

DATA MINING OF NIGERIANS' SENTIMENTS ON THE ADMINISTRATION OF FEDERAL GOVERNMENT OF NIGERIA

¹L. B. Amusa, ¹W. B. Yahya, ²A. O. Balogun

¹Department of Statistics, University of Ilorin, Ilorin, Nigeria

²Department of Computer Science, University of Ilorin, Ilorin, Nigeria

Corresponding Author: L. B. Amusa, amusasuxes@gmail.com

ABSTRACT: The opinions and sentiments expressed by citizens of a country on the policies of the government of such country are very vital to the overall running of the affairs of such a government. This paper therefore explored data mining tools to evaluate peoples' sentiments (positive or negative) towards the administration of the Federal Government of Nigeria (FGN) under President Muhammadu Buhari (PMB). Data were collected through a popular social medial network (Twitter) on various tweets by Nigerians with respect to their perceptions about the current administration PMB. The simple but powerful Naïve Bayes (NB) classifier was adopted to classify the various tweets submitted by Nigerians through this medium into positive and negative sentiments. For polarity, it was trained on the combination of Janyce Wiebe's subjectivity lexicon and Bing Liu's subjectivity lexicon which polarized the submitted words as being negative or positive. Out of about 13,000 features (peoples' sentiments) considered, 4,770 of them were used after data cleaning. The results showed that the proportion of positive and negative sentiments, as obtained from the data, were 45.2% and 54.8% respectively. However, the data were randomly partitioned into 80:20 training and testing parts respectively and the NB classifier was learned on the training set while its goodness was assessed on the test set. The prediction accuracy, misclassification error rate, sensitivity and specificity of the classifier were 78.3%, 21.7%, 82.5% and 88.1% respectively. All analyses were carried out in the environment of R statistical package (version 3.2.2).

KEYWORDS: Naive Bayes, Sentiment, Twitter, Text Mining, Polarity.

1. INTRODUCTION

With the explosive growth of printed data from the electronic archives and World Wide Web, legitimate arrangement of such tremendous measure of data to meet our requirements is a basic stride towards many business achievements. As of late, various examination exercises have been directed in the field of report grouping, especially applying in spam separating, messages ordering, site characterization, development of information archives and metaphysics mapping.

In any case, the time has come devouring and work escalated for human to peruse over and effectively order an article physically. In an attempt to address

this, programmed record characterization studies are increasing and more interests are now in the direction of content mining research as of late.

Subsequently, an expanding number of methodologies have been produced for fulfilling such reason, including k-nearest neighbor (KNN) ([Yah12]), Naive Bayes (NB), Support Vector Machines (SVM) ([B+15]), decision tree, neural network ([YOJ12]) and maximum entropy classifiers. Among these methodologies, the Naive Bayes content classifier ([Ris01]) has been generally utilized on account of its simplicity as a part of both the preparation and ordering stage. In spite of the fact that it is less exact than other discriminative techniques, (for example, SVM), various specialists demonstrated that NB is sufficiently powerful to characterize the content in numerous spaces ([CN06]). Naive Bayes models allow each attribute to contribute towards the final decision equally and independently from other attributes, in which it is more computational efficient when compared with other text classifiers.

The use of electronic media is increasing daily. For example, about 1.8 million Nigerians are using the Twitter platform to express their feelings on a lot of issues concerning the affairs of the government at all levels.

Recently, there have been many researches to monitor public opinion and social trends ([A+10]). They include election prediction using Twitter data ([BKY10]), monitoring of consumer on a certain brand ([HL04]), movie performance prediction using Twitter ([BAO14]), disease and disaster tracking using Internet information ([SOM10]), and unemployment benefit prediction using Internet search information ([DM09]).

The opinions and sentiments expressed by citizens of a country on the policies of the government of such country are very vital to the overall running of the affairs of such a government. As the adage says that time is money or even more valuable than money, therefore instead of spending times in reading and figuring out the positivity or negativity of text comments as submitted by individuals on social medial networks, a number of automated techniques can be explored for sentimental analysis.

This research work therefore aims at exploring data mining tools using the Naïve Bayes Algorithm to analyze the sentiments of Nigerians towards the current administration of President Muhammadu Buhari using their various tweets from Twitter (a popularly known social media platform) with respect to their feelings about the activities and performance of the government.

2. Materials and Methods

The data used for this study were tweets of Nigerians from Twitter with respect to their perceptions about the current administration of President Muhammadu Buhari. A total of 10,000 tweets were retrieved from the Twitter Application Programming Interface (API) from which 13,000 features were generated. A total of 4,770 features were used for analysis after data cleaning.

Given the nature of the data generated, some of the words might be useless (i.e. do not have predictive strength of the sentiments' group) while some words conveyed similar meanings (e.g the words "bank" and "banks"), therefore the dataset has to be preprocessed to filter out the useless features and unnecessary duplications.

After the data preprocessing phase, critical attributes have to be selected. In this study, *critical* means the importance of such attribute towards the response class. For example, the term "bank" categorized in business class has the highest score in term of *term frequency*, therefore it is conjectured that "bank" is one of the critical attributes to represent the documents that fell in the business class. Thus, less important features can be removed and so the computational time can be improved upon significantly.

As for the classification phase, different classifiers (such as SVM, K-NN, and Maximum Entropy) are usually employed to generate the model ([G+13]). However, this study only focused on the use of Naïve Bayes to classify the documents. Given the probabilistic characteristic of Naïve Bayes, each training document is vectorized by the trained Naïve Bayes classifier through the calculation of the posterior probability value for each member of the response classes. Finally, the goodness of the fitted model is evaluated on a set of test data. However, in order to test the classification ability of the model, several evaluation measures such as precision, Sensitivity and Specificity were adopted. Without loss of generality, the following phases were adopted to build the Naïve Bayes classifier on the data.

Phase 1: Preprocessing

This is the data clean up phase at which several useless attributes (words) like 'a', 'the' and so on were removed using a stopword removing algorithm.

To initialize the algorithm, a set of stop words such as *a, a's, able, about, above, according, accordingly, across*, etc. was set beforehand and hence stored in a text file. Therefore, the model can simply match the attributes with those preset stop words. After the stop word algorithm, a missing data checking algorithm was adopted. This algorithm was used to identify any missing data and hence interpret a value to it.

The third algorithm applied in the preprocessing phase is the *stemming*. Since some words carry similar meanings but in different grammatically form (such as "bank" and "banks"), therefore it is desirable to combine them into one attribute. In this way, the documents can show a better representation (with stronger correlations) of these terms. This would reduce the dimension of the data and the model would be more efficient for achieving faster processing time.

Phase 2: Feature Selection

Feature selection is an important phase in information mining. It is a successful dimensionality reduction system to expel noise feature.

All in all, the essential thought of the feature selection algorithm seeks through every conceivable blend of credits in the information to discover which subset of elements works best for forecast. In this manner, the quality vectors can be lessened in number by which the most important ones are kept and the immaterial or excess ones are expelled and erased. Vectors can be diminished in number by which the most significant ones are kept and the superfluous or excess ones are expelled and erased.

Phase3: Building the Naïve Bayes Classification Model

Due to preprocessing and feature selection, the numbers of attributes in the data will be significantly reduced and are more precise for use in building the classification model. The data were randomly partitioned into 80% training and 20% test sets. The Naïve Bayes classification model was built using the training set while its goodness in document and text classification was assessed on the test data. Each test samples in the new documents were classified into their right categories according to the highest posterior probability.

For polarity, it is trained on the combination of *Janyce Wiebe's subjectivity lexicon* ([WWH05]) and *Bing Liu's subjectivity lexicon* ([MB04]), and will polarize words as being 'negative' or 'positive' sentiments as expressed by individuals. Based on the mixture of positive and negative words, each tweet was assigned value within the range of -5 (*very negative*) and +5 (*very positive*).

Phase 4: Model Evaluation

To test and evaluate the fitted model, the remaining 20% of the dataset were used. By comparing the actual class of the instance with the predicted one (i.e. generated by the classification model), system performance was measured in terms of precision. Precision can be defined as:

$$\frac{\text{Number of correctly classified categories}}{\text{Total number of classified categories}} \times 100$$

Other performance measures such as sensitivity and specificity of the classifier were computed.

2.1. Naïve Bayes Methodology

For a document d and class c , by Bayes theorem we have that

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (1)$$

$P(c|d)$ is the posterior probability of class (target) given predictor (feature).

$P(c)$ is the prior probability of class c .

$P(d|c)$ is the likelihood which is the probability of predictor given class.

$P(d)$ is the prior probability of predictor

Thus, the Naïve Bayes classifier can be represented by the relation:

$$c^* = \arg \max_c P(c|d) \quad (2)$$

For Text Classification

Naive Bayes works as follows for classification of texts into positive (+) or negative (-) sentiments as follows:

1. Represent each document by vector of words.
2. Use training examples to estimate $P(+)$, $P(-)$, $P(\text{doc}/+)$ and $P(\text{doc}/-)$.

Naive Bayes conditional independence assumption

$$P(\text{doc}/v_j) = \prod_{i=1}^{\text{length}(\text{doc})} P(a_i = w_k/v_j) \quad (3)$$

where $P(a_i = w_k/v_j)$ is the probability that word in position i is in w_k , given v_j

Another assumption is:

$$P(a_i = w_k/v_j) = P(a_m = w_k/v_j), \text{ for all } i, m \quad (4)$$

A new sentence is then classified according to:

$$V_{NB} = \arg \max_{w \in \text{words}} P(v_j) \prod P\left(\frac{w}{v_j}\right) \quad (5)$$

where (w/v_j) stands for predicted probability value or class.

3. Data Analysis and Results

The goal of this study is to classify the given specified experimental dataset into two categories (i.e. positive and negative) correctly. In the data generation phase, tweets including term “Buhari Administration” in English are generated. The tweets search was restricted to the period from 29th May, 2015 to 29th May, 2016 with tweets being equally sampled from Twitter users from all the 36 states of Nigeria. A snapshot of the sampled dataset generated from Twitter social media network is presented as Figure 1.

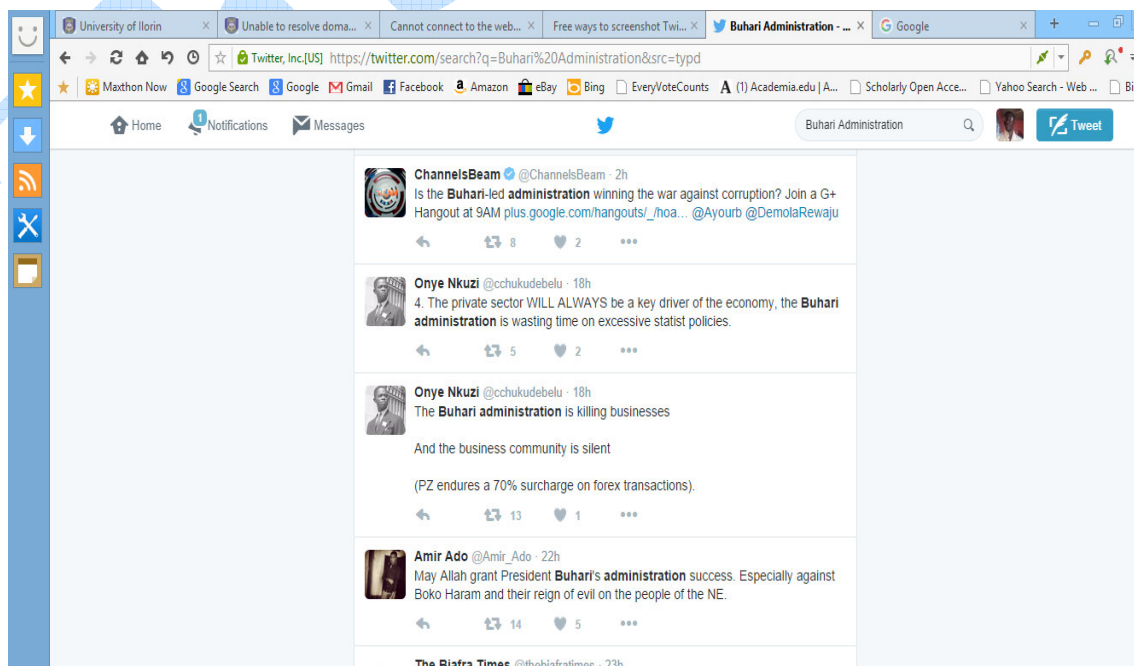


Fig. 1: Sample of tweets on Twitter Social Media Network System

Fig. 2 is the model of Entity Extraction. Data from stream is the input for the Entity Extraction module. There is a list of entities specified with respect to this particular case. The data from stream is then matched with the list and if there is a match then that entity becomes the entity for that particular data after which the sentiment analysis is performed. For this, the tweet goes through refining processes namely, stop word removal and noise cleaning. In stop word removal, words which do not add any meaning to the sentence, for example *of, the, is,* etc. were removed from the tweet. In noise cleaning,

characters such as *_@', _#*hashtags, extra white spaces and repeated characters are removed. After the refining process the features are extracted. The Naïve Bayes Classifier was trained for sentiment analysis with a training data set labeled with sentiments positive, negative or neutral. Then the sentiment classifier model labels the tweet with a sentiment. The Naïve Bayes classifier was built on the training data after the stop word removal, noise cleaning and feature extraction processes were completed on the input tweets.

text	polarity
changetochainsfulani herds men are the new killer squad of the bokoharam	negative
gmb has never address the nation regarding the killings carried out by fulani herdsmen dai	negative
a govt of propaganda by propagandist and rice eaterschangetochains	negative
i no go talk now sha but comemake una carry una sef comot aso rock rubbish changetochai	negative
nigeria according to apcs chrisngige those complaining of hardship areloyalists changetoch	negative
tyrantbuhari has lost control of the economy lai mohammed has lost control lies	negative
they have blamed everything for their administrative failures and now its time to dump it	negative
the best way to solve nigerias problems is to be recolonized again for anotheryears by brit	negative
oga when a flower doesnt grow you fix the environment in which it grows not the flower c	negative
nigerians were warnedhunger and hardship don setchangetochains	negative
its obvious that pmb is a curse to nigeria things are going from bad to worse daily changeto	negative
nnpc pengassas nupeng are all on strike at the same time	negative
even thetoday agreed that there is lawlessness in the land how can u invade a state house	negative
only in changetochains era can a hoa be invaded by the military	negative
buhari promised to build four new refineries in four years one year is almost coming to an	negative
changetochainspalestinians agitation for statehood must be respected but back home agit	negative
changetochains gdp growth ratefalls to a year low ofpercent – the lowest since the return	negative
the level of hardship experienced by nigerians now is alarming pple cant affordmeals per	negative
rd day of darkness ran out of generator fuel last night car fuel below half tankinto your har	negative
changetochains nigerians lashyou need to read these reactions	negative
notmyintention but what shall it profiteat a plat of rice and loss her consciencechangetoch	negative
under buhariapcled government millions of nigerians are suffering amp frustrated changeto	negative

Fig. 2: Sample of cleaned tweets with their respective polarity classification

The most important part of text analysis is to get the feature vectors for each document. Next is to create a document-term matrix. In the document-term matrix, as shown in Fig. 3, each document (on the

rows) is for a respondent, while each term (on the columns) are the unique words as identified from individual tweets. The number in each cell is the frequency of counts for each word per document.

Doc	love	fuel	Abacha	power	darkness	loot	corruption	dasuki	suffer	hate	change	refinery	unemployment	hope	chains	hardship	economy	forex	nnpc
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2	0	1	0	1	1	0	0	0	0	0	1	0	0	0	1	0	0	1
3	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	11	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
12	12	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
13	13	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
14	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	15	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
16	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	17	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	19	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
20	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	21	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
22	22	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
23	23	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
24	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 3: Snap-Shot of Document-Term matrix

A natural question to ask at this point is: how good is this prediction. This question cannot be answered with only a single run of the model; we need to do many runs and look at the spread of the results. The summary of statistics for correct predictions over 200 runs is shown in Table 2:

Table 2: Summary Statistics of Correct Predictions

Mean	80.5%
Minimum	72.7%
Maximum	82%
Standard Deviation	0.0258

4. DISCUSSION OF RESULTS

Based on the objective of performing sentiment analysis on drone regulation, we submitted the query “Buhari Administration” using twitter API. The API has a parameter that specifies the language to retrieve tweets in. We set the parameter to “English” to retrieve tweets in English only. We polled a total of 10,000 tweets for this study.

We have used the tweets obtained from Twitter to analyze sentiments expressed by Nigerians about the administration of President Muhammadu Buhari. At the end of the analysis, it was discovered that about 45% of the peoples’ sentiments were positive towards the policy and programs of the current administration of PMB while about 55% of these sentiments were negative.

The word cloud representation of the entire as shown by Fig 4 revealed that amongst other things, the words that kept creeping into people’s thoughts on President Buhari’s administration were on issues of *fuel* situations, *blame* game on many government policies, yet unfulfilled electoral *promises* as well as her *change* mantra. This result is not surprising because the issue of fuel scarcity for example generated a lot of controversies during the earlier periods of the current administration of PMB in Nigeria. Consequently, the blame-game started in which the government keeps apportioning the blame on some people including previous administrations. Ironically, many Nigerians are very curious and seriously yearning to experience the promised change amidst the obvious challenges that are confronting the current government.

The Naïve Bayes classifier has proven to be very efficient in text mining as shown by the results in this work. Results of classification of various sentiments as expressed by Nigerians showed that the NB classifier has overall average prediction accuracy of about 80% with a standard deviation of 0.0258. The estimated average overall sensitivity and specificity of this classifier were about 83% and 88% respectively which further justified the goodness of the NB model for text mining.

5. CONCLUSION

In this paper, application of a data mining methodology, the Naive Bayes, to monitor and classify the opinions of Nigerian citizens on the administration of President Muhammadu Buhari was undertaken. The procedure consist of four phases which include (1) generation of related tweets, (2) extraction of potential sentimental terms, (3) building of sentiment dictionary, and (4) tweets sentiment classification and evaluation of the classifier.

Naive Bayes, though simple, has once again proven to be a very good classifier for a text data like the one engaged here by achieving a good prediction accuracy of the sentiments class of about 80% relative to some other complex classifiers. It can be deduced from the results generally that although, more Nigerians were having a negative opinion about current administration of President Buhari, there are still a considerable number of people who have faith in his administration given a number her policies and programs.

In future study, the efficiency of more sophisticated classification methods like SVM, Maximum entropy and k-NN for text mining shall be examined and their relative performances shall be compared with the NB classifier adopted in this work.

REFERENCES

- [A+10] **C. G. Akcora, M. A. Bayir, M. Demirbas, H. Ferhatosmanoglu** - *Identifying Breakpoints in Public Opinion*. In: 1st Workshop on Social Media Analysis, pp. 62--66. Washington, DC, 2010.
- [BAO14] **H. M. Baek, J. H. Ahn, S. W. Oh** - *Impact of Tweets on Movie Sales: Focusing on the Time when Tweets are Written*. J. ETRI, 2014.
- [BKY12] **A. Boutet, H. Kim, E. Yoneki** - *What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election*. In: The International AAAI Conference on Weblogs and Social Media, 2012.
- [B+15] **A. W. Banjoko, W. B. Yahya, M. K. Garba, O. R. Olaniran, K. A. Dauda, K. O. Oloredo** - *Efficient Support Vector Machine Classification of Diffuse Large B-Cell Lymphoma and Follicular Lymphoma mRNA Tissue Samples*. Annals. Computer Science Series, 13(2): 69-79, 2015.

- [CN06] **R. Caruana, A. Niculescu-Mizil** - *An empirical comparison of supervised learning algorithms*. Proc. 23rd International Conference on Machine Learning, 161-168. [CiteSeer X10.1.1.122.5901](#). 2006.
- [DM09] **F. D'Amuri, J. Marcucci** – “*Google it!*” *Forecasting the US Unemployment Rate with a Google Job Search Index*. In: Conference on Urban and Regional Economics, 2009.
- [G+13] **G. James, T. Hastie, D. Witten, R. Tibshirani** - *An introduction to statistical learning* (Vol. 112). New York: Springer, 2013.
- [HL04] **M. Hu, B. Liu** - *Mining and summarizing customer reviews*, in Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, pp. 168–177, ACM, 2004.
- [MN03] **A. McCallum, K. Nigam** - *A comparison of event models for naïve Bayes text classification*, Journal of Machine Learning Research, Vol. 3, 2003, pp. 1265–1287.
- [Ris01] **I. Rish** - *An empirical study of the naive Bayes classifier*. IJCAI Workshop on Empirical Methods in AI, RC 22230 (W0111-014): 1-7, 2001.
- [SOM10] **T. Sakaki, M. Okazaki, Y. Matsuo** - *Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors*. In: 19th International Conference on World Wide Web, pp. 851--860. ACM, 2010.
- [WWP05] **T. Wilson, J. Wiebe, P. Hoffmann** - *Recognizing contextual polarity in phrase-level sentiment analysis*, in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLP/EMNLP), pp. 347–354, ACL, 2005.
- [Yah12] **W. B. Yahya** - *Genes selection and Tumour Classification in Cancer Research: A new approach*. Lambert Academic Publishing, Saarbrücken, Germany, 2012.
- [YOJ12] **W. B. Yahya, M. O. Oladiipo, E. T. Jolayemi** - *A fast algorithm to construct neural networks classification models with high-dimensional genomic data*. Annals. Computer Science Series, 10(1): 39-58, 2012.