# Using Natural Language Processing to Observe and Understand Public Opinion of the President of Nigeria

## Abstract

This paper explores a method of assessing public opinion of a public figure through the use of natural language processing techniques. Specifically, it employs sentiment analysis and targeted topic modeling using Latent Dirichlet Allocation to model 9,385 news articles heavily featuring Muhammadu Buhari, the 15th president of Nigeria. The methods employed in this paper were successful in assessing public opinion of Buhari in the year of his inauguration. The same methods applied to news articles spanning the four years of his presidency are successful in observing the trend in public opinion of the president but less successful in definitively understanding the causes of the observed trend.

# Introduction

It is generally accepted that transparency during periods of major elections has significant implications on the results of the elections themselves. This notion came to the forefront of political discussion following the aftermath of two major political events in 2016. The first of these was the Brexit referendum of June 2016 where the population of British voters who wished to stay in the EU had a lower turnout than expected due to expectations on the result of the election borne from opinion polls and media representation. The second of these was the United States presidential election of November 2016 where voter turnout for the democratic party was much lower than expected for similar reasons.

The Nigerian political system does not have an effective and accurate means of assessing public opinion at times of general elections, and as a result Nigerian voters are often left to guess where 'public opinion' regarding election candidates lies. This was especially the case in the Nigerian general election of 2015 which was particularly significant in the country's democratic history because an incumbent president was removed from office solely through the voter power.

In practice, the task of accurately assessing public opinion is one which requires significant capital and labor investment. As a result, the Nigerian general election process remains opaque due to a lack of resources to carry out this task. Indeed, even after-the-fact analysis of the election process is often inaccurate due to a lack of key numerical data. Furthermore, wide held scrutiny of the accuracy of easily manipulated data made public by the government limits the impact that analysis using such data can make due to issues of legitimacy.

This project, then, aspires to remedy this problem by proposing a method to bypass both the issue of a lack of numerical data and that of easily misrepresented data by instead drawing insight into the general election process through analysis of qualitative data on a large scale.

I attempt to do this by using two Natural Language Processing techniques on a large number of news articles heavily featuring the 15th and current president of Nigeria, Muhammadu Buhari. My analysis will center around observing and understanding public opinion of him in the months running up to and during his term as president of Nigeria, firstly, by using sentiment analysis to observe the trend in emotional valence over time, and secondly, by using topic modelling in conjunction with the results derived from sentiment analysis to identify the events associated with the fluctuations in sentiment. I now present some background on the topic that will aid in understanding the results of the analysis.

Buhari won the Nigerian general election in March 2015 and was inaugurated as the 15th President of Nigeria in May 2015. When he was sworn in, there was a predominantly positive public opinion of him due to a number of factors. One of the most relevant was that his election coincided with the ousting of an incumbent president solely through voter power. This, in turn, came about due to large global media attention over on the elections which aided in preventing misrepresentation of the election results.

However, after he came into power in 2015, sentiments towards Buhari became increasingly negative as a consequence of numerous incidents that the public viewed negatively. These incidents include the ineffective handling of a recession which began in 2016, the president's several-month sick leave during his term, the lack of impact of his promises to eradicate corruption in the country's political system, and the handling of the kidnapping of schoolgirls from the town of Chibok in northeastern Nigeria by Boko Haram.

I propose that through sentiment analysis on news articles featuring Buhari I will be able to observe the trend of increasingly negative sentiments over time. Moreover, I expect that the observed trend in sentiment will reflect the timing of some of these major incidents. Additionally, I shall employ the use of topic modelling to summarize the articles in particular timeframes - centered around changes in sentiment - in the hope of identifying these incidents as they correlate with the negative sentiment trends that become apparent.

It is important to note two key assumptions on which the success of this project heavily relies. The first is that the sentiment analysis procedure I use will be effective in assessing the opinions present in the articles, and the second is that news articles represent a suitable proxy for public opinion on a public figure.

In completing the project, I found that the analytical process described was successful in observing and understanding public opinion of Buhari over time; although the process was notably more succesful with the smaller timescale of 2015 compared to the larger timescale of November 2014 to November 2018. I was also able to find strong support for my claim that public opinion of Buhari declined over time.

# Data

In this section, I will explain the process of acquiring the data I used to carry out my analysis.

## Overview of the dataset

The first dataset that I use in the "Analysis of 2015 Articles" section is comprised of 2270 observations each representing a single news article, and 10 columns each representing a key detail for each article. The average length of articles in this data set is 1619 words and the timespan of the articles in this dataset is from 2015-01-02 to 2015-12-31.

Similarly, the final dataset which I use in the "Analysis of All Articles" chapter is comprised of 9385 observations and 9 columns. The average length of articles in this data set is 1454 words and the timespan of the articles in this dataset is from 2014-11-01 to 2018-11-01.

Initially, before I cleaned the data, there were a number of non-helpful variables in my data frame originally, so I subsetted the data into 9 variables with information that might prove useful in my analysis. The 9 variables that I retained in each dataset were the following:

- FileName: The name of the original html file downloaded from Factiva
- Section: The news section that the article was taken from
- Headline: The headline of the article
- WordCount: The word count of the pre-processed article
- PubDate: The date that the article was published
- SerialName: The serial name of the article publisher
- TD: The text body of the article
- Pub: The full name of the article publisher
- DocumentNumber: The ID of the article according to the Factiva labeling format

## Data Acquisition

I after a long stint in API querying, I reached out to the Yale University Data Librarian, Barbara Esty, who informed me that as a Yale student I had access to the university's **Factiva** subscription Factiva is a data acquisition and research tool that aggregates content from both licensed and free news sources. It provides organizations with search, alerting, dissemination, and other information management capabilities.

With Barbara's assistance, I learned how to query the Factiva database for the news articles I desired by inputting my search terms into the Factiva search window using the appropriate syntax. The code I used to search for articles was:

> **(hlp=("Muhammadu Buhari") or hlp=("Buhari")) and atleast5 Buhari and wc>1000 and date from 01/01/2015 to 12/31/2015**

This code looks for appearances of "Muhammadu Buhari" appearing together or simply "Buhari" appearing at least five times in articles published between January 1st, 2015 and December 31st, 2015. I opted to limit my search terms to articles where the president's name appears at least five times to remove the influence of articles that may not have been about Buhari himself but may mention his name once or twice. Additionally, I restricted my articles to those with word counts greater than 1000 to ensure that each article used is substantive in content and to remove the influence of article snippets that may only contain a small number

Table 1: Top article sources

| Source | Number of Articles |
|---|---:|
| All Africa Global Media | 3244 |
| Other | 1563 |
| The Sun Publishing Ltd. | 1070 |
| Vanguard Media Limited | 876 |
| AllAfrica, Inc. | 705 |
| Leaders & Company Limited | 551 |
| Punch Nigeria Limited | 407 |
| African Newspapers of Nigeria Limited | 387 |
| Media Trust Limited | 358 |
| Daily Independent | 224 |

of words. This search code resulted in a total of 3791 articles, but after using the Factiva duplicate removal feature that omitted 1521 similar articles, I was left with 2270 articles to work with.

Similarly, the search code that I used to search for the articles of the "All Articles" chapter dataset was:

> **(hlp=("Muhammadu Buhari") or hlp=("Buhari")) and atleast5 Buhari and wc>750 and date from 11/01/2014 to 11/01/2018**

The only differences between the code above and the previous chunk are the word count and timeframe. Here, I relaxed my restriction on the wordcount to allow for all articles greater than 750 words in length, and I searched for articles published between November 1st, 2014 and November 1st 2018. I relaxed the wordcount restriction in order to obtain a larger number of articles for the main analysis if the full presidential term. This search code resulted in a total of 16,323 articles, but after using the Factiva duplicate removal feature that omitted 6938 similar articles, I was left with 9385 articles to work with.

After having obtained the articles on Factiva, I was confronted with the challenge of aggregating the thousands of articles I had acquired in a way that I allowed me to work with them in R. For both Factiva searches, I had to download a maximum of 100 articles at a time; creating a compiled HTML file for each one. I then needed to find a way to compile the many HTML files into one readable CSV file. Fortunately, this was a roadblock that Barbara Esty had encountered before and she was able to give me access to her **SAS**[1] program that creates an excel file with the aggregated article data after receiving the many HTML files as input.

## Sources

I obtained the articles I used in this project by downloading and compiling news articles from various sources in Factiva.

Table 1 presents the top of sources from which I gathered news articles:

## Packages

The packages that I use in my analysis are the following:

- dplyr (Wickham, François, Henry, & Müller, 2018)
- syuzhet (Jockers, 2015)
- lubridate (Grolemund & Wickham, 2011)

[1] A software suite developed by SAS institute for advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics

- ggplot2 (Wickham, 2016)
- tm (Feinerer, Hornik, & Meyer, 2008)
- topicmodels (Grün & Hornik, 2011)
- textreg (Miratrix, 2018)
- kableExtra (Zhu, 2018)
- gridExtra (Auguie, 2017)
- wordcloud (Fellows, 2018)
- tidytext (Silge & Robinson, 2016)

## Data Cleaning

To clean up the article text in the `TD` column I used the `tm` (Feinerer et al., 2008) package to create corpus containing the 2270 articles. A corpus is a collection of documents containing natural language text. The purpose of this data cleaning was to remove all parts of the text that do not aid the natural language processing techniques in extracting semantic, syntactic or contextual information from the articles. I removed all punctuation and numbers from the documents, transformed all words to lower case, and removed all common stopwords.

# Analysis of 2015 News Articles

The purpose of this section is to explain in detail the sentiment analysis and topic modelling procedures as well as to take a close look at the first year that president Muhammadu Buhari came into power. The analysis in this chapter was done using a dataset of 2270 articles featuring Buhari that were published in 2015.

## Sentiment Analysis of 2015 Articles

My main source for learning how to perform sentiment analysis was Jonathan D. Fitzgerald's article on the topic (Fitzgerald, 2018).

The sentiment analysis method that I used is a **lexicon-based** approach which calculates sentiment according to each word's semantic orientation towards the following 8 emotions and 2 sentiments:

**Emotions**

- Anger
- Anticipation
- Disgust
- Fear
- Joy
- Sadness
- Surprise
- Trust

**Sentiments**

- Negative
- Positive

The first 8 variables represent the 8 basic emotions as defined by the National Research Council of Canada.

I carried out sentiment analysis using a word-emotion association lexicon and Matthew Jockers' "Syuzhet" package (Jockers, 2015). Sentiment and emotion word-association lexicons are databases which have captured word to sentiment and word to emotion associations that enable the analysis of human emotions present in natural language text.

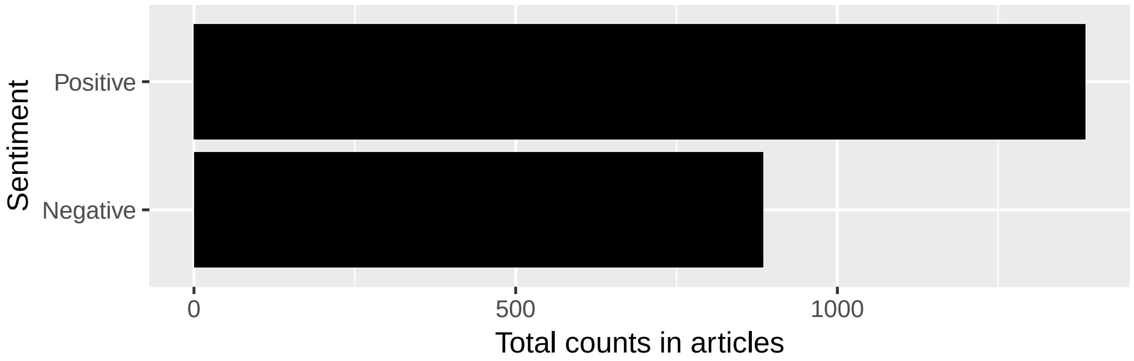

Figure 1: Sentiments present in 2015 articles

This project makes use of the 14,182-word **NRC Word-Emotion association Lexicon** (Mohammad & Turney, 2010, 2013) developed by Saif M. Mohammad, a senior research officer at the National Research Council Canada, which was manually created through crowdsourcing techniques. In the lexicon structure, each of the 14,182 English words listed is associated with one of 8 emotions or 2 sentiments. For each entry of a given word-emotion/word-sentiment combination, the lexicon lists a value of 0 or 1 to indicate whether the word is associated with that particular emotion/sentiment or not.

> The sentiment analysis procedure is carried out by scraping through the text input, recognizing words that appear in the NRC lexicon, and for each recognized word, adding to the tally of emotion or sentiment that the lexicon associates with it.

I illustrate this with a few examples below.

1. The word "*abandon*" has associations of 1 with fear, sadness, and negative, and associations of 0 with all other emotions/sentiments.
2. The word "*junk*" has an association of 1 with negative and 0 with everything else.
3. The word "*overture*" has an association of 1 with anticipation and 0 with everything else.
4. The word "*vicar*" has associations of 1 with positive and trust and 0 with everything else.
5. The word "*vote*" has associations of 1 with anger, anticipation, joy, negative, positive, sadness, surprise, and trust, and 0 with disgust or fear.

The output is thus a dataframe with as many rows as articles (i.e. 2270 in this case), and 10 columns which list the counts of positive, negative, anger, anticipation, disgust etc. related words in each article.

Note that not all words have a positive/negative association, some words have both positive and negative associations (such as *vote*), and not all words have an emotion association.

### Visualizing the data

Before I could begin to interpret the results, I realized that the lexicon-based sentiment analysis method that I used puts additional weight on articles with a greater wordcount. To account for this, I normalized the emotion and sentiment values to give each article the same weight as all the others by recording each emotion as a proportion of the sum of all emotions; and each sentiment in proportion to the sum of the two.

Figure 1 displays an aggregated count of the positive and negative sentiments in each of the articles. It is clear here that the overall sentiment towards Buhari in 2015 was mostly positive, although not overwhelmingly so.

Figure 2 is a barplot which displays aggregated counts of the 8 emotions identified in the articles. Of all the emotions, "Trust" is the most present by a wide margin, appearing more than twice as frequently as most of the other emotions.

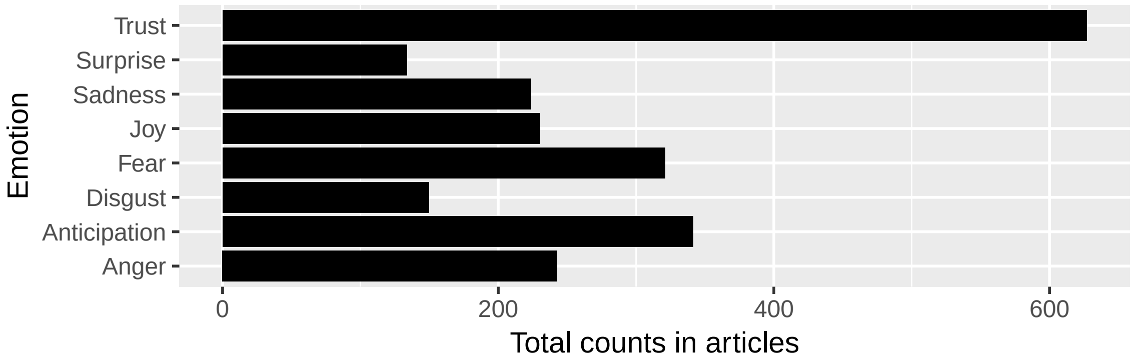

Figure 2: Emotions present in 2015 articles

Table 2: Precise emotion values for 2015 articles

| Emotion/Sentiment | Counts |
|---|---:|
| Anger | 242 |
| Anticipation | 341 |
| Disgust | 149 |
| Fear | 321 |
| Joy | 230 |
| Sadness | 223 |
| Surprise | 133 |
| Trust | 626 |
| Negative | 884 |
| Positive | 1385 |

This result is in line with my hypothesis that as Buhari came into power in May 2015, feelings towards him were still largely positive. Additionally, it is understandable that public emotion towards a newly democratically elected president would indicate a large presence of trust.

Table 2 displays the precise aggregated values of each of the emotions and sentiments.

### Observing emotional valence

**Emotional valence** is is the term used to define the sentiment tendency of a body of text - here, a news article. It is calculated using the following formula:

$$\text{Emotional Valence} = \text{Positive word count} - \text{Negative word count}$$

For example, if after running sentiment analysis a particular article comes up with a positive count of 38 and a negative count of 17 the emotional valence would be 21. Conversely, if the article's positive count was 17 and its negative count 38, its emotional valence would be -21. I use the terms *emotional valence*, *public sentiment*, and *sentiment*[2] interchangeably and emotional valence is the proxy I use for public opinion.

To observe emotional valence of the articles, I create a vector of emotional valence for each of the articles using the above formula using the non-normalized sentiment values to allow for more natural interpretation of the graphs that follow.

[2] When not referring to one of 'Positive' or 'Negative' but instead referring to how an individuals or a group feels towards something or someone

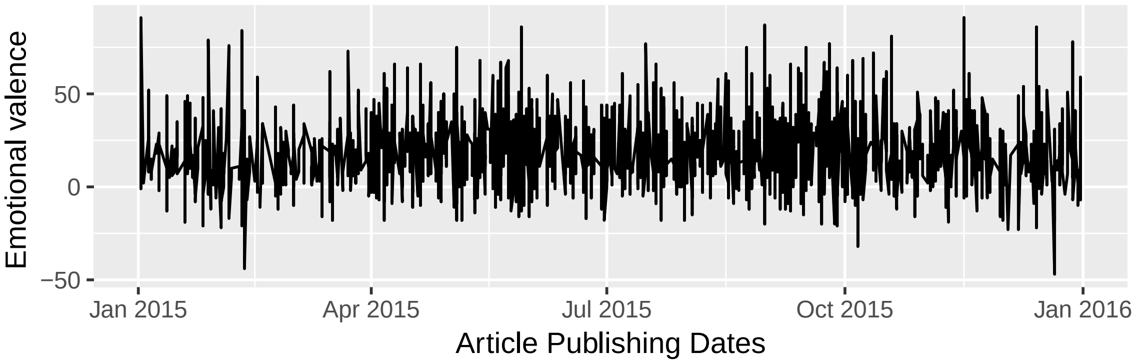

Figure 3: Line graph of emotional valence over 2015

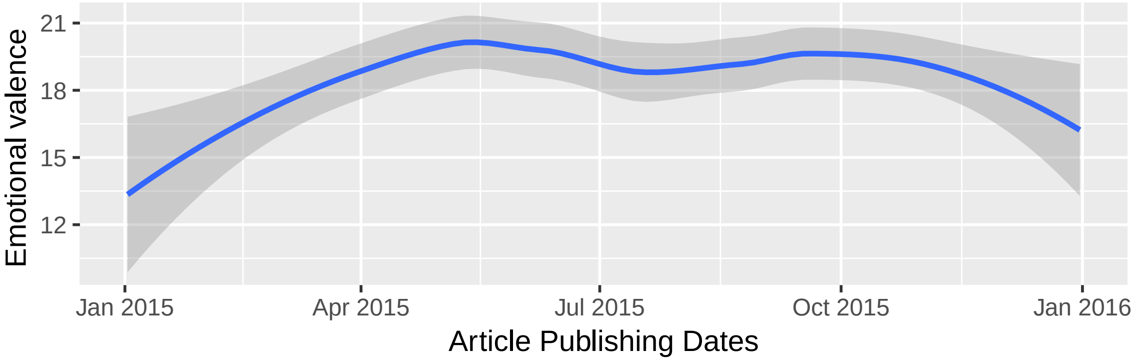

Figure 4: Loess-smoothed line graph of 2015 emotional valence

Below, I display the summary statistics of the articles' emotional valences to give a broad sense of the distribution of sentiment in the text.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -47.00    9.00   18.00   18.95   28.00   91.00
```

In the following section, I give a more thorough account of this distribution.

### Tracking emotional valence across time

The emotional valence values of the previous section provide a sense of the sentiment within the 2015 articles on Buhari but fail to give us an idea of how sentiments developed over time. In this section, I will use the emotional valance values along with the publishing dates of each article to track public sentiments towards Buhari over the course of 2015.

Figure 3 is a line graph of emotional valence over time. As is clear from the graph, there doesn't seem to be an observable trend in emotional valence over the year due to the noise in the data. Simplifying the plot using smoothing methods makes the trend more observable. Below, I explain how **LOESS Smoothing** is performed, then apply it to the data.

LOESS is a nonparametric method for estimating regression surfaces. In the LOESS method, weighted least squares are used to fit linear or quadratic functions of the predictors at the centers of neighborhoods. The radius of each neighborhood is chosen so that the neighborhood contains a specified percentage of the data points. The fraction of the data, called the smoothing parameter, in each local neighborhood controls the smoothness of the estimated surface. Data points in a given local neighborhood are weighted by a smooth decreasing function of their distance from the center of the neighborhood.

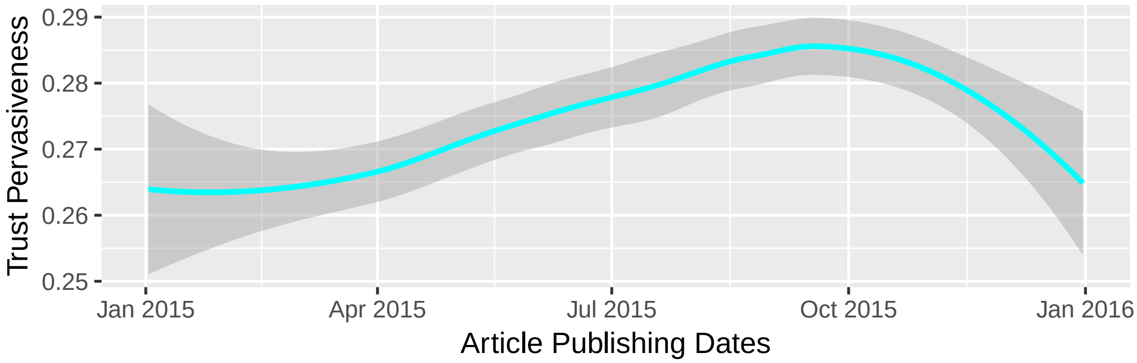

Figure 5: Presence of Trust emotion over 2015

Figure 4 displays the 2015 emotional valence after having been smoothed using LOESS. Already, the trend in emotional valence over 2015 becomes a lot more clear. Sentiments fluctuate over the course of the year; with the lowest point being at the start of 2015. There is a slight dip in February before sentiments become notably more positive and peak in May. They then decrease again to a local minimum in late July/early August before peaking again in October and finally becoming much less positive towards the end of the year but still rounding off slightly greater than at the start of the year.

Looking at the graph as a whole, the most interesting feature is the highest peak that occured in May. I hypothesize that this peak is linked to the president's inauguration that occurred that same month. Later in this chapter, I will attempt to identify the events behind the *May Peak* with more certainty using **topic modelling**.

### Tracking public emotion over time

In this section, I display loess-smoothed line graphs of each of the eight emotions as they fluctuate over 2015 and draw insights from them.

First I show the plots of Trust, Anticipation, and Fear because they were the most present of all eight emotions.

**Trust:** Figure 5 displays the pervasiveness of trust over time. Trust experiences the same dip in February as shown in the emotional valence plot, peaks in October then decreases significantly until the end of the year. From this, one can interpret that the event(s) that caused the second peak and rapid decline in sentiments are very closely correlated with trust.

It is interesting to note, however, that the May peak seems to have had no correlation with trust. If the peak is in fact caused by the president's inauguration, this might imply that the inauguration itself had little bearing on how much the public trusted him. This makes sense given that while a presidential inauguration is a significant event, it does not actually indicate a president's capabilities and as a result would not affect how much he or she is trusted.

**Anticipation:** Figure 6 displays the pervasiveness of anticipation over time. With anticipation - the second most present emotion in the articles - I noticed a stark difference in comparison to the trust plot. While trust was highly correlated to the October peak, the February dip, and the end-of-year decline, anticipation is hardly correlated with anything but the May peak.

This strengthens my hypothesis that the May peak was caused by Buhari's inauguration because it is reasonable to assume that a presidential inauguration would cause a significant increase in public anticipation leading up to the event.

**Fear:** Figure 7 displays the pervasiveness of fear over time. This plot is interesting because it seems to closely negatively mirror the anticipation plot, though with less extreme values.

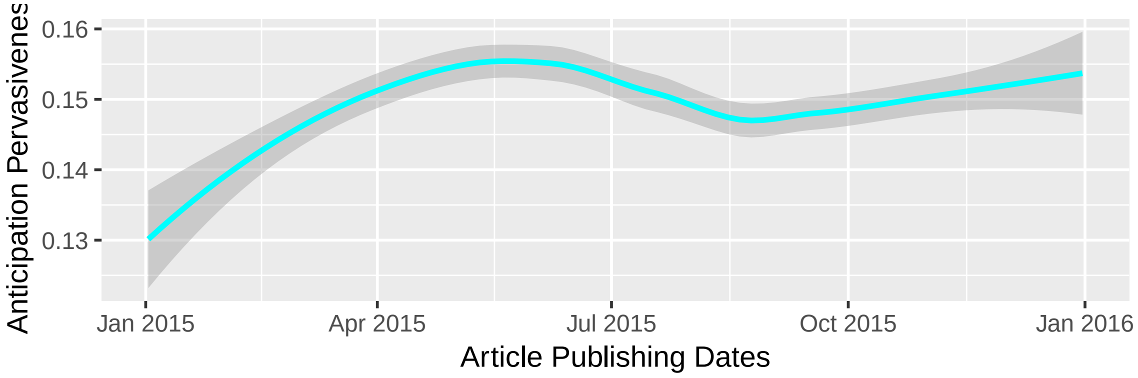

Figure 6: Presence of Anticipation emotion over 2015

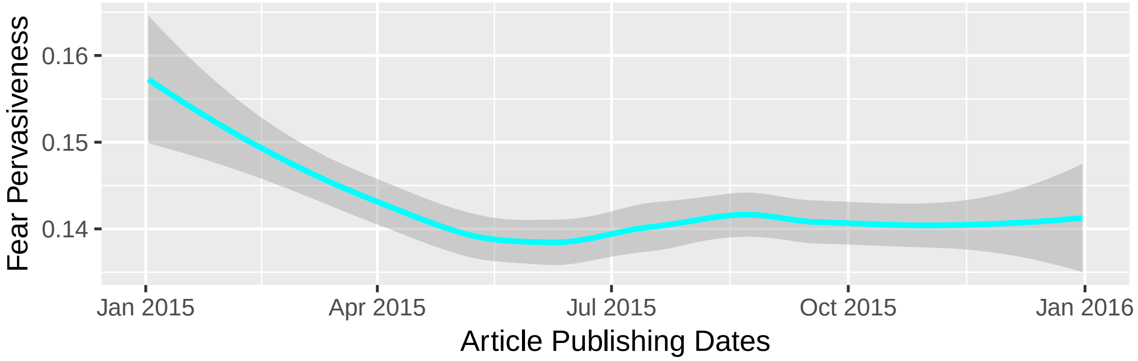

Figure 7: Presence of Fear emotion over 2015

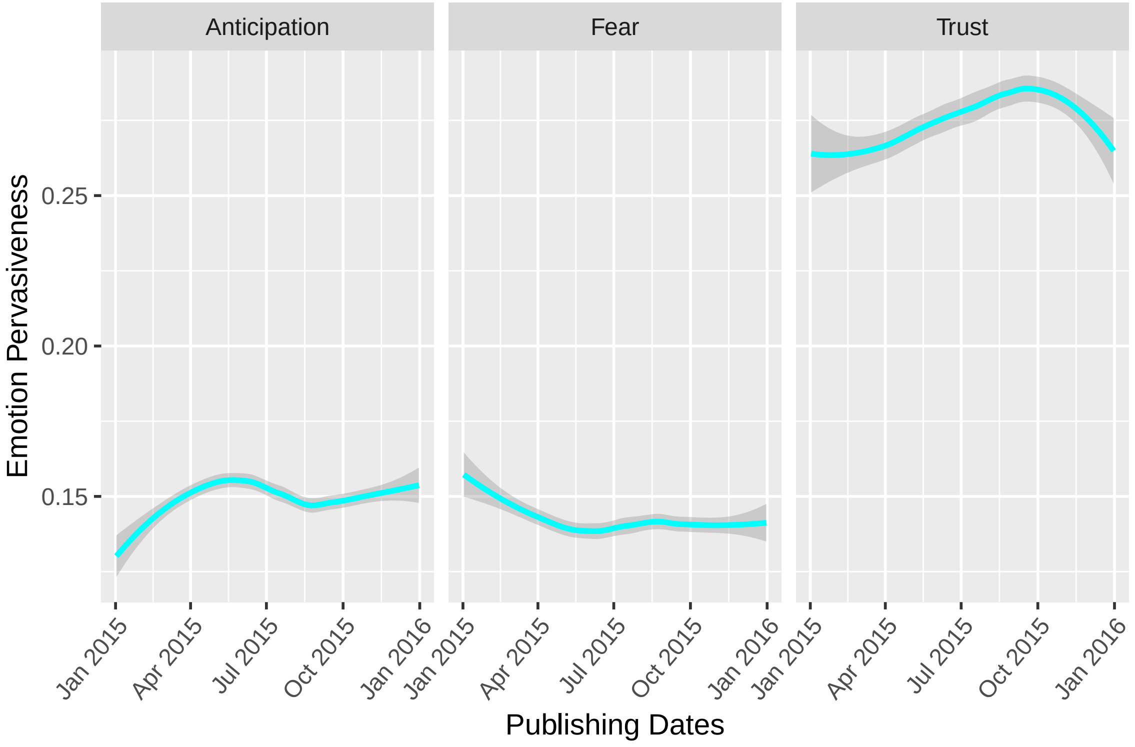

Figure 8: Presence of Trust, Anticipation, and Fear emotions over 2015

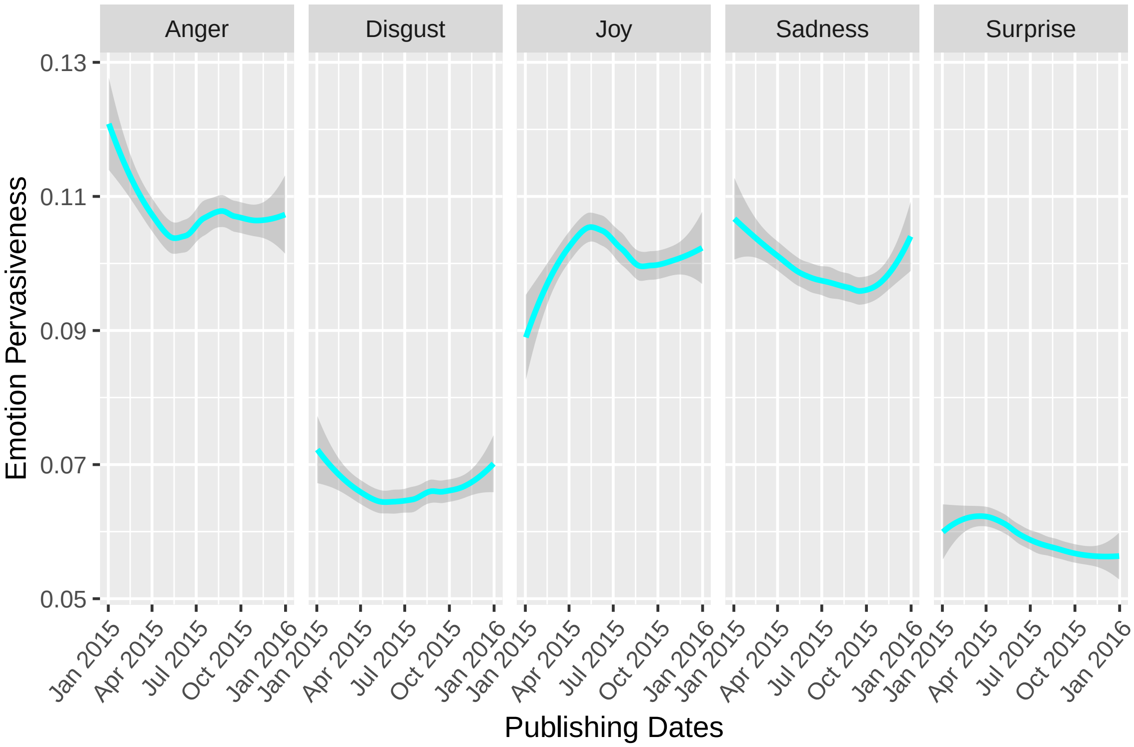

Figure 9: Presence of Anger, Disgust, Joy, Sadness and Surprise emotions over 2015

Figure 8 displays the plots of Trust, Anticipation, and Fear together for ease of comparison. Figure 9 displays the plots of the other five emotions mainly for interest to the reader.

## Topic Modelling of 2015 Articles

### Topic Modelling and Latent Dirichlet Allocation

In this section, I will apply topic modelling to the set of 2015 articles.

### Description

**Topic modelling** is an unsupervised learning tool for organizing and summarizing text data. The specific topic modelling method that I shall use to analyze the articles is **Latent Dirichlet Allocation (LDA)**.

A topic model is a particular type of statistical model used for discovering the abstract topics contained within a collection of documents. LDA, which is one of the most common topic models currently in use, was first presented by David Blei, Andrew Ng, and Micheal I. Jordan in 2003 (Blei, Ng, & Jordan, 2003).

### Assumptions

LDA assumes the text to be organized in a **Bayesian hierarchical model**. In general, this is a statistical model that estimates the parameters of the posterior distribution using the Bayesian method. When referring to a Bayesian hierarchical model in the context of text analysis, the following assumptions must hold:

1. Each document is a mixture of topics
2. Each topic is a distribution over words
3. Each word is sampled from one of those topics

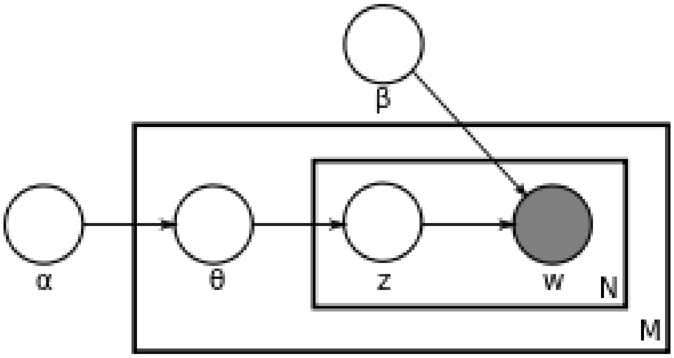

Figure 10: LDA Plate Notation

Bayesian hierarchical modelling is written in multiple levels which estimate the parameters of the posterior distribution using the Bayesian method. In short, LDA assumes that documents are a probability distribution over latent topics and topics are probability distributions over words.

LDA also assumes that individual words contain semantic information and that words with related semantic information will typically occur together in a given document. In the LDA modelling process, latent topics are identified by observing groups of words in the corpus that frequently occur together within documents.

### LDA Model in Plate Notation

The various dependencies among the LDA model parameters can be consisely represented using plate notation. Figure 10 displays the plate notation diagram for Latent Dirichlet Allocation.

### Model Parameters

The model parameters are the following:

- $\alpha$ is the parameter of the Dirichlet prior on the per-document topic distribution
- $\beta$ is the parameter of the Dirichlet prior on the per-topic word distribution
- $\theta_m$ is the topic distribution for document $m$
- $z_{mn}$ is the topic for the n-th word in document $m$
- $w_{mn}$ is the specific word

### The Dirichlet Distribution

The probability density function of a Dirichlet distribution is as follows:

$$p(x|\alpha) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

### The Generative Process

In order to understand how LDA derives a collection of latent topics through observations of the words, it is important to first understand the assumed generative process whereby documents are created (Blei et al., 2003).

For a corpus $D$ consisting of $M$ documents each of length $N_i$:

1. Choose $\theta_i \sim \text{Dir}(\alpha)$, where $i \in \{1, \ldots, M\}$ and $\text{Dir}(\alpha)$ is a Dirichlet distribution with a symmetric parameter $\alpha$ which typically is sparse (i.e. $\alpha < 1$)

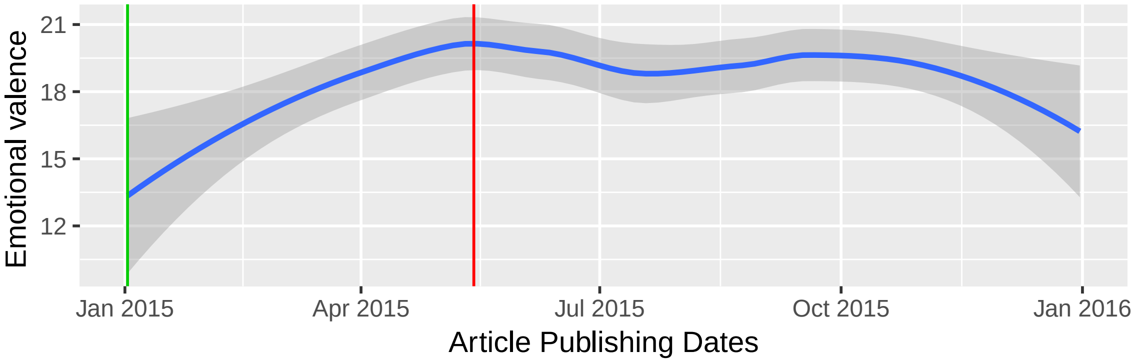

Figure 11: Emotional valence of 2015 articles with timeframe-based subset indicators

2. Choose $\varphi_k \sim \mathrm{Dir}(\beta)$, where $k \in \{1, \dots, K\}$ and $\beta$ which typically is sparse (i.e. $\beta < 1$)
3. For each of the word positions $i, j$ where $i \in \{1, \dots, M\}$ and $j \in \{1, \dots, N_i\}$

a. Choose a topic $z_{i,j} \sim \mathrm{Multinomial}(\theta_i)$
b. Choose a word $w_{i,j} \sim \mathrm{Multinomial}(\varphi_{i,j})$

### The LDA process using R

In this section, I explain the process of Latent Dirichlet Allocation and use it to attempt to identify the causes of the positive peak that occurred in May 2015 (see Figure 4).

### Selecting the Data

The first step was to subset the data within my desired timeframe and create a separate dataframe with which I could carry out topic modelling. As the goal here is to identify topics that may have lead to the positive spike that occurred around May 2015, I restricted my analysis to 612 articles published between 2015-01-02 and 2015-05-14.

Figure 11 displays again the emotional valence of the articles over 2015 but with indicators to show how I have subsetted the data for topic modelling. The green line indicates where the new dataset starts and the red line indicates where the new dataset stops.

### Additional Pre-processing

Before carrying out the topic modelling process, I had to do some additional pre-processing of the data. I removed any additional whitespace from the documents and stemmed all words in the documents to ensure that words that have the same meaning or different verb forms of the same word aren't duplicated. It would not have made sense to stem the words prior to carrying out sentiment analysis because it would have inhibited the identification of words and corrected attribution of correct emotions and sentiments. However, because Latent Dirichlet Allocation is a *bag of words*[3] model, the actual meaning of individual words is irrelevant, so word distortion does not affect the efficiency of the algorithm.

### LDA Modelling

The LDA pre-processing step is an important one because it will significantly affect the quality of the results.

[3] In a bag of words model, text is represented as a "bag" of its words, which disregards grammar, word order, and even word meaning but takes frequency of word occurrence into account

First, I used the corpus to create a *document term matrix (dtm)*, which is a large matrix for which the rows represent each document and the columns represent each unique word in all documents. This document term matrix is of size 612 x 3966. An entry $x_{ij}$ indicates the number of times word $j$ appears in document $i$

After creating the document term matrix, I removed words which occured too frequently between articles because they would skew the algorithm's topic-identifying procedure, and I remove the articles that don't have any of the frequently occuring words. I then remove articles with words that occur too rarely. I infer the rareness of words by calculting the **Term frequency-inverse document frequency (tf-idf)** which is defined as

$$\text{tf-idf}_{dt} = \text{tf}_{dt} * \log\left(\frac{D}{\text{df}_t}\right)$$

where:

- $\text{tf}_{dt}$ is relative frequency of term $t$ in document $d$
- $\text{df}_t$ is the number of documents containing term $t$

The tf-idf measures the importance of a term and as such, I exclude terms that have a low tf-idf to exclude rare words. After all the LDA pre-processing was done, I used the `topicmodels` (Grün & Hornik, 2011) package to create multiple LDA models with differing numbers of topics.

### Picking the Number of Topics for the LDA Model

In running LDA, one needs to decide on the number of topics beforehand. As a result, the question of how many topics to include needs to be addressed. This can be done by creating several models, each with a different number of topics $k$, then choosing the optimal model based on its ability to predict an unseen test set. A typical measure to do this for language models is **Perplexity**, which is the inverse probability of the test set normalized by the number of words. It is given by the following formula:

$$\text{Perplexity}(M_{\text{test}}) = \exp\left[\frac{\sum_{d=1}^{D} \log(p(\boldsymbol{w}_d))}{\sum_{d=1}^{D} N_d}\right]$$

This measure decreases monotonically in the likelihood of the test data, thus lower values indicate better modeling performance. Put differently, a good model will give a high probability to a real sentence, a lower perplexity implies a high probability for the particular model, and thus a lower perplexity implies a better model (Hörster, Lienhart, & Slaney, 2007).

where:

- $M$ is the model being evaluated
- p($M$) is the likelihood of the model $M$
- $D$ is the number of documents in the test data set
- $N_d$ is the number of words in document $d$
- $\boldsymbol{w}_d$ is the set of words in document $d$

The process of picking the optimal model is as follows:

1. Split the dataset into a training data set and a test data set
    a. The training set contains 90% of the full[4] data set
    b. The test set contains 10% of the full data set
2. Run LDA for several values of $k$ using the training data set
3. For each model calculate the perplexity using the test data set
4. Pick the optimal model with $k^*$ topics based on its perplexity. This could be through one of the two following methods
    a. Pick the model with lowest perplexity
    b. Pick the model with a low perplexity but without too many topics
5. Run LDA with $k^*$ topics using the full data set to obtain the final model

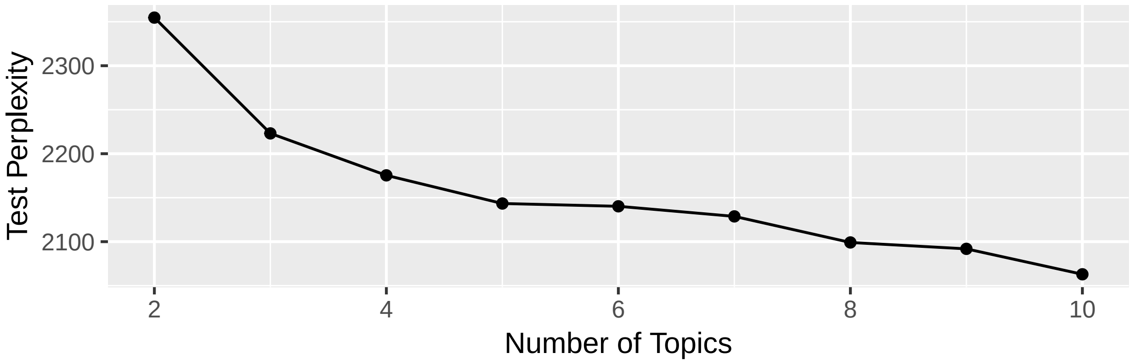

Figure 12: Perplexity of May Peak LDA models with 2-10 topics

Figure 12 displays the perplexities of 9 LDA models with 2-10 topics. The decrease in perplexity seemed to plateau after 5/6 topics so I opted to use an LDA model of $k^* = 6$ topics. It's important to pick the process by eyeballing as well as by observing the reduction in perplexity because the greater the number of topics, the lower the perplexity but also the less detailed the topics themselves are. A common strategy in this type of situation is to use the *elbow method* which requires one to eyeball the plot and decide on the value where the "elbow" of the exponentially decreasing line graph occurs.

**Identifying the 2015 Topics**

Figure 13 displays the 6 topics identified through LDA and the top 20 most frequently occurring words in each of those topics. Below, I attempt to identify what some of these topics are using the words displayed for each topic and my general understanding of the subject area.

**Topic 1:** This topic seems to be the closest to what I was hoping to identify through LDA. The topic frequently features words such as "voter", "card", "defeat", and "won" which are are closely related to the subject of a general election. Additionally, "abuja"[5] also frequently appears in this topic, which might be because it is where the inauguration took place.

**Topic 2:** Words such as "peac-*e*", "delta" and "niger"[6], "ethnic", "violence", and "insurg-*ency*" imply that this topic is related to tensions around the Niger Delta region where militant groups formed to oppose the exploitation of the people and land of that region by large oil companies which mine crude oil in region.

**Topic 3:** The words of this topic seem to be related to Nigeria's economic state, and might cover Buhari's claims regarding economic stimulation of the country, or how the country's economic situation would change when he becomes president. Frequently appearing words such as "nnpc"[7], "crude", "petroleum" , "price", "revenu-*e*", "naira"[8] all relate to money and oil. Nigeria is one of the largest petroleum exporters in the world and these exports account for a significant proportion of the country's GDP. As such, oil and oil revenues are usually a particularly pervasive topic in Nigeria's political discourse.

**Topic 5:** It is difficult to discern the broad topic here. However, the frequency of the words "christian" and "muslim" suggest that this topic covers discourse around religion and politics. The vast majority of Nigeria's population identify as christian or muslim, with a slightly greater muslim population overall. Moreover, the religious identity of the president is often very relevant in political discourse due to it's affect on constantly shifting power dynamics between the Nigeria's predominantly muslim north and predominantly christian south.

[4] The subsetted dataset with 612 articles

[5] The political capital of Nigeria

[6] The Niger Delta is an area in the southern region of Nigeria where most of the country's crude oil is extracted

[7] Nigerian National Petroleum Corporation

[8] The currency used in Nigeria

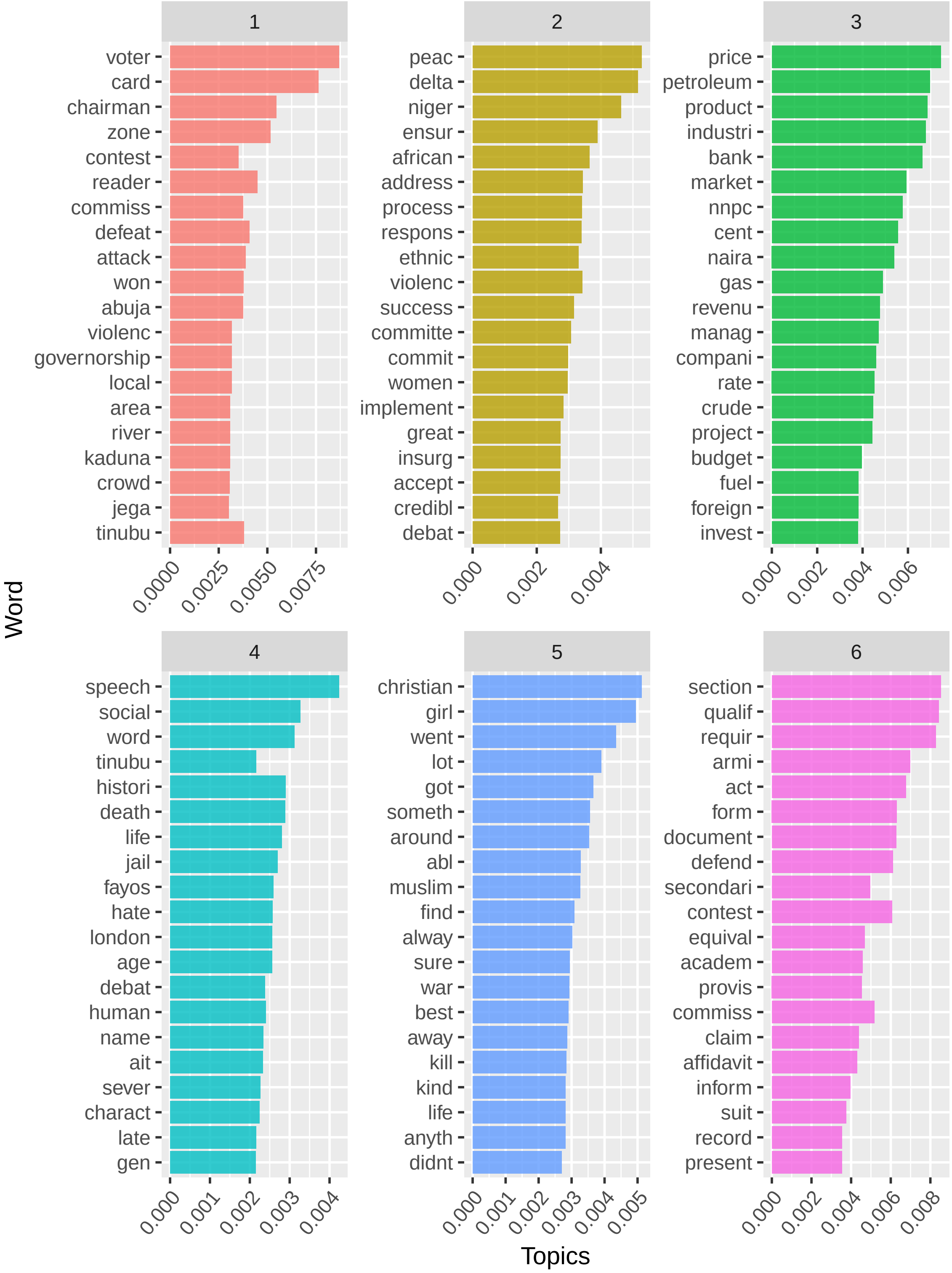

Figure 13: Topics of articles between January 2nd 2015 and May 14th 2015

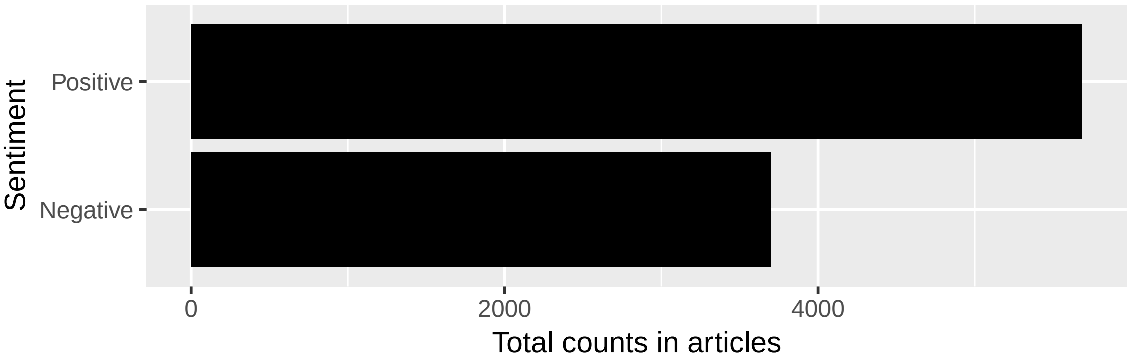

Figure 14: Sentiments present in 2014-2018 articles

**Topic 6:** The context of this topic is not entirely clear, however the frequency of words such as “document”, “contest”, “claim”, “affidavit” and “suit” imply a relation to the law or a law case of some sort.

# Analysis of All News Articles

In this section, I show the results from carrying out the same process of sentiment analysis and topic modelling that I use for the 2015 articles in Chapter Two with the dataset of 9385 articles featuring Buhari that were published in the four years between November 2014 and November 2018.

The motivation behind this section is to test whether the same process I used to understand public opinion of Buhari during his election year can be used to track sentiments towards him over the course of his presidency thus far. I hoped to identify a decreasing trend in emotional valence over time, with negative fluctuations appearing around the times of some of the significant incidents mentioned in the introduction.

## Sentiment Analysis

### Visualizing the data

Figure 14 displays an aggregated count of the of the positive and negative sentiments in each of the articles. It seems that the relative presence of positive sentiment to negative sentiment is similar to that which was shown in the 2015 data. In both cases, positive sentiment makes up about 60% of total sentiment counts.

At first glance, the data would suggest that I may have to revise my hypothesis that sentiments towards Buhari became more negative with time because the aggregated data still implies that public opinion was largely positive. However, observing the trend in emotional valence over time gave some interesting results; this I elaborate on in the following section.

### Tracking emotional valence across time

In this section, I will use the emotional valance values along with the publishing dates of each article to track public sentiments towards Buhari over the course of his presidency so far.

Figure 15 is a line graph of the emotional valence over time. As is evident, there doesn’t seem to be any observable trend in emotional valence over the year due to the noise in the data. Simplifying the plot with LOESS smoothing will help make the trend more understandable.

Figure 16 displays a loess-smoothed line graph of emotional valence over the four years that my analysis is centered around. Now notice that although the aggregated data that I displayed in the previous section

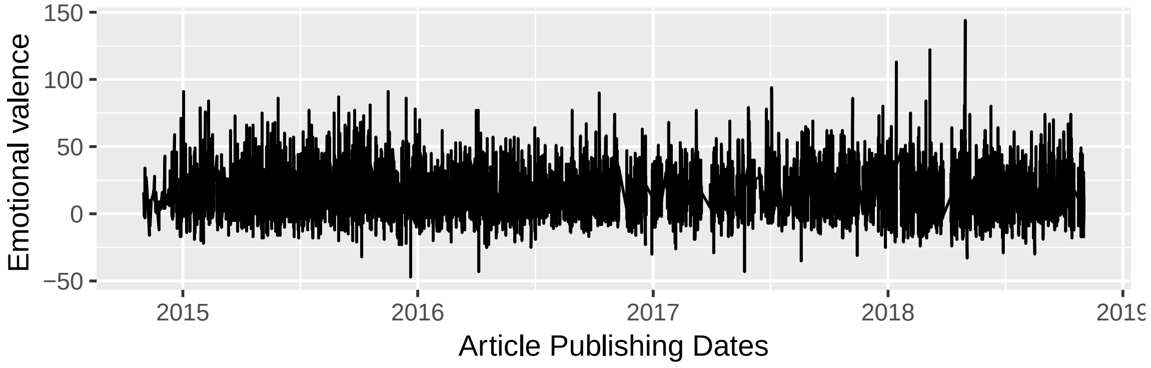

Figure 15: Line graph of emotional valence from 2014-2018

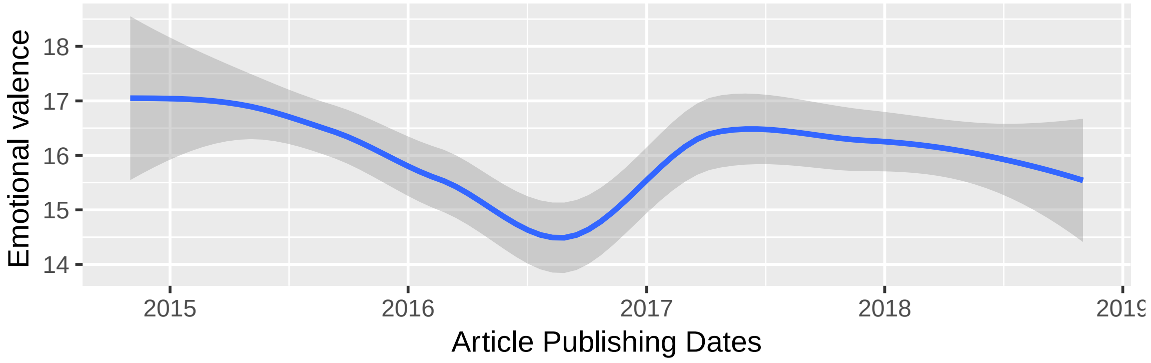

Figure 16: Loess-smoothed line graph of 2014-2018 emotional valence

suggested that the overall presence of the 10 emotions and sentiments were relatively similar in comparison with the 2015 article data, the trend over time tells a completely different story. Looking at the full timescale, there seems to be a consistent downward trend in the data, which is consistently lower than the emotional valence of the 2015 articles in Figure 4. This confirms my hypothesis. It is important to note, however, that the confidence intervals of the loess plot are rather wide and as a result one has to be careful about how precisely one interprets the graph. Despite this, the data does support my hypothesis that sentiments towards Buhari became increasingly negative as his presidency progressed.

## Topic Modelling of 2014-2018 Articles

### The LDA Process using R

In this section, I show the results of applying LDA to the 2014-2018 data in an attempt to identify the causes of the negative trough that occurred in 2016 (see Figure 16).

### Selecting the data

The first step again was to subset the data within my desired timeframe and create a separate dataframe with which I could carry out topic modelling. Here, I restricted my analysis to 4536 articles published between 2015-01-02 and 2016-08-20.

Figure 17 displays again the emotional valence of the articles from 2014-2018 but with indicators to show how I have subsetted the data for topic modelling. The green line indicates where the new dataset starts and the red line indicates where the new dataset stops.

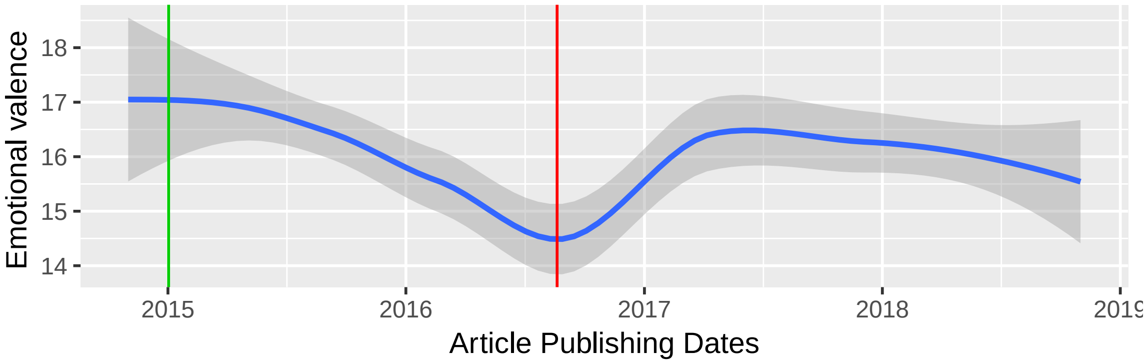

Figure 17: Emotional valence of 2014-2018 articles with timeframe-based subset indicators

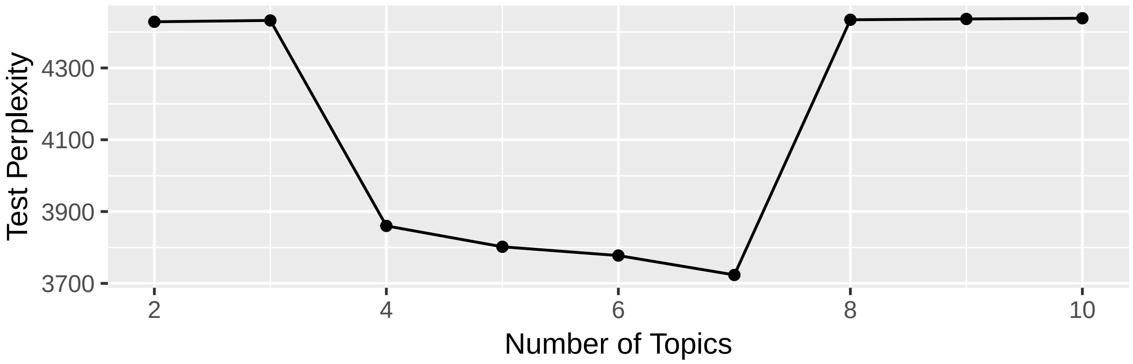

Figure 18: Perplexity of LDA models with 2-10 topics

**Number of Topics to use**

Figure 18 displays the perplexities of 9 LDA models with 2-10 topics. Clearly, the model with 7 topics has the lowest perplexity of the set of models, so I opted to use an LDA model of $k^* = 7$ topics. Note, however, that the average perplexities of these models are almost double those of the LDA models displayed in Figure 12. This implies that the topics of the models created from the 2014-2018 data may not be as succinct as those of the 2015 data.

**Identifying the 2014-2018 Topics**

**Figure QQQ** displays the 7 topics identified through LDA and the top 20 most frequently occurring words in each of those topics. Below I attempt to identify what these topics are using the words displayed for each topic and my general understanding of the subject area.

**Topic 1:** Although it is difficult to ascertain what the topic covering all these words might be, the words do show signs of several important subjects in Nigeria's political sphere. "osinbajo" is the last name of Yemi Osinbajo, the vice president of Nigeria under Buhari. As was explained in the topic modelling section of the previous chapter, "christian" and "muslim" tensions often appear in political discourse. Lastly, "biafra" was a secessionist state which existed from 30 May 1967 to January 1970 and was made up of the states in the Eastern Region of Nigeria.

**Topic 3:** Frequently appearing words such as "restructure-$e$" "contract", "refiner-$y$", and "energ-$y$" imply that this topic is vaguely related to Nigeria's energy infrastructure.

**Topic 4:** The appearance of words such as "presidentelect" and "inaugur-$ation$" imply that this collection of words somehow related to the run-up to Buhari's inauguration after he had been elected as president in March 2015.

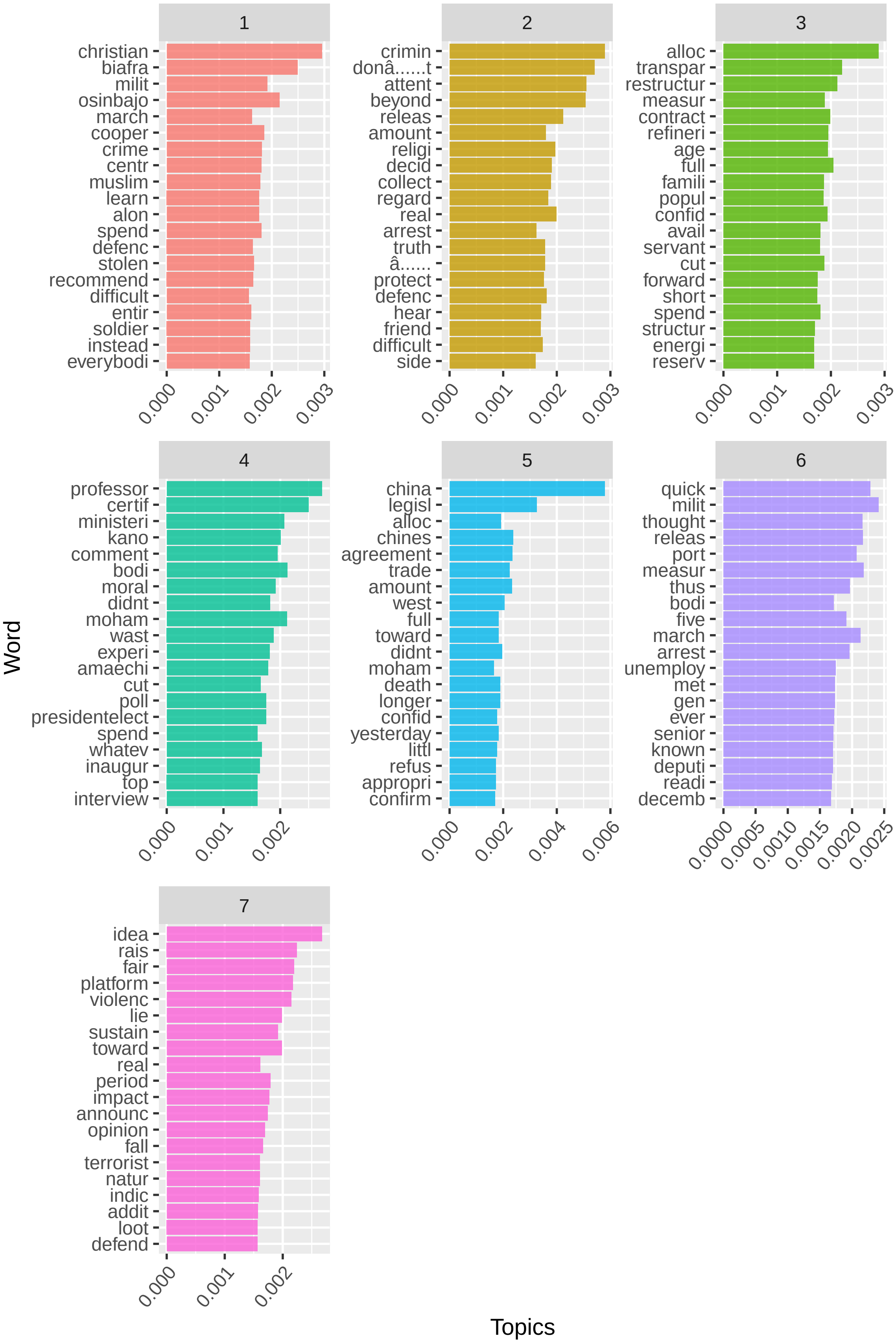

Figure 19: Topics of articles between January 2nd 2015 and August 20th 2016

**Topic 5:** This topic features words such as “china”, “chines-*e*”, “trade” and “agreement” so it likely related to loan and trade agreements that the Buhari administration may have had with China.

It seems that the topics identified in this LDA model are difficult to ascertain. This may be because the timeframe that I subset the data into for topic modelling is too wide and thus the topics covered are particularly sparse. Note that the perplexities of the LDA models created through the subset of all articles in Figure 18 are all much greater than the perplexities of the 2015 articles dataset displayed in Figure 12 . It is likely that the large perplexity of these LDA models indicate that the topics identified would be less concise.

# Conclusion

This paper set out to employ a method of assessing public opinion towards a public figure through the use of sentiment analysis and targeted topic modelling using Latent Dirichlet Allocation. The first task of the project was to observe how public opinion towards Buhari fluctuated over the year 2015 and to understand the underlying events that caused the observed fluctuations. In this task, the project was largely successful. The sentiment analysis procedure produced results which - after having been smoothed using LOESS regression - displayed an observable trend with a particularly interesting spike in positive sentiment around May 2015. I hypothesized that the peak would be related to Buhari’s inauguration, which took place on the 29th of May 2015, because it was the most significant event relating to the president to occur during that period. My hypothesis was supported by the trend in emotions that I observed; in particular, the ‘anticipation’ present in news articles was highly positively correlated with the May peak and itself peaked around the date of the inauguration. The topic modelling procedure was also relatively successful as I was able to identify a topic in the LDA model which featured discourse around the president’s inauguration

The second task of the paper was to gain insight into how public opinion of Muhammadu Buhari fluctuated over the course of his presidency. Sentiment analysis of the data was informative and greatly supported my initial claim that sentiments towards the president became more negative over time. Additionally, there was a significant trough in emotional valence in mid-2016 which I believed to have been caused by a surge in negative press resulting from the country entering recession around that time. It is generally accepted that the Buhari administration mismanaged the country’s economic situation as it entered recession and exacerbated the issue. Thus, it was not surprising to identify a dip in sentiment around that time. The topic modelling procedure, however, was largely unsuccessful in proving my hypothesis that the trough was caused by Nigeria’s recession.

Overall, my research supported prevailing theory that natural language processing techniques offer a sufficient means of drawing information from text. However, the topic modelling procedure has its limitations and there are steps I could take to improve the LDA performance. Most notably, the time subset on which I carried out the LDA procedure was particularly large and thus may have covered too wide a range of events to have been concisely summarized into 7 topics. Future analysis may benefit from creating an LDA model with a larger number of topics, or from running the analysis using a subset of data from a smaller timeframe.

The paper concludes that while the methods of assessing public opinion used in this project have their limitations, they were succesful in deriving useful information on the subject. Although further refinement of the techniques is neccessary, more thorough employment of sentiment analysis and topic modelling may become and effective proxy for opinion polls in countries that do not provide such information to the public, yet have large written media coverage on elections and on public officials.

# References

Auguie, B. (2017). *GridExtra: Miscellaneous functions for "grid" graphics.* Retrieved from https://CRAN.R-project.org/package=gridExtra

Bkkbrad. (2008, February). Latent dirichlet allocation diagram in plate notation.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022. Retrieved from http://dl.acm.org/citation.cfm?id=944919.944937

Feinerer, I., Hornik, K., & Meyer, D. (2008). Tm: Text mining infrastructure in r. *Journal of Statistical Software*, *25*(5), 1–54. Retrieved from http://www.jstatsoft.org/v25/i05/

Fellows, I. (2018). *Wordcloud: Word clouds.* Retrieved from https://CRAN.R-project.org/package=wordcloud

Fitzgerald, J. D. (2018, January). Sentiment analysis of (you guessed it!) donald trump's tweets. *Storybench.* Storybench. Retrieved from http://www.storybench.org/sentiment-analysis-of-you-guessed-it-donald-trumps-tweets/

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. Retrieved from http://www.jstatsoft.org/v40/i03/

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30. http://doi.org/10.18637/jss.v040.i13

Hörster, E., Lienhart, R., & Slaney, M. (2007). Image retrieval on large-scale image databases. In *Proceedings of the 6th acm international conference on image and video retrieval* (pp. 17–24). New York, NY, USA: ACM. http://doi.org/10.1145/1282280.1282283

Jockers, M. L. (2015). *Syuzhet: Extract sentiment and plot arcs from text.* Retrieved from https://github.com/mjockers/syuzhet

Knispelis, A. (2016). LDA topic models. Retrieved from https://www.youtube.com/

watch?v=3mHy4OSyRf0

Miratrix, L. (2018). *Textreg: N-gram text regression, aka concise comparative summarization*. Retrieved from https://CRAN.R-project.org/package=textreg

Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the naacl hlt 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 26–34). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1860631.1860635

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon, *29*(3), 436–465.

Silge, J., & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, *1*(3). http://doi.org/10.21105/joss.00037

Sklar, M. (2014). Tutorial on dirichlet distribution by max sklar. Retrieved from https://www.youtube.com/watch?v=6k7IzONQOzM

Sullivan, S. (2017). LDA algorithm description. Retrieved from https://www.youtube.com/watch?v=DWJYZq_fQ2A

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H., François, R., Henry, L., & Müller, K. (2018). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Zhu, H. (2018). *KableExtra: Construct complex table with 'kable' and pipe syntax*. Retrieved from https://CRAN.R-project.org/package=kableExtra