 doi:10.20944/preprints202210.0238.v1

From Twitter to Aso-Rock: A Natural Language Processing Spotlight for Understanding Nigeria 2023 Presidential Election

[1,2]Olusola OLABANJO, [3]Ashiribo WUSU, [2]Mauton ASOKERE, [4]Rebecca PADONU, [5]Olufemi OLABANJO, [6]Oluwafolake OJO, [7]Oseni AFISI, [6]Olusegun FOLORUNSHO, [8]Manuel MAZZARA, [2]Benjamin ARIBISALA

[1]Department of Mathematics, Morgan State University, USA

[2]Department of Computer Science, Lagos State University, Ojo, Lagos Nigeria

[3]Department of Mathematics, Lagos State University, Ojo, Lagos Nigeria

[4]Department of Theatre Arts, Ahmadu Bello University, Zaria, Nigeria

[5]Department of Computer Science, Oduduwa University, Ile-Ife, Nigeria

[6]Department of Computer Science, Federal University of Agriculture, Abeokuta, Nigeria

[7]Department of Philosophy, Lagos State University, Ojo, Lagos Nigeria

[8]Institute of Software Development and Engineering, Innopolis University, Innopolis, Russia

Abstract

Introduction: Social media platforms such as Facebook, LinkedIn, Twitter, among others have been used as a tool for staging protests, opinion polls, campaign strategy, medium of agitation and a place of interest expression especially during elections. Past studies have established people's opinion elections using social media posts. The advent of state-of-the-art algorithms for unstructured text processing implies tremendous progress in natural language processing and understanding.

Aim: In this work, a Natural Language framework is designed to understand Nigeria 2023 presidential election based on public opinion using Twitter dataset.

Methods: Raw datasets concerning discourse around Nigeria 2023 elections from Twitter of 2,059,113 χ 18 dimensions were collected. Sentiment analysis was performed on the preprocessed dataset using three different machine learning models namely: Long Short-Term Memory (LSTM) Recurrent Neural Network, Bidirectional Encoder Representations from Transformers (BERT) and Linear Support Vector Classifier (LSVC) models. Personal tweet analysis of the three candidates provided insight on their campaign strategies and personalities while public tweet analysis established the public's opinion about them. The performance of the models was also compared using accuracy, recall, false positive rate, precision and F-measure.

Results: LSTM model gave an accuracy, precision, recall, AUC and f-measure of 88%, 82.7%, 87.2% , 87.6% and 82.9% respectively; the BERT model gave an accuracy, precision, recall, AUC and f-measure of 94%, 88.5%, 92.5%, 94.7% and 91.7% respectively while the LSVC model gave an accuracy, precision, recall, AUC and f-measure of 73%, 81.4%, 76.4%, 81.2% and 79.2% respectively.

Conclusion: The experimental results show that sentiment analysis and other Natural Language Processing tasks can aid in the understanding of the social media space. Results also revealed the leverage of each aspirant towards winning the election. We conclude that sentiment analysis can form a general basis for generating insights for election and modeling election outcomes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0238.v1

I. Introduction

Forecasting and analysis of election results have gained a wide popularity in the field of political methodology[1], a subfield of political science and political analysis concerned with the study quantitative and qualitative approaches used to understand politics and political systems[2]. These methods often combine statistics, mathematics and formal theory for the understanding of politics. Election results have usually been predicted in the past using analytical and statistical techniques where the methodology has relies on surveys and qualitative methods, and analyzing political party manifestos while observing the trends in the mainstream media[3, 4]. With an increased contest and opposition in the government, particularly in nations with multi-party system like Nigeria[5, 6], it has become more difficult but interesting to predict elections.

The power of social media and how greatly it has been established to affect elections and electoral systems motivated this study. The influence of social media on elections became noticeable first in the early year 2000.[6]. In his first presidential campaign, Barack Obama made use of social media to mobilize the public and win the 2008 election. Statistics revealed that 74% of internet users—or 55% of the adult population—looked for election news online during Obama's first campaign. Another famous example is when Beto O'Rourke came dangerously close to unseating Senator Ted Cruz in 2018. The 2018 Texas Senate race broke the record for the most money spent in a U.S. Senate election, according to the Center for Responsive Politics, spending $93 million, a large portion of which was raised and used for social media events and advertisements. The social media is, no doubt, a key tool to understanding and analysis of public political opinion and prediction of election outcomes.

Before and even after the social media revolution, surveys have been used frequently to identify opinions prior to elections; however, this approach has been limited by the difficulty in constructing an appropriate sampling procedure, hence making it difficult to obtain a representative sample of political viewpoints. Social media platforms such as Twitter, Facebook, Reddit and Instagram have helped in some way to overcome the "systematically appropriate" sampling procedure in survey studies and have been the prominent tools of political campaigns and activisms during elections. The recent advancements in the fields of Natural Language Understanding (NLU) and Natural Language Processing (NLP) have improved the reliability of prediction models built for unstructured and unsupervised dataset. The social media has radically upended the traditional campaign norms and tactics in national elections vis-à-vis its volume, scope and tactics of use[7, 8]. Studies also showed that even if it cannot be categorically said that social media singlehandedly elected Donald Trump, his social media campaign strategies changed the way social media will be used in elections in the future[8, 9].

There has been a growing interest in the use of NLP and other artificial intelligence techniques to predict election results in recent years with social media datasets. These techniques include sentiment analysis and topic modeling in addition to more advanced models that incorporate deep learning and fundamental statistical techniques. Two important aspects of NLP are sentiment analysis and opinion mining, two processes which aid in classifying and investigating the behavior and approach of social media users with regards to brands, events, companies, customer services as well as elections. They are algorithms for automating the process of extracting emotions from user's posts by processing unstructured texts and preparing model that will extract knowledge from it. In this study, we consider Twitter; one of the prominent social network sites, to analyze the upcoming 2023 Nigeria election. Given the statistics, Twitter has an active monthly user of 316 million and, on the average, about 500 million tweets are posted daily. A review of related works in subsequent section shows that Twitter is one of the most powerful tools in political analysis. In the current fast-paced global network system, information spreads digitally between users and shapes the way these users feel about an event, thereby making it crucial to understand the thought polarity, emotions and sentiment of the masses.

This study has been conducted to understand the national and international opinions on Nigeria 2023 presidential election using Twitter dataset. It aims to capture, process and evaluate the public opinion from three main perspectives: sentiment and timeline analysis of the personal accounts of contesting candidates; sentiment and general tweet analysis of the public on the three candidates and sentiment and general tweet analysis of the public on the Nigeria 2023 elections. The study would therefore concentrate on the following:

i. Identifying keywords, hashtags and accounts which wholistically explains the subject matter as well as captures the top three contestants for Nigeria 2023 presidential elections.
ii. Scraping the tweets through Twitter API using Python programming.
iii. Preprocessing the tweets by data cleaning (removing white spaces, links, punctuations, stop words, tokenization, retweet).
iv. Developing the three machine learning models (LSTM, BERT and LSVC) and training with existing annotated IMDB dataset
v. Using Natural Language Understanding techniques such as topic modeling, entity extraction, word frequency, word cloud etc. to understand the personal profiles of the three candidates.
vi. Analyzing the result to predict the direction of the election to be conducted in 2023.

The study is further structured into sections: the first section discusses briefly the political and electoral system in Nigeria since democracy became the rule of law; the next discusses existing works related to our study; the next section which covers methodology explains in detail the technique, models construction and evaluation as well as other analyses done in this to achieve our main aim and objectives, the next section discusses the results obtained in the various experimental setups while, finally we discuss the results and concludes with our findings and direction for future work.

II. The Political and Electoral Systems in Nigeria

Nigeria is a Federation consisting of 774 Local Governments, 36 States and a Federal Capital Territory in ascending order of hierarchy and each headed by Chairman, Governors and President-appointed minister respectively. Subject to the provisions of the 1999 Constitution, the executive power of the Federation is vested in the President and the executive rule of law vested in a State under the Constitution shall be so exercised as not to impede or prejudice the exercise of the executive powers of the Federation. The system of local government by democratically elected local government councils headed by the Chairman is under this Constitution guaranteed[10]. Contextually, the current Constitution (1999) governing the Nigerian Political and Electoral systems provide for separation of power among the three branches of government. The 1999 constitution which was amended in January 2011 clearly stipulates that the President is eligible for two four-year terms alongside the bicameral National Assembly 109-member Senate and 360-member House of Representatives[11]. The President appoints members of the supreme court (Judiciary) subject to senate approval, while the citizens exercise their sovereignty through voting at the polls based on the universal suffrage at age 18, administered by the Independent National Electoral Commission (INEC) as stipulated via the Electoral Act of 2006. Based on this Federalism, citizens are eligible to vote in leaders at both National and Sub-national levels[12].

Democracy in the developed and developing countries today was given birth to as a result of embracing party systems in politics and Nigeria has chosen democracy as her political administrative principle[13]. Party systems remain an integral part of the political system and most developing democracies like Nigeria will be shaped by the complexities of their party systems, whereby citizens have the right to partake in public affairs, the right to vote and be elected. Nigeria's elections have in the past witnessed acts such as tarnishing the image of oppositions, intimidating the citizens, engaging in political violence and buying the votes of eligible citizens through deliberate distribution of palliatives to the poorest of the poor and most vulnerable populace.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0238.v1

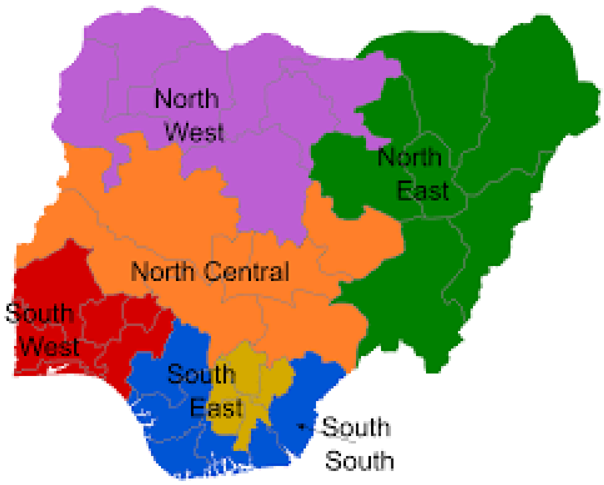

Figure 1: Nigeria's Six Geopolitical Zones on the Map (adopted from https://en.wikipedia.org/wiki/Geopolitical_zones_of_Nigeria).

There are six geopolitical divisions (as in Figure 1) within the Federal Republic of Nigeria, which are frequently referred to as zones. They are a form of administrative division that President General Sani Abacha's administration used to group the nation's states. Resources in Nigeria's economy, politics, and education were frequently divided within these zones. The six zones were not wholly divided based on geography; rather, states with comparable ethnic populations and/or shared political histories were grouped together. There are about 400 ethnic groups and 450 languages in Nigeria. For efficient resource distribution, the government needed to combine comparable groups, hence the need for the zoning[14-16]. The population density of these geopolitical zones as at the last National Census is given in Table 1.

Table 1: Geopolitical Zones, their States, Ethnicity and Total Population(from https://infomediang.com/largest-geopolitical-zone-by-population/)

Zone	States	Ethnicity	Population
North Central	Benue, Kogi, Kwara, Nasarawa, Niger, and Plateau States, as well as the Federal Capital Territory.	Hausa & Fulani	29,252,408
North East	Adamawa, Bauchi, Borno, Gombe, Taraba, and Yobe States.		26,263,866
North West	Jigawa, Kaduna, Kano, Katsina, Kebbi, Sokoto, and Zamfara States.		48,942,307
South East	Abia, Anambra, Ebonyi, Enugu, and Imo States.	Igbo and Niger Deltans	21,955,414
South South	Akwa Ibom, Bayelsa, Cross River, Delta, Edo, and Rivers States.		28,829,288
South West	Ekiti, Lagos, Ogun, Ondo, Osun, and Oyo States.	Yoruba	38,257,260

The dominance of the North in Nigeria's election cannot be overlooked as shown in Figure 2. The Northern Nigeria constitutes an estimate of 60% of the entire Nigerian population and this is noteworthy for the electorates in mapping out their strategies in terms of pre-election campaigns and in their choice of Vice-Presidential candidates.

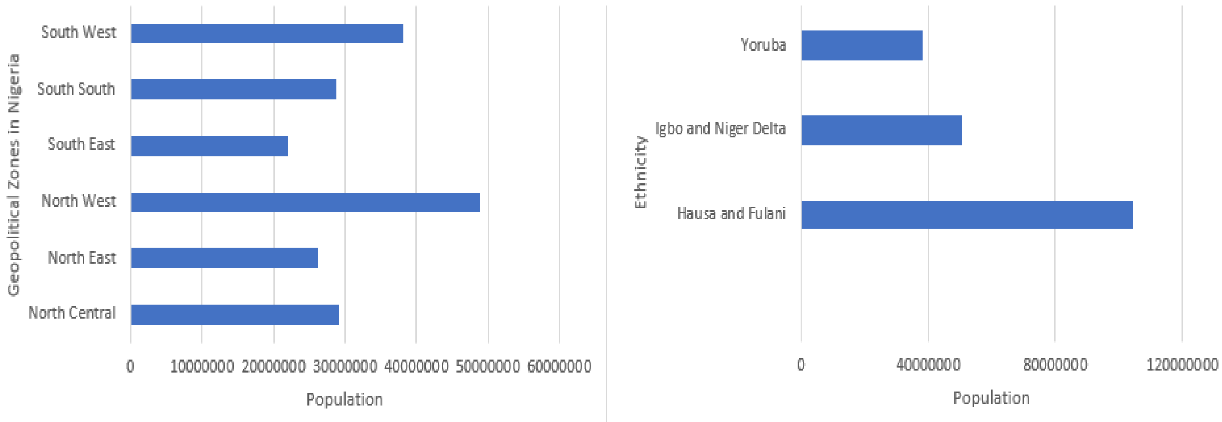

Figure 2: Population Distribution of Nigeria by Zone (Left: Grouped by Geopolitical Zones. Right: Grouped by Ethnicity)

It is noteworthy, however, that in the Nigeria's constitution, all incumbents have a four-year term and can be re-elected only once. The Presidency race since 1999 has been a two-party show, although there were several other parties in the contest. Table 1 shows an overview of the race for Presidency since the adoption of democracy in Nigeria. Nigeria held its first presidential elections on February 27, 1999; these were the first democratic elections since the 1993 military coup. The election results have since been bipolarized as the larger percentage of the total election results. Table 2 and Figure 3 show how the race to Aso-Rock (the White House of Nigeria) since democracy has been dominated by two parties even though there are about 30 contesting parties.

Table 2: The Presidency Election Statistics since 1999 as published online by Independent National Electoral Commission – INEC https://www.inecnigeria.org

Year of Ruling	Winner Count	First Runner Up Count	Second Runner Up Count	Candidate (Party)
1999 – 2003	18,738,154	11,110,287	-	Olusegun Obasanjo (People's Democratic Party – PDP)
2003 – 2007	24,456,140	12,710,022	1,297,445	Olusegun Obasanjo (People's Democratic Party – PDP)
2007 – 2011	24,638,063	6,605,299	2,637,848	Umaru Yar'Adua (People's Democratic Party – PDP)
2011 – 2015	22,616,416	12,250,853	2,079,151	Muhammadu Buhari (All Progressives Congress – APC)
2015 - 2019	15,424,921	12,853,162	53,537	Muhammadu Buhari (All Progressives Congress – APC)

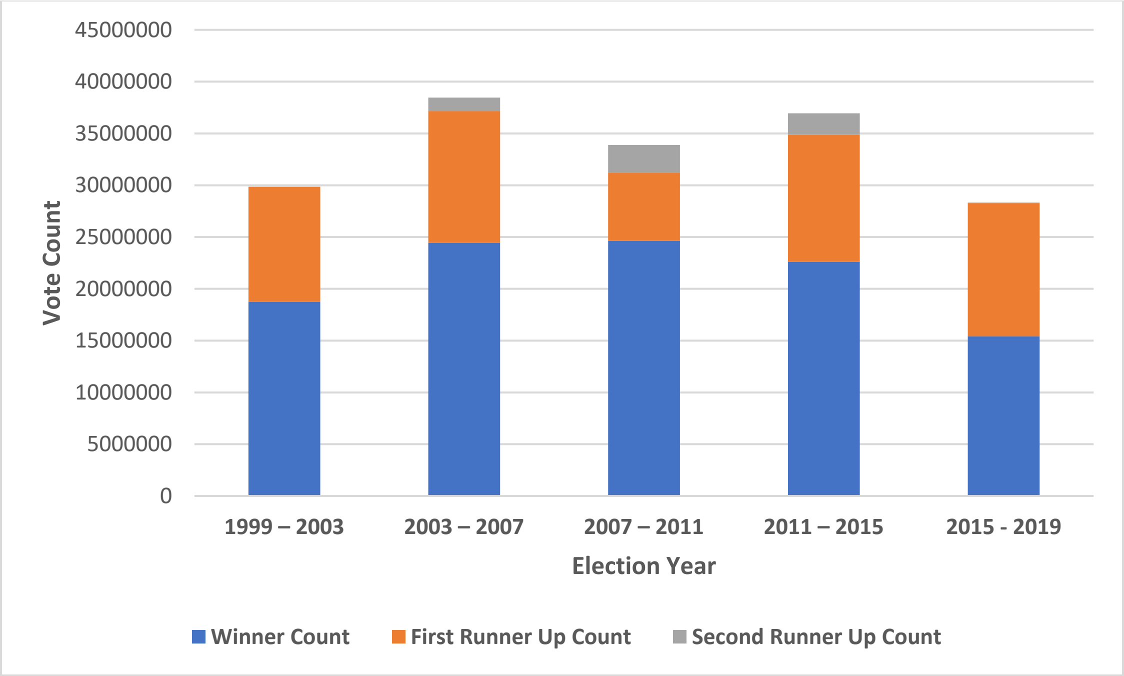

Figure 3: Electoral Vote Count of Nigeria's Presidential Election since 1999.

Before now, Nigerians pay more attention to political parties rather than the candidates. Interestingly, at the time of this study, there seems to be a change of mindset, where a larger youthful population continuously show interest in the presidential candidate from a minor political party (Labour Party), putting the candidate above the political party, which somewhat led to an emergence of another party and has, inadvertently, placed the 2023 presidential elections on an unprecedented pane. There are three candidates at the campaign forefront of the forthcoming 2023 Nigeria's Presidential Election: Atiku Abubakar of the Peoples' Democratic Party (PDP) from the Northern region, Bola Tinubu of All Progressives' Congress (APC) from the Southern region and Peter Obi of the Labour Party (LP) from the Easter region.

III. Related Works

Sentiment analysis has been used to predict the opinions of the citizens on US election using Twitter data [17]. The authors used 17,000 tweets to train their model (Naïve Bayes) and the model achieved a less than 60% prediction accuracy by classifying the tweets into positive, negative, neutral and not-sure and this assisted them in analyzing real time tweets from the people which gave great insights about public opinions on each candidate. Another study[18] used Twitter as a source of data to analyze tweets to get international opinions concerning a protest that happened in India conducted by the farmers and about 20,000 tweets was scraped from Twitter which was analyzed and categorized into positive, negative and neutral. Bag of words and TF-IDF was used to conduct the analysis and Bag of words outperformed TF-IDF. Other authors[19, 20] also established that Twitter can be used as an election indicator, which has the ability to predict the favorite candidate of the people which intend to emerge as the winner before the election is conducted. The people gave positive opinions about Donald Trump in almost the whole states which gave a high signal of emerging the winner of the election which he won. 1,000,000 tweets were collected from various users from different states and sentiment analysis was conducted. A study[21] used Twitter data to check and conduct how different countries affected by the Corona virus coped with the situations, in which the tweets posted in English were analyzed to give the opinion and emotion of the people concerning the pandemic in their respective countries and 50,000 tweets were used in the study.

Studies combined different features and used ensemble models to increase the detection of polarity of tweets for sentiment analysis and established that unsupervised and ensemble machine learning models performed better in

 doi:10.20944/preprints202210.0238.v1

detecting opinions from text[22, 23]. Authors have used Twitter data to examine political homophily in American Presidential Elections in 2016 where 4.6 million tweets were collected for analysis[24].

In some studies, authors used the ratio of positive message rate and negative message rate to predict the likely winner of the forth coming election using US presidential election using twitter data and it shows that these opinions could be used to predict candidate that will emerge as the winner[25]. Twitter data has been used to establish that social media is not only used to express opinions, but it is also used to share ideas and opinions among other users. Authors used 100,000 tweets to predict German federal election in 2009 which could serve as a political indicator for the election[26] and another study [27] also predicted German presidential election 2021 using 58,000 tweets which they established that traditional machine learning methods like Naive Bayes did worse than transformer-based models like the bidirectional encoder from transformers (BERT).

Authors[28] performed social media sentiments about political parties to study and forecast Pakistan's general election. They used supervised machine learning algorithms to classify tweets into positive, negative, and neutral. The findings of their experiment shows that social media content can be a useful indication for identifying political behavior of various parties. In another study[29], 90154 tweets was analyzed and the results were compared with the actual election results, their model predicted the party that took first and fourth accurately.

Our overall goal in this study is to determine the sentiments in tweets with mentions of election-related words for the upcoming Nigeria 2023 elections and find if these tweets can provide meaningful insights regarding the election outcome. Our strengths over existing works in reviewed literature include the size of dataset used, further NLU tasks employed and the timeliness of the study. We used 2 million datasets in our study which is unprecedented. Also, this study has not been conducted in Nigeria to predict elections. Finally, and most significantly, our studies pre-date the actual Presidential election, hence the timeliness of this study.

IV. Methodology

This study seeks to analyze tweets of the political discourse around the upcoming 2023 Nigeria presidential election and its prominent contesting candidates and this section describes in detail the approach used in discovering the sentiments of people around this area of public interest. Figure 4 depicts the framework used in the tweet analyses done in this study.

Every tweet with a political content either contains a neutral, positive or negative sentiment for or against a party or candidate. The sentiments contained in tweets especially when it is specific to a candidate is not easy to compute with algorithms since emotion expression varies with the personality, region and cultural background of each person. Since this is an unstructured (unlabelled) dataset, sentiment analysis is often challenging because of the features, context and semantics peculiarity of each tweet.

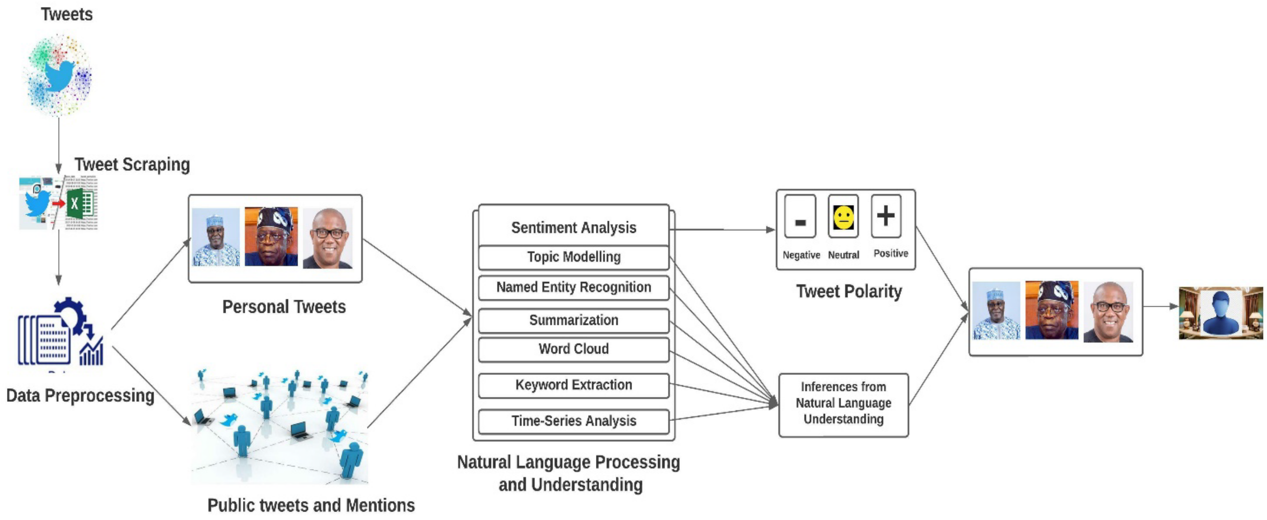

Figure 4: Framework for twitter-based NLP analysis for Nigeria 2023 elections.

This process started with identification of tweets, trends, keywords and hashtags which wholistically represents the discourse around Nigeria 2023 elections. This is used in a logical structure to scrape tweets from Twitter. The tweets are grouped into the personal tweets of each candidate and the general tweets of the public about the election. The preprocessed dataset was passed into our NLP pipeline to detect sentiments in the posts which is in turn used to perform analyses and provide insights concerning the election.

i. Dataset

A total of 2,059,113 raw tweets was extracted[30] over a period of four months (June to September 2022). This was collected using tweepy[31], an open-source python library to access the Twitter API using the provided authorization and access tokens. The chosen start date is justified by the fact that the declaration of intent to contest for Presidency began fully in June 2022. The exact search terms used for scraping was carefully constructed to capture tweets about each of the three candidates as well as the general view on the upcoming elections. The dataset contained 360387 unique users from 27 uniquely identifiable countries and the attributes are described in Table 3. Only tweets in English were considered for this study.

Table 3: Feature Description of Dataset

Feature	Description
TweetId	Unique identifier for each tweet
TweetDate	Date tweet was posted
Username	Twitter handle of the poster
Tweet	The twitter post
UserLocation	Coordinates (if available) of user
UserVerifiedStatus	True if the post is from a Twitter-verified account and False otherwise
UserFollowerCount	Number of followers the user has
FollowingCount	Number of accounts the user follows
AccountCreationTime	The date the user joined Twitter
TweetLocation	Location name of where the tweet was posted
ReplyCount	Number of replies of the tweet
FavoriteCount	Number of likes of tweet
Retweet	Number of retweets

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0238.v1

ii. Data Preprocessing

To fully prepare the scraped dataset for computation and to reduce bias in the sentiment models, the dataset was preprocessed to remove noisy, inconsistent and irregular patterns[32]. The tweets were originally unstructured and unlabelled, full of noise and unwanted texts and emojis. In this study, the preprocessing was done in a precise and systematic manner as given in Table 4. Our data cleaning approach also included duplicates removal, @-mentions removal, RT which implies retweet and hashtag removal, since all of these essentially do not affect the sentiment of the actual tweet.

Table 4: Algorithm for Preprocessing our Twitter Dataset

Input: Twitter comments or text data
Output: Pre-processed text data

For each comment in the Twitter data file

Initialize temporary empty string processedTweet to store result of output.

1. Replace all URLs or https://links with the word 'URL' using regular expression methods and store the result in processedTweet.
2. Replace all '@username' with the word 'AT_USER' and store the result in processedTweet.
3. Filter All #Hashtags and RT from the comment and store the result in processedTweet.
4. Look for repetitions of two or more characters and replace with the character itself. Store result in processedTweet.
5. Filter all additional special characters (: n | [ ] ; : {} – + ( ) < > ? ! @ # % *,) from the comment. Store result in processedTweet.
6. Remove the word 'URL' which was replaced in step 1 and store the result in processedTweet.
7. Remove the word 'AT_USER' which was replaced in step 1 and store the result in processedTweet.

Return processedTweet.

To fully achieve this algorithm, we leveraged on Python's string manipulation, pattern and regular expressions capabilities such that, as a result, the filtered data can achieve high performance in analyzing data when the machine learning algorithm is trained. With the knowledge that there are unrelated tweets with just anchor tags of the election trend, the data was further checked for the initial keywords of interest and unmatched tweets were eliminated. The final dataset is then separated into personal tweets of the candidates and public tweets. The processedTweet was then tokenized. Tokenization is a term which describes separating a corpus into smallest, syntactically meaningful, units. It is a fundamental step in modeling text data which aids in understanding the meaning behind the text by analyzing the sequence of the words. Porter stemmer was used to reduce the inflection towards their root forms. This was done by stripping the suffix to produce stems[33]. This is then passed into the NLP pipeline for processing.

iii. Natural Language Processing (NLP) and Natural Language Understanding (NLU) Tasks

Natural Language Processing (NLP) and Natural Language Understanding (NLU) describe a comprehensive set of standard tasks and techniques for automatic generation, manipulation and analysis of natural human languages[34]. These enable us to process a huge amount of unstructured corpus through sentence- and token-based analysis as well sentence- and document-level polarity via state-of-art linguistic and lexical processing tools such as WordNet, SentiWordNet and Treebanks[35-37]. The NLP techniques involved in this study for tweet analysis includes processes such as word tokenization, word stemming and lemmatization, topical modeling, named-entity recognition, summarization, word cloud, sentiment analysis and keyword extraction. NLP tasks use both linguistics and mathematics to connect the language of humans with the language of computers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0238.v1

a. *Word Tokenization:* The raw tweets after preprocessing and cleaning is broken down into smallest recognizable words and punctuations known tokens[38], the goal of which is generate the list of words which eventually is used for word cloud, summarization and sentiment analysis. The accuracy of this task is often influenced by the training vocabulary, unknown words and our-of-vocabulary (OOV) words.

b. *Word Stemming and Lemmatization:* This transforms our tokens to its base – dictionary – form by filtering the affixation or by changing a vowel from the word[39]. Stemming and lemmatization aim to reduce the inflectional forms of a word and occasionally related derivational forms to a root form.

c. *Topic Modeling:* Topic modeling is a technique for unsupervised categorization of the twitter documents which helps to identify natural groups of words even when we are not certain what the outcome will be. This gives us a general understanding of the discourse around our corpus. In this study, we used the Latent Dirichlet Allocation (LDA), a particularly popular algorithm for achieving this task[40, 41]. This is a mathematical method for finding the mixture of words associated with each topic while also determining the mixture of topics that describes each document. LDA is a generative probabilistic model of a corpus which follows a generative process for our twitter document,τ as given in Table 5.

Table 5: Algorithm for LDA Modeling of our Twitter Dataset

Input: processedTweet **Output:** n-Gram Combinations
1. Choose $\boldsymbol{N} \sim \boldsymbol{Poisson}(\boldsymbol{\xi})$. 2. Choose $\boldsymbol{\theta} \sim \boldsymbol{Dir}(\boldsymbol{\alpha})$. 3. For each of the $\boldsymbol{N}$ words $\boldsymbol{w_n}$: (a) Choose a topic $\boldsymbol{z_n} \sim \boldsymbol{Multinomial}(\boldsymbol{\theta})$. (b) Choose a word $\boldsymbol{w_n}$ from $\boldsymbol{p}(\boldsymbol{w_n} \mid \boldsymbol{z_n}, \boldsymbol{\beta})$, a multinomial conditioned on the topic $\boldsymbol{z_n}$.

A $n-$dimensional Dirichlet random variable θ takes values in the $(k-1)$-simplex, that is, a k-vector θ lies in the $k-1$ simplex $iff\ \theta_i \geq 0, \Sigma_{\{i=1\}}^{k}\theta_i = 1$ and has the probability density function on this simplex as given in Equation (1)

$$p(\theta|\alpha) = \frac{\Gamma(\Sigma_{\{i=1\}}^{k}\alpha_i)}{\Pi_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\{\alpha_1-1\}} \ldots \theta_k^{\{\alpha_k-1\}} \quad \ldots\ldots(1)$$

where α is a $k-$vector with component $\alpha_i > 0$, $\Gamma(x)$ is the Gamma function.

d. *Word Cloud:* We used WordCloud, also called TagCloud to visually represent our Twitter corpus. Tags are tokens, the importance of which is represented with font size or color as a depiction of word significance and word co-occurences. The size of each word in our WordCloud is given in Equation (2).

$$s_i = \left\lceil \frac{f_{max} \cdot (t_i - t_{min})}{t_{max} - t_{min}} \right\rceil \forall\, t_i > t_{min}\ else\ s_i = 1 \ldots\ldots(2)$$

where s_i is display font size

f_{max} is the maximum font size

t_i is the count

t_{min} is the minimum count

t_{max} is the maximum count

Words in our WordCloud appear bigger the more often they are mentioned and are great for visualizing unstructured text data and getting insights on trends and patterns[42, 43].

e. *Sentiment Analysis:* We used sentiment analysis as a core component of this study. Sentiment analysis, often regarded as opinion mining, is a natural language processing (NLP) method for identifying the positivity, negativity, or neutrality of data. We used sentiment analysis on our processed textual data to track the perception and reviews of twitter users as regards each of the three prominent contestants of the Presidential election and to summarize the view of the masses in general. A tokenized tweet is sent to our prediction models for processing and its performance was measured. LSTM, a complex area of deep learning, belongs to a class of Recurrent Neural Network (RNN) which can learn order dependence in sequence prediction class of problems[44-46]. This

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0238.v1

behavior is essential for solving complicated problems in areas like speech recognition and machine translation, among others.

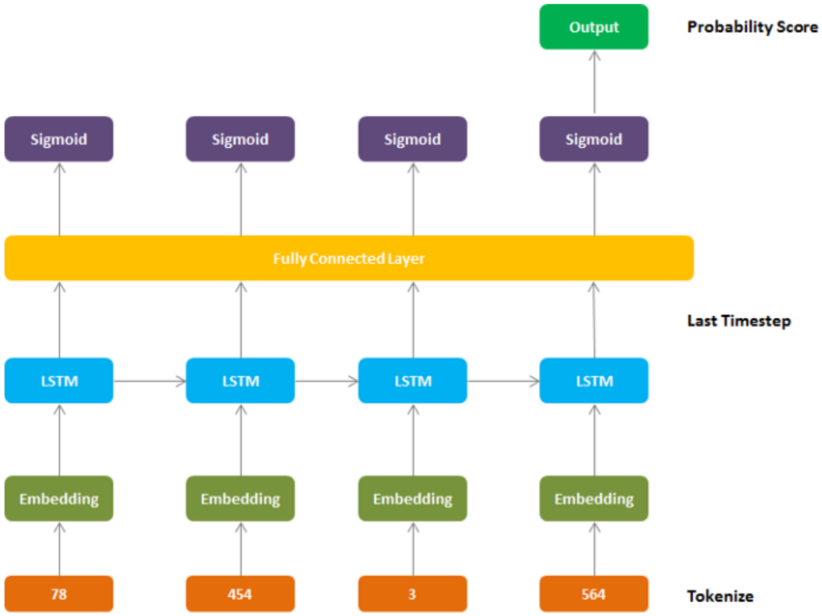

Figure 5: LSTM Architecture for Sentiment Analysis

The architecture of our proposed model is given in Figure 5. The tokenize layer has been achieved outside the LSTM model. The LSTM layer is defined by the hidden state dimensions and number of layers. The full connected layer maps the output of the LSTM layer to a desired output size. The sigmoid layer is the activation layer which turns all output values into the map from a closed and bounded interval [-1,1] while the output, the final, layer gives the probability score.

The Bidirectional Encoder Representations from Transformers (BERT) model is a Google AI Language masterpiece which has been widely applied in a wide range of NLP tasks[47, 48]. The main strength of the BERT model is its application of bidirectional training of Transformer which is a popular attention model to language modeling. Earlier similar models looked at text sequences from either a left-to-right or a combined left-to-right and right-to-left training perspective. BERT models have demonstrated that bidirectionally trained language models can comprehend context and flow of language more deeply than single-direction language models[49]. Classification tasks such as sentiment analysis are done similarly to Next Sentence classification, by adding a classification layer on top of the Transformer output for each token. The Linear Support Vector Classifier was used to test the accuracy of our deep models. SVM has been one of the most robust prediction techniques which is based on a statistical learning framework[50]. Its workings and model development strategies are expounded in some existing works[51-53].

The sentiment models were assessed using accuracy, area under curve (AUC), recall, false positive rate (FPR), precision and F-measure as seen in Equations 3 - 7. The area under the ROC curve (AUC) is a plot of true positive rate (TPR, or specificity) against false positive rate (FPR, or sensitivity). The true positive rate (TPR), sensitivity, or recall is the number of sentiment labels predicted positive that are actually positive. Recall is defined by Equation (4). The false positive rate (FPR) is the number of sentiment labels predicted positive that are actually negative and is defined in Equation (5). Precision is the proportion of the predicted positive cases that were correct. The precision can be calculated using Equation (6). The F-Measure score is the harmonic mean of the precision and recall. This evaluates the equivalency between the sensitivity (recall) and the precision (correctness) of the data. This gives us the interpretation of how the measure recall and precision values behave for the dataset. The F-measure can be calculated using Equation (7).

Let TP_+ = true positive for all positive tweets, TP_- = true positive for all negative tweets, TP_ϕ = true positive for all neutral tweets. Then:

$$Accuracy = \frac{TP_+ + TP_- + TP_\phi}{\Sigma all\ instances} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

$$TPR, Sensitivity, Recall = \frac{\Sigma TP_+}{\Sigma \text{All } Actual\ Positives} \ldots\ldots\ldots\ldots (4)$$

$$FPR = \frac{\Sigma TP_-}{\Sigma \text{All } Actual\ Negatives} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

$$Precision = \frac{TP}{TP + FP} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

V. Experimental Setup

In this study, we were interested in what goes on in the Twitter space as regards the upcoming Nigeria's 2023 elections. We sought to achieve this via tweet scraping concerning the prominent aspirants, the election in general as well as through the analysis of the personal tweets of the aspirants. For these individual datasets, we analyzed their tweet sentiments, tweet frequency, tweet popularity, hashtag and wordtag outputs as well as their social media opinion polarity. This helped us to gain understanding of the relative social media power of each candidate, their strategies, strengths, weaknesses, opportunities and threats of each of the three candidates. We also performed social network analysis of the social media circles kept by them. Sentiment analysis gave the perspective of each tweet, WordCloud gave a summarization of entities concerning them while social network analysis shows the connections maintained by each of the candidates. Personal and public tweet analyses give understanding of trends, patterns, keywords and sentiments of each candidate as well as the masses respectively. In building and validating our sentiment model, we used the IMDB dataset which contained 50000 movie reviews and has widely been used for sentiment analysis tasks[54]. We passed the preprocessed IMDB dataset into our LSTM, BERT and LSVC models and measured their accuracies. We then passed our preprocessed Twitter dataset into the pretrained models. The performance of our models on the IMDB dataset is our confidence in the sentiments returned for each tweet in this study.

VI. Results

a. Dataset

Our Twitter dataset was extracted from June 1 to September 30, 2022 containing tweets related to the candidacies of Tinubu, Atiku and Obi and generally about the upcoming 2023 Nigeria Presidential elections. There is a global interest in the upcoming Nigeria 2023 elections from many prominent countries of the world including Nigeria, United States of America, United Kingdom, Canada, Germany and Dubai. Table 6 shows the statistics of the countries with the highest tweets. The location of majority of users are marked as Not Available (NA).

Table 6: Tweets By Countries

Countries	*Number of Tweets*
Nigeria	873,111
USA	45,918
UK	30,042
Canada	17,898
Others	230,229
NA	861,915
Total	2,059,113

Figure 6 is a representation of the change in the size of dataset during each preprocessing task. The raw dataset was reduced from 804MB to 603MB after the preprocessing tasks were completed.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0238.v1

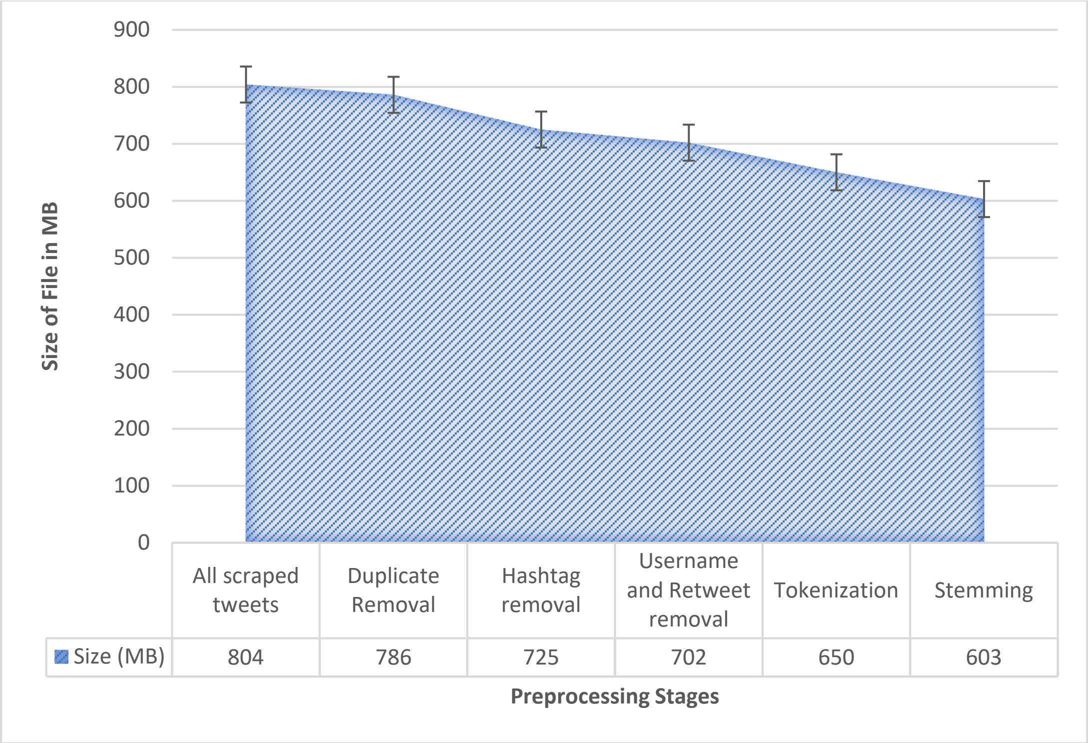

Figure 6: File size variation at each preprocessing task

b. Personal Tweet Analysis

In this section, we examined the personal Twitter accounts (@atiku, @officialABAT and @PeterObi) of each of the candidates. The first dataset contained the personal tweets of the candidates and Table 7 shows their tweet summarization. Although the date each one joined Twitter varies, Figure 7 shows their tweet frequency and impressions made by the personal tweets of the three candidates. It shows that Atiku joined tweeter the earliest but highest tweets during the time covered by this study was from Peter Obi. Peter Obi came first in total number of retweets, mentions and favorited tweets. Atiku, however, has the highest number of followers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0238.v1

Table 7: Personal Tweet Summaries of the Presidential Candidates

	Atiku Abubakar	***Peter Obi***	***Tinubu***
Handles	@atiku	@peterobi	@officialABAT
Total Tweets	6,654	1,226	2,156
Tweet since June 1st – September 30th 2022	490	753	199
Total Retweet	2,233,022	3,096,402	304,607
Tweet Retweet June 1st – September 30th 2022	189,493	2,444,221	117,452
Total Favorited	7,557,994	11,685,304	1,036,233
Tweet Favorited June 1st – September 30th 2022	904,423	9,396,626	546,369
Number of Followers	4,592,032	2,211,370	1,458,411

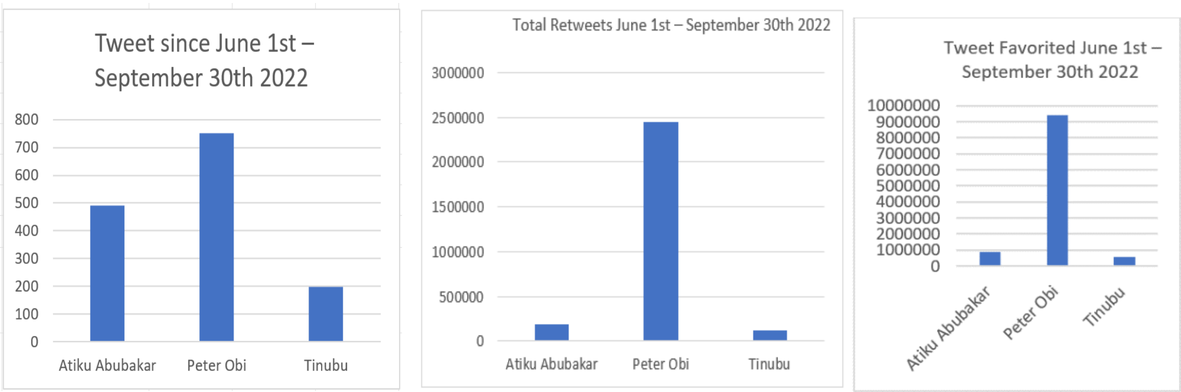

Figure 7: Impressions of Personal Tweets of the Three Candidates

The sentiment analysis of tweets posted by the three candidates is summarized in Table 8. Figure 8 shows the sentiment analysis of the favorite words of the candidates. It shows that Peter Obi has more favorite words as well as more positive favorite words than either of his peers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0238.v1

Table 8: Personal Tweet Sentiments

	Atiku Abubakar			Peter Obi			Tinubu		
	Negative	*Neutral*	*Positive*	*Negative*	*Neutral*	*Positive*	*Negative*	*Neutral*	*Positive*
All Tweets	696	3,920	2,035	109	803	312	309	1368	477
Tweets Since June 1 – September 31	34	346	110	59	537	156	6	126	67
Total **Favorited** Tweets Since June 1 – September 31	33,549	593,389	277,485	653,356	6,385,453	2,353,280	7,262	398,440	140,667
Total **Retweeted** Tweets Since June 1 – September 31	7,853	137,640	44,000	183,661	1,729,385	529,879	1,429	91,078	24,945

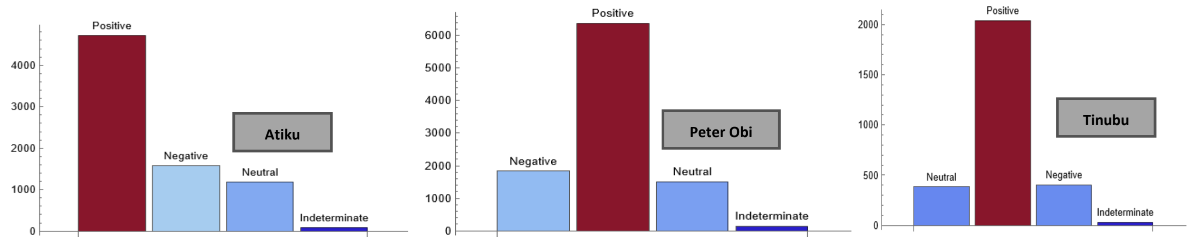

Figure 8: Sentiments of the Favorite Words of the Three Candidates

Figures 9, 10 and 11 depict the flow of tweets of each candidate on a hourly, weekly and yearly basis. The daily tweet pattern shows that Atiku tweets most on Thursdays, while Tinubu and Peter Obi tweets most on Wednesdays. The hourly tweet pattern shows that Atiku and Peter Obi spend the early hours of the day tweeting while Tinubu tweets in the afternoon around 2pm and 7pm. The yearly pattern shows that Peter Obi has not been active on Twitter until the year he started vying for the seat of Presidency. However, a similar pattern is evident with Atiku and Tinubu who tweet most only during election years.

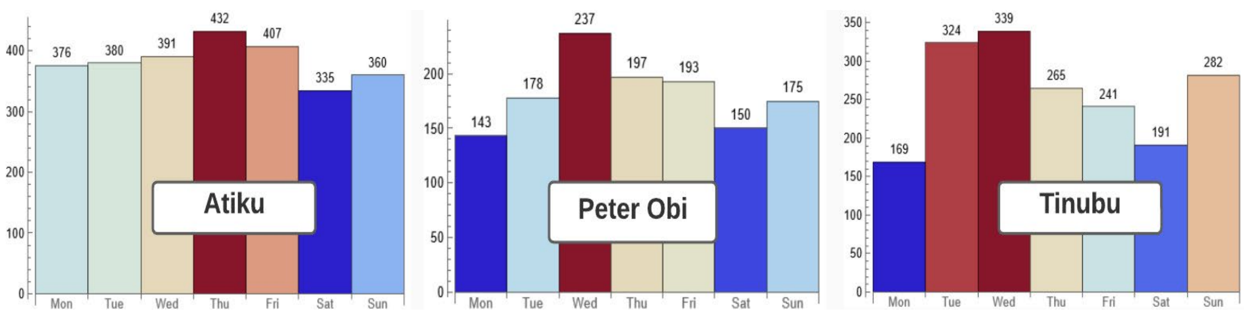

Figure 9: Daily tweet pattern of the Presidential candidates

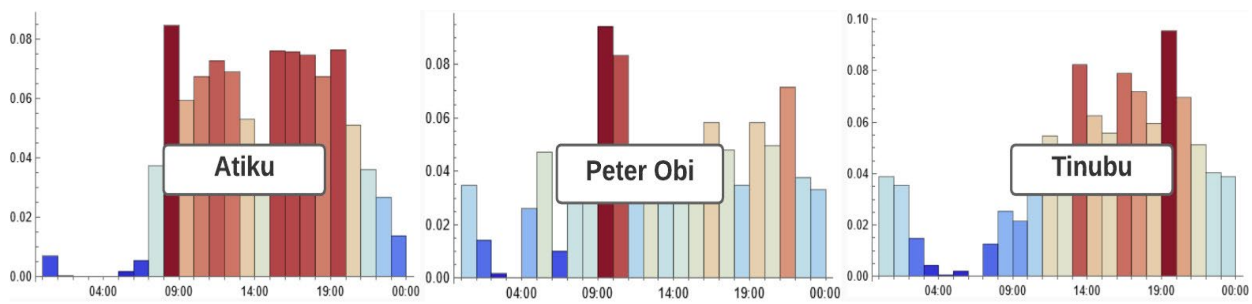

Figure 10: Hourly tweet pattern of the Presidential candidates

 doi:10.20944/preprints202210.0238.v1

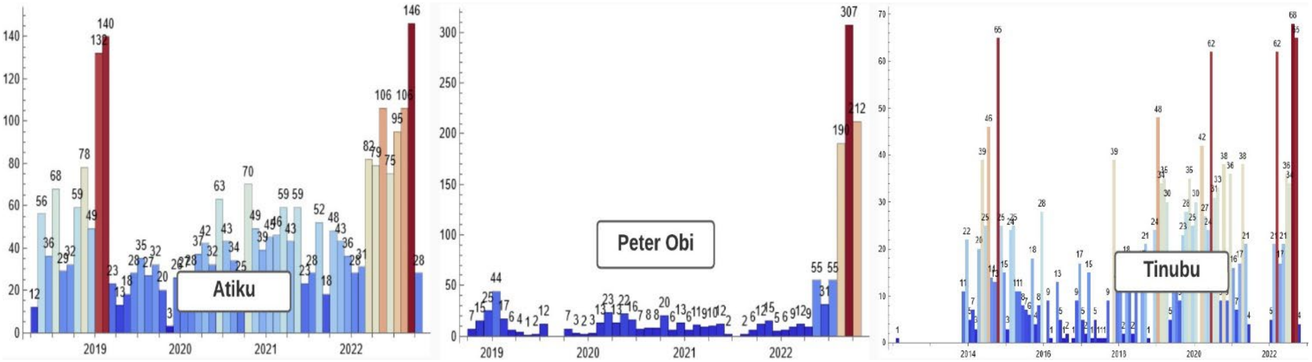

Figure 11: Yearly tweet pattern of the Presidential Candidates

Figures 12, 13 and 14 show the word clouds of the favorite words, mentioned users and popular hashtags of the three vying candidates. These figures reveal that Atiku was big on words such as Nigeria, government, family, health and democracy; Peter Obi is frequent with words such as Nigeria, government, economy, leadership, diaspora, education, security, etc. while Tinubu's tweets are centered around Nigeria, people, government, party and Buhari. Figure 14 shows that Atiku and Peter Obi used hashtags of their tours to various locations particularly local and international places respectively while Tinubu prefers to use Colloquium, ReadytoLead and ReadytoServe.

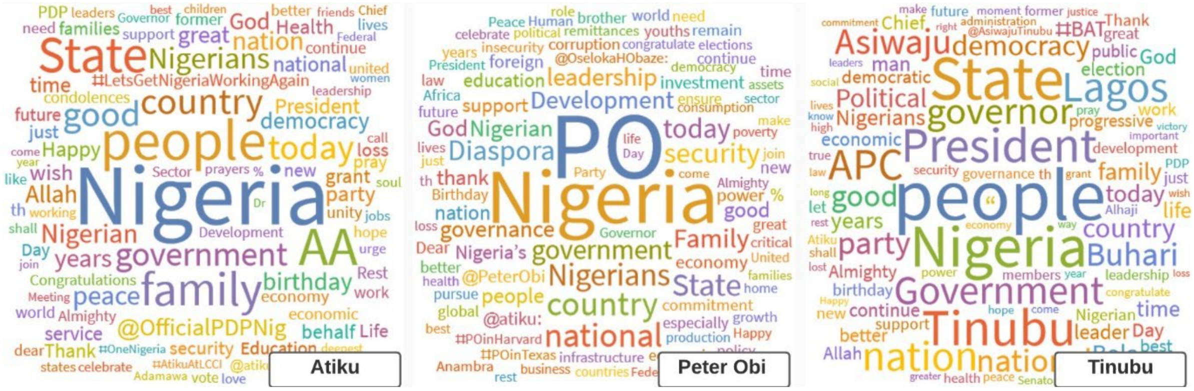

Figure 12: WordCloud of the favorite words of the Presidential Candidates

Figure 13: WordCloud of the mentioned users in tweets of the Presidential Candidates

 doi:10.20944/preprints202210.0238.v1

Figure 14: WordCloud of the mentioned hashtags in tweets of the Presidential Candidates

The mentioned hashtags show that Atiku showcases his travels to some non-Northern states such as Bauchi, Enugu, Oyo, Osun etc. and is an advocate of *#oneNigeria* as part of his campaign strategies. Peter Obi seems to be an international traveler with visits to Texas, Michigan, Washington DC, Havard, Atlanta etc. He also traveled to Jos as part of his campaign to voters outside his territory. Tinubu, although may have strategies outside the Twitter space, is not clear in his tour and campaign zone strategies.

Our social network analysis on each of the three candidates produced Figure 15. It shows that Tinubu maintained the strongest and most active network of Twitter friends, followed by Peter Obi. Atiku, on the other hand, has many friends who are dormant and inactive to his posts. Tinubu has prominently active friends including the incumbent President Muhammadu Buhari, Babatunde Raji Fashola (minister for Works and Housing of Nigeria), Babajide Sanwoolu (current Governor of one of the most populous states in Nigeria) etc. Atiku has @RenoOmokri as a very strong and active member. Peter Obi has more networks than Atiku but less than Tinubu.

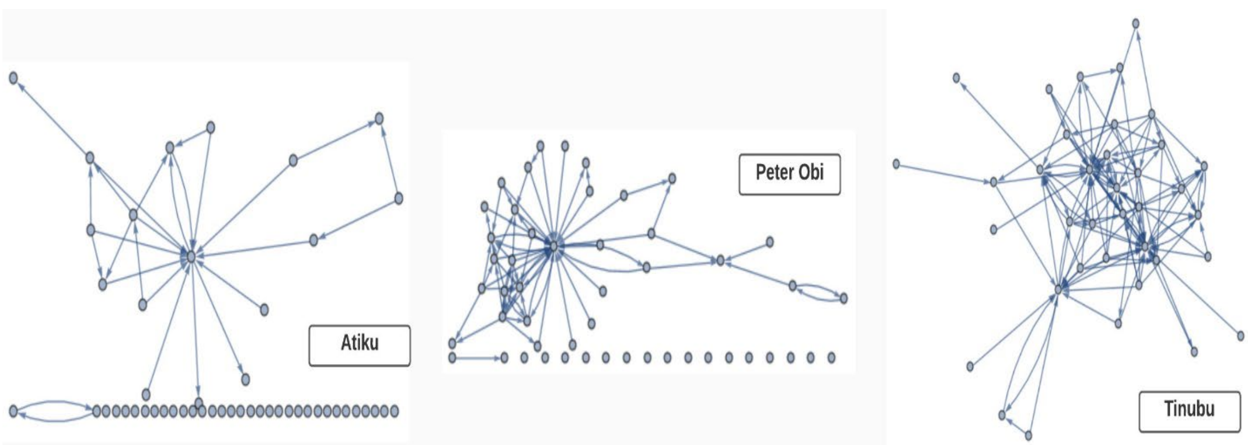

Figure 15: Social Network of the Presidential Candidates

c. Public Tweet Analysis

The public tweets concerning Atiku, Peter Obi and Tinubu were analyzed to obtain their respective public opinions. Table 9 shows the tweet summaries of the public tweets as obtained in this study. Table 10 shows the result of sentiment analysis of the public tweets categorized into each of the three candidates.

Table 9: Public Tweet Summaries of the Presidential Candidates

	Atiku Abubakar	***Peter Obi***	***Tinubu***

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0238.v1

Search Term	ATIKU OR ATIKULATED	PETER OBI OR OBIDIENT	TINUBU OR JAGABAN OR BATIFIED
Tweet since June 1st – September 30th 2022	50,688	980,336	1,028,058
Tweet Retweet June 1st – September 30th 2022	444,903	5,122,375	5,265,670
Total Favorited June 1st – September 30th 2022	1,366,512	17,784,008	16,430,587

Table 10: Summarization of Public Tweet Sentiments

	Atiku Abubakar			**Peter Obi**			**Tinubu**		
	Negative	***Neutral***	***Positive***	***Negative***	***Neutral***	***Positive***	***Negative***	***Neutral***	***Positive***
Tweets Since June 1 – September 31	6,599	33,960	10,105	35898	457,019	504,665	894,766	409,643	587,993
Total **Favorited** Tweets Since June 1 – September 31	162,497	1,026,372	177,438	1,150,954	6,976,112	8,008,726	1,765,443	4,004,534	9,309,876
Total **Retweeted** Tweets Since June 1 – September 31	53,120	348,001	43,729	534,775	1,856,223	2,998,766	859,882	2,034,598	2,650,986

d. Sentiment Analysis Model Performance

Three models were used in this study and their performance was measured. Table 11 and Figure 16 show the performance for the models which were developed to determine the sentiments in the datasets. The BERT model gave the best prediction with the highest accuracy, AUC, precision, recall and F-Measure. The split ratio was experimentally selected as 80:20 and 10-fold cross-validated. The BERT model, however, spent the most time in its model building and predictions. This is due to the fact that is a deep neural network model with a deeper intermediate layer.

Table 11 Performance of Sentiment Analysis Models

	Recall	Precision	F-Measure	Accuracy	AUC	Processing Time(s)
LSTM	0.872	0.827	0.829	0.88	0.876	122
SVC	0.764	0.814	0.792	0.73	0.812	95
BERT	0.925	0.885	0.917	0.94	0.947	322

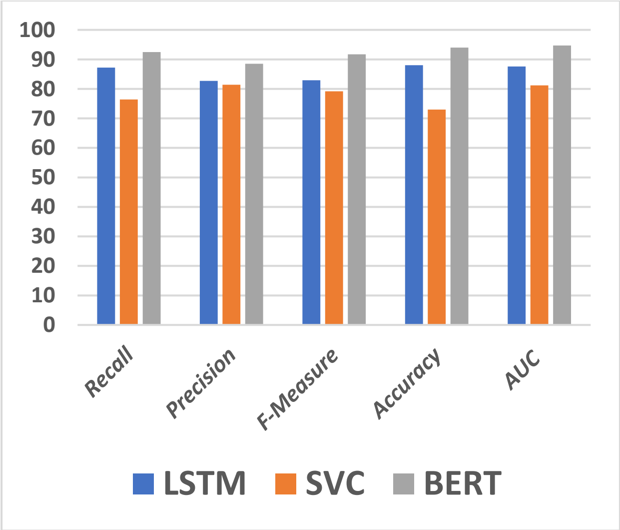

Figure 16: Performance Accuracy of the Three Sentiment Analysis Models

VII. Discussion

This work is primarily focused on the analysis of the Twitter space as regards the discourse around Nigeria 2023 presidential elections. We took it further by, first, analysis the individual tweets of the three candidates as well as the public opinion about them. The aim is to generate insights on each candidate and discover their strengths and weaknesses as well as suggest areas of improvements regarding their road to Aso-Rock. Our major analysis was centered on tweets on the election vis-à-vis sentiments expressed, frequently mentioned terms and impressions via mentions, favoriting and retweets. We observed that three candidates have their specific strategies and positives:

i. Atiku concentrates more on local networks and conferences.
ii. Peter Obi recently recognized the Twitter space as a viable medium of manifesto in recent times and keyed into it. He also has the most international outfits of the three candidates.
iii. Tinubu looks more like a network strategist who maintained an active Twitter network. This will help in creating a ripple effect of his campaign strategies.
iv. Peter Obi leads in the overall impressions and engagements since the three candidates declared their intents to run. He owns the highest tweets, retweets, mentions and favorited count among the three electorates.

It is interesting, but somewhat disappointing, to note that the three candidates did not pay enough attention to discussing the power, unemployment, education and security of the nation. As at the time of this writing, there is a chaotic dwindling of the nation’s economy, University education has been shut down for nine months, there has been an epileptic power supply due to incessant collapse of the national grid and the security system has not found solution to bombings, killings and kidnappings across the Federation. This, indeed, should be the major campaign points of each of these candidates; but unfortunately, not.

The tweet frequency and total impressions of the public tweets concerning the three candidates say a lot.

Our exploratory data analysis reveals that while Twitter is a good platform for political discussion and debate, a very small percentage of people control it. Also, majority of Twitter users (who disclosed their geo-location) only use Twitter to follow trends and conversations. During the height of the declaration of intent to run for presidency, it appears that most users who published their geolocation are passive and do not actively participate in dialogues nor express their opinions.

One major strength of this work is that sentiment analysis and NLP, to the best of our knowledge, has not been applied to understand elections in Nigeria. This shows that this work is novel in the Nigerian context. Another strength is the fact that most works do not consider the personal profiles of the candidates they examine. In this work, we provide an omnibus study which wholistically examines the internal and external profiles of the candidates to project a clearer and more reliable picture of the forthcoming Presidency election. Also, most studies were conducted retrospectively. We, in this work, conducted a scientific study prior to the election and present the result as is. This means that this study is independent and bias-free. Social network analysis on Twitter profiles were uncommon in reviewed works. In this study, we understood each candidate in the light of the friends they keep on the social media space. Finally, the size of our dataset is a major relative advantage to existing works. Authors took time to scrape the data comprehensively which necessarily and sufficiently covers Twitter discourse on the subject matter from the declaration of intent to run to the date this study was closed.

One limitation of this study is the time covered in the study. Although three months gave a very huge number of tweets, we know that strategies of campaign and manifestoes may change as events unfold. Since the result of the actual election outlive this study, we may not have a sufficient ground truth to validate the results of our experiments. Another limitation is that eligible voters may not have Twitter accounts or contribute to the election discourse. This may have a lot to say about comparing the outcome of the results in this study with the actual results of the actual election in 2023. Although the authors attempted further classify the sentiments by states using geocordinates, we were limited by the fact that most users prefer not to share their locations for obvious reasons.

VIII. Conclusion and Future Work

Social media platforms have proven to be an excellent tool which gives us opportunities to share thoughts, idea and opinions. Due to the upsurge in the number of internet users, social media networks have grown in popularity. As a result, there has been a tremendous increase in the quantity of tweets from individuals who expressed their opinions about the Nigeria 2023 elections all over the world. Users have shown their personal agitation on this discourse.

Finally, it is noteworthy that:

i. Peter Obi leads the chart in terms of Twitter impressions and engagements.
ii. Tinubu shows the strongest connection of active friends. This is a hidden winning strategy towards the presidential election.
iii. Atiku, although with the most followers, made the least impression on Twitter.
iv. If Nigeria 2023 presidential election is a two-horse race (as in Table 2), then Tinubu and Peter Obi are the real candidates to beat in the forthcoming presidential race.

In this work, we have investigated scientific and analytical methods for understanding the opinion polarity of people by developing a sentiment analysis model and determining the course that the Nigeria 2023 election campaign is taking. We found the relative strengths and weaknesses of each candidate and presented them as is. This study will be helpful for each candidate in consolidating on their strengths and intensifying efforts where they lag. Although LSTM, a class of deep neural network, is a powerful tool in NLP [55, 56], future work may look at applying other sentiment classification models for polarizing the public opinions. This may help in improving the accuracy of our models.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0238.v1

References

1. Roberts, M.E., *What is political methodology?* PS: Political Science & Politics, 2018. **51**(3): p. 597-601.
2. Achen, C.H., *Toward a new political methodology: Microfoundations and ART.* Annual review of political science, 2002. **5**(1): p. 423-450.
3. Ascher, W., *Political forecasting: The missing link.* Journal of Forecasting, 1982. **1**(3): p. 227-239.
4. Remmer, K.L., *The political economy of elections in Latin America, 1980–1991.* American Political Science Review, 1993. **87**(2): p. 393-407.
5. Oluro, M.J. and J.O. Bamigbose, *Legislative Cross-carpeting, Multiparty System and the Challenges of Democratic Good Governance in Nigeria.* Journal of Public Administration and Governance, 2021. **11**(1): p. 2639-2639.
6. Cameron, M.P., P. Barrett, and B. Stewardson, *Can social media predict election results? Evidence from New Zealand.* Journal of Political Marketing, 2016. **15**(4): p. 416-432.
7. Allcott, H. and M. Gentzkow, *Social media and fake news in the 2016 election.* Journal of economic perspectives, 2017. **31**(2): p. 211-36.
8. Hendricks, J.A. and D. Schill, *The social media election of 2016.* The 2016 US presidential campaign, 2017: p. 121-150.
9. Wani, G. and N. Alone, *A survey on impact of social media on election system.* International journal of computer science and information technologies, 2014. **5**(6): p. 7363-7366.
10. Paul, O.I., *Legitimacy, legitimation and succession in Nigeria: An appraisal of Section 42 (2) of the constitution of the Federal Republic of Nigeria 1999 as amended on the rights of inheritance.* Journal of Law and Conflict Resolution, 2012. **4**(3): p. 31-44.
11. FRHD, B.N., *Legislative efficiency and democratic stability in the fourth republic governance and politics of Nigeria: A re-appraisal of national assembly.* Kuwait Chapter of the Arabian Journal of Business and Management Review, 2012. **1**(9): p. 116.
12. Fawole, W.A., *Voting without choosing: interrogating the crisis of'electoral democracy'in Nigeria.* Liberal Democracy and its Critics in Africa: Political Dysfunction and the Struggle for Social Progress. Zed Books, London/New York, 2005: p. 149-150.
13. Adeniran, A.O., *Analytical study of political party systems in Nigeria.* Public Policy and Administration Research, 2015. **5**(12): p. 1-6.
14. Udemezue, S.C., *A Compendium of the Historical, Legal and Institutional Framework for the Practice of Company Law and Corporate Governance in Nigeria.* Journal of Commercial and Property Law, 2021. **8**(2).
15. Ayodele, N.C., *Party Politics and Local Government Elections: Implications for Third-Tier Governance in Ogun State, Nigeria (1999-2018).* FUDMA JOURNAL OF POLITICS AND INTERNATIONAL AFFAIRS, 2020. **3**(3): p. 148-162.
16. Udemezue, S., *A Compendium of the Historical, Legal and Institutional Framework for Company and Corporate Law Practice in Nigeria.* Legal and Institutional Framework for Company and Corporate Law Practice in Nigeria (April 28, 2021), 2021.
17. Wang, H., et al. *A system for real-time twitter sentiment analysis of 2012 us presidential election cycle*. in *Proceedings of the ACL 2012 system demonstrations*. 2012.
18. Neogi, A.S., et al., *Sentiment analysis and classification of Indian farmers' protest using twitter data.* 2021. **1**(2): p. 100019.
19. Somula, R., et al., *Twitter sentiment analysis based on US presidential election 2016*, in *smart intelligent computing and applications*. 2020, Springer. p. 363-373.
20. Chandra, R. and R.J.I.A. Saini, *Biden vs trump: Modeling US general elections using BERT language model.* 2021. **9**: p. 128494-128505.

21. Kausar, M.A., et al., *Public sentiment analysis on Twitter data during COVID-19 outbreak.* 2021. **12**(2).
22. Carvalho, J. and A.J.A.I.R. Plastino, *On the evaluation and combination of state-of-the-art features in Twitter sentiment analysis.* 2021. **54**(3): p. 1887-1936.
23. Bibi, M., et al., *A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis.* 2022. **158**: p. 80-86.
24. Caetano, J.A., et al., *Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 American presidential election.* 2018. **9**(1): p. 1-15.
25. Yavari, A., et al., *Election Prediction Based on Sentiment Analysis using Twitter Data.* 2022. **35**(2): p. 372-379.
26. Tumasjan, A., et al. *Predicting elections with twitter: What 140 characters reveal about political sentiment*. in *Proceedings of the International AAAI Conference on Web and Social Media*. 2010.
27. Schmidt, T., et al. *Sentiment Analysis on Twitter for the Major German Parties during the 2021 German Federal Election*. in *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*. 2022.
28. Razzaq, M.A., A.M. Qamar, and H.S.M. Bilal. *Prediction and analysis of Pakistan election 2013 based on sentiment analysis*. in *2014 IEEE/ACM International conference on advances in social networks analysis and mining (ASONAM 2014)*. 2014. IEEE.
29. Singh, P., R.S. Sawhney, and K.S. Kahlon. *Predicting the outcome of Spanish general elections 2016 using Twitter as a tool*. in *International Conference on Advanced Informatics for Computing Research*. 2017. Springer.
30. Al Walid, M.H., D. Anisuzzaman, and A.S. Saif, *Data analysis and visualization of continental cancer situation by Twitter scraping.* International Journal of Modern Education and Computer Science, 2019. **11**(7): p. 23.
31. Kusumasari, B. and N.P.A. Prabowo, *Scraping social media data for disaster communication: how the pattern of Twitter users affects disasters in Asia and the Pacific.* Natural Hazards, 2020. **103**(3): p. 3415-3435.
32. Hemalatha, I., G.S. Varma, and A. Govardhan, *Preprocessing the informal text for efficient sentiment analysis.* International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2012. **1**(2): p. 58-61.
33. Jivani, A.G., *A comparative study of stemming algorithms.* Int. J. Comp. Tech. Appl, 2011. **2**(6): p. 1930-1938.
34. Cambria, E. and B. White, *Jumping NLP curves: A review of natural language processing research.* IEEE Computational intelligence magazine, 2014. **9**(2): p. 48-57.
35. Fellbaum, C., *WordNet*, in *Theory and applications of ontology: computer applications*. 2010, Springer. p. 231-243.
36. Baccianella, S., A. Esuli, and F. Sebastiani. *Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining*. in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. 2010.
37. Taylor, A., M. Marcus, and B. Santorini, *The Penn treebank: an overview.* Treebanks, 2003: p. 5-22.
38. Mielke, S.J., et al., *Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP.* arXiv preprint arXiv:2112.10508, 2021.
39. Lovins, J.B., *Development of a stemming algorithm.* Mech. Transl. Comput. Linguistics, 1968. **11**(1-2): p. 22-31.
40. Yu, H. and J. Yang, *A direct LDA algorithm for high-dimensional data—with application to face recognition.* Pattern recognition, 2001. **34**(10): p. 2067-2070.

41. Wei, X. and W.B. Croft. *LDA-based document models for ad-hoc retrieval*. in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 2006.
42. Heimerl, F., et al. *Word cloud explorer: Text analytics based on word clouds*. in *2014 47th Hawaii international conference on system sciences*. 2014. IEEE.
43. Jin, Y., *Development of word cloud generator software based on python.* Procedia engineering, 2017. **174**: p. 788-792.
44. Yu, Y., et al., *A review of recurrent neural networks: LSTM cells and network architectures.* Neural computation, 2019. **31**(7): p. 1235-1270.
45. Huang, Z., W. Xu, and K. Yu, *Bidirectional LSTM-CRF models for sequence tagging.* arXiv preprint arXiv:1508.01991, 2015.
46. Staudemeyer, R.C. and E.R. Morris, *Understanding LSTM--a tutorial into long short-term memory recurrent neural networks.* arXiv preprint arXiv:1909.09586, 2019.
47. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.
48. Ravichandiran, S., *Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT*. 2021: Packt Publishing Ltd.
49. Lee, J.-S. and J. Hsiang, *Patent classification by fine-tuning BERT language model.* World Patent Information, 2020. **61**: p. 101965.
50. Noble, W.S., *What is a support vector machine?* Nature biotechnology, 2006. **24**(12): p. 1565-1567.
51. Suthaharan, S., *Support vector machine*, in *Machine learning models and algorithms for big data classification*. 2016, Springer. p. 207-235.
52. Pisner, D.A. and D.M. Schnyer, *Support vector machine*, in *Machine learning*. 2020, Elsevier. p. 101-121.
53. Widodo, A. and B.-S. Yang, *Support vector machine in machine condition monitoring and fault diagnosis.* Mechanical systems and signal processing, 2007. **21**(6): p. 2560-2574.
54. Tripathi, S., et al. *Analyzing sentiment using IMDb dataset*. in *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*. 2020. IEEE.
55. Yao, L. and Y. Guan. *An improved LSTM structure for natural language processing*. in *2018 IEEE International Conference of Safety Produce Informatization (IICSPI)*. 2018. IEEE.
56. Zhang, X., M.H. Chen, and Y. Qin. *NLP-QA Framework Based on LSTM-RNN*. in *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*. 2018. IEEE.