

DESIGN AND IMPLEMENTATION OF A TWITTER SENTIMENT ANALYSIS TOOL: 2019 NIGERIA GENERAL ELECTION AS A CASE STUDY

Okeke, A. C.

Department of Computer
Science,
University of Ibadan, Nigeria
trent9002@gmail.com

Asoro, O. B.

Department of Computer
Science,
University of Ibadan, Nigeria
raymondblessing5@gmail.com

Osunade, O.

Department of Computer
Science,
University of Ibadan, Nigeria
seyiosunade@gmail.com

ABSTRACT

Public and private opinions of Nigerians were extensively expressed about the two leading presidential candidates in the Nigeria 2019 general election through Social media. Twitter is one of the social media platforms used to voice opinion about the suitability of these candidates. It is therefore necessary to analyze the opinions expressed so as to predict the attitude and feelings of Nigerians to each of these candidates. In this work, a Twitter Sentiment Analysis Tool was developed using the Naïve Bayes classification algorithm to determine if the perspectives of Nigerians to the two candidates were positive, neutral or negative. A single web-page application was developed using Python, Twitter API and REST API. The results for each candidate were presented as graphs and a pie chart. The accuracy of the sentiment analysis was high, but improvements are possible.

Keywords: Twitter, Sentiments, opinion mining, social media, Natural Language Processing (NLP)

1. INTRODUCTION

Nigeria, a large and populous country in Africa recently conducted elections into the three tiers of government in February and March 2019 respectively. As stated by the country's Independent and National Electoral Commission (INEC, 2019), a total of 67 political parties fielded candidates for the elections. Two presidential candidates: Mohammed Buhari of the All Progressive Congress (APC) and Abubakar Atiku of the People's Democratic Party (PDP) were however leading in terms of mobilization, publicity and profile. Buhari is the incumbent president while Atiku was a former Vice President. The contest appeared tough and undecided thus a scientific approach was developed to determine the likely outcome of the results. This work thus sets out to develop a sentiment analysis tool for comments and opinions shared as tweets on Twitter.

Twitter represents a social platform where opinions are given freely and can be used as a data source.

Kalia (2013) defined social media as a group of Internet-based applications that are created in the ideological and technological foundations of Web 2.0 which allowed users to generate and share content. The benefits of using social media applications like Instagram, Twitter, WhatsApp, Pinterest and Facebook includes wider geographical coverage, reduced cost of content production and fast transmission of content between users. Social media is being used for product promotion, recruiting, learning online and electronic commerce.

According to Statista (2017), twitter is one of the popular social media platforms that currently has over 300 million users who share millions of opinions on a daily basis. Access to its data has made Twitter the leading source of customer data for marketing purposes. Twitter is a micro-blogging website and ranks third amongst the present social media websites (Go et al, 2009). Tweets are used by the public to express opinions about various topics ranging from elections to shopping by individuals, events, products and services. Osimo and Mureddu (2010) agreed that Twitter is an ideal platform for the extraction of general public opinion on specific issues. In order to retrieve tweets a user needs to have a twitter account and obtain credentials (API key, API secret, Access token and Access token secret) on the developer's site to access the Twitter API (Crockford, 2006). A collection of tweets on a particular opinion can be used as the primary corpus for sentiment analysis (Pak and Paronbek 2010).

Alsaedi et al., (2017) said sentiment analysis also known as opinion mining aims at determining the attitude, feeling or bias of a speaker or a writer with respect to a topic or the overall tone of a document. Liu et al. (2009) defined a sentiment or opinion as a quintuple $\langle o_j, f_{jk}, s_{ijkl}, h_i, t_l \rangle$, where o_j is a target object, f_{jk} is a feature of the object o_j , s_{ijkl} is the sentiment value of the opinion holder h_i on feature f_{jk} of object o_j at time t_l , s_{ijkl} is +ve, -ve, or neutral, or a more granular rating, h_i is an opinion holder, t_l is the time when the opinion is expressed". Sentiment Analysis can also be performed either at the document level, sentence level and aspect/feature level where phrases in a sentence are classified according to their polarity: negative, positive or objectivity (Sharma et al, 2014)

Sentiment analysis on data classification can be achieved using Rule-based and Automatic Natural Language Processing

techniques to parse text and give meaning to it via classification. Rule-Based system, gives a set of pre-defined rules (stemming, tokenization, part of speech tagging and parsing. In some cases, lexicons: list of words and expressions that identify subjectivity, polarity, or the subject of opinion). Automatic System, makes use of machine learning to assign sentiments.

Some classification techniques used by Sentiment Analysis includes;

- [1] Machine Learning Approach (ML): This approach relies on the famous ML algorithms to solve the sentiment analysis as a regular text classification problem that makes use of syntactic and/or linguistic features. It is divided into *unsupervised learning and supervised learning*: that depends on the existence of labelled training documents such as Probabilistic classifiers (Naive Bayes classifiers, Bayesian Network, Maximum Entropy Classifier), Rule based classifiers, decision tree classifiers and linear classifiers.
- [2] Lexicon-Based Approach: it is a collection of known and precompiled sentiment terms. It is divided into the dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity (Murphy, 2006).
- [3] Naïve Bayes Classifier (NB): The Naïve Bayes classifier is one of the simplest and most commonly used classification algorithms. Naïve Bayes classification model computes the posterior probability of a class, based on the distribution and ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set/text belongs to a particular label/category.
- [4] Theory Behind Bayes' Theorem:
 $P(H)$: is the probability of hypothesis H being true (regardless of any data). This is known as the prior probability of h.
 $P(D)$: is the probability of the available data (regardless of the hypothesis). This is known as the prior probability.
 $P(H|D)$: is the probability of hypothesis H given the data D. This is known as posterior probability.
 $P(D|H)$: is the probability of data D given that the hypothesis H was true. This is known as posterior probability.

This theory is the gave birth to the Naive Bayes Classifier Machine Learning algorithm. From research, it has been observed that for any statistic and probabilistic studies, the Bayes' Theorem is your best bet due to the fact that it helps the user to have logical perceptive on the general front.

2. RELATED WORKS

Alec et al (2009) provide one of the first studies on sentiment analysis on micro-blogging websites. They described a distant supervision-based approach for sentiment classification by using hash tags in tweets to create training data and implement a multi-class classifier with topic-dependent clusters.

Go et al (2009) classified Tweets for a query term into negative or positive sentiment. To collect positive and negative tweets, they query twitter for happy and sad emoticons. They tried various features unigrams, bigrams and Part-of-Speech and trained their classifier on various machine learning algorithms: Naive Bayes, Maximum Entropy and Scalable Vector Machines and compare it against a baseline classifier by counting the number of positive and negative words from a publicly available corpus. They report that Bigrams alone and Part-of-Speech Tagging are not helpful and that Naive Bayes Classifier gives the best results.

Kouloumpis et al, (2011) investigated the utility of linguistic features for detecting the sentiment of Twitter messages. They evaluated the usefulness of existing lexical resources as well as features that capture information about the informal and creative language used in micro blogging. They took a supervised approach to the problem, but leverage existing hashtags in the Twitter data for building training data.

Saif, et al, (2012) discussed a semantic-based approach using Naïve Bayes and the extracted semantic features to identify the entity being discussed in a tweet, like a person, organization etc. They also demonstrated that removal of stop words is not a necessary step and may have an undesirable effect on the classifier.

Tripathy et al. (2017) made use of Naïve Bayes, SVM, random forest and discriminant analysis to classify reviews with regards to their polarity. They made use of four different phases that generated an output for two different datasets that were subjected to different performance metrics (accuracy classification, precision, recall).

3. METHODOLOGY

This research is quantitative in nature and experimental in approach. The sentiment analysis tool implemented uses the

Naives Bayes algorithm for the classification of the preprocessed data into appropriate polarity. The tool was developed in the Python language because of its support for natural language processing.

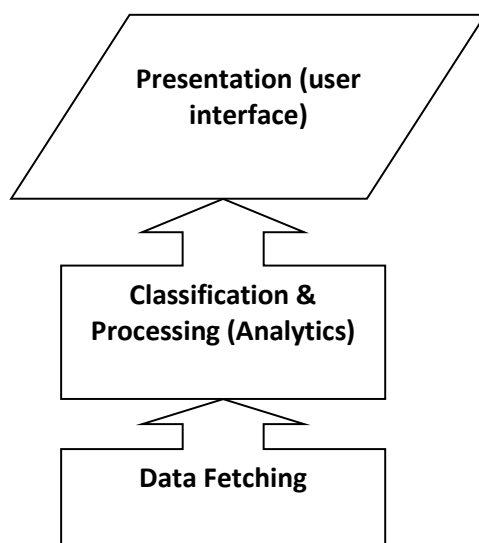


Figure 1: Architecture for Sentiment Analysis Tool

Figure 1 presents the sentiment analysis tool architecture. The architecture has 3 components and was adapted from the Django framework (Django, 2016). The first component is the web interface that accepts input text from the user. The second component performs the preprocessing of text and classification. The tweet mining engine was designed using Python’s natural language processing toolkit. The third component is the view or interface that displays the result of the analysis.

The sentiment analysis tool developed is called Twitter Sentiment Classifier. The user enters the keyword to be analyzed and clicks on the “Classify Tweets” button. The user is presented with a result page that displays a pie chart of the expressed sentiments and charts for each type of sentiment.

The Python Natural Language Processing toolkit was used to process the data downloaded through the API. The tweets used for this study were about two presidential candidates in Nigeria’s 2019 general elections and the electoral body.

The following steps were observed in building the Classification Model for the tweets:

- **Corpus Data Set:** Twitter Search API wrapper written in Python was used to fetch about 500,723 tweets related to the election. These tweets consist of tweet handles, topics, sentiment labels, tweet creation date for each tweet ID. It offered a more streamlined and flexible way for making search queries to the Twitter API query of the twitter API python library.
- **Pre-Processing:** This involves the removal of unwanted variances/duplicates, which eventually helps to down-size the original data so as to make it more suitable for learning by the Naive Bayes Machine Learning algorithm using the python Language. Conversion of tweet labels, punctuations, hash tags, emoticons and replacing them with other regular expressions, tokenization of the data into single words and then stemming which helped to normalize the data for synonyms words (words with same meaning) with the
- aid of the Potter Steemer was done to enhance the preprocessing stage.
- **Feature Extraction:** The Unigram model was used to further classify the tweets after the preprocessing stage. This helped to prevent frequent repetition of words and in turn, data scaling.
- **Classifier Model:** After the feature extraction stage, the data were transformed into occurrences, which was then used as the features that was fed into the model. The python processing Natural Language Toolkit library was used to build models and classify training /text preprocessing.

The Training phase: models were classified and trained by the classifier using a specified dataset while the *Testing phase;* involved the evaluation of the classifier based on parameters like accuracy, error and recall.

On completion of the training and testing of the classifier model, the Naive Bayes Classifier was then initialized with the data fitted into it and analyzed based on whether a tweet was positive, neutral or negative.

Naive Bayes Classifier was chosen for this work because of its statistical and probabilistic nature based on the Bayes Theorem. Also, apart from the fact that it is one of the simplest supervised learning algorithms, its speed, accuracy and reliability especially in cases of large datasets cannot be overlooked.

- **Logical Interface:** A REST API access point to the application was designed using:
 - Natural Language Toolkit*, a popular python library for Natural language processing to handle text processing and classifications
 - Scikit-learn*, A Machine Learning Tool in Python
 - Twitter Search*, A Python library to easily iterate tweets found by the Twitter Search API
 - Flask - a python web micro-service framework. It generates and responds with the sentiment classifications of each search results as specified by its input.
- **User Interface:** A web application interface was built using:
 - React.js*, a JavaScript library that was used to build user interfaces that allowed user to interact (generate results analysis in form of either texts or graphs) with the REST API.
 - Axios*, an HTTP AJAX client that allowed the browser easy communication (it handles requests for the analysis result from the web application to the REST API server) with the REST API.
 - Chart.js*, a JavaScript library that was used for data visualization (with HTML, SVG and CSS) of the sentiment results on the web application.

4. RESULTS AND DISCUSSION

The software developed was run to analyze tweets with specific words such as Trump, Buhari, Atiku and INEC. The user types in a keyword at the web interface. The application accepts the keyword, connects to the Twitter Search API for retrieval of tweets. The retrieved tweets are processed and classified for opinions. The result of the analysis is sent to the web interface for display. The results of the analysis are presented in the Figures below.



Figure 2: user interface for Sentiment Analysis tool

The user interface shown in Figure 2 is the presentation layer of the sentiment analysis tool developed. The keyword to be analyzed is the only input into the tool.

Sentiment Statistical Analysis for keyword "trump". (1001 tweets analyzed)

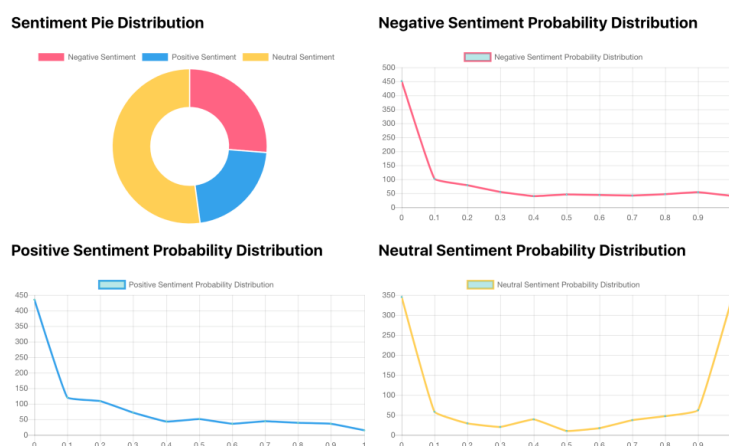


Figure 3: Sentiment Analysis for “Trump”

The result for a sentiment analysis for the keyword “trump” is given in Figure 3 comprising of four charts. The pie chart gives a visual representation of the classification of the tweets into polarity. The other three charts give a graph of the polarity count against the probability of its occurrence.

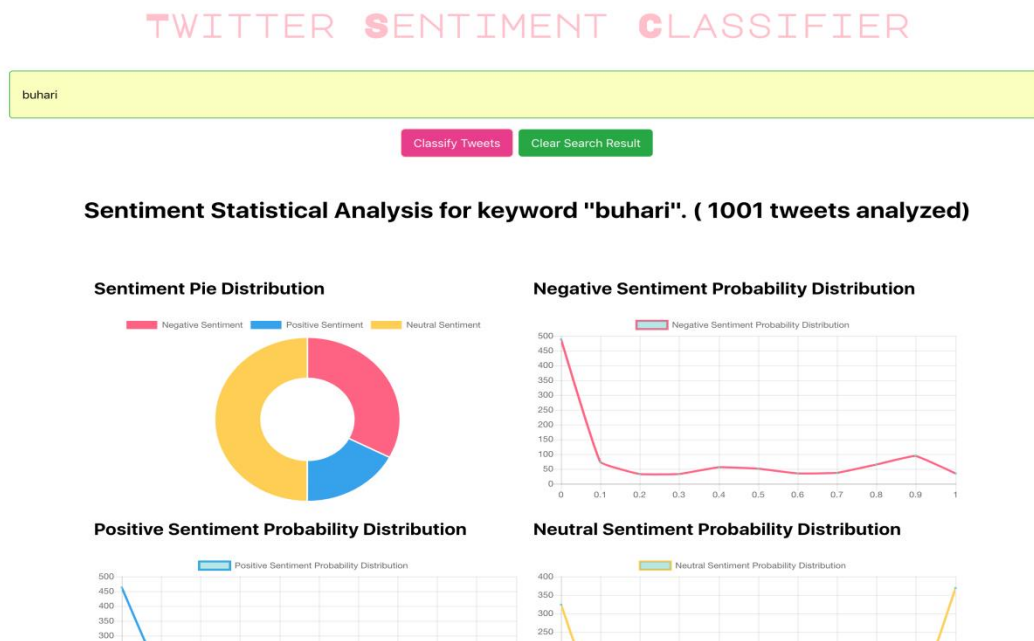


Figure 4: Sentiment Analysis results for “Buhari”

In Figure 4, the search bar and the results are both displayed. The results for the word “buhari” show that the posts were mostly neutral. The number of positive tweets are very minimal.

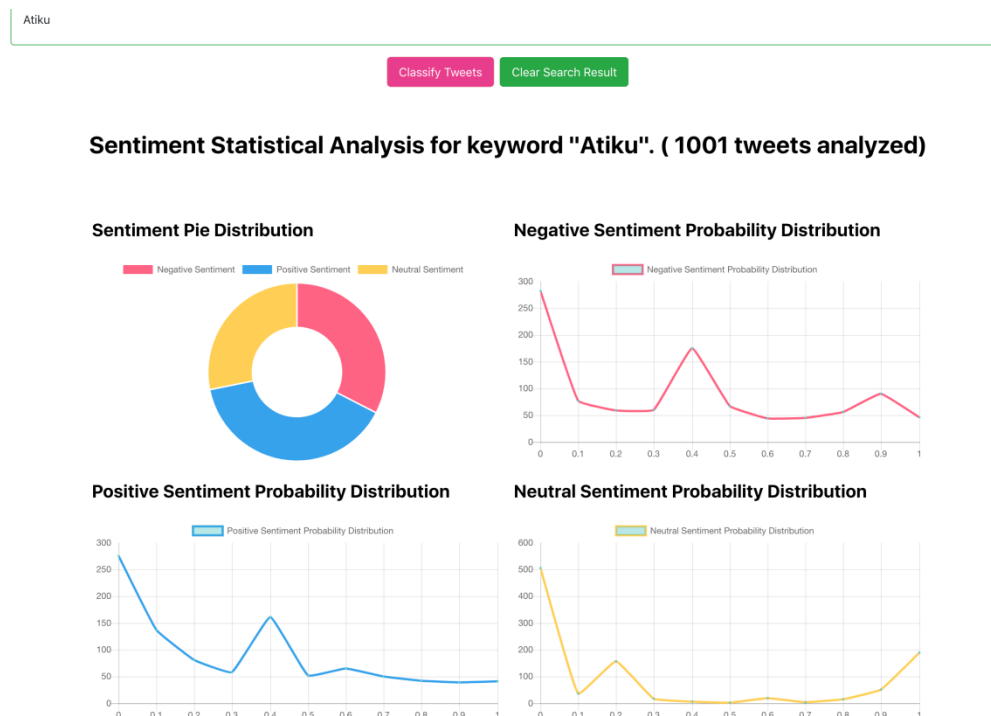


Figure 5: Sentiment Analysis results for “Atiku”

The results for the keyword search of “Atiku”, in Figure 5, shows that there was almost an equal number of positive, negative and neutral tweets.

The results for the electoral body “INEC” involved in the election process showed that there were more negative tweets than positive tweets about the organization. These are provided by the charts in Figure 6.

From the results presented above it could be inferred that tweets represent the opinions of individuals. It however does not indicate their voting preferences as shown by the final results of the 2019 elections in Nigeria.

Sentiment Statistical Analysis for keyword "INEC". (894 tweets analyzed)

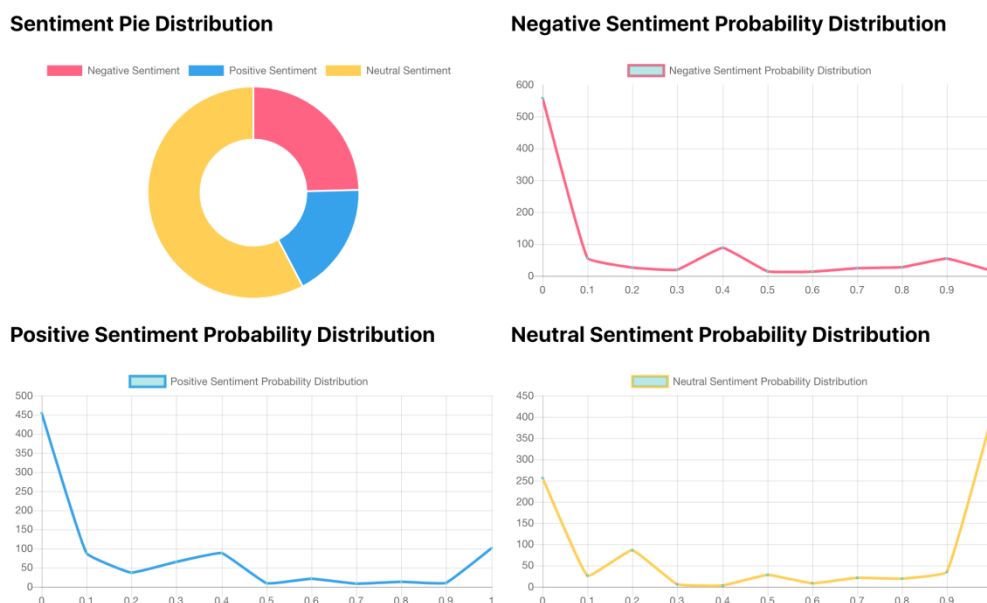


Figure 6: Sentiment Analysis results for “INEC”

The sentiment analysis for the two leading contestants in the 2019 Nigeria general elections indicates that there were more neutral tweets about “Buhari” than “Atiku.” This might be attributable to the fact that people were self-moderating their tweets to avoid any backlash from the government or other citizens. The negative sentiments about both contestants for the presidential position are about the same. This is in tandem with the number of visible supporters of each candidate who have chosen to be identified with the contestant of their choice. The number of positive tweets for candidate “Atiku” is more than that of contestant “Buhari.” This indicates that more citizens were positively disposed to having Atiku as the president.

5. CONCLUSION

This work developed a functional web application, called Twitter Sentiment Classifier, which analyzes tweets from a social media platform called Twitter. The Naïve Bayes classification algorithm was used to classify tweets into positive, neutral and negative. The tool was tested using tweets about the two leading candidates in the 2019 Nigeria general elections. The results showed a preference towards one candidate even though the final election results indicated otherwise. The accuracy of the developed Sentiment Analysis Tool was 77%. The authors suggest the use of multiple algorithms to classify tweets so as to improve the accuracy. Images and emojis were not analyzed so further work can be done along this area.

6. REFERENCES

- [1] Abdullah, A. and Mohammad, Z. K. (n.d). A Study on Sentiment Analysis Techniques of Twitter Data. *International Journal of Advanced Computer Science and Applications (IJACSA)*. 10(2), 48
- [2] Alec, G., Richa Bhayani, and Lei Huang. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1, 1-6.
- [3] Tripathy and Rath, S. K. (2017). Classification of sentiment of reviews using supervised machine learning techniques," *International Journal of Rough Sets and Data Analysis (IJRSDA)*, 4(1), 56-74.

- [4] Crockford, D. (2006). The application/json Media Type for JavaScript Object Notation (JSON). *Technical report, IETF*.
- [5] Kalia, G. (2013). A Research Paper on Social Media: An Innovative Educational Tool. Chitkara University. (1) 43-50.
- [6] Go, A., Bhayani, R. and Huang, L. (2009). Twitter sentiment classification using distant supervision. *Processing*, pp 1-6.
- [7] Murphy, K. P. (2006). Naive Bayes classifiers. University of British Columbia, vol. 18.
- [8] Kouloumpis, E.; Wilson, T. and Moore, J. (2011). Twitter sentiment analysis: The good the bad and the OMG! *ICWSM*. 11:538-541.
- [9] Osimo, D. and Mureddu, F. (2010). Research Challenge on Opinion Mining and Sentiment Analysis. *Proceeding of the 12th Conference of FRUCT Association, United Kingdom*.
- [10] Pak, A. and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pp.1320-1326.
- [11] Sharma, R., Nigam, S. and Jain, R. (2014). Polarity detection at sentence level. *International Journal of Computer Applications*. 86(11).
- [12] Saif, H., He, Y. and Alani, H. (2012). Alleviating Data Scarcity for Twitter Sentiment Analysis. *Association for Computational Linguistics*.
- [13] Statista. (2017). <https://www.statista.com/statistics/282087/number-ofmonthly-active-twitter-users/>