**Chapter Three**

**Methodology**

**3.1 Introduction**

In this study, we conduct sentiment analysis on the Nigerian 2023 Presidential Election using VADER and RoBERTa sentiment analysis models. We will try to find out whether there is a correlation between tweets of Nigerians on X and the outcome of the Presidential Elections. We will further infer whether in future, we can depend on measuring sentiments on X (formerly known as Twitter or any other social media) to predict the outcome of elections. This study will also look into suggesting the best model performs better so that more focus will be in applying the appropriate model in the future.

The process of conducting sentiment analysis on the Nigerian 2023 Presidential Election involves using Natural Language Processing (NLP) techniques and machine learning algorithms to classify sentiments expressed in text data related to the election.

**3.2 Research Design**

In this study, weadopt the Correlational Research Design. This research design method is focused on finding relationship between two or more variables and perform statistical analysis of the correlation between such variables. Whereas, there are various research designs such as: Descriptive Research Design, Experimental Research Design, Quasi-Experimental Research Design, Longitudinal Research Design, Cross Sectional Research Design etc that we could pick from, but the Correlational Research Design proved to be the best for this study.

In this study, we intend to understanding public sentiment on social media (X formerly known as Twitter) regarding the election outcome. We want to know whether the overall sentiment expressed in tweets is positive, neutral or negative. With this we collect tweet data using relevant hashtags, keywords and user mentions related to the election.

**3.3 Participants or Population**

The total population of this study is 78,946. The shape of our entire dataset is presented on the figure below on a jupyter notebook cell using python programming language. This dataset contains 78,946 rows and 16 columns.

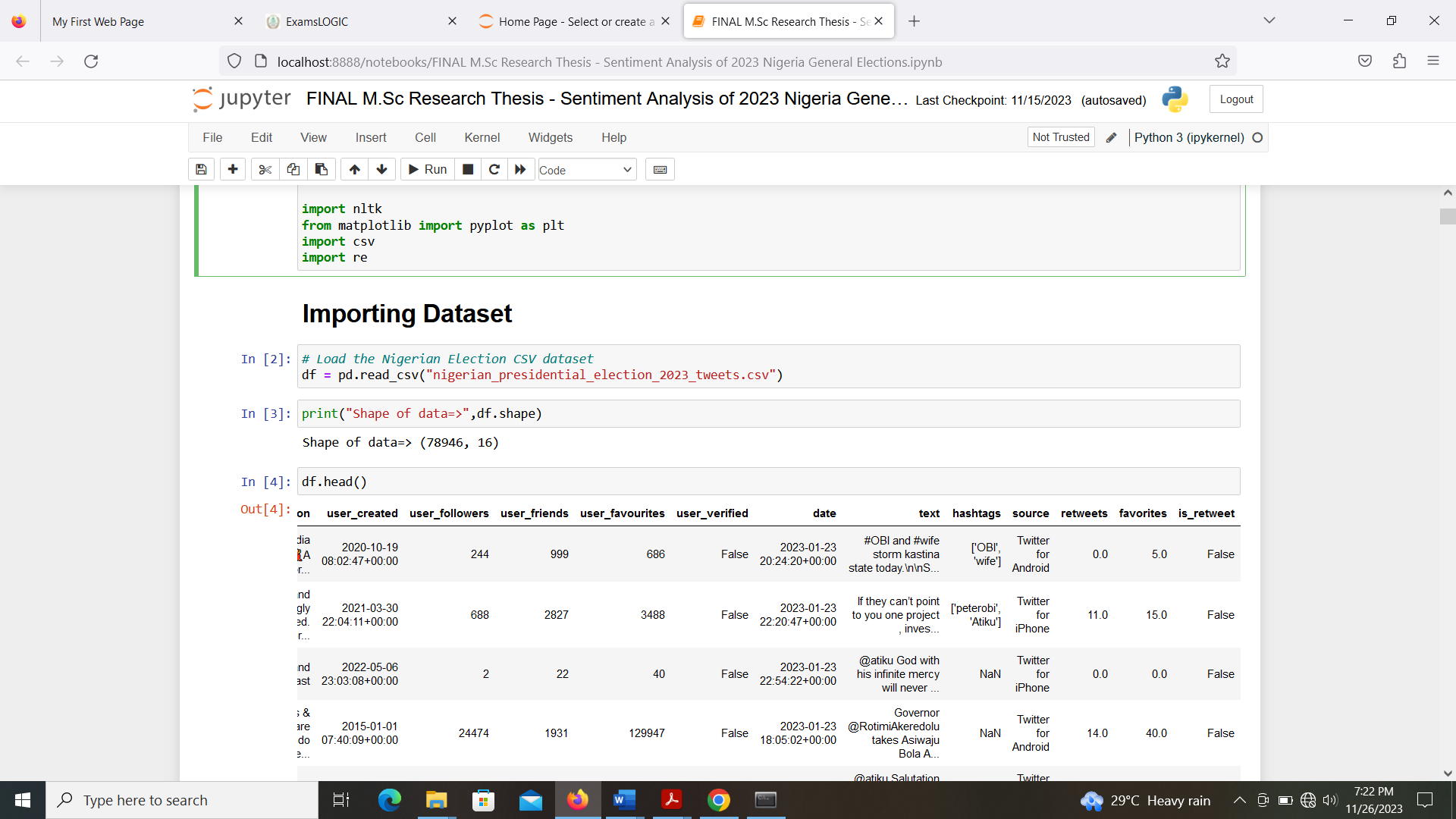


Figure 2: Shape of Data

The follow certain criteria to scrape relevant tweets for our study among which is, all tweets by the three (3) major candidates for the 2023 Nigerian presidential elections between January to March 2023 i.e. Bola Ahmed Tinubu, Atiku Abubakar and Peter Obi. The presidential elections was took place on 25th February, 2023 and we gather data at the peak of the election campaign period and a month after the election.

We collected data on the popular hashtags like, atikulated, jagaban, po, obidatti. Tweets from the official handles of the major political parties was collected and comments and retweets of the three major candidates form our sample for this study.

**3.4 Sampling Technique**

Considering the kind of study we are embarking, appropriate sampling technique is important to ensure that the data collected is relevant to the study and adequately represent the population of the study. Looking at the available sampling techniques, most of the popular sampling techniques (like simple random sampling, stratified random sampling, systematic sampling, cluster sampling, purposive sampling etc) will not satisfy our strategy of collecting samples from the pool of tweets available on X. However, Quota Sampling stand out and proven to meet our desire for extracting the particular data we need for this study.

Quota Sampling enable us set quotas for certain characteristics like particular hashtags and handle, then non-random sampling was used to meet these quotas. For instance, we search the hashtag atikulated, then we collect all the tweets that contain this hashtag.

**3.5 Data Collection**

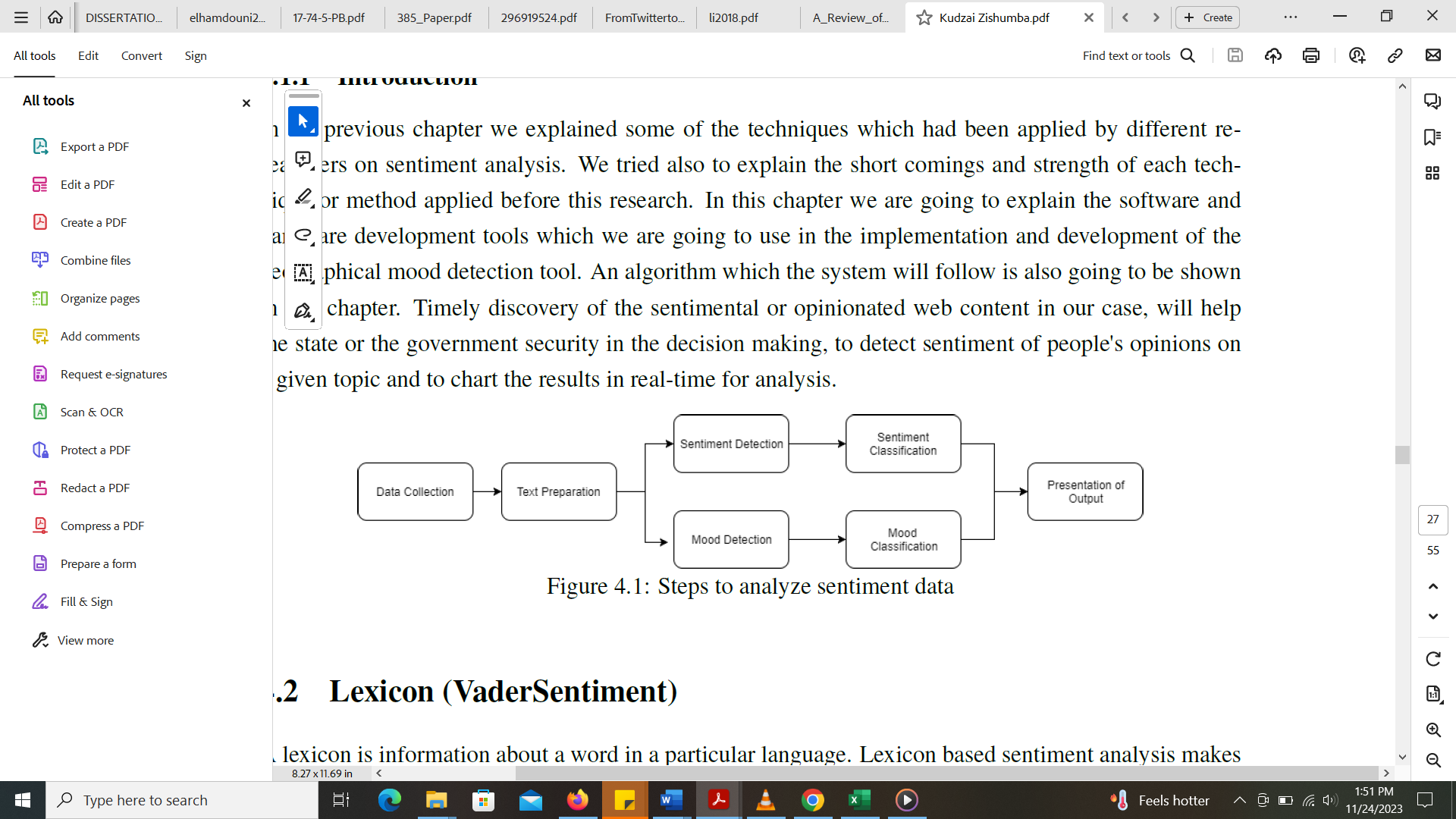
We used Tweepy (Python library for the Twitter API) to scrape our data. We first of all pip install tweepy then we create Twitter Developer Account by going to twitter developer portal then we create a new application to obtain API keys and access tokens needed for authentication.

**3.6 Data Analysis**

There are different stages of data analysis of our data. We follow the following processes; data preprocessing, sentiment analysis, visualization and word clouds.

**3.6.1 Data Preprocessing**,

* 1. Clean the data to remove irrelevant information, such as URLs, special characters, and irrelevant symbols.
  2. Tokenize the text into words or phrases. Remove stop words (commonly used words that may not contribute much to sentiment).
  3. Normalize the text by converting it to lowercase
  4. Classification of the tweets according to each candidate was done by filtering each tweet text based on the popular hashtags and keywords relating to each candidate. For instance, tweets with words like, atikulated, pdp, officialpdpnig and okowa etc are classified as atiku related tweets. Tweets that contain words like, obidient, po, lp, obidatti, labour party etc are classified as obi related tweets. While tweets with words like, batified, asiwaju, jagaban, bat and officialabat etc are classified as tinubu related tweets.
  5. The classified tweets are added into pandas dataframe as a new column on our dataset with the name candidate.



Steps to analyze sentiment data

**3.6.2 Sentiment Analysis:**

Utilize VADER and RoBERTa sentiment analysis tools or natural language processing (NLP) algorithms to categorize each piece of text into positive, negative, or neutral sentiments. NLTK VADER model and RoBERTa model adopted for this study have proven to perform better than other machine learning models for sentiment analysis, such as Support Vector Machines (SVM), Naive Bayes, or deep learning models like recurrent neural networks (RNN) or convolutional neural networks (CNN). We generate sentiment scores (e.g., on a scale of -1 to 1, where -1 is negative, 1 is positive, and 0 is neutral). Finally, we apply the trained model to the entire dataset to classify sentiment as positive, negative, or neutral for each text. Determine the overall sentiment distribution in the dataset. Explore sentiment patterns over time.

**3.6.3 Visualization**

To have a better understanding of patterns on our dataset, we create visualizations to better understand the sentiment distribution. We use bar charts, pie charts, or line graphs to illustrate the proportion of each candidates total positive, negative, and neutral sentiments. We also plot sentiment trends over time using line charts.

**3.6.4 Word Clouds**

We generate word clouds to visually represent the most frequent words associated with each candidate representing the three (3) popular political parties i.e. PDP, LP and APC and their respective rate of positive, negative, or neutral sentiments. Larger words in the cloud represent higher frequency.

**3.7 Ethical Considerations**

Big tech companies are obligated to implement measures to ensure data security such as encryption and authentication. Twitter (X) on the other hand ensures that developers or researchers adhere to ethical guidelines and respect user privacy. In other to access data using Tweepy API, Twitter collects your name, verify your email address, ensure that you are using the data for research or for academic purpose and mandate you to sign Developer Agreement and Policy as well as Twitter Rules.

We clearly communicated our intent for collecting this data, which is strictly for academic research purpose. We prioritize the anonymization of individual’s data to protect individual’s privacy. We will not share this data with anyone for any reason as we are liable to any breach that may occur; hence we will only share the result of our analysis but not the data itself.

**3.8 Validity and Reliability**

It is important to know the extent to which the data will accurately analyze the 2023 Nigerian presidential elections. Social media platforms have provided a level playing ground for all political parties to share, inform and educate the electorates. The electorates in turn use the platforms to express their concerns without fear or biases. The politicians or political parties cannot induce social media users to their favour, they cannot force a narrative on social media and are forced transparently carry out their campaigns. Hence, data generated are trustworthy, reliable and valid to measure the sentiment of users.