Election Data Analytics: Analysis of 2015 General Elections in Nigeria

Steve A. Adeshina, Collins E. Onyemaobi, and Adegboyega Ojo,

Abstract—2015 General Elections conducted in Nigeria is seen to be the most credible election since the inception of the democratic governance in Nigeria. During this process, the Voters Register (VR), Permanent Voters Card (PVC) and Smart Card Reader (SCR) were introduced into the system to identify, authenticate and verify individual voters. Twitter and other social media platforms were highly exploited by both candidates and voters to air their views on the elections. This work adresses the design and implemention of a tool that delivers election data analytics from multiple data sources, which includes twitter, Voter Register (VR) and SCR Smart Card Readers (SCR) from 2015 general elections. The tweets are being processed by the extraction of metadata, geocoding the physical addresses and analyzing the sentiments of the content of each tweets. These results are further compared and combined with the data from VR and SCR into a front-end visualization applications. Ultimately we show relationship between the above data analytics and subsequent outcome of elections.

Index Terms—Data Analytics; Twitter API; Nigerian Elections; Independent National Electoral Commission

1 INTRODUCTION

The recent dispensation of democratic governance in Nigeria commenced again from May 29, 1999. Since then the Independent National Electoral Commission (INEC) has conducted five consecutive general elections in 1999, 2003, 2007, 2011 and 2015 respectively. Citizens engagement has played a vital role in different ways for these elections. The increasing penetration of smartphones has reduced traditional barriers of access to broadband Internet. Candidates, Voters and Observers alike have debated the role social media has played in voter participation between 2011 and 2015. Discussions on social media adoption and its impact towards the electoral outcome have been stressed and promoted during 2011 general election. Social media redefined citizens engagement, not just with themselves, but with the INEC in ways never before experienced. Platforms like Twitter, Facebook, Youtube and SMS portal became the prominent spheres for engagement. INEC, which in earlier years relied on government media outlets and on few private ones took to the social media [1].

Twitter ultimately proved to be the most efficient way to interact with INEC. The commission's use of social media led to its website receiving a record 25 million hits in three days during the presidential elections. By using social media to inspire voters, the electoral commission has redefined elections in Nigeria. In addition, INECs introduction of PVC with embedded chip for only valid registered voters as the only acceptable means of identification for voting and the SCR; a system designed to authenticate and verify authentic PVCs has greatly influenced the electoral outcome.

With the innovative systems and processes introduced by INEC, and greater number of Twitter adoption by voters which defeated the incumbent president and has been referred to as Muhammadu Buhari's historic electoral victory. This has significantly played out on social media and innovative technology for 2015 general Nigeria elections. Yet there is no such system that provides insight into the different sources available data and its historical correlation. One is compelled to ask a number of questions with these; what is the impact of this social media (Twitter in this case) and the newly innovative electoral technology (SCR) on the electoral outcome? How can the available data from these

- *Steve A. Adeshina and Collins Onyemaobi are with the Department of Computer Sciences, Nile University of Nigeria, Abuja, Nigeria.*
  E-mail: steve.adeshina@nileuniversity.edu.ng
- *Adegboyega Ojo is with Insight Centre, University of Galway, Ireland*

sources be analyzed to determine relationships using a software tool?

This work looks retroactively into the part played by Social Media, and in particuler Twitter in the outcome of Nigeria's 2015 General elections. A tool has been developed and implemented. This tool will be useful in predicting the outcome of Elections not only in Nigeria and sub-sahara Africa, but also in the entire developing countries. It could also be adapted to use other socia media and be used for other purposes other than elections.

2 Related Work

Several works have attempted to answer the question; could activity on sites like Twitter be a predictor of election results? Senator Barack Obamas successful 2008 presidential campaign established social media as an integral part of the campaign toolbox [2]. The 2010 European Digital Competitiveness Report (cited in Aparaschive, 2011 [10]) states that over 60% of Romanians do not have any kind of knowledge or skill to participate in the digital era. With such a low penetration of voters being able to be reached by social media, it would be worth noting if those few who did use social media were influenced by the most recent player in our media ecology. Tumasjan *et al.*[2] discovered that the relative volume of tweets closely mirrored the results of the German federal elections. The researchers concluded that Twitter was being used as a platform for political deliberation, and that the number of tweets reflected voter presence, which in turn closely resembled the live political debate. However, elections are about deciding change by either rejecting it or choosing to move in a different direction [11].

The size and availability of data on Twitter can be used to spot trends in attitudes and perceptions before it appears in almost any other medium. Researchers are finding that Twitter can measure public sentiment, track political activity, and monitor events in the population at large [3]. For example, in 2009 and 2010, tracking flu-related keywords allowed a Southeastern Louisiana University researcher to predict future flu outbreaks. Conversely, it can take up to two weeks for the U.S. Centers for Disease Control and Prevention to collect data on influenza and disseminate the information. While Twitter reports are less precise, they are available in real time and cost less to collect. Another emerging trend is to examine how Twitter can gauge sentiments and supplement traditional polling [3]. Researchers admit, however, that Twitter data is noisy, not always consistent as to the meaning of words used in messages (the differences between sick for illness and sick meaning good), and has sample bias, as it is established that certain segments of the population use Twitter more than others [9] [8].

There is also the use of Twitter for unfiltered information that is abusive, especially in politics. Two researchers at Wellesley University found that during a special U.S. Senate election in Massachusetts, the democratic candidate, Martha Coakley, was the subject of a Twitter bomb attack. A conservative group sent out 929 tweets in just over two hours linking to a Website that attacked Coakley. With retweets, the potential audience could have been seen by more than 60,000 people [3]. This type of message propagation can be either natural or pushed by artificial means to manipulate the system. Ultimately, the winner in a political contest, no matter what means are used to sway opinion, is decided by the number of votes cast. Getting a registered voter to participate in voting is the final payoff to the campaign. Additionally, Gayo-Ayello *et al.*[4] warns that non-responses often play a more important role than collected data. If, for example, the lack of information affects one group, the results might differ considerably from reality if that group were included. Finally, Gayo-Avello states that researchers should carefully evaluate positive reports from social media before assuming that the reported methods are applicable to any similar scenario with identical results. Researchers should also identify the various users based on age, income, gender and race to properly weigh their opinions according to the percentage of the population to assure the integrity of the results.

3 System Methodology

The adopted methodology is based on prototyping and rapid application development (RAD) life cycle.

3.1 Data

3.1.1 Twitter data source 1

For this research work, we sourced thousands of messages on Twitter that are related to the Nigerian Presidential Elections (Random Sampling). Only messages or tweets that are generated within 50km radius of the borders of Nigeria and geolocate it to Nigerian states. The Twitter data between January 14 to March 27, 2015. Keywords such as Jonathan Goodluck, JEG, Jonathan, Goodluck, Buhari, GMB

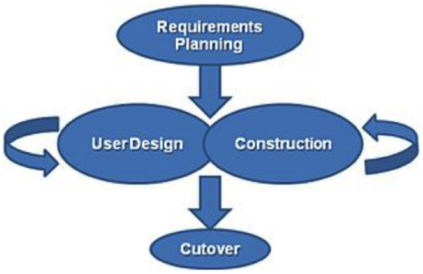


Fig. 1. Phases in the James Martin approach to RAD - Rapid application development (Wikipedia, 2016)

were explored to understand social opinions / sentiments about a candidate.

3.1.2 Data Source 2 Voters Register/PVC

National Voters Register comprises of eligible voters for any election to be conducted in Nigeria. The PVC which is then printed as an identity to authentic voters and as the only instrument for voting. INEC provided the 2015 National Voters Register on these fields (State,LGA,RA,PU).

3.1.3 Data Source 3 : SCR verification system

Smart Card reader was deployed by INEC for use in the 2015 general elections. The SCR verification system provides data showing the accreditation done on Election Day prior to voting. The data provided will also be useful as it gives insight into the relationship between the voters register and the final election outcome

3.2 Data model

Figure 2 shows our data model. This is a conceptual representation of the data structures that are required by a database. The first step in designing a database is to develop an Entity-Relation Diagram (ERD). The ERD serves as a blueprint from which a relational database may be deduced.

Entity $[scr_2015]$ matches exactly one record in entity $[vr_2015]$ and every record in $[scr_2015]$ matches exactly one record in $[vr_2015]$. And both $[scr_2015]$ and $[vr_2015]$ matches one record to entity $[states]$. Entity $[states]$ one has one record that matches many records in entity $[tweets]$ and entity $[tweets]$ has many records matching exactly one record in entity $[states]$ and likewise entity $[geo_zone]$ has one record matching many records in entity $[states]$. In the Relational Database model, each of the entities will be transformed into a table.

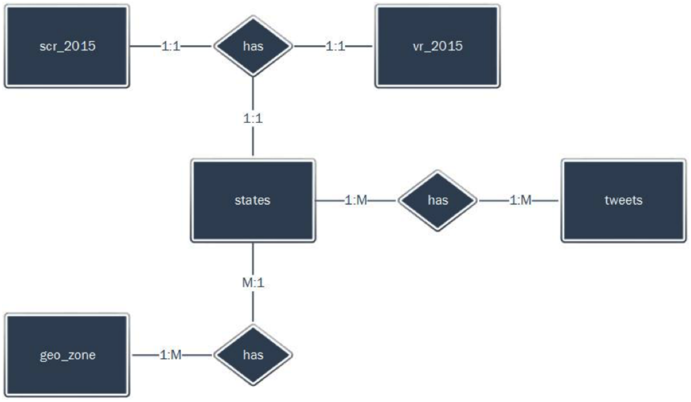


Fig. 2. Entity Relationship Diagram (ERD)

3.3 Archtectural Framework

Figure 3 shows the system architecture of the framework and the data flow through different components. The first step is to retrieve data from sources which includes static INECs dataset and twitter. Twitter, in particular, provides multiple levels of interfaces to access the Twitter feeds. A tweet crawler was developed based on the Twitter search API to collect tweets posted within a 50km radius of the borders of Nigeria and geo-locating the source of the tweets to the states in Nigeria. In the second step, a text mining method was applied to the messages, including sentiment analysis to understand how effective both campaigns have been in getting their messages across and how their respective candidates match up. It should be noted that, depending on the application scenarios, other data mining methods could be plugged into this step to gain the information of interest from each tweet. The resulted tweets are then cleaned, organized and loaded into a database for the construction and presentation of election data analytics system

3.4 Sentiment Classification Techniques

This research uses Naive Bayes, a commonly used supervised machine-learning algorithm. Naive Bayes classifier is based on Bayes theorem of probabilistic model. In this we tried to estimate the probability of a twitter text based on whether it belongs to positive, negative or neutral class.

3.5 Naive Bayes classifier

Bayesian classifiers are based around the Bayes rule, a way of looking at conditional probabilities that allows you to flip the condition around in a convenient way. A conditional probability is a probability

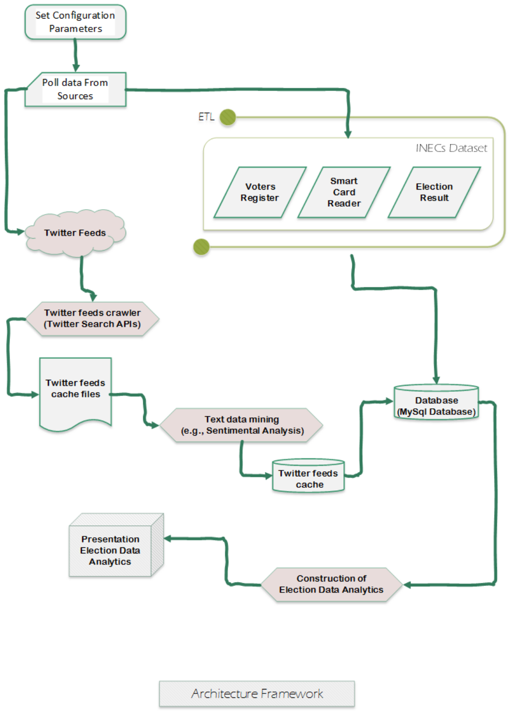

Set Configuration Parameters
Poll data From Sources
ETL
INECs Dataset
Voters Register
Smart Card Reader
Election Result
Twitter Feeds
Twitter feeds crawler (Twitter Search APIs)
Twitter feeds cache files
Text data mining (e.g., Sentimental Analysis)
Twitter feeds cache
Database (MySql Database)
Presentation Election Data Analytics
Construction of Election Data Analytics
Architecture Framework

Fig. 3. Architectural Framework

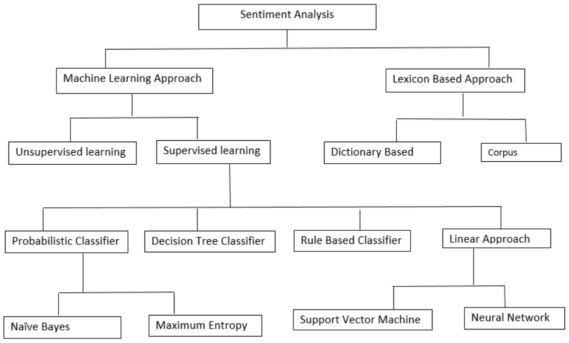


Fig. 4. Sentiment Classification Techniques - Bhardwaj *et al.*. (2015) [6] [7]

that event $\mathbf{X}$ will occur, given the evidence $\mathbf{Y}$. That is normally written

$$\mathbf{X/Y} \tag{1}$$

The Bayes rule allows us to determine this probability when all we have is the probability of the opposite result, and of the two components individually:

$$\mathbf{P(P/Y)} = \frac{\mathbf{P(X)P(Y/X)}}{\mathbf{P(Y)}} \tag{2}$$

This restatement can be very helpful when we're trying to estimate the probability of something based on examples of it occurring. In this case, we're trying to estimate the probability that a document is positive or negative, given its contents. We can restate that in terms of the probability of that document occurring if it has been predetermined to be positive or negative. This is convenient, because we have examples of positive and negative opinions from our data set above. The thing that makes this a naive Bayesian process is that we make a big assumption about how we can calculate at the probability of the document occurring: that it is equal to the product of the probabilities of each word within it occurring. This implies that there is no link between one word and another word. This independence assumption is clearly not true: there are lots of words which occur together more frequently that either do individually, or with other words, but this convenient fiction massively simplifies things for us, and makes it straightforward to build a classifier. We can estimate the probability of a word occurring given a positive or negative sentiment by looking through a series of examples of positive and negative sentiments and counting how often it occurs in each class. This is what makes this supervised learning - the requirement for pre-classified examples to train on. So, our initial formula looks like this.

$$\mathbf{P(Senti/Senten)} = \frac{\mathbf{P(Senti)P(Senten/Senti)}}{\mathbf{P(Senten)}} \tag{3}$$

where Sent is sentiment and senten is sentence

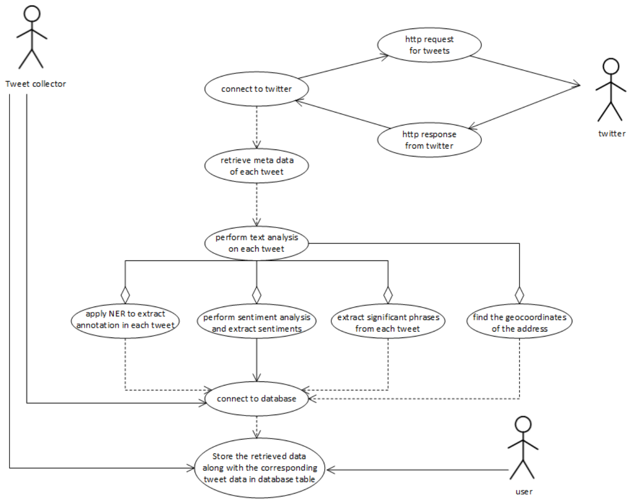
We can drop the dividing $\mathbf{P(line)}$, as it's the same for both classes, and we just want to rank them rather than calculate a precise probability. We can use the independence assumption to let us treat $\mathbf{P(sentence - sentiment)}$ as the product of $\mathbf{P(token - sentiment)}$ across all the tokens in the sentence. So, we estimate

$\mathbf{P(Sentiment/Sentence)}$ as

$$\frac{\mathbf{count(thistokeninclass) + 1}}{\mathbf{count(alltokensinclass) + count(alltokens)}} \tag{4}$$

The extra $\mathbf{1}$ and count of all tokens is called 'add one' or Laplace smoothing, and stops a $\mathbf{0}$ finding its way into the multiplications. If we didn't have it any sentence with an unseen token in it would score zero.

To achieve the above, *phpInsight* was adopted. *phpInsight* is a PHP implementation of sentiment classifier. It uses a dictionary of words that are categorized as positive, negative or neutral, and a naive Bayes algorithm to calculate sentiment. To improve accuracy, phpInsight removes 'noise' words.


Tweet collector
http request for tweets
connect to twitter
twitter
http response from twitter
retrieve meta data of each tweet
perform text analysis on each tweet
apply NER to extract annotation in each tweet
perform sentiment analysis and extract sentiments
extract significant phrases from each tweet
find the geocoordinates of the address
connect to database
Store the retrieved data along with the corresponding tweet data in database table
user

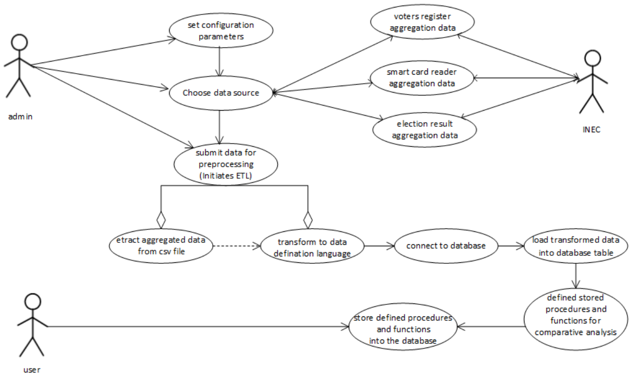
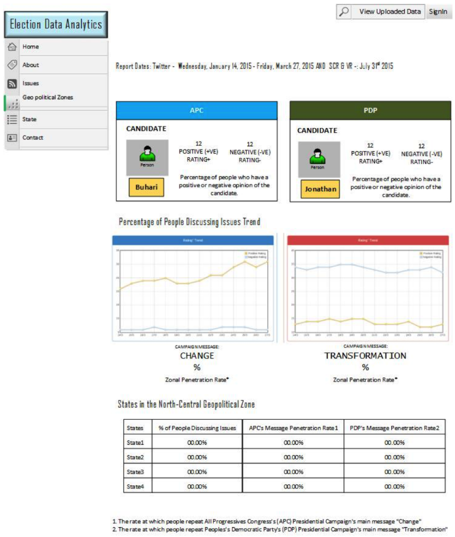

set configuration parameters
voters register aggregation data
admin
Choose data source
smart card reader aggregation data
INEC
election result aggregation data
submit data for preprocessing (Initiates ETL)
etract aggregated data from csv file
transform to data defination language
connect to database
load transformed data into database table
defined stored procedures and functions for comparative analysis
store defined procedures and functions into the database
user

Fig. 5. Use case diagram

3.6 Use Case

Figure **??** represents the use case diagram of the system. The interaction between different components of the system and control flows is represented in the above figure. Tweet collector is a main component that interacts with twitter to get access based on search criteria. Tweet collector is dependent of processing which involves the information extraction using Text analysis of four modules annotations, sentiment extraction, significant phrases and geocoding. The main component interacts with all the modules and aggregates the data and get stored in the database while preprocess of extract, transform and load is initiated data import from data sources. The extracted data from raw tweet delivers users a clear understanding of dependencies in a tweet; also it simplifies the work of filtering information from data.



Report Dates: Twitter - Wednesday, January 14, 2015 - Friday, March 27, 2015 AND SCR & VR -: July 31st 2015

APC

CANDIDATE

Buhari

12 POSITIVE (+VE) RATING+

12 NEGATIVE (-VE) RATING-

Percentage of people who have a positive or negative opinion of the candidate.

PDP

CANDIDATE

Jonathan

12 POSITIVE (+VE) RATING+

12 NEGATIVE (-VE) RATING-

Percentage of people who have a positive or negative opinion of the candidate.

Percentage of People Discussing Issues Trend

CAMPAIGN MESSAGE:
CHANGE
%
Zonal Penetration Rate*

CAMPAIGN MESSAGE:
TRANSFORMATION
%
Zonal Penetration Rate*

States in the North-Central Geopolitical Zone

States	% of People Discussing Issues	APC's Message Penetration Rate1	PDP's Message Penetration Rate2
State1	00.00%	00.00%	00.00%
State2	00.00%	00.00%	00.00%
State3	00.00%	00.00%	00.00%
State4	00.00%	00.00%	00.00%

1. The rate at which people repeat All Progressives Congress's (APC) Presidential Campaign's main message "Change"
2. The rate at which people repeat Peoples's Democratic Party's (PDP) Presidential Campaign's main message "Transformation"

Fig. 6. Public View data by geo political zones

4 Experiments and Results

The results and evaluation of this research work was done by harvesting and analyzing Twitter data for 2015 presidential election campaigns of the two major parties that participated, the All Peoples Congress (APC) and the People Democratic Party (PDP). For close to 3 months, we harvested thousands of messages everyday generated within a 50km radius of the borders of Nigeria and geolocating the source of the tweets to states, and applying big data mining techniques, including sentiment analysis on the data to understand how effective both campaigns have been in getting their messages across and how their respective candidates match up. Below are the key things learnt from monitoring tweets regarding the presidential campaigns.

4.1 Output Screen

Figure 6 shows the output screen.

4.2 The Incubent President fate

In the last election the presidential candidate for the PDP, President Goodluck Jonathan rode on a lot of goodwill from the people, and some would argue good luck as well, to clinch the presidency. However this time around, the reverse happened. I used sentiment analysis to get an idea of how many people saying positive or negative things about the candidates and found that for the President, more people tweeted negative messages than positive messages about him, it is found exactly the opposite

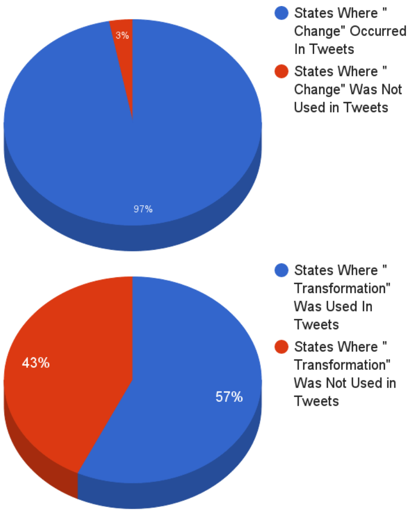


Fig. 7. Campaign message CHANGE Capaign message TRANSFORMATION

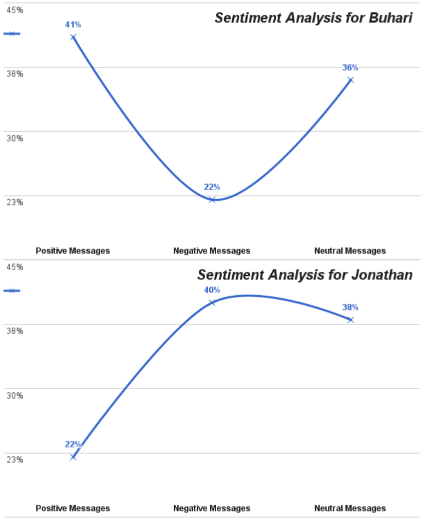


Fig. 8. Campaign message CHANGE Capaign message TRANSFORMATION

for his main opponent, Muhammadu Buhari, with more people posting positive messages about him than negative ones.

4.3 Political issue discussed

When it came to discussing issues around the present electoral process, the most prickly subject for people was the issue of corruption. It was by far the most talked about political issues, being mentioned more than twice as much as the next talked about issue, security.

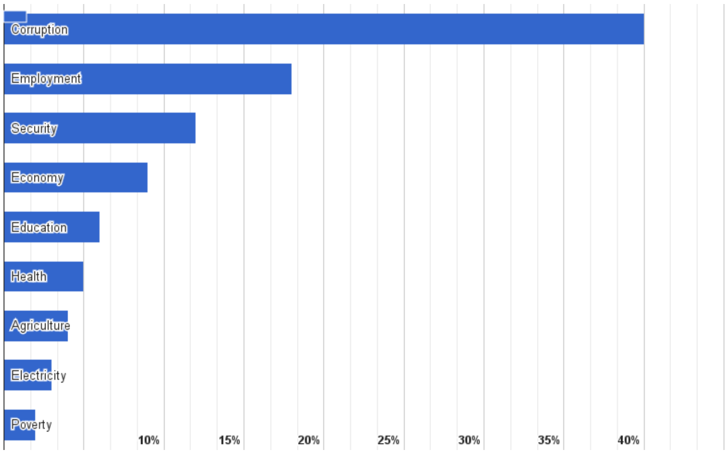


Fig. 9. Political issues discussed

4.4 Postponement of the Inevitable

It was speculated than the real reason behind the postponement of the elections from 14th of February by 6 weeks to the 28th of March was to blunt the momentum of the Buharis campaign and claw back some advantage by for President Jonathans campaign. However analysis of the tweets posted during the period leading to 14th February and those posted after that to the new date of the elections showed very little change in the support for either candidate.

5 Discussions and Conclusions

A common thread amongst those reporting about the elections is those who cast their votes wanted to ensure that their votes counted. There were stories written about people waiting for long hours in the rain and sun to cast their votes. In some cases counting of the votes were done at night using touch lights or generator power to ensure that the electoral process was not disrupted and that the results of the polls were not manipulated. Independent observers generally described the 2015 general election as free and fair. One of the reasons behind this project was to see if we could determine how people would vote in the upcoming elections by analyzing Twitter data and see if social media could be a good yardstick for predicting subsequent elections in Nigeria.

While the results and findings proved the proposed reference architecture relevant and provides good utility, there are still a number of limitations, which should be noted. As previously discussed, an important limitation in this research project is in establishing absolute relationship between twitter and

INECs dataset (SCR and VR). Due to the following reasons

- It was difficult to find geographical coordinates of registration areas in Nigeria and in some case tweets were not found in some of the local governments. Reasons being that some twitter users do not share their locations on Twitter.
- Not exactly those who registered to vote in an area actually voted during election. Secondly, significant phrases and annotation extraction on lexical variations of the word(s) in tweets is not fully implemented. Lexical variations simply means different variant of words on tweets. This was seen in some tweets as twitter has 140 character limit

Thirdly, this work was not designed to answer the research question by investigating other defining characteristics of the population such as political affiliation, gender orientation, or marital status, to name just a few. Therefore, the results should not infer opinions of these groups outside the scope of this study. While these are important designations that presidential political campaigns rely on in formulating their message and targets, focusing to address these characteristics was outside the scope of this study.

In conclusion, from some of the results that we have reported it was clear who Nigerians on Twitter wanted to be their President. On a national level, more people were likely to use the APCs main message of Change in their tweets about the election, more people were likely to post positive messages about the APCs candidate, Buhari, as well as post less negative messages in general. While more people had negative views of Jonathan than they did for Buhari. Buhari ran an anti-corruption campaign, which incidentally happened to be most important issue to those discussing the elections on Twitter. Along with the next 3 top issues Employment, Security and Economy.

This work delivers a mechanism to facilitate NER in extracting the annotation, Sentiment analysis and significant phrase identification in a text within geographic coordinates (Nigerian states) on messages accessed from twitter, which is readily available to use as a framework for future elections.

REFERENCES

[1] Collins O Oneamaobi, Election data Analytics: Analysis of 2015 elections in Nigeria, Masters degree thesis, Nile University of Nigeria, 2016

[2] Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M., 2011. Election forecasts with Twitter How 140 characters reflect the political landscape. Social Science Computer Review.

[3] Savage, N., 2011. Twitter as Medium and Message. Communications of the ACM, 54(3).

[4] Gayo-Avello, D., 2011. Don't Turn Social Media into another 'Literary Digest' Poll. Communications of the ACM, 54(10).

[5] Bilal, M., Israr, H., Shahid, M., Khan, A., 2015. Sentiment classification of Roman-Urdu opinions using Naive Bayesian, Decision Tree and KNN classification techniques. Journal of King Saud University Computer and Information Sciences, pp. 1319-1578.

[6] Fersini, E., Messina, E. and Pozzi, F., 2014. Sentiment analysis: Bayesian Ensemble Learning. Elsevier, Volume 68, pp. 26-38.

[7] Bhadane, C., Dalal, H., Doshi, H., 2015. Sentiment analysis: Measuring opinions. International Conference on Advanced Computing Technologies and Applications (ICACTA-2015), pp. 808-814.

[8] Evelien D'heer, P. V., 2014. Conversation about the elections on Twitter: Towards a structural understanding of Twitter's relation with the political and the media field. European Journal of Communication, 29(6) (ejc.sagepub.com), pp. 720-734.

[9] Heli Aramo-Immonen, J. J. J. H., 2015. Exploring co-learning behavior of conference participants with visual network analysis of Twitter data. ELSEVIER, 51(Computers in Human Behavior), pp. 1154-1162.

[10] Aparaschivei, P. A., 2011. The Use of New Media in Electoral Campaigns: Analysis on the Use of Blogs, Facebook, Twitter and YouTube in the 2009 Romanian Presidential Campaign. Journal of Media Research, 2(10), pp. 39-60.

[11] Caroline J. Tolbert, D. A. S. J. C. G., 2009. Strategic Voting and Legislative Redistricting Reform: District and Statewide Representational Winners and Losers. Political Research Quarterly, 62(Number 1), pp. 92-109.

Steve A. Adeshina Biography text here.

Collins Onyemaobi Biography text here.

Adegboyega Ojo Biography text here.