OPEN ACCESS Journal

International

Of Modern Engineering Research (IJMER)

A Simple Opinion Mining On Some Likely Nigerian Presidential Candidates in the 2019 Elections

[1]Baale A. Adebisi, [2]Adegun A. Adekanmi, [3]Chinedu M. UchechukwuAnd [4]Adelodun F. Ojiyovwi

1 Computer Sc. &Engr. Dept., LadokeAkintola University Of Technology, Ogbomoso, Nigeria:
2,3 Computer Science Dept., Landmark University, Omuaran, Nigeria.
4 Computer Studies Dept., The Polytechnic, Ibadan, Nigeria.

ABSTRACT: *Opinion can be defined as a view or judgement formed about something or someone, not necessarily based on fact or knowledge. Opinion mining, sometimes called sentiment mining or sentiment analysis is a type of natural language processing for tracking the mood of the public on a particular object. Most of the time, the success or failure of a candidate in an election to a public position is a direct reflection of the polarity of the opinions by the public involved. Using twitter data set, this paper attempts to analyze the opinions of Nigerians on some likely presidential candidates (Muhammadu Buhari, AtikuAbubakar, RabiuKwankwaso and Ayo Fayose) in the country's 2019 presidential elections. Tweetswere manually annotated as positive, negative or neutral by human evaluators for better classification speed and accuracy as described by Mozetic, Grcar and Smailovic, 2016. The labelled tweets were used to train the Naïve Bayes Classifier which was then used to classify new tweets for the sentiment analysis. The classified twitter data is displayed using pie charts. The results of the analyses showed Buhari had the highest % of tweets over the period. Regarding the 2019 presidential elections, Atiku had the lowest % of negative opinions and the highest % of positive opinions.*

KEYWORDS: *Opinion mining, Naïve Bayes classifier, Sentiment Analysis, Nigeria, Elections, Politics*

Date of Submission: 27-07-2018 Date of acceptance: 11-08-2018

I. INTRODUCTION

Opinion Mining or Sentiment Analysis can be said to be a series of methods, techniques, and tools about detecting and extracting subjective information, such as opinion and attitudes, from language (Liu, 2009). The explosion of the use of the social media all over the world has made the automatic collection and analyses of the social media data for the discovery of interesting patterns and information over these data sets possible. The impact of the social media has never been felt greater than in the 2014 presidential elections in Nigeria when it was used as a major tool in the presidential campaigns of one of the candidates, Muhammadu Buhari who later won the elections.This has created the awareness of the importance of the various use of the social media as a means to an end, especially in politics.

The term Sentiment Analysis or Opinion Analysis was originally coined by Nasukawa and Yi in 2003. The topic in question can either represent an individual, event or topic. Sentiment Analysis generally involves the process of relating syntactic structures from the levels of phrases, clauses, sentences, and paragraphs to the level of the writing and their language-independent meanings. Traditionally, Sentiment Analysis has been about opinion polarity, i.e., whether someone has positive, neutral, or negative opinion towards something (Dave, Lawrence, andPennock, 2003). Opinion Mining could also be defined as the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes (Liu and Zhang, 2005). The process of Opinion Mining requires an understanding of lexical hierarchy including but not limited to synonyms, antonyms, and sarcasm. It also relates to concepts like connotation and collocation, which is the combination of words that can be or frequently are surrounding a single word.

Sentiment Analysis is used in the context of computer science with data mining processes to gather data, process it and receive information arranged on a set of parameters (positive, negative or neutral). Social media is a very important tool for engaging people and gathering personal opinions due to its popularity and the rising growth of internet users. Social media sites such as Facebook, Twitter, Instagram, and Snapchat have become popular mediums for opinion analysis due to massive user interaction. Twitter's microblogging platform is being used in this project due to itshaving a large amount of relevant data, better public opinion, and overall better user interaction as compared to other conventional blog sites and social networking sites. Due to its limit to 120 characters, classification can be done easily and more accurately to deliver better results.

II. LITERATURE SURVEY

The evolutionary process of "social networks" can be traced back to the late 1970s when Bulletin Board Systems (BBS) were in vogue. Popular platforms for BBS included vBulletin and PhP BB. These BBS allowed data and software to be uploaded and downloaded; reading news and bulletins and exchanging messages with other users through emails and public message boards were also possible. Forums were the direct descendants of BBS and they, in turn, played an important role in the evolution of social networks (Edosomwam, 2011).

A social network can be defined as "a group of Internet-based applications that build on the ideological and technological foundations of web 2.0 and allow the creation and exchange of user-generated content" (Kaplan, Haenlein 2010).

Twitter, founded in 2006 by Jack Dorsey, is a social networking service, which allows users to communicate with each other by messages consisting of a maximum of 140 characters, called Tweets (Bryl, 2014). Twitter users can do a retweet or resend messages sent by other users and write messages by topic using the sign # (hashtag)*topic*. Growth in the number of Twitter users has risen rapidly, as there are now over 500 million Twitter users. Its use in general increases dramatically during popular events, so it appears at the Trending Topic feature. The high popularity of Twitter has made the service to be used for various purposes in various aspects, for example as a tool to convey news, political campaigns, learning, and as an emergency communication media.

Sentiment Analysis on Twitter in general aims to determine users' response to a topic which can either be positive or negative (Luce, 2012). To determine the polarity of the user's response, Text Classification is performed on the tweet using various techniques. Existing notable applications include the use of text mining techniques (Supervised or Unsupervised) to predict or give certain insights into the twitter data. Some of the common techniques used include;

Support Vector Machines (SVM): They are discriminate classifiers formally defined by a separating hyperplane which is efficient for text categorization. SVMs were developed from statistical learning theory by (Vapnik and Vapnik, 1998) on the basis of the structural risk minimization principle. The algorithm classifies opinionated text vectors by separating it into positive and negatives classes with a hyperplane, which can be further extended to non-linear decision boundaries using the kernel trick (Sharma andDey, 2012).

Naïve Bayes (NB): The Naïve Bayes classifier is a probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naïve) independent assumptions. NB classifiers have worked unexpectedly well in many complex problems or situations. The Naïve Bayes Classifier equation is shown below;

Likelihood　　Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\, P(c)}{P(x)}$$

Posterior Probability　　Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Explanation of the equation:

- P (c | x) is the class posterior probabilities (c, target) given predictor (x, attributes).
- P (c) is the class prior probability.
- P (x | c) is a likelihood that indicates the probability of a certain class predictor.
- P (x) is the predictor prior probability.

Naive Bayes models are easy and quick to predict the class of the test data set and have also performed well in a multi-class prediction. When there are independent assumptions, a Naive Bayes classifier performs better than the other models such as logistic regression and the need training data less and have also performed well in the case of the input category variable compared with a numerical variable.

Maximum Entropy: This makes no assumption regarding the relationship in between the features extracted from dataset. The classifier always tries to maximize the entropy of the system by estimating the conditional distribution of the class label (Khard and Sonawane, 2016)

Decision Tree: A decision tree is a tree in which each branch node represents a choice between many alternatives and each leaf node represents a decision (Dunham, 2006). Decision trees are easily interpretable because the tree structure can be represented graphically and we can follow branches down the tree according to the input variables, requiring less training time.

K-Nearest Neighbor Classifier: The K-Nearest Neighbor (KNN) classifier is an instance-based classifier that relies on the class labels of training documents which are like the test document. Thus, it does not build an explicit declarative model for the class. KNN classifications proceed in two stages; the first determines the nearest neighbor's and the second determines the class using those neighbor's (Cunningham and Delanay, 2007).

Zhang, Fuehres and Gloor, 2011 predicted the results of stock market indicators such as the Dow Jones, NASDAQ and S&P 500 by analyzing Twitter posts.

Rao, Yarowsky, Shreevats and Gupta, 2010 did an exploratory study of Twitter user attribute detection which uses simple features such as n-gram models, simple sociolinguistic features such as the presence of emoticons, statistics about the user's immediate network such as the number of followers/friends and retweet frequency as communication behavior.

Kaul, 2015 used the Agenda Detector to label tweets with 'has and has not' political policy agenda using the Decision Tree (J48), Naïve Bayes and Support Vector Machine techniques.

Parikh and Movassate, 2009classified tweets using two Naive Bayes unigram models, a Naive Bayes bigram model and a Maximum Entropy model and concluded that the Naive Bayes classifiers worked much better than the Maximum Entropy model.

Kiruthika, Sanjana and Priyanka Giri, 2016 analyzed tweets about movies into negative, neutral and positive using supervised learning approach.

Tighe and Cheng, 2018 modelled Personality Traits (Five Factor Model) of Filipino Twitter Users using multiple learning algorithms including Linear Regression, Ridge Regression, linear Support Vector Machines and Logistic Regression.

III. THE PROPOSED SYSTEM

The proposed system models a semantic analysis system based solely on sentiment classification to group tweets into three different categories (positive, negative and neutral) using the Naïve Bayes classification technique. The system hybridizes manual classification and automated training sets. Data from the Twitter REST API collected over a period of 30 days were processed and passed into a set of comma separated values (.csv) format files which were used as a training set for the classifier. Information derived from the training set (through Naïve Bayes algorithm) is then used to classify tweets based on their sentiments. The proposed system was implemented with the Python programming language. The Django python framework was used to port the system into a web application.

3.1 Data Acquisition

The tweets used (30,000) for the training set were acquired from the Twitter REST API between December, 2017 and January, 2018usingOAuth. The tweets used (65,547) for the analysis were collected between January and March, 2018. The dataset consists of tweets regarding or related to the topics e.g. *Buhari, Fayose, Atiku + 2019 Electionse.t.c.*

3.2 Data Preprocessing

The data pre-processing was done using *Tweepy*and NLTK python libraries. Here, HTML Links, punctuations, unnecessary slangs, tweets written in different language, emoticons, hashtags, stops are eliminatede.t.c.This is done in order to make the data ready for mining by presenting the data in terms of words that can easily classify the tweets. Anexample is shown in Table 1 below;

Table 1: Table describing a pre-processed tweet

Tweet type	Result
Original tweet	@abc I hate how during each dec. period dere is always fuel scarcity #scarcity
Pre-processed tweet	Hate, period, always, scarcity, angry

3.3 Human Labelling

The contents of the tweets were labelled manually without bias or assumption, grouped into sentiments (positive, negative), neutral, ambiguous and irrelevant categories. The ambiguous and irrelevant tweets were discarded while the sentiments i.e. positive and negative are then grouped into two separate files. Table 2 below shows some examples of human labelling;

Table 2: Table describing manual labeling

Extracted Tweets	Class
We Can No Longer Laugh With Buhari -Governor Nasir ElRufai	NEGATIVE
Pls act now rescue Nigerians The suffering is too much Thank God wizkid helped one yesterday	NEUTRAL
FUEL SCARCITY Petroleum Minister Muhammadu Buhari Hides In Shame As Marketers Finally Speak Out Set The Record	NEGATIVE
NigeriaNewsdesk Ex-VP AtikuAbubakar hails UcheSecondus emergence as PDP national chairman	POSITIVE
Loving myself even in the hash economic situation under President Buhari	POSITIVE
Pres Buhari "I thought I was 74 years old but I was told I am 75	NEUTRAL
What Buhari's birthday messages to Saraki and Dogara say about their relationship	NEUTRAL

3.4 Polarity Testing

Polarity test involves using one or multiple classification techniques with the trained dataset to test new sets of tweets and determine positive, negative, neutral and ambiguous tweets. It is used to train the system by automating the classifier (Naïve Bayes) with the dataset. Polarity test involves the following processes:

- Extracting word features from a text file that contains a set of positive and negative tweets.
- Training dataset with the results of feature and tweets extraction.
- Classifying the training set using the Naïve Bayes classifier.
- Extracting the features of new tweets.
- Classifying the new tweets based on a training set that has already been classified.
- Automated classification into either positive or negative sentiments. Figure 1 below shows the workflow for the training set.

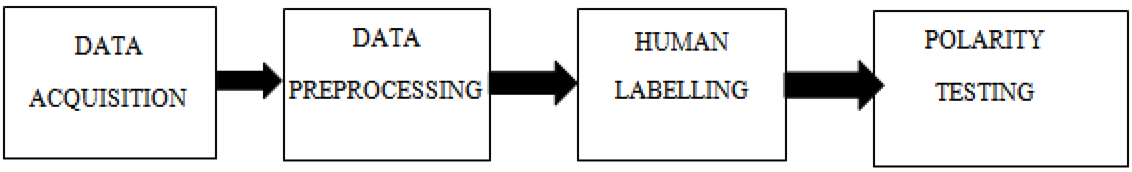

Figure 1: Workflow for the training set

3.5 Classification

The Naïve Bayes classifier was trained with the training set from the human labelled tweets and the NLTK library, after which it was used to classify the new tweets into positive and negative tweets to produce overall sentiment opinions for each of the proposed candidate. Classification is done by performing semantic processing on tweets whereby tweets are categorized based on their overall average score.

3.6 System Architecture

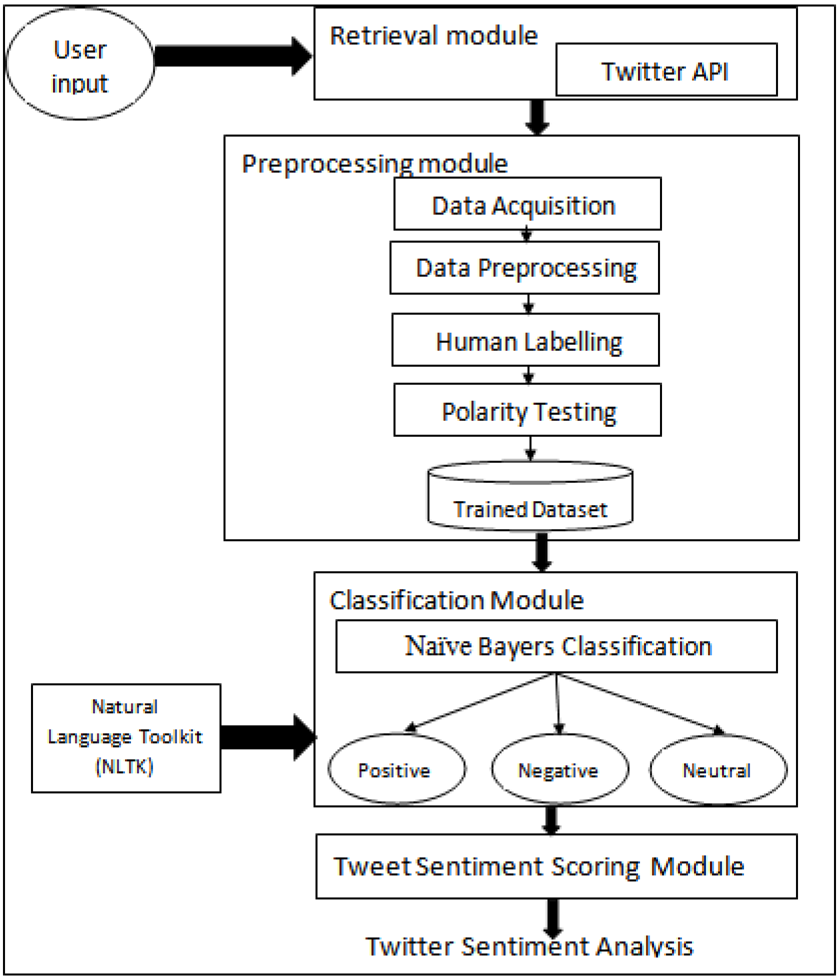

Figure 2: System architecture diagram

The proposed system architecture as depicted in figure 2 above is divided into four modules which include:
i. Retrieval module.
ii. Preprocessing module.
iii. Classification module.
iv. Sentiment scoring module.

Retrieval module: the retrieval module describes all functions relating to the acquisition of data which would be used by the sentiment analysis system. This module receives input from the user by way of topic, keyword or hashtag and sends it to the twitter analysis system (i.e. using the *tweepy* python library with collaboration with OAuth). This system then receives a set of twitter data based on the user's preference and sends it to the preprocessing module in JSON () format using the Twitter RESTful (Representational State Transfer) API.

Preprocessing module: the preprocessing module performs the basic function of processing twitter data before the classification process can begin. This module can be broken further down into five submodules. Data acquisition collects the twitter data generated from the retrieval module in JSON format and converts it into a user readable format (.csv). The next step involves manual labeling of preprocessed tweets, this is done by grouping tweets into four categories; positive, negative, neutral and ambiguous. Polarity testing is then done by comparing the manually labeled tweets with a new set of tweets to test for accuracy. After this is done, the preprocessed dataset is stored in a database.

Classification module: This module describes the process of classifying new tweets generated based on the trained dataset from the classification module using the Naïve Bayes classification technique which uses a probabilistic learning method to compare new tweets with those in the trained dataset. The natural language toolkit is used for natural language processing through stemming and tokenization of retrieved tweets. It is used in collaboration with the Naïve Bayes classification algorithm to reduce language processing time and increase accuracy.

Sentiment scoring module: This is the final module from which the twitter sentiment analysis system is built on. The scoring model assigns a base score between 0 to 1 to each word whereby 0 indicates a complete negative word while 1 indicates a complete positive word. Several words in each tweet are then analyzed based on this module and the overall sentiment of the tweet can then be gotten.

IV. RESULTS

4.1 Screenshots of the application

The homepage is shown in figure 3 below: this is the landing page of the sentiment analysis system. The search bar provides a space whereby topics, personnel, hashtags, and words can be analyzed by the system. The count of Twitter data is also specified on this page.

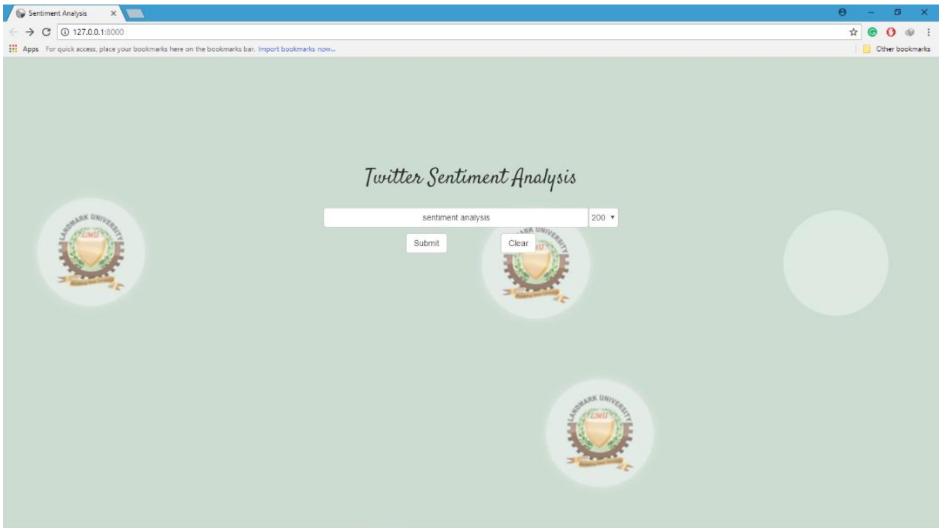

Figure 3: Homepage illustration

The sentiment graph window is shown in figure 4 below: After sentiment analysis has been performed based on user requirements, the application program then transfers the analyzed data into a new page with a pie chart describing the overall sentiment of the analyzed tweet.

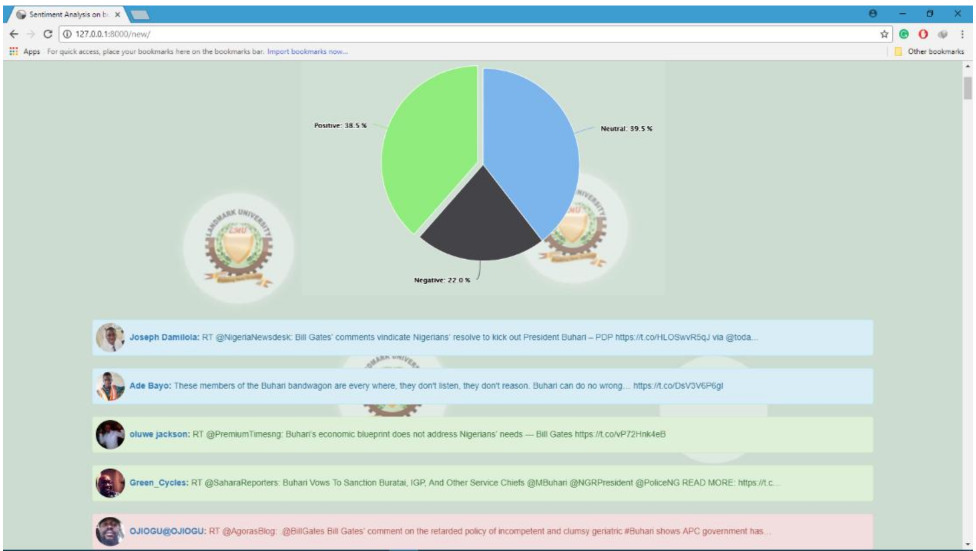

Figure 4: Sentiment graph window

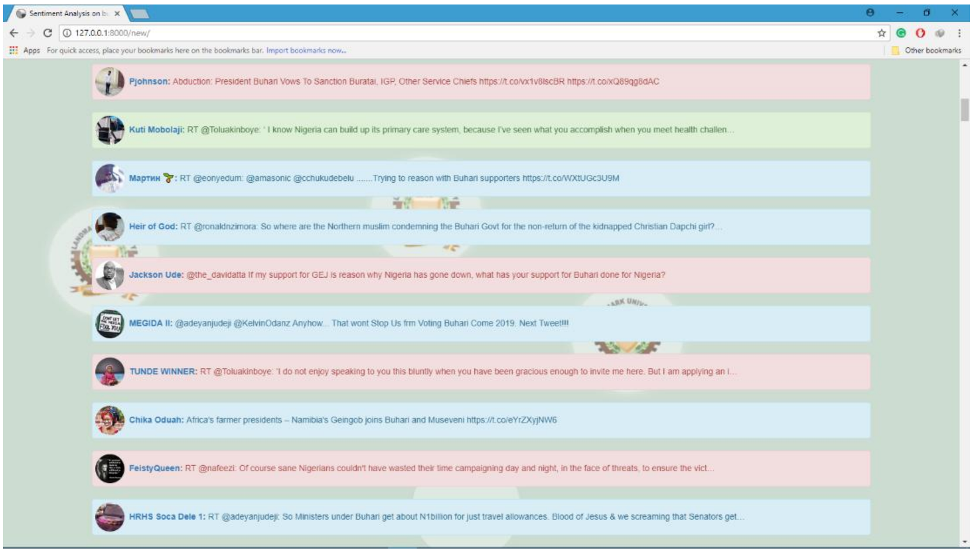

Figure 1: Analyzed tweet window

Analyzed tweet window: This page displays the analyzed tweets processed by the twitter sentiment classifier. It also groups each tweet according to its sentiment which can either be positive, negative or neutral/ambiguous.

4.2 Analyses

Over fifteen thousand tweets over a course of 30 days were gathered on the topic. From those tweets, four major candidates were derived based on text analysis. They include;

i. Muhammadu Buhari.
ii. RabiuKwankwaso.
iii. AtikuAbubakar.
iv. Ayo Fayoshe.

Muhammadu Buhari, AtikuAbubakar, RabiuKwankwaso and Ayo Fayose had 40%, 31%, 19% and 10%, respectively of the total tweets. Sentiment analysis was then performed on the processed tweets based on their individual personality and in relation with the 2019 Presidential Elections and results for each candidate shown in figure 6 below:

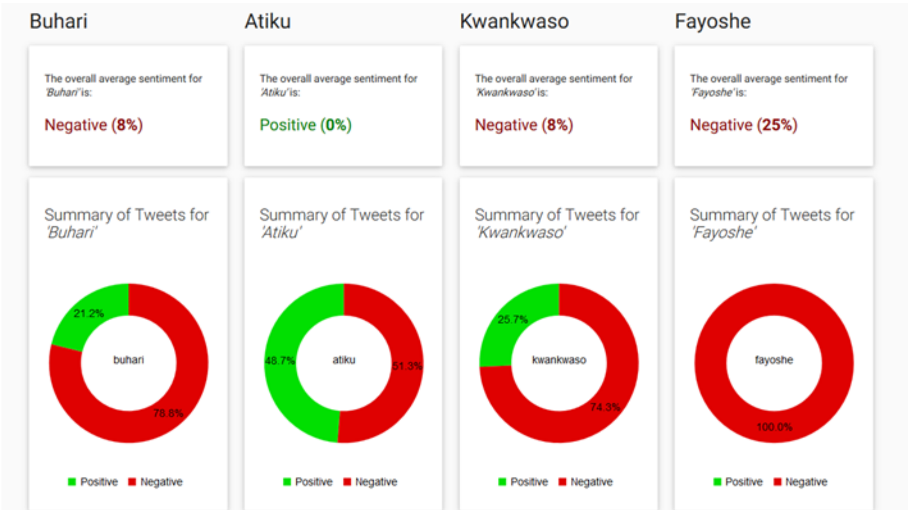

Figure 6: Chart comparing all candidates together

Table 3: table describing the result of the sentiment analysis system

S/N	NAME	Overall positivity score	Tweets regarding 2019 elections		Tweets regarding each candidate		Percentage of tweet count
			NEG.	POS.	NEG.	POS.	
1.	Muhammadu Buhari	-8.0%	78.8%	21.2%	22.0%	38.5%	40.0%
2.	AtikuAbubakar	0%	51.3%	48.7%	10.0%	18.0%	31.0%
3.	RabiuKwankwaso	-8%	74.3%	25.7%	6.0%	19.0%	19.0%
4.	Ayo Fayoshe	-25%	100%	0%	10.0%	22.0%	10.0%

V. DISCUSSION

From the summarized results in table 3, the following information concerning the proposed four major candidates of the 2019 Nigerian presidential elections can be derived;

1. **Muhammadu Buhari:** From the sentiment analysis performed on topics concerning the presidential elections, Muhammadu Buhari has the highest number of tweet mentions (40%) over the other candidates. When compared with the other presidential aspirants, he has a positive-negative ratio of 21.2% to 78.8% with a general positivity score of -8.0% (8% negative). Although he has a general negative sentiment score, Buhari manages to account for almost half of tweet mentions concerning the presidential election which makes him a slight favorite over the other presidential candidates. Muhammadu Buhari has a personal twitter positive to negative ratio of 38.5% to 22.0%.
2. **AtikuAbubakar:** Based on the result table above, Abubakar accounted for about 31.0% of tweet mentions in relation to the 2019 general presidential elections with a positive to negative ratio of 48.7% to 51.3% with a general positivity score of 0%. Given that he has the second largest amount of tweet mentions with a more positive sentiment ratio than that of Muhammadu Buhari, Abubakar has a very good chance of putting up a challenge against the incumbent president. AtikuAbubakar has a personal twitter positive to negative ratio of 18.0% to 10.0%.
3. **RabiuKwankwaso:** He accounts for about 19.0% of all tweet mentions in regard to the 2019 presidential elections in Nigeria. RabiuKwanwaso has a general positive to negative ratio of 25.7% to 74.3% with a general positivity score of -8% (8% negative). In spite of the fact that he has the second highest positivity score in relation to the elections, he still stands a distant third due to the fact that he has a relatively low tweet mention count with a general negative ratio of 74.3%. However, RabiuKwanwaso has a personal positive to negative sentiment score of 19.0% to 6.0%.
4. **Ayo Fayoshe:** Ayo Fayoshe represents only 10.0% of all tweets mention in relation to the 2019 presidential elections. He also has the lowest positive to negative ratio of 0% to 100% with a general positivity score of -25% (25% negative). Although he has a personal positive to negative sentiment ratio of 22.0% to 10%, he is still the underdog in the presidential race due to his low tweet count number and negative sentiment percentage, especially when compared to the other candidates.

VI. CONCLUSION

From the results shown above, we can deduce that Muhammadu Buhari along with AtikuAbubakarare the major frontrunners for the position of the president of the Republic of Nigeria in the next election period with RabiuKwanwaso close behind them. However, Ayo Fayoshe is the lame duck in the race with him having the lowest chance of winning the elections.

REFERENCES

[1]. Bryl S. (2014): Twitter Sentiment Analysis With R. Diperoleh 30 September 2016. Retrieved from http://analyzecore.com/2014/04/28/twitter-sentiment-analysis.

[2]. Cunningham P. and Delany S. J. (2007): “k-nearest neighbor classifiers”, Technical Report UCD-CSI-2007-4, pp. 1–17.

[3]. Dave K., Lawrence S. and Pennock D. M. (2003): “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews”. In Proceedings of the 12th International Conference on World Wide Web (WWW), pp. 519–528.

[4]. Dunham M. H. (2006): Data mining: Introductory and advanced topics. Pearson Education

[5]. India.

[6]. Edosomwan S., Prakasan S. K., Kouame D., Watson J. and Seymour T. (2011): “The history of

[7]. social media and its impact on business”. Journal of Applied Management and Entrepreneurship, vol. 16, no. 3, pp. 79–91.

[8]. Kaplan A.M. and Haelein M. (2010): “Users of the world, unite! The challenges and opportunities of social media”. Business Horizons, vol. 53, no. 1, pp. 59-68.

[9]. Kaul S. (2015): Agenda detector: Labelling tweets with political policy agenda. An M.Sc. Theses submitted to the Program of Study Committee, Iowa State University, Ames, Iowa.

[10]. Kharde V. A. and Sonawane S. S. (2016): “Sentiment Analysis of Twitter Data: A Survey of Techniques”. International Journal of Computer Applications: vol. 139, no.11, pp. 5-15.

[11]. Kiruthika M., Sanjana W. and Priyanka G. (2016): “Sentiment Analysis of Twitter Data”. International Journal of Innovations in Engineering and Technology. vol. 6, no. 4, pp. 264-273.

[12]. B. Liu (2009): Handbook Chapter: Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing. Marcel Dekker, Inc. New York, USA.
[13]. Liu B. and Zhang L. (2005): “A survey of opinion mining and sentiment analysis”. In Proceedings of Recent Advances in Natural Language Processing.
[14]. Luce L. (2012): Twitter Sentiment Analysis using Python and NLTK. Retrieved from http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk
[15]. Mozetic I., Grcar M. and Smailovic J. (2016): “Multilingual Twitter Sentiment Classification: The Role of Human Annotators. PLoS ONE. vol. 11, no. 5, pp. 1–26.
[16]. Nasukawa T. and Yi J. (2003): “Sentiment Analysis: Capturing Favourability using Natural Language Processing”. In Proceedings of the Second International Conference on Knowledge Capture: pp. 70-77.
[17]. Parikh R. and Movassate M. (2009): “Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques", CS224N Final Report, 2009.
[18]. Rao D., Yarowsky D., Shreevats A. and Gupta M. (2010): “Classifying latent user attributes in
[19]. twitter”, In Proceedings of the 2nd International workshop on Search and Mining user-generated contents, ACM, 2010: pp. 37–44.
[20]. Sharma A. and Dey S. (2012): “A comparative study of feature selection and machine learning
[21]. techniques for sentiment analysis”, In Proceedings of the 2012 ACM Research in Applied
[22]. Computation Symposium: pp. 1–7.
[23]. Tighe E. P. and Cheng C. K. (2018): “Modeling Personality Traits of Filipino Twitter Users”. In Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality and Emotions in Social Media, New Orleans, Louisiana: pp. 112-122.
[24]. Vapnik V. N. and Vapnik V. (1998): Statistical Learning theory: vol. 1. Wiley, New York.
[25]. Zhang X., Fuehres H. and Gloor P. A. (2011): “Predicting Stock Market Indicators through Twitter “I hope it is not as bad as I fear””. Procedia-Social and Behavioral Sciences: vol. 26, pp. 55–62.

1Baale A. " A Simple Opinion Mining On Some Likely Nigerian Presidential Candidates in 2019 Elections. “International Journal Of Modern Engineering Research (IJMER), vol. 08, no. 07, 2018, pp.60-68.