STUTERN GRADUATE ACCELERATOR

[Data Science/Machine Learning Track]

SOCIAL MEDIA AND SENTIMENT ANALYSIS: THE PAST 2 NIGERIAN ELECTIONS

Final Project Report - 2020 Due Date: May 29, 2020

Kwentua Hilary Chijindu

Email: kwentuajindu@gmail.com

Submitted to:

Akinlabi Ajelabi

Joseph Oladokun

SOCIAL MEDIA AND SENTIMENT ANALYSIS: THE PAST 2 NIGERIAN ELECTIONS

ABSTRACT

Twitter undoubtedly contains a diverse range of political insight and commentary. But, to what extent is this representative of an electorate? Can we analyze political sentiment effectively enough to capture the voting intentions of a nation during an election campaign? In this present day, social media platforms are playing a vital role in influencing people's sentiment in favor or against a government or an organization. Twitter-based data is not inherently a representative sample of society. However, opinion mining using machine learning techniques can categorize a tweet as positive, negative, and neutral in such a way that the election winner can be predicted almost quite accurately based on the ratio of positive tweets to the total tweet mentions. This project aims to identify and analyze public sentiments towards the top presidential candidates within the past 2 Nigerian elections, with the aim of determining their chances of being elected into the highest position of authority in Nigeria based on social media comments. We perform sentiment analysis on election related posts from Twitter using supervised machine learning (ML) techniques with the aim of detecting their sentiment polarity (i.e. negative or positive or neutral).

Keywords: Social Media, Sentiment Analysis, Twitter, Election Prediction, Machine Learning, Natural language processing.

INTRODUCTION

Nigeria has the largest democracy in Africa. It is a government formed by the people, of the people and for the people. The people elect their representatives to take decisions in the parliament and run the country. Citizens try to elect the candidate they think will bring some changes. The general elections are held in a gap of 4 years where the citizens above the age of 18 are eligible to cast vote and choose the suitable candidate. A peaceful and transparent election in Nigeria has implications to the advancement of democracy in Africa. As the most populous country and largest economy in Africa, Nigeria is the most important country on the continent and has the potential to influence developments not only in West Africa but, indeed, the entire continent.

On 28 and 29 March 2015, general elections were held in Nigeria, the fifth quadrennial election to be held since the end in 1999. The 2015 elections were historic, with the opposition winning for the first time since the transition from military rule in 1999, and with the incumbent presidential candidate, Goodluck Jonathan, conceding defeat and thus paving the way for a peaceful handover of power. Although there were 14 candidates for the presidency, the real contest was between the incumbent President Goodluck Jonathan of the People's Democratic Party (PDP) and General Muhammadu Buhari of the All Progressives Congress (APC) party.

Nigeria's presidential election scheduled for February 16, 2019, was the country's sixth since May 1999, when the military handed over power to a democratically elected civilian government. Though there are 73 political parties fielding candidates in the election, the race is generally believed to be between the incumbent President Muhammadu Buhari of the All Progressives Congress (APC) and Alhaji

Atiku Abubakar, a former vice president of the country from 1999 to 2007, of the People's Democratic Party's (PDP's).
Twitter is a popular microblogging service in which users post messages that are very short: initially less than 140 characters, since September 2017, there are 280 characters for each post and available as public data, averaging 20 words per message. It is convenient for research because there are a very large number of messages, many of which are publicly available, and obtaining them is technically simple compared to scraping blogs from the web. Microblogging today has become a very popular communication tool among Internet users. Millions of users share opinions on different aspects of life every day. Therefore, microblogging web-sites are rich sources of data for opinion mining and sentiment analysis. Because microblogging has appeared relatively recently, there are a few researches that are devoted to this topic. In this project, we focus on using Twitter, the most popular microblogging platform, for the task of sentiment analysis. I show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. We perform sentiment analysis of the collected corpus and explain discovered phenomena. Using the corpus, we build a sentiment classifier that is able to determine positive, negative and neutral sentiments for a document. In this research, I worked with English, however, the proposed technique can be used with any other language.
Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinion, sentiment, evaluation, appraisal, attitude, and emotion towards entities such as product, service, organization, individual, issue, event, topic, and their attributes. However, analysis of social media streams is usually restricted to just basic sentiment analysis and count based metrics. With the recent advances in deep learning, the ability of algorithms to analyze text has improved considerably. Creative use of advanced artificial intelligence techniques can be an effective tool for doing in-depth research. Sentiment analysis is a new field of research born in Natural Language Processing (NLP), aiming at detecting subjectivity in text and/or extracting and classifying opinions and sentiments. Sentiment analysis studies people's sentiments, opinions, attitudes, evaluations, appraisals and emotions towards services, products, individuals, organizations, issues, topics events and their attributes. In sentiment analysis text is classified according to the following different criteria:

- the polarity of the sentiment expressed (into positive, negative, and neutral);
- the polarity of the outcome (e.g. improvement versus death in medical text)
- agree or disagree with a topic (e.g. political debates)

NEED OF THE PROJECT

The scientific community has turned its interest in analyzing web data, such as blog posts or social networks' users' activity as an alternative way to predict election outcomes, hopefully, more accurately. Furthermore, traditional polls are too costly, while online information is easy to obtain and freely available. This is an interesting research area that combines politics and social media which both concern today's society. It is interesting to employ technology to solve modern-day challenges. Social media has become the most popular communication tool on the internet. Hundreds of millions of messages are being posted every day in the popular social media sites such as Twitter and Facebook. Social media websites become valuable sources for opinion mining because people post everything. Users post each and every thought

that they have about the current policies as well as political views. The internet is shifting from the quality and lengthy blog posts to much more numerous short posts that are posted by a lot of people. This trait is very valuable as now we can collect a different kind of people's opinions or sentiments from the social web. he prediction could be derived by comparing the number of tweets mentioning each candidate party or by comparing the number of tweets that have positive sentiments towards each candidate party. From the sentiments, we try to find out the political sentiments of the masses and have a general idea towards whom the public views are inclined and draw a conclusion.

LITERATURE REVIEW

Real human languages provide many problems for Natural Language Processing such as ambiguity, anaphora, and vagueness. Existing research has leveraged social media for election purposes in several ways, such as analyzing public sentiments towards each candidate and predicting election results. The last decade has seen an increase in sentiment analysis and opinion mining for several uses. This may be partly due to the increased availability of data expressing personal opinions. This led to the application of sentiment analysis across a variety of disciplines for different purposes.

For example, Tumasjan et al. [1] performed three research studies within the context of 2009 German federal election. First, they investigated whether Twitter really facilitates political deliberations by collecting tweets that either mention the six political parties or popular politicians in those parties. Second, they evaluated whether tweets reflect offline political sentiments. Finally, they analyzed whether volume of tweets reflects the popularity of parties in the real world and predicts election results. Their findings validate the popular belief that social media provides a platform for discussing political issues, and that social messages strongly reflect offline sentiments.

Monti et al. [2] modeled the political disaffection on twitter, they randomly selected 50,000 Italian Twitter users, and collected their followers. The dataset analyzed contained 261,313 users and more than 35 million tweets from those users. The authors classified tweets as political (related to politics), negative (has negative sentiment expressed) and general (tweet do not mention any candidate). They considered as politically disaffected only tweets, political, negative, and general. They applied different classifiers to automatic identify political dis- affection in tweets. Their results showed that Random Forest presented the best result on classification. They validated and compared their results with public opinion surveys regarding vote intentions and political topics. They found out a strong relationship between their classifier and the public opinion surveys. Thus, we see that predictive systems which utilize social media are both promising and challenging. The contention of our research is that the development of techniques for political public sentiment monitoring and election prediction is a promising direction requires more research work before we fully understand the limitations and capabilities of such an approach.

Ibrahim et al. present approach for predicting the results of Indonesia Presidential Election using Twitter as the main resource. First, they collected Twitter data during the campaign period. Second, they performed automatic buzzer detection on Twitter data to remove those tweets generated by computer bots, paid users, and fanatic users that usually become noise in data. Third, they performed a fine-grained political sentiment analysis to partition each tweet into several sub-tweets and subsequently assigned each sub-tweet with one of the candidates and its sentiment

polarity. Their study suggests that Twitter can serve as an important resource for any political activity, specifically for predicting the final outcomes of the election itself [12].

Razzaq et al. [4] also analyzed and predicted Pakistan general election using public sentiments expressed towards political parties on social media. They applied supervised machine learning techniques in classifying tweets into positive, negative, or neutral sentiments. They compared the average accuracies of several ML algorithms, including Naïve Bayes and SVM. Naïve Bayes performed best with an average accuracy of 70% for binary classification and about 55% for multiclass classification.

The use of twitter in prediction often follows either of two approaches: The use of historical tweets over months or years in predicting sentiment and popularity of a party and thereby extrapolate their chances of winning, such as the work of Sharma and Moh [5]; or the use of real-time streaming data for a short period of time, often dates very close to the elections to predict public sentiment and likelihood of winning, as is the case with Shah et al [6] and Wang et al [7].

TOOLS USED IN THIS PROJECT

A. PYTHON

Python is a widely used general-purpose, high level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code. Python is a programming language that lets you work quickly and integrate systems more efficiently. Some of the libraries and packages used include Pandas, Seaborn, GetOldTweets, NLTK, etc.

B. TWEEPY

Tweepy is open-sourced, hosted on Github and enables Python to communicate with Twitter platform and use its API. Tweepy supports accessing Twitter via Basic Authentication and the newer method, OAuth. Twitter has stopped accepting Basic Authentication so OAuth is now the only way to use the Twitter API. Tweepy provides access to the well documented Twitter API. With tweepy, it's possible to get any object and use any method that the official Twitter API offers. One of the main usage cases of tweepy is monitoring for tweets and doing actions when some event happens. Key component of that is the Stream Listener object, which monitors tweets in real time and catches them.

C. GETOLDTWEETS

Twitter basically has 2 APIs for developers who have accounts with them to retrieve data, the Representative State Transfer (REST) API and the Streaming API. The REST API allows you to go back in time to retrieve tweets, often this includes going 7 days back, as it always comes with a limit, unless you want to subscribe to premium access, where you pay a certain amount of money. The streaming API as the name implies looks into the future, and capture tweets as they arrive in real time. Twitter API's has some implementation which were not in favor of the direction of the project. You can't get tweets older than 7 days, also, the maximum number of tweets you could request for with the API is just 100.

I used and modified GetOldTweets3 to suit our purpose, which was an improvement fork of the GetOldTweet-Python web scraping script originally written by Jefferson

Henrique. We specified the user name of interest, the search keyword of interest, the duration of search (in form of start data and end date), the tweets language to be returned in the output, and the number of tweets. At the end of the mining, I had 40,853 tweets saved in different csv files. It is worthy to note that these tweets were accompanied with some underlying metadata, such as replies, username, permalink, favorites, and retweets.

D. NLTK

NLTK is one of the leading platforms for working with human language data and Python, the module NLTK is used for natural language processing. NLTK is literally an acronym for Natural Language Toolkit. NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum. Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project. NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language." NLTK provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more.

METHODOLOGY

A.DATA COLLECTION

The data collection step is the initial phase in the project, where data is collected from twitter. There are two methods on how to connect and collect tweets from Twitter. The first method is by searching tweets matching to the keywords. The second method is by collecting all the tweets provided by Twitter through streaming API, or all the tweets in a specific language, or all the tweets in a specific location, then putting all of them into our database. Both methods have their own advantages and disadvantages. For example, the first method requires only small storage as the data is relatively small. The downside is that we cannot get data from other keywords (if we need to) from an earlier time. Twitter allows the search API only for 7 days backwards. This data collection method is suitable if the focus is on the feature extraction or the prediction method. With the second method, we can apply any set of keywords to get the best result. As we are going for prediction and future analysis the twitter API extraction process is more helpful. Although, I encountered a problem in fetching tweets, I found a package called GetOldTweets which is able to get data from twitter as far back as necessary. Combined with the twitter API, I was able to get tweets as far back as 2015. We collected tweets using a manually built web scraper to gather tweets from the twitter website's archives. The tweets contained either the usernames of the contesting candidates or included the usernames of the political parties, that is '@MBuhari' or '@atiku'. This was done for the last 3 months (November to February) preceding the last 2 Nigerian election cycles: 2015 and 2019.

The total number of gathered tweets was above 80000 and is classified between the two major political parties that is shown in the table below:

PARTY	APC	PDP	TOTAL
2015	40100	41100	81100
2019	40000	40100	80100

The table above shows the number of tweets gathered.

B. DATA PREPROCESSING

The resulting dataset was very messy, unstructured and required extensive preprocessing before analysis could be carried out. This involved removing website URLS, usernames, 'likes' count, special characters, emoticons and punctuations. We carried out a version of analysis that removed the duplicates (in the form of retweets) and one without removal of duplicates. While the results were similar, here was a significant reduction in the number of tweets about the democrats after duplicates were removed. Tokenization is the process of splitting individual strings in this case, sentences into tokens. This was done to facilitate the process of stemming, which is removing unnecessary suffixes such as '-ly' or '-ness' in order to get the root meaning and accurate context of a word. After stemming, the tokens were joined again for accurate data labelling.

C. FEATURE EXTRACTION

This refers to the extraction of relevant features from cleaned data. To do this, we used Term Frequency-Inverse Document Frequency (TF-IDF) to convert the text into numbers for entry into the machine learning algorithm.

D. DATA LABELLING

The data was labelled based on the sentiment of each tweet: whether positive, negative or neutral. To do this, we made use of a library of words and their respective sentiment polarity values. We applied the merged datasets of APC and PDP related tweets on each of them to classify and label accordingly which tweets about the PDP were positive, neutral or negative as well as carry out the same approach on the APC dataset.

E. ALGORITHMS USED

After categorizing the sentiment of the tweets, we applied a machine learning approach to building a suitable model. This involved the use of classifying algorithms for training the model to correctly predict future tweet sentiment. The following algorithms were tried in the building of this prediction model.

i) Random Forest Classifier:

This is an algorithm used for classification problems. It operates by creating many decision trees (hence the term 'forest') at the training stage and outputting the class that is the most frequent or mean prediction of the individual trees. It employs feature randomness when building each individual tree to create an uncorrelated forest of trees whose prediction is more accurate than that of an individual tree. While it is a popular classifier, it is not always well suited to text mining problems.

ii) Naïve-Bayes Multinomial:

This algorithm is an adaptation of the classic Naïve-Bayes algorithm, which is used for classification problems in machine learning and is based on the Bay's Probability

Theorem, a theory recognized because of its simplicity as well as for its efficacy. The multinomial Naïve-Bayes classifier is however suitable for classification of discrete features or categorical data, which is the nature of text classification problems. They are better suited to analysis of textual data which is the reason the algorithm was being considered for inclusion into the model.

iii) Logistic Regression:
The logistic regression algorithm is a supervised classification algorithm that works by assigning its observations into a discrete set of classes. It will then transform its output using some non-linear 'sigmoid' function to return a probability value. It is also useful for predictive modelling because it is based on probability. It is commonly used with binary classification problem (where there are only 2 possible outcomes), though it can be adapted for multinomial (3 or more outcomes) and ordinal (outcomes with ordered categories).
The accuracy measures using 5-fold cross validation method for the different algorithms are presented in the table below:

2015

Random Forest Classifier	Naïve Bayes Multinomial	Logistic Regression
56.9%	64.4%	67.1%

2019

Random Forest Classifier	Naïve Bayes Multinomial	Logistic Regression
57.2%	66.9%	68.7%

The tables above showing the accuracy score of algorithms for the 2015 and 2019 election years respectively. As such, the logistic regression algorithm was selected for training and testing because it yielded the highest level of accuracy.

RESULTS
After building the model using the method described above, we go on to record the results of predictions. This follows a 2-step process where each dataset is first split 70% - 30% for testing and training. The model is trained with 4 different datasets which consists of 2 different set of tweets for APC for an election year, PDP for each election year and a combined dataset of the 2015 and 2019 election using the historical tweets; and tested with the same dataset. The results displayed below contain each of the predicted tweets as a percentage of the total tweets:

The table below showing positive tweets percentages;

	2015	2019
APC	26.4	30.6
PDP	24.4	32.0

The table below showing negative tweets percentages;

PARTY	2015	2019
APC	9.0	17.3
PDP	9.5	17.7

The table below showing neutral tweets percentages;

PARTY	2015	2019
APC	64.6	52.1
PDP	66.1	50.3

CONCLUSION

To predict the results of the past 2 Nigerian presidential election, I analyzed over 80,000 tweets from the past 2 election cycles in 2015 and 2019. By carrying out a sentiment analysis of the tweets, I was able to determine the number of tweets for or against the party for each election year. By using a supervised machine learning model (logistic regression algorithm), I trained the model to perform public sentiment analysis and predict the election outcome based on the party with the higher ratio of positive tweets. I was able to successfully predict the winner of election of the past 1 out of 2 election cycles. This model is best used in conjunction with other statistical models and offline techniques like conventional surveys and exit polls.

However, I recognize that twitter alone is insufficient to predict the results because the entirety of the electorate is not actively sharing their opinions on Twitter, especially the older generation or the masses. The metric of positive tweets ratio that this model relies on may also be an insufficient indicator of popularity as it ignores the ratio of neutral tweets or level of online participation of the party and candidate. It should also be noted that different people opt for other social media platforms besides Twitter to air their opinions such as Facebook or Instagram, whose data is not considered in this model. In addition, unexpected events such as economic shocks may occur which can swing political sentiment that this model does not account for.

Based on this, an area for future research would be improving this model by creating an automated framework which mines and analyzes data for months in real time since election result prediction is a continuous process that requires analysis over long periods of time. A larger dataset consisting of data from other social media platforms is also an issue to be considered when attempting to answer this research question in the future. The model could also be modified to incorporate emotion classification and the use of emoticons, as well as detect and account for sarcasm in tweets, all of which will result in greater precision and accuracy.

ADVANTAGES OF THE PROJECT

The advantages of using tweets as a data source are as follows;

- the number of tweets is very huge and they are available to the public.
- tweets contain the opinion of people including their political view.

DISADVANTAGES OF THE PROJECT

The disadvantages of the project are as follows:

- The long retweets couldn't be retrieved fully, as a result it was represented by "…", so the algorithm analyses it to be of neutral sentiment.
- Sarcasm was not detected in some sentences due to misuse of the semantics.
- Cannot get 100% accuracy in analyzing the tweets.

LIMITATIONS

Limitations faced in applying sentiment classification approaches and tools for sentiment analysis of posts in social media is to overcome the ambiguity that actually represents particular problem since it is not easily make use of coreference information. Typically, the analyzed posts contain irony and sarcasm, which are particularly difficult to detect. So, an evolution of approaches and tools is required to overcome this limitation.

FUTURE USE

With the help of sentiment analysis, we can not only predict the outcome of the election but also use it in different fields:

- It can be used in the field of education to know what the response of the students and parents is towards the change in the curriculum and what improvements are expected by the people
- It can be used in the field of medicine to know the use and side effects caused due to the introduction of a particular drug in the market. It also helps to know about the new diseases and the cures that can be found by the companies.
- The governments can know the response of the people about the new schemes and plans introduced by them and how it is helping the society to grow.
- It can be used by companies from IT sector to know the inclination of the present youth and develop their schemes accordingly.

REFERENCES

[1] Tumasjan A, Sprenger TO, Sandner PG, Welpe IM. Predicting elections with twitter: What 140 characters reveal about political sentiment. Icwsm.2010;10(1):178–185.

[2] Monti, C.; Zignani, M.; Rozza, A.; Arvidsson, A.; Zappella, G.; Colleoni, E. Modelling political disaffection from twitter data. In Proceedings of the 2nd International Workshop on Issues of Sentiment Discovery and Opinion Mining, Chicago, IL, USA, 11 August 2013.

[3] Ibrahim, M.; Abdillah, O.; Wicaksono, A.F.; Adriani, M. Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation. In Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015. [CrossRef]

[4] M. A. Razzaq, A. M. Qamar, and H. S. M. Bilal, “Prediction and analysis of Pakistan election 2013 based on sentiment analysis,” ASONAM 2014 - Proc. 2014 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., no. Asonam, pp. 700–703, 2014.

[5] Sharma, P., & Moh, T. S. (2016, December). Prediction of Indian election using sentiment analysis on Hindi Twitter. In 2016 IEEE International Conference on Big Data (Big Data) (pp. 1966-1971). IEEE.

[6] Ramteke, J., Shah, S., Godhia, D., & Shaikh, A. (2016, August). Election result prediction using Twitter sentiment analysis. In 2016 international conference on inventive computation technologies

[7] Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). System for real-time twitter sentiment analysis of 2012 us presidential election cycle. Proceedings of the ACL 2012 system demonstrations, 115-1

[8] D. J. S. Oliveira, P. H. de S. Bermejo, and P. A. dos Santos, “Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls,” J. Inf. Technol. Polit., vol. 14, no. 1, pp. 34–45, 2017.
[9] O'Connor, B. Balasubramanyan, R. Routledge, B. R., & Smith, N. A. (2010, May). From tweets to polls: Linking text sentiment to public opinion time series. In Fourth international AAAI conference on weblogs and social media.
[10] Skoric, M.M.; Poor, N.D.; Achananuparp, P.; Lim, E.P.; Jiang, J. Tweets and votes: A study of the 2011 Singapore General Election. In Proceedings of the 45th Hawaii International Conference on System Sciences, Maui, HI, USA, 4–7 January 2012. [CrossRef]
[11] DiGrazia, J.; McKelvey, K.; Bollen, J.; Rojas, F. More tweets, more votes: Social media as a quantitative indicator of political behavior. PLoS ONE 2013, 8, e79449. [CrossRef] [PubMed]
[12] Metaxas, P.T.; Mustafaraj, E. Social media and the elections. Science 2012, 338, 472–473. [CrossRef] [PubMed]
[13] Franch, F. Wisdom of the Crowds: 2010 UK election prediction with social media. J. Inf. Technol. Politics 2013, 10, 57–71. [CrossRef]
[14] Gayo-Avello, D.; Metaxas, P.T.; Mustafaraj, E. Limits of electoral predictions using twitter. In Proceedings of
the ICWSM, Barcelona, Spain, 17–21 July 2011.
[15] T. H. Nguyen, K. Shirai, and J. Velcin, “Sentiment analysis on social media for stock movement prediction,” Expert Syst. Appl., vol. 42, no. 24, pp. 9603–9611, 2015.
[16] Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." LREc. Vol. 10. No. 2010. 2010
[17] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, “Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens’ political preferences with an application to Italy and France,” New Media Soc., vol. 16, no. 2, pp. 340–358, 2014.