Predicting Sentiment in *Yorùbá* Written Texts: A Comparison of Machine Learning Models

Abimbola Rhoda Iyanda(✉) and Omolayo Abegunde

Obafemi Awolowo University, Ile-Ife, Nigeria
abiyanda@oauife.edu.ng, omolayoabegunde@gmail.com

Abstract. Sentiment Analysis (SA) provides a rich set of tools and techniques for extracting, evaluating subjective information from large dataset. Users opinions concerning an event are what determines the user perspective of such event whether it is good or bad. This study compared three machine learning models (Logistic Regression, Naïve Bayes and Support Vector Machine) with a view to identifying the best model for predicting sentiment in *Yorùbá* written texts at sentence-level. Corpus of *Yorùbá* records was created from several online and offline sources such as dictionaries, experts, Bible, social media as well as awayoruba blog and processed using *Tákàdá*. The system was implemented using the Python programming language and evaluated using mean opinion score and receiver operating characteristics. The research concludes that Naïve Bayes (NB) outperforms other algorithm for analysis of sentiments for *Yorùbá* sentences.

Keywords: Sentiment analysis · Opinion mining · Machine learning models · *Yorùbá* · Natural Language Processing · Sentence Level

1 Introduction

Sentiment analysis (SA) is the computational study of how opinions, attitudes, emotions and perspectives are expressed in a language[1]. SA provides a rich set of tools and techniques for extracting this evaluative and subjective information from large datasets and summarizing them. The application of SA does provide information to the companies on how good or bad their product is[2] stated that recent breakthroughs mean that this analysis can go beyond a general measure of positive vs. negative, isolating a fuller spectrum of emotions and evaluations and controlling for different topics and community norms. According to[3], everyone

[1] http://gen.lib.rus.ec/book/index.php?md5=5409ACC88F5C7209D8D8DEA5E06D3176.

[2] http://sentiment.christopherpotts.net/overview.html.

[3] https://www.goodreads.com/quotes/1109069-1-everyone-is-entitled-to-their-opinion-about-the-things.

Y. Bi et al. (Eds.): IntelliSys 2019, AISC 1038, pp. 416–431, 2020.
https://doi.org/10.1007/978-3-030-29513-4_30

is entitled to his or her opinion about the things read, watch, listen to, taste, or whatever contribution needed to be made by any individual towards an event. Due to this freedom of expression of ones thought, it is highly required to carry out a study on how people feel about the social or economic development of the society.

SA is an exciting and new field of research in Artificial Intelligence combining Natural Language Processing, Machine Learning and Psychology [1]. Companies across the world have implemented machine learning to do this automatically. SA is widely used for getting insight into customers opinion as the customers' comments or reviews are analyzed to identify what people like or dislikes. Sometimes, sentiment analysis is referred to as opinion mining, thus, the basic task in opinion mining is the determination of a sentence or phrase that are extracted from the corpus as an opinion or just described as a fact. It is an established fact that opinion mining helps in getting feedback about what influences the opinions of people. According to [2], opinion mining enables one to discover the reasons why people like or do not like something by learning relationships among the traits/products via semantic rules and the factors that lead to change on the opinions such as from positive to negative.

The study reported in this thesis addressed the development of a system that detects and analyse peoples opinion in *Yorùbá*. [3] stated that facts are vital for individual or governmental organization for judgment and decision making. These are the opinions or believes that an individual has concerning an event since they are influenced by their subjective feelings and beliefs. The decision-making day by day can be influenced by others' perceptions of the world to some certain degree.

The rest of this paper is organized as follows: Sect. 2 summaries the related work done while Sect. 3 focus on the methodology; Sect. 4 discusses the implementation; Sect. 5 presented the result and discussion of the study. Section 6 presented the system evaluation and Finally, Sect. 7 concludes.

2 Machine Learning Classifiers

The classifier is an algorithm that is used in classification, it is a mathematical function that map inputs data to the category. A classifier has a set of variables that need to fit data (trained data). All the classifier mention below have different algorithm to optimize the process. The model generated fits on the data it was trained on and new data is completely strange to the fit data[4].

2.1 Naïve Bayes Classifier

This is a probabilistic classifier that is based on Bayes theorem with the assumption of independence between predictors[5]. In Naïve Bayes (NB) classifier, the

[4] https://www.quora.com/What-is-a-training-data-set-test-data-set-in-machine-learning-What-are-the-rules-for-selecting-them.

[5] https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/.

presence of one feature in a class is unrelated to any other feature in the dataset. One of the benefits of Naïve Bayes is that it is simple to implement and it involves a bunch of counts. NB algorithm will converge quicker compare to discriminative models such as Logistic regression when subjected to less training data. Therefore, NB conditional independence holds. If the NB assumption did not hold, NB classifier will still perform well in practice. NB can be used when there are limited resources in terms of CPU (Central Processing Unit) and Memory[6]. One of the reasons why NB has been considered apart from aforementioned is that NB training time is quicker in relation to Maximum Entropy and it is widely used in many of the researches.

There are numerous NB variations and it is worthy of note that the variations of this algorithm deliver different results. In this work, Multinomial Naïve Bayes was used. Multinomial Naïve Bayes (MNB) is used when the multiple occurrences of the words matter a lot in the classification problem. For example, in the case of SA, where it does not matter how many times a particular word is mentioned e.g. bàjẹ́ (bad) can appear at any time in a sentence. The model of NB is described as follows:

$$P(label|features) = \frac{P(label) * P(features|label)}{P(features)} \tag{1}$$

$$P(Y|X_1, ..., X_n) = \frac{P(Y) * P(X_1, ..., X_n|Y)}{P(X_1, ..., X_n)} \tag{2}$$

- $P(label)$: is the prior probability of the label occurring
- $P(features|label)$: is the prior probability of a given feature being classified as that label.
- $P(features)$: is the prior probability of a given feature set occurring
- $P(label|features)$: the probability that the given features should have that label.

To compute the prior probability:

$$TP(c) = \frac{N_c}{N} \tag{3}$$

- $P(c)$ is the probability of the class
- N_c is the total count of a particular class in the training set
- N is the total count of class in the training set.

To compute the conditional probability/Likelihood of each word attribute:

$$P(w|c) = \frac{count(w, c) + 1}{count(c) + |V|} \tag{4}$$

- $P(w|c)$: is the conditional probability / likehood. where

[6] http://blog.datumbox.com/machine-learning-tutorial-the-naive-bayes-text-classifier/.

  - w is the word attribute and
  - c is the class
- Count w, c is the total count of word attribute occurs in c class.
- $+1$ is Laplace smoothing
- $count(c)$ is the total count of word attribute in a particular class occurs in the training set.
- $|V|$ is the vocabulary. The total count of different word attribute in the training set.

Equation 5 compute the posterior probability:

$$C_{MAP} = argmaxP(X_1, ..., X_n) * P(c)c \in C \tag{5}$$

MNB is a specialized version of NB that is suited for more of text documents. While the ordinary Naïve Bayes would model a document as the presence and absence of particular words. MNB explicitly models the word counts and regulate the underlying calculations. A multinomial model, observation (samples or feature vectors) represents the occurrence at which events have been generated by a multinomial $\theta_y = (\theta_{y1}, \theta_{y2}, ...\theta_{yn})$, for each class y, where n is the number of features in text classification and the size of the vocabulary, θ_y is the probability $P(x_i|y)$ of feature i appearing in the observation belonging to y label.

The parameters θ_y is the estimated by a smoothed version of the maximum likelihood which is the relative frequency of the counter which is express in the below equation:

$$\hat{\theta_{yi}} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \tag{6}$$

where $N_{yi} = \sum_{x \in T} x_i$ is the number of times feature i appears in a sample class y in the training set T, and $N_y = \sum_{i=1}^{|T|} N_{yi}$ is the total of all features for class y. The smoothing priors $\alpha \geq 0$ accounts for features that are not present in the learning samples and prevents zero probabilities in further computations. If $\alpha = 1$ is called Laplace smoothing, while $\alpha < 1$ is called Lidstone smoothing. The equation for this Multinomial equation was gotten from the scik-learn library for machine learning.

2.2 Support Vector Machines

Support Vector Machine (SVM) is a type of machine learning algorithm that is used to analyze, recognize and classify data [4]. It uses different loss function from Logistic regression. They can express meaning in different form i.e. maximum-margin. Practice shows that SVM with a linear kernel is not much different from Logistic Regression. Although the majority of the researchers have reported that SVM is good in text classification. But SVM may be preferred over Logistic Regression (LR) if the problem is not linearly separable. Another reason for SVM to be preferred is when considering high dimensional space. One of the

disadvantages of using SVM is that they are in inefficient in training[7]. Therefore, there is a need to provide the system with many samples.

The goal of the SVM is to find the best separating hyperplane (Decision boundary). It classifies the training vectors into two classes. The equation for the classifier is stated as follows:

$$f(x_i) = sign(w^T x_i + b) \tag{7}$$

Functional margin of x_i is:

$$Y_i(w^T x_i + b) \tag{8}$$

To increase the margin w and b need to be scaled, where w is the decision hyperplane normal vector, x_i is the data point i, Y_i is the class of data point $i(+1$ or $-1)$.

Since the study deals with linear data, that is, not a data that can be clustered into different groups, this leads to the use of linear kernel in such as SVM algorithm. Therefore, the following two constraints follow for a training set $\{(x_i, y_i)\}$:

$$w^T x_i + b \geq 1 \quad if \quad y_i = 1 \tag{9}$$

$$w^T x_i + b \leq -1 \quad if \quad y_i = -1 \tag{10}$$

To maximize the margin (M):

$$M = \frac{2}{|w|} \tag{11}$$

To minimize the margin (m):

$$m = \frac{1}{2} * w^t w \tag{12}$$

SVM produces high accuracy and amusing theoretical guarantees regarding overfitting if the appropriate kernel is applied, the system can work well even if the data is not linearly separable in the vector space.

One of the disadvantages of SVM is that it is sensitive to noise, a small number of mislabelled examples can decrease the system performance and this classifier considers two labels (classes).

2.3 Logistic Regression

According to[8], Logistic Regression (LR) is one of the most popular machine learning algorithm for binary classification. It is also be referred to as "logit" regression that is used to estimate the discrete values like "0" or "1", "yes" or "no", "true" or "false" based on the set of the independent variables. It performs

[7] https://www.quora.com/What-are-the-advantages-of-different-classification-algorithms.

[8] https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/.

well on a wide range of problems and its predictive probability lies between "0" and "1". The logistic function is defined as:

$$T = \frac{1}{(1 + e^{-x})} \tag{13}$$

Where T is the transformed, e is the Euler's number, x is the input that will be plugged into the function.

LR can be trained as long as the features are linear and there is a problem of linearly separable (See footnote 6). For some features that are not linear, some machine learning techniques such as feature engineering to turn non-linear into linear can be applied. It is robust to noise and overfitting can be avoided in the feature selection process. LR output can be interpreted as a probability, which is advantageous to this work.

Some of the weak points of LR is that it cannot be used for continuous outcome, that is it is based on classification. If observation is related to one another, the model will tend to overweight the significance of those observations. The model is vulnerable to overconfidence, that is, the model can appear to have more predictive power than they actually do as a result of sampling bias.

The three algorithms are compared by (See footnote 6) as follows:

i. For Naïve Bayes algorithm
   (a) There is no compulsory requirement for it
   (b) It is good few categorical variables
   (c) It compute the multiplication of independent distributions
   (d) It suffer col-linearity
ii. For SVM
   (a) There is no distribution of requirements
   (b) It computes the hinge loss (loss function) which are used to train the classifiers and maximum-margin classification.
   (c) It has flexible kernel selection for nonlinear correlation
   (d) It has no problem with col-linearity
iii. Logistic regression
   (a) No requirement distribution required
   (b) It behaves well with few categories of categorical variables
   (c) Good in the computation of logistic distribution
   (d) It suffers collinearity

2.4 Ensemble Methods

Ensemble methods are the techniques employed in ML for creating multiple models and then merge these models together to form a single model in other to produce an accurate model. Ensemble methods do produce a precise result than using single model[9]. This approach was adopted in this work.

[9] https://www.toptal.com/machine-learning/ensemble-methods-machine-learning.

There are metrics used in evaluating classifiers with respect to accuracy [5] and these are Precision, Recall and Receiver operating characteristics (ROC). These are the most commonly used metrics in text classification. There are some facts and errors that justify these metrics which[10] are:

i. False Positive (FP): this error happens when the classifier classifies a feature set with a label it shouldn't have gotten.
ii. False Negative (FN): this error happens when a classifier doesn't assign a label to the feature set that should have it.
iii. True Positive (TP): the number of records in the datasets that belongs to the positive class.
iv. True Negative (TN): the number of records in the datasets that belongs to the negative class.
v. Precision: Precision is the size of the intersection of both sets divided by the size of the test set. That is the percentage of the test set that was guessed correctly.
vi. Recall: is the size of the intersection of both sets divided by the size of the reference set, or the percentage of the reference set that was guessed correctly.
vii. F_Measure: is weighted harmonic mean of precision and recall.

3 Review of Related Works

In *Yorùbá* language, natural language processing is not rich compared with the English language. In the field of sentiment analysis, there is paucity of the work done in *Yorùbá* language. There are no available datasets or dependable API's like SentiWordNet, SenticNet, WordNet-Affect, which can be used to carry-out sentiment analysis task, this necessitate the research.

Author in [6] worked on Sentence-Level Arabic Sentiment Analysis. The study focused on sentiment classification in the Arabic language at the sentence level by classifying sentence from blog, review and tweet. Twitter Application Package Interface (API) was used in the collection of the tweets by setting the Arabic language to (lang=ar). 4000 tweets were collected and 1000 tweets were extracted to form the datasets. Two machine learning approaches were used for the classification which is: Naïve Bayes (NB) and Support Vector Machine (SVM). The features used are unigram and bigram. Each of the tweets was annotated to be 500-positive and 500-negative. The feature in each tweet was extracted and presented in feature vector format. The feature vector form the training phase of each of the classifier. Weka application was used for the implementation. Two results were generated: one for the stop-words removed and the other for the stop-words present. Four metrics (Accuracy, Precision, Recall and F-Measures) were used to evaluate the classifier.

[10] http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/.

With stop words included in the dataset: SVM classifier (Accuracy, Precision, Recall and F-Measure = 72%). Using Naïve Bayes (Accuracy, Precision, Recall, and F-Measure = 65%). Removing the stop-words from the dataset: Using SVM classifier (Accuracy, Precision, Recall and F-Measure = 73%). Using Naïve Bayes (Accuracy = 65%, Precision = 66%, Recall = 65% and F-Measure = 65%). The work is different from the work in this study in that the work consider sentiment detection in Arabic at sentence level using unigram and bi-gram feature extraction techniques and two classifiers was used, whereas the work in this study considers sentiment detection in *Yorùbá* at sentence level using unigram techniques with three classifiers.

Author in [7] worked on sentimentality extraction of Bengali language. The study focused on novel approach using Naïve Bayes classification to model the Bengali Language. The dataset used in the work was retrieved from Facebook social media. Above 1000-positive comments and 1000-negative comments were collected for the train sets, where 500-comments was used as the test set. Data processing was performed on the collected corpus. The feature used are unigram and bigram. The extracted features were fed into the Naïve Bayes algorithm to generate the data model. The machine learning was later evaluated for unigram and bigram. The result for unigram are Precision = 65%, Recall = 56% F-Measure = 60% and for bigram, Precision = 77%, Recall = 68%, F-Measure = 72%. It was observed that the bigram features performed better than unigram. The work is different from the work in this study in that, the work focused on Bengali language and it only based on one machine learning classifier.

Author in [8] presented extraction of emotions from the multilingual text. The study focused on the use of advanced framework for detection of emotions of users in multilanguage text data. Three domain area were chosen (Political election, Health-care, and Sports) for real-time empirical study. The technique used in the data collection was Rich Site Summary (RSS) feeds through the headline news from Twitter. Text preprocessing was carried out on the tweets such as normalization, tokenization, stop-words removal, stemming, lemmatization, etc. The dataset contained 1085721 tweets and it was split into three parts according to the selected domain. The dataset was annotated into emotion and non-emotion. Two machine learning algorithms SVM and Naïve Bayes. The result obtained in the study was for SVM = 72% accuracy and for NB = 69% accuracy. The work is different from the work in this study in that, the work focused on emotion extraction from multi-language focused on three domains, while the work in this study considers a single language with four domains.

Author in [9] presented sentiment analysis framework in implicit opinions for the Thai language. The study focused on sentence-level implicit opinion in the Thai language. The framework used consists of three modules for data preparation (Knowledge, Construction and Data preprocessing). The experiment was conducted in one of the mobile leading domain for data collection and all reviews were collected from Siamphone website. The total number reviews are 1090 sentences. The data were analyzed using pattern matching approach. The evaluation

of the result was based on three metrics (Precision, Recall and F-Measure). For positive opinion, the result gotten was 83.33% and for negative opinion, the result gotten was 88.24%. The work is different from the work in this study in that, the work focused on implicit opinion in the Thai language, no machine learning algorithm was used, while the work in this study considers the use of four machine learning metrics (Precision, Recall, F-Measure, and ROC_AUC) and mean opinion score (MOS).

Author in [10] worked on idiomatic expression for Chinese sentiment analysis. The study focused on the improve precision and performance of the emotion classifier. A web crawler with Plurk search API was applied to collect Chinese short text messages from the Plurk platform. The collected messages end with emotion icons with positive or negative. The data collected for the trained phase was 52694 sentences and 1000 sentence for the test phase. The collected data were segmented with Jieba and bigram feature extraction approach was applied on the datasets while Naïve Bayes classifier was applied on the feature-sets. Five trained sets were used to construct the corpus. Using the size of the datasets (larger than 40000), Combined Jieba Customary Language Model (CJCLM) based sentiment classifier has higher precision than Combined CKIP Language Model (CCLM) based. The work is different from the work in this study in that, the work focused on the improved precision performance for emotion which compares four models using Naïve Bayes classifier, while the work in this study considers one model with three classifiers.

Author in [11] worked on sentence-level sentiment analysis in Persian. The study addressed the problem of resource scarcity. SPerSent which contains 150000 sentences was used for binary classification. Lexicon based method was applied manually on the corpus to build the sentiment words and unigram feature extraction method was applied. The feature selection conducted in the study were occurrence filter and stop-word filter and the selected feature-sets were applied on the Naïve Bayes algorithm. In the training and the validation, 8-fold cross-validation was used to make the process independent of training data.

4 Methodology

4.1 Data Collection

The research corpus was built from four different domain which includes government parastatal, schools, health sectors, and marketing sector. About 1039 observations were collected through *Yorùbá* dictionary and online sources with 639 being negative and 400 being positive, the remaining being objective sentences were discarded. The data collected in this work represent the human feelings about an event in the specified domain. The feelings here shows the behavioural characteristics or attitude towards an event and the data that lacks this feature was discarded from the datasets. The corpus was annotated at sentence level only and two classes of polarity (positive and negative) were considered.

4.2 Data Preparation and Processing

Feature engineering techniques were applied to the dataset which made machine learning algorithm able to give good results. The data was converted to a useful scale format (CSV) with all the meaningful features extracted from the collected corpus; preprocessing and data cleaning were performed on the data before being fed into the machine learning algorithm. Furthermore, the study employed the use of data iterations, exploration, and analysis for the performance of the classifier.

Examining the data is a good way to detect abnormalities or irregularities and peculiarities within the datasets. In data gathering and preparation, some precautions were taken into consideration which are: (i) to determine if the sentences or phrase gathered is subjective or not; (ii) to determine the polarity of the texts whether they are positive or negative; (iii) to identify texts with missing letters and diacritics.

Data were collected and processed using *Tákàdá*, a *Yorùbá* processing editor. Table 1 shows the opinionated example of the data collected and it indicates positive and negative observation of some users. In preprocessing the text, the following techniques used: (i) Tokenization of the texts (ii) Stop words removal and (iii) Feature extraction

Table 1. Simple sentence that expresses sentiment

S/N	Positive Sentence	Negative Sentence
1	*Ọ̀mọ̀wé wa ni ilé ọgbà wa kò sí ẹlẹgbẹ́ wọn ní ìlú yìí.* (There are scholars in our school, and there is none like them in this city).	*Ọjọ́ wo ni a ó bọ́ kúrò ní ọwọ́ àwọn òfófònà tí mbá ilẹ̀ wa fàá.* (When are we going to be delivered from the thieves troubling us in this city).
2	*Ohun tí àwọn ọlọ́jà wa ń tà ní ìsisìyí wúni lórí lọ́pọ̀lọpọ̀.* (What our marketters are selling are so impressive).	*Ìbàjẹ́ ọjà lóhún tà ó kàn tilẹ̀ pọ̀ lójú ni.* (She's is selling quantity and not quality)
3	*Ìwa rẹ̀ ni ìwúlò ni àwùjọ.* (His good attitude is important for the society)	*Nǹkan tó burú ni kí ẹnu àwọn àgbá ìlú ma kò àti ṣe nnkan èyẹ sí ìlú yìí mba mí lérù.* (It's a bad thing for the elders not to be united, I am even scared to do good thing in this land)
4	*Adúpẹ́ pé akò ba ti ara wa tì ní ilẹ̀ yí.* (We appreciate our effort towards the success recorded in this land.)	*Ọmọ ìgboro ni, ó kàn n fi ọjà bojú ni.* (He is a street boy, he uses his market as a cover)
5	*Modúpé mò sì tún júbà fún olùrànlọ́wọ́ wa* . (I thank and appreciate all our supporters)	*Ìwà ìbàjẹ́ ti wá dàṣà ní àárín wa àwọn ajẹgúdújẹrá n pọ̀ si.* (Evil act has become a normal thing in our society, embesslers are multiplying).

5 Implementation

There are two columns in the dataset, the first one consists of the target variables, while the second column is the opinion collected from users. The second column was the variables found in the problem set that helps to build the model. Figure 1 shows the screenshot of the extracted features by the system from the datasets. The classification is done by learning from labeled features sets which is a key-value pair that map each feature names to the feature values. The feature names are words and the values are "True". Figure 2 shows the screenshot for the extracted input features. The extracted features shows what the system uses to relate with the model in order for the system to detect the opinion of the user. The user enters a sentence in *Yorùbá* e.g.: *ìlú yìí dára púpọ̀ kódà mo fẹ́ máa gbé ibẹ̀* (This city is so beautiful, I will like to live here). The extracted feature is fitted into the models and the result is shown in Fig. 3.

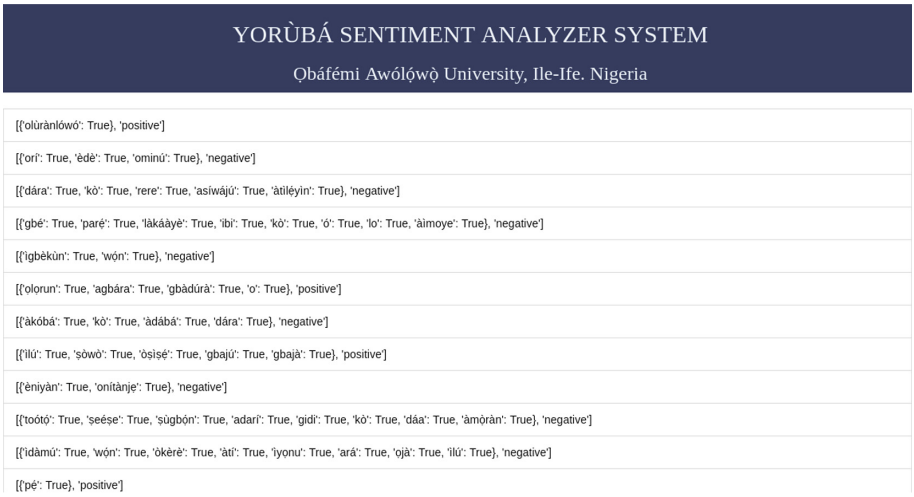

Fig. 1. Screenshot for feature extracted by the system

The result shows "positive" which denotes that the sentence the user entered has a positive opinion. To increase the performance of the system, the three classifiers (SVM, Logistic Regression and Naïve Bayes) were combined and the labels with the maximum vote were selected as the result by the system using the derived logic table shown in Table 2.

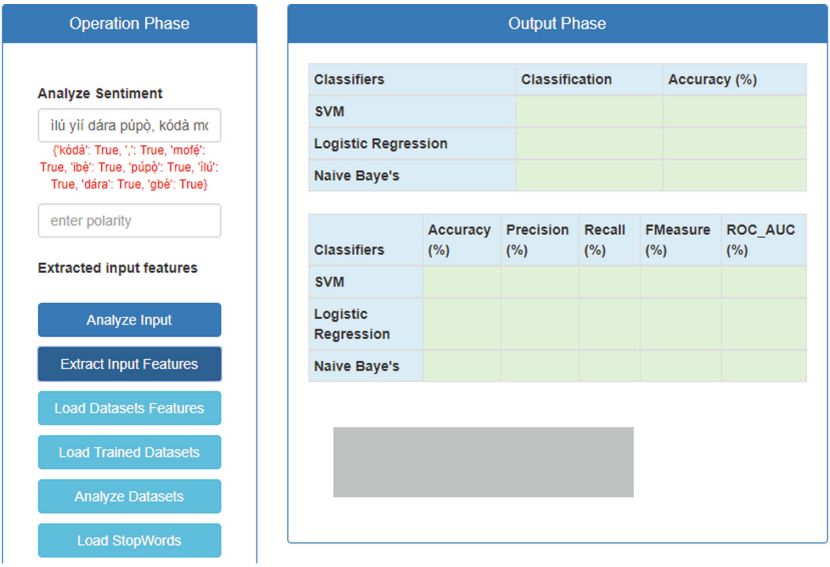

Fig. 2. Screenshot for the extracted input features

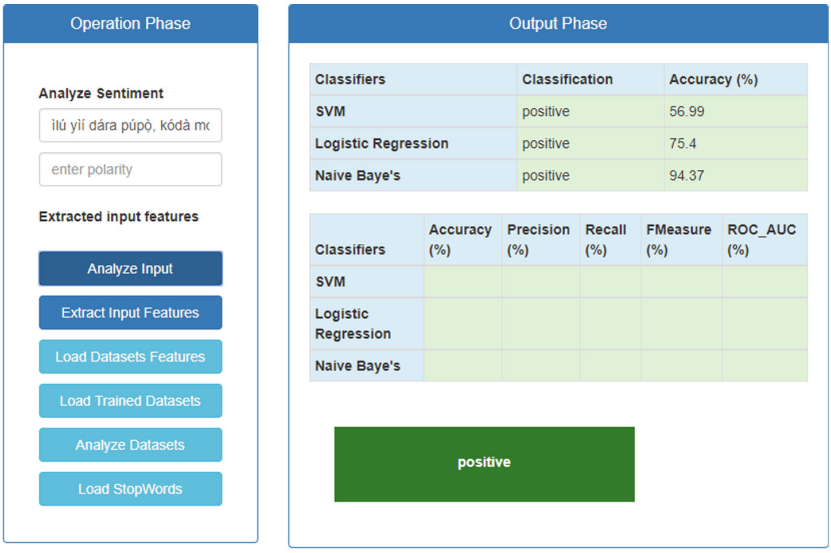

Fig. 3. Screenshot showing positive opinion for (*ìlú yíí dára púpọ̀ kódà mo fẹ́ láti máa gbé ní ìbẹ̀*)

Table 2. Derived Logic table for result analysis

S/N	Logistic Regression	Naïve Bayes	SVM	Result
1	Negative	Negative	Negative	Negative
2	Negative	Negative	Positive	Negative
3	Negative	Positive	Negative	Negative
4	Negative	Positive	Positive	Positive
5	Positive	Negative	Negative	Negative
6	Positive	Negative	Positive	Positive
7	Positive	Positive	Negative	Positive
8	Positive	Positive	Positive	Positive

6 Result and Discussion

The performance of the classifiers were compared by calculating the accuracy, precision, recall and area under the curve. Receiver operation characteristics (ROC) curves help in the visualizing the performance of the classifier. It can only be used with a binary classification, i.e. a problem with two classes. The $x - axis$ represents the predicted probabilities while the $y - axis$ represent the observations. Accuracy rate is the percentage of correct predictions.

From the ROC curve, the true positive rate (TPR) is plotted on the $y - axis$ while the false positive rate (FPR) is plotted on the $x - axis$ for every possible classification threshold. The ROC graph enables the user to benchmark the result obtained from the classifier. Table 3 shows a SVM, Logistic Regression and Naïve Bayes classifiers with Area under curve (AUC) of 0.56, 0.58 and 0.61 respectively with positive polarity while Table 4 shows SVM, Logistic Regression and Naïve Bayes classifiers with AUC of 0.44, 0.42 and 0.39 respectively with negative polarity. The graphs of these three classifiers is shown in Figs. 4, 5 and 6.

The area under the curve for both positive and negative curves show the measure of text accuracy. The positive results go above the line to the left for the accuracy and the negative curve move under the line which is a worse guess for the classifier (The higher the curve the better the accuracy). From Table 3 and Fig. 6, it can be established that the system gives a good result and NB out-performs other classifiers.

Table 3. Screenshot for the system evaluation result with positive polarity

Classifier	Accuracy (%)	Precision (%)	Recall (%)	FMeasure (%)	ROC_AUC
SVM	59.77	46.99	39.8	43.09	0.56
Logistic regression	64.06	54.84	34.69	42.5	0.58
Naïve Bayes	65.23	56.34	40.82	47.34	0.61

Table 4. Screenshot for the system evaluation result with negative polarity

Classifier	Accuracy (%)	Precision (%)	Recall (%)	FMeasure (%)	ROC_AUC
SVM	59.77	65.99	72.15	68.88	0.44
Logistic regression	64.06	67.01	82.28	73.86	0.42
Naïve Bayes	65.23	68.65	80.38	74.05	0.39

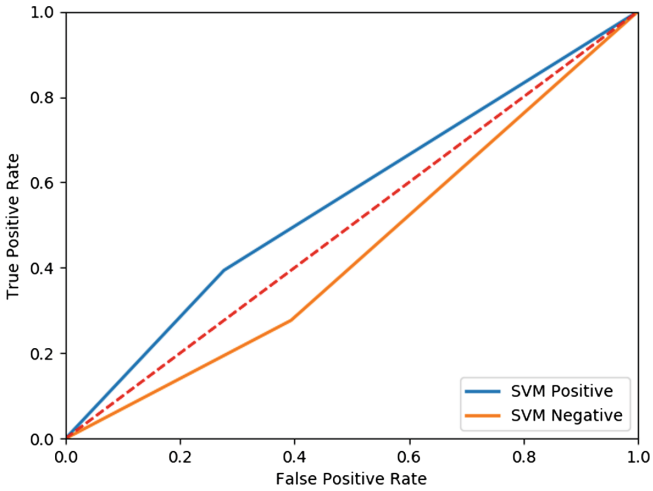

Fig. 4. Screenshot for the ROC curve with SVM classifier

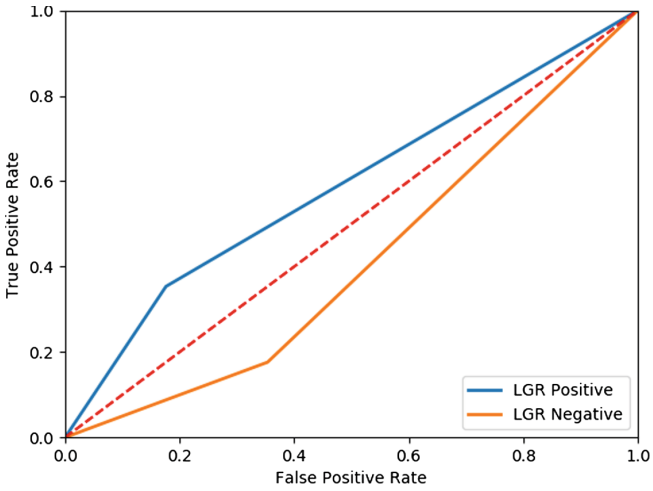

Fig. 5. Screenshot for the ROC curve with LGR classifier

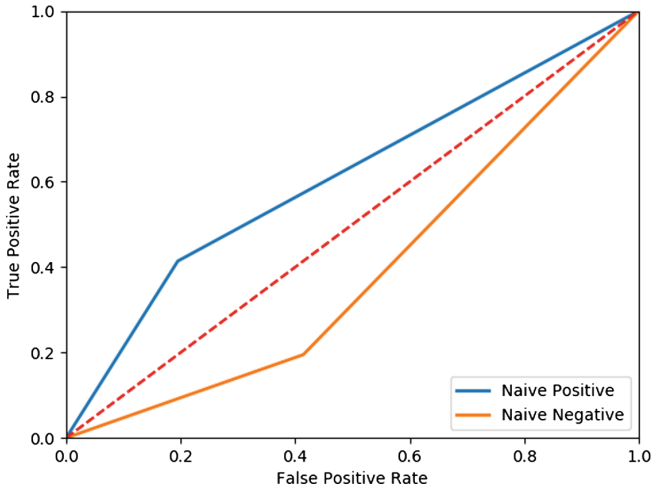

Fig. 6. Screenshot for the ROC curve with NB classifier

7 System Evaluation

Mean opinion score which is a quantitative method was also used for the evaluation of the system. The questionnaire comprises of twenty-five (25) statements with eleven negative (11) opinions and fourteen (14) positive opinions. The results based on the expert, system and respondents' responses is shown in Table 5 where R1 to R14 represents respondent 1 to 14. It can be deduced that the system has 100 % and 79 % for negative opinions and positive opinions correctly classified respectively while compared with the expert's classification.

Table 5. MOS evaluation result

Respondent	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	Expert	System
No. of correctly classified negative statements	3	11	11	11	0	5	9	11	9	9	11	11	10	5	11	11
No. of correctly classified positive statements	8	12	13	8	4	6	13	12	10	6	12	12	12	7	14	11
% Negative opinion correctly classified	27	100	100	100	0	46	82	100	82	82	100	100	91	46	100	100
% Positive opinion correctly classified	57	86	93	73	29	43	93	86	71	43	86	86	86	50	100	79

8 Conclusion

In this study, a system was developed to detect sentiment from *Yorùbá* sentences at sentence level with a view to extract the users' opinion from the sentence. The study has pointed out various ways in which the work can be achieved using machine learning with different algorithms. The specific challenges encountered during the implementation of the SA for *Yorùbá* language have been pointed out. The result shows that users' opinion in *Yorùbá* sentences can be mined and the stand on an event can be determined. This study provides a holistic assessment of the SA system in *Yorùbá*, however, some sentences that express sarcasm were not captured in the study. The research concludes that NB outperform other algorithm for analysis of sentiments for *Yorùbá* sentences. Our future goal is to explore sentiment analysis in the area of name entities recognition and sarcasm for *Yorùbá* language.

References

1. Mukherjee, S., Bhattacharyya, P.: Sentiment analysis in Twitter with lightweight discourse analysis. In: 2012 Proceedings of COLING, pp. 1847–1864 (2012)
2. Bilici, E., Saygın, Y.: Why do people (not) like me? Mining opinion influencing factors from reviews. Expert Syst. Appl. **68**, 185–195 (2017)
3. Qin, Z.: A framework and practical implementation for sentiment analysis and aspect exploration. Ph.D. University of Manchester (2016)
4. Schrauwen, S.: Machine learning approaches to sentiment analysis using the Dutch netlog corpus. Computational Linguistics and Psycholinguistics Research Center. Accessed 21 Mar 2018
5. Perkins, J., Hardeniya, N.: Natural Language Processing: Python and NLTK. Packt Publishing Ltd, Birmingham (2016)
6. Shoukry, A., Rafea, A.: Sentence-level Arabic sentiment analysis. In: 2012 International Conference on Collaboration Technologies and Systems (CTS), pp. 546–550. IEEE (2012)
7. Islam, M.S., Islam, M.A., Hossain, M.A., Dey, J.J.: Supervised approach of sentimentality extraction from Bengali Facebook status. In: 2016 19th International Conference on Computer and Information Technology (ICCIT), pp. 383–387. IEEE (2016)
8. Kumar, J.V., Shishir, K., Lawrence, F.S.: Extraction of emotions from multilingual text using intelligent text processing and computational linguistics. J. Comput. Sci. **21**, 316–326 (2017)
9. Masdisornchote, M.: A sentiment analysis framework in implicit opinions for Thai language. In: IECON 2015-41st Annual Conference of the IEEE Industrial Electronics Society, pp. 000357–000361. IEEE (2015)
10. Su, Y.-J., Huang, H.-W., Hu, W.-C.: Using idiomatic expression for Chinese sentiment analysis. In: 2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media), pp. 1–4. IEEE (2017)
11. Basiri, M.E., Kabiri, A.: Sentence-level sentiment analysis in Persian. In: 2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA), pp. 84–89. IEEE (2017)