. (2017). Effective headlines of newspaper articles in a digital environment. Digital Journalism, 5(10), 1300-1314.

Kumar, S., & Karthika, R. (2014). A survey on text mining process and techniques. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 3(7), 2279-2284.

Kwartler, T. (2017). Text mining in practice with R. John Wiley & Sons.

Li, B., Chan, K. C., Ou, C., & Ruifeng, S. (2017). Discovering public sentiment in social media for predicting stock movement of publicly listed companies. Information Systems, 69, 81-92.

Mitchell, R. (2018). Web Scraping with Python: Collecting More Data from the Modern Web. " O'Reilly Media, Inc.".

Mittermayer, M. A. (2004, January). Forecasting intraday stock price trends with text mining techniques. In 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the (pp. 10-pp). IEEE.

Nwabueze, C., Okafor, I., & Obiakor, C. Exposure to Economic News on News Tickers and its Influence on Perception of Nigeria Economy by Television Audience in Awka.

Nwidobie, B. M. (2014). The random walk theory: An empirical test in the Nigerian capital market. Asian Economic and Financial Review, 4(12), 1840-1848.
Olley, O. W. (2015). Readers' perception of Nigerian newspapers on the internet. In Journal of Philosophy, Culture and Religion (Vol. 4, pp. 26-34).

Osamwonyi, I. O., & Evbayiro-Osagie, E. I. (2012). The relationship between macroeconomic variables and stock market index in Nigeria. Journal of Economics, 3(1), 55-63.

Osisanwo, B. G., & Atanda, A. A. (2012). Determinants of stock market returns in Nigeria: a time series analysis. African Journal of Scientific Research, 9(1).

Ou, P., & Wang, H. (2009). Prediction of stock market index movement by ten data mining techniques. Modern Applied Science, 3(12), 28-42.

Shapiro, A. H., & Wilson, D. J. (2017). What's in the News? A New Economic Indicator. FRBSF Economic Letter, 2017, 10.

Srivastava, A. N., & Sahami, M. (2009). Text mining: Classification, clustering, and applications. Chapman and Hall/CRC.

Uzuke, C. A., & Daniel, J. (2016). Timeseries Analysis of All Shares Index of Nigerian Stock Exchange: A Box-Jenkins Approach. International Journal of Sciences, 5(06), 23-38.

Yogapreethi, N., & Maheswari, S. (2016). A Review On Text Mining in Data Mining. International Journal on Soft Computing (IJSC), 7(2), 1-8.

Zubair, A. (2013). Causal relationship between stock market index and exchange rate: Evidence from Nigeria. CBN journal of Applied Statistics, 4(2), 87-110.

Appendix 1: R Code for Web scraping

```
library(tidyverse)
library(rvest)
site_to_scrap =
read_html("https://www.vanguardngr.com/category/business/page/1/")
content = site_to_scrap %>%
  html_nodes(".entry-title a") %>%
  html_text()

#scraping multiple pages
for (i in 2:10){
  site_to_scrap=
read_html(paste0("https://www.vanguardngr.com/category/business/page/", i))
  temp = site_to_scrap %>%
    html_nodes(".entry-title a") %>%
    html_text()
  content = append(content, temp)
}

write.csv(content, file="content1.csv",
row.names = F)
```

Appendix 2: R Code for Text Analysis

```
library(tidyr)
library(tidyverse)
library(tidytext)
library(ggplot2)
library(SnowballC)
library(stringr)
library(wordcloud)
library(reshape2)
library(topicmodels)
library(igraph)
library(knitr)
#importing the data  into R
news <- read.csv('Vanguard3.csv', header = T,
stringsAsFactors = F)
#ommiting all missing values
news <- na.omit(news)
#a sample of our imported data
kable(news[1:20,])
#carrying out Tokinanization. that is,making the letters stand independent.this also cleans the data by removing all puntuations and special characters in the text
train = mutate(news, text = gsub(x = headline,
pattern = "[0-9]+|[[:punct:]]|\\(.*\\)", replacement = "  ")) %>% unnest_tokens(input = text, output = word)
#removing stop words that do not impact the on the analysis in any way
train1 <- train%>% anti_join(stop_words)
#counting the individual words by groupig them into the Label of increase or decrease in stock index
train1 <- train1 %>%count(word,Label, sort = TRUE) %>%ungroup()
#viewing the the groupd data
kable(train1[1:20,])

total_words <- train1 %>%
  group_by(Label) %>% summarise(total = sum(n))
train1 <- left_join(train1, total_words)

freq_by_rank <- train1 %>%
  group_by() %>%
  mutate(rank = row_number(),
         `term frequency` = n/total)
kable(freq_by_rank[1:20,])


tf <- train1 %>%
  bind_tf_idf(word,Label,n)
kable(tf[1:20,])

tf1 <- tf %>%
  select(-total) %>%
  arrange(desc(tf_idf))
kable(tf1[1:20,])

bigram_graph <- tf1 %>%
  filter(n > 200) %>%
  graph_from_data_frame()
bigram_graph

set.seed(1234)
worldcloud<-train1%>%with(wordcloud(word, n,
max.words = 50))

graph1<-tf1 %>%
  arrange(desc(tf_idf)) %>%
  mutate(word = factor(word, levels =
rev(unique(word)))) %>%
  group_by(Label) %>%
  top_n(15) %>%
  ungroup %>%
  ggplot(aes(word, tf_idf, fill=Label)) +
```

```
 geom_col(show.legend = FALSE) +
 labs(x = NULL, y = "tf-idf") +
 facet_wrap(~Label, ncol = 2, scales = "free") +
 coord_flip()
graph1

bigrams <- bigrams1<-mutate(news, text = gsub(x
= headline, pattern = "[0-9]+|[[:punct:]]|\\(.*\\)",
replacement = "  ")) %>%
unnest_tokens(bigram,text,token = "ngrams", n =
2)

bigrams %>%
 count(bigram, sort = TRUE)

bigrams_separated <- bigrams %>%
 separate(bigram, c("word1", "word2"), sep = " ")
kable(bigrams_separated[1:20,])

bigrams_filtered <- bigrams_separated %>%
 filter(!word1 %in% stop_words$word) %>%
 filter(!word2 %in% stop_words$word)
# new bigram counts:
bigram_counts <- bigrams_filtered %>%
 count(word1, word2, sort = TRUE)
kable(bigram_counts[1:10,])

#joining the two words together
bigrams_united <- bigrams_filtered %>%
 unite(bigram, word1, word2, sep = " ")
kable(bigrams_united[1:20,])

#carrying out their TF-IDF analysis for two pairs of
words
bigram_tf_idf <- bigrams_united %>%
 count(Label, bigram) %>%
 bind_tf_idf(bigram, Label, n) %>%
 arrange(desc(tf_idf))
head(bigram_tf_idf)

bigram_tf_idf <- bigrams_united %>%
 count(Label, bigram) %>%
 bind_tf_idf(bigram, Label, n) %>%
 arrange(desc(tf_idf))

graph2 <- bigram_tf_idf%>%
 arrange(desc(tf_idf)) %>%
 mutate(bigram = factor(bigram, levels =
rev(unique(bigram)))) %>%
 group_by(Label) %>%
 top_n(10) %>%
 ungroup %>%
 ggplot(aes(bigram, tf_idf, fill=Label)) +
 geom_col(show.legend = FALSE) +
 labs(x = NULL, y = "tf-idf") +
 facet_wrap(~Label, ncol = 2, scales = "free") +
 coord_flip()
graph2

bigram_graph <- bigram_tf_idf %>%
 filter(n >20) %>%
 graph_from_data_frame()
bigram_graph


library(ggraph)
set.seed(2017)
graph3 <-ggraph(bigram_graph, layout = "fr") +
 geom_edge_link() +
 geom_node_point() +
 geom_node_text(aes(label = name), vjust = 1,
hjust = 1)
graph3

set.seed(2016)
a <- grid::arrow(type = "closed", length = unit(.15,
"inches"))
graph4<- ggraph(bigram_graph, layout = "fr") +
 geom_edge_link(aes(edge_alpha = n),
show.legend = FALSE,
              arrow = a, end_cap = circle(.07, 'inches'))
+
 geom_node_point(color = "lightblue", size = 5) +
 geom_node_text(aes(label = name), vjust = 1,
hjust = 1) +
 theme_void()
graph4

igrams2 <- bigrams1<-mutate(news, text = gsub(x
= headline, pattern = "[0-9]+|[[:punct:]]|\\(.*\\)",
replacement = "  ")) %>%
unnest_tokens(bigram,text,token = "ngrams", n =
3)

bigrams2 %>%
 count(bigram, sort = TRUE)

bigrams_separated2 <- bigrams2 %>%
 separate(bigram, c("word1", "word2","word3"),
sep = " ")
kable(bigrams_separated2[1:20,])

bigrams_filtered2 <- bigrams_separated2 %>%
 filter(!word1 %in% stop_words$word) %>%
 filter(!word2 %in% stop_words$word)%>%
 filter(!word3 %in% stop_words$word)
# new bigram counts:
bigram_counts2 <- bigrams_filtered2 %>%
 count(word1, word2,word3, sort = TRUE)
kable(bigram_counts2[1:10,])

#joining the two words together
bigrams_united2 <- bigrams_filtered2 %>%
 unite(bigram2, word1, word2,word3, sep = " ")
```

```
kable(bigrams_united2[1:20,])

#carrying out their TF-IDF analysis for two pairs of
words
bigram_tf_idf2 <- bigrams_united2 %>%
  count(Label, bigram2) %>%
  bind_tf_idf(bigram2, Label, n) %>%
  arrange(desc(tf_idf))
head(bigram_tf_idf2)

bigram_tf_idf2 <- bigrams_united2 %>%
  count(Label, bigram2) %>%
  bind_tf_idf(bigram2, Label, n) %>%
  arrange(desc(tf_idf))

bigram_graph2 <- bigram_tf_idf2 %>%
  filter(n >2)
graph22 <- bigram_graph2%>%
  arrange(desc(tf_idf)) %>%
  mutate(bigram2 = factor(bigram2, levels =
rev(unique(bigram2)))) %>%
  group_by(Label) %>%
  top_n(10) %>%
  ungroup %>%
  ggplot(aes(bigram2, tf_idf, fill=Label)) +
  geom_col(show.legend = FALSE) +
  labs(x = NULL, y = "tf-idf") +
  facet_wrap(~Label, ncol = 2, scales = "free") +
  coord_flip()
graph22

bigram_graph2 <- bigram_tf_idf2 %>%
  filter(n >2) %>%
  graph_from_data_frame()
bigram_graph2

library(ggraph)
set.seed(2017)
graph32 <-ggraph(bigram_graph2, layout = "fr") +
  geom_edge_link() +
  geom_node_point() +
  geom_node_text(aes(label = name), vjust = 1,
hjust = 1)
graph32

set.seed(2016)
a <- grid::arrow(type = "closed", length = unit(.15,
"inches"))
graph42<- ggraph(bigram_graph2, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n),
show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches'))
+
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1,
hjust = 1) +
  theme_void()
graph42

bigrams3 <-mutate(news, text = gsub(x = headline,
pattern = "[0-9]+|[[:punct:]]|\\(.*\\)", replacement
= "  ")) %>% unnest_tokens(bigram,text,token =
"ngrams", n = 4)

bigrams3 %>%
  count(bigram, sort = TRUE)

bigrams_separated3 <- bigrams3 %>%
  separate(bigram, c("word1",
"word2","word3","word4"), sep = " ")
kable(bigrams_separated3[1:20,])

bigrams_filtered3 <- bigrams_separated3 %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word)%>%
  filter(!word3 %in% stop_words$word)%>%
  filter(!word4 %in% stop_words$word)
# new bigram counts:
bigram_counts3 <- bigrams_filtered3 %>%
  count(word1, word2,word3,word4, sort = TRUE)
kable(bigram_counts3[1:10,])

#joining the two words together
bigrams_united3 <- bigrams_filtered3 %>%
  unite(bigram, word1, word2,word3,word4, sep =
" ")
kable(bigrams_united3[1:20,])

#carrying out their TF-IDF analysis for two pairs of
words
bigram_tf_idf3 <- bigrams_united3 %>%
  count(Label, bigram) %>%
  bind_tf_idf(bigram, Label, n) %>%
  arrange(desc(tf_idf))
head(bigram_tf_idf3)

bigram_tf_idf3 <- bigrams_united3 %>%
  count(Label, bigram) %>%
  bind_tf_idf(bigram, Label, n) %>%
  arrange(desc(tf_idf))

bigram_graph3 <- bigram_tf_idf3 %>%
  filter(n >2)
graph23 <- bigram_graph3%>%
  arrange(desc(tf_idf)) %>%
  mutate(bigram = factor(bigram, levels =
rev(unique(bigram)))) %>%
  group_by(Label) %>%
  top_n(10) %>%
  ungroup %>%
  ggplot(aes(bigram, tf_idf, fill=Label)) +
  geom_col(show.legend = FALSE) +
  labs(x = NULL, y = "tf-idf") +
  facet_wrap(~Label, ncol = 2, scales = "free") +
```

```
  coord_flip()
graph23

bigram_graph3 <- bigram_tf_idf2 %>%
  filter(n >2) %>%
  graph_from_data_frame()
bigram_graph3

library(ggraph)
set.seed(2017)
graph33 <-ggraph(bigram_graph3, layout = "fr") +
  geom_edge_link() +
  geom_node_point() +
  geom_node_text(aes(label = name), vjust = 1,
hjust = 1)
graph33

set.seed(2016)
a <- grid::arrow(type = "closed", length = unit(.15,
"inches"))
graph43<- ggraph(bigram_graph3, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n),
show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches'))
+
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1,
hjust = 1) +
  theme_void()
graph43

dtm <- train1%>%cast_dtm(Label,word, n)
ap_lda <- LDA(dtm, k = 2, control = list(seed =
1234))
ap_lda
ap_topics <- tidy(ap_lda, matrix = "beta")
ap_topics

set.seed(1234)
dtms <- bigram_tf_idf %>%
cast_dtm(bigram,tf_idf,n)

ap_top_terms <- ap_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
ap_top_terms

graph5<-ap_top_terms %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
graph5
```

support vector machine codes

```
library(tm)
library(NLP)
library(SnowballC)
library(wordcloud)
library(ggplot2)
library(e1071)
library(caret)
spam1 <- read.csv('vanguard3.csv', header = T)
spam1 <- na.omit(spam1)
kable(spam1[1:20,])

corpus <- Corpus(VectorSource(spam1$headline))
corpus <- tm_map(corpus,
content_transformer(tolower))
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, removeWords,
stopwords())
corpus <- tm_map(corpus, removePunctuation)
corpus2 <- tm_map(corpus, stripWhitespace)
dtm <- DocumentTermMatrix(corpus2)
features <- findFreqTerms(dtm, 10)
summary(features)
dtm2 <- DocumentTermMatrix(corpus2, list(global
= c(2, Inf),
                                         dictionary = features))
inspect(dtm2)
set.seed(1234)
train_idx <- createDataPartition(spam1$Label,
p=0.75, list=FALSE)
train1 <- spam1[train_idx,]
test1 <-  spam1[-train_idx,]
train2 <- corpus2[train_idx]
test2 <- corpus2[-train_idx]

dict2 <- findFreqTerms(dtm2, lowfreq=10)

sms_train <- DocumentTermMatrix(train2,
list(dictionary=dict2))
sms_test <- DocumentTermMatrix(test2,
list(dictionary=dict2))
convert_counts <- function(x) {
  x <- ifelse(x > 0, 1, 0)
  # x <- factor(x, levels = c(0, 1), labels = c("Absent",
"Present"))
}

sms_train <- apply(sms_train,MARGIN=2,
FUN=convert_counts)
sms_test <-  apply(sms_test,MARGIN=2,
FUN=convert_counts)

sms_train <- as.data.frame(sms_train)
sms_test <- as.data.frame(sms_test)
str(sms_train)
```

```
sms_train1 <- cbind(cat=factor(train1$Label),
sms_train)
sms_test1 <- cbind(cat=factor(test1$Label),
sms_test)

sms_train1<-as.data.frame(sms_train1)
sms_test1<-as.data.frame(sms_test1)
set.seed(1234)
svm.nokernel <- svm(cat~., data=sms_train1)
set.seed(1234)
pred.nokernel <- predict(svm.nokernel,
na.omit(sms_test1))

nokernel <- confusionMatrix(pred.nokernel ,
sms_test1$cat, positive="1")
nokernel

svm.linear <- svm(cat~., data=sms_train1,
scale=FALSE, kernel='linear')
pred.linear <- predict(svm.linear, sms_test1[,-1])
linear <-
confusionMatrix(pred.linear,sms_test1$cat)
linear

svm.radial <- svm(cat~., data=sms_train1,
scale=FALSE, kernel='radial')
pred.radial <- predict(svm.radial, sms_test1[,-1])
radial <-
confusionMatrix(pred.radial,sms_test1$cat)
radial

svm.poly <- svm(cat~., data=sms_train1,
scale=FALSE, kernel='polynomial')
pred.poly <- predict(svm.poly, sms_test1[,-1])
poly <- confusionMatrix(pred.poly,sms_test1$cat)
poly

svm.sigmoid <- svm(cat~., data=sms_train1,
scale=FALSE, kernel='sigmoid')
pred.sigmoid <- predict(svm.sigmoid,sms_test1[,-
1])
sigmoid <-
confusionMatrix(pred.sigmoid,sms_test1$cat)
sigmoid

Kernels <- c("No
Kernel","Linear","Polynomial","Radial
Basis","Sigmoid")
Accuracies <-
round(c(nokernel$overall[1],linear$overall[1],poly
$overall[1],radial$overall[1],sigmoid$overall[1]),4)
acc <- cbind(Kernels,Accuracies)
kable(acc,row.names=FALSE)
```