Quest Journals
Journal of Research in Applied Mathematics
Volume 8 ~ Issue 9 (2022) pp: 07-16
ISSN(Online) : 2394-0743 ISSN (Print): 2394-0735
www.questjournals.org

Research Paper

Sentiment Analysis of Hausa Language Tweet Using Machine Learning Approach

Muhammad Sani[1], Abubakar Ahmad[2] and Hadiza S. Abdulazeez[2]
[1]*Department of Mathematical Science, Federal University Dutsin-Ma Katsina state, Nigeria.*
[2]*Department of Computer Science and IT Federal University Dutsin-Ma Katsina state, Nigeria.*

ABSTRACT: *The most appropriate way to gather information is to focus on how and what people think. Sentiment Analysis deals with identifying and classifying opinions expressed in a piece of text. Hausa language is one of the most widely spoken indigenous Lingua Franca in west and central Africa. It's being used as language of trade and spoken either as first or second language by more than 149million people. Therefore, investigation into such a huge population to understand opinions/emotions is not only essential in business, politics, education etc. but also have significant influence on social, economic and security setting. Sentiment Analysis to extract attitudes, appraisals, opinions, emotions and product reviews from Hausa text have a great challenges to researchers and decision makers because such comments are written in an informal language with unstructured format. This paper used both machine learning and lexicon-based approaches to get better classification results. Two machine learning based classification algorithms were employed (Multinomial Naive Bayes MNB and Logistic Regression LR) using Count Vectorizer and TF-IDF method on Hausa dataset generated from BBC Hausa twitter handle using twitter API. The results indicate significant performance of LR over MNB in Hausa language text categorization.*
KEYWORDS: *Machine learning, Multinomial Naive Bayes, Logistic Regression, Count Vectorizer, TF-IDF*

R*eceived 22 August, 2022;* ***R****evised 02 Sep., 2022;* ***A****ccepted 04 Sep., 2022 © The author(s) 2022.*
P*ublished with open access at* ***www.questjournals.org***

I. INTRODUCTION

Sentiment Analysis (SA) is the process of computationally identifying and categorizing opinions expressed in a piece of text especially in order to determine whether the writer's opinion towards a particular topic or product is positive, negative or neutral [9]. Among various methods and tools, SA is one of the key processes that interprets and classify emotions [2]. The adoption of social media platforms, such as Twitter has made SA of tweets an important area of research in customer feedback, public opinion polls, advertisement etc. For enterprises, sentiment analysis is used to identify whether their products are being liked by the customers or not [7]. The reviews, ratings and recommendations for the products can be generalized into positive or negative as well as neutral categories [4]. Before a product is purchase, people look for reviews online from other customers, the amounts of reviews, ratings and recommendations generated are often too large for a normal user to analyze. In order to automate this, various sentiment analysis techniques are used.

There are three basic approaches used for SA on linguistic data. First, is the Lexicon based techniques which is an approach based on a dictionary prepared to store the polarity values of each lexicon. It determines a score for each word in the sentence and annotates using the feature from the lexicon database that is present. Second is the Machine learning technique which requires creating a model by training the classifier with labeled examples such as a positive, negative or a neutral class. Extract the features (words) from the dataset and then train the algorithm based on the examples. The third is the Hybrid technique which requires the two techniques combined together to get better classification results. The approach to be used depends on the nature of the type of data and the platform. Most research carried out in the field of SA employs either lexicon-based analysis or machine learning techniques.

Application of these techniques in analysing sentiment has become an essential research field applied in various domains such as politics, education, tourism, entertainment, commerce, Health etc. Most organizations face numerous issues concerning data extraction. The available software to extract data regarding a person's sentiment on a specific product or service is readily not sufficiently utilized for African languages such as Hausa language. There are relatively little available resources for Hausa language sentiment analysis,

*Corresponding Author: Muhammad Sani
Mathematical Science Department FUDMA, Nigeria

mostly because of the limited scholarly work and research funding on the language when compared to other languages such as English and Arabic.

The aim of this research is to carry out sentiment analysis on Hausa language dataset extracted from BBC Hausa. For this we used both machine learning and lexicon-based approaches to get better classification results. Multinomial Naive Bayes (MNB) and Logistic Regression (LR) was utilized on Hausa dataset from twitter generated via BBC Hausa tweeter handle using twitter API.

II. RELATED WORK

In recent years there is an increasing awareness of Hausa language processing resources and digital information access and storage facilities. [6] introduced Hausa stemmer by modifying the Porter algorithm and used some rules to handle exceptional cases that occur in Hausa language. [3] shortly followed with similar research on the subject utilizing 1500 Hausa root words. The study used 78 affix stripping rules applied in 4 steps and a reference lookup. The use of reference lookup helped in reducing stemming errors. [10] investigate the interaction of syllable structures and syllable weight of Hausa Language showing the phonological interactions that occur within the morphological process of reduplication using Optimality Theory. Another significant study was the design and development of corpora for Hausa language carried out by [8]. A bag of words was created from 268 samples of Hausa language text and consists of a million plus Hausa words in corpus. The words were from a wide range of genres similar to that of widely used English corpora. The study contributes towards addressing the existing gaps for Hausa NLP resources [8]. The first attempt for Hausa sentiment analysis was the proposed model for Part-of-Speech (POS) tagging of Hausa sentences towards the realization of sentiment analysis of Hausa web content [1]. Additionally, the technique for POS tagging of Hausa sentences using the Hidden Markov Model was proposed by [1]. Corpus of Hausa-Based texts from Freedom Radio and Afri Hausa were collected and trained with a model using text collected annotated on the eight basic POS with the addition of processes in the form of a number and tense maker as independent POS tagged sets.

In literature, very few researches have been conducted on sentiment analysis in Hausa, particularly on classifying messages into positive, negative or neutral and for prediction purposes. [8] proposed a text classification framework for Sentiment Analysis based on Multinomial Naive Bayes (MNB) classification Algorithm and Method TF - IDF. A framework concept oriented towards the MNB algorithm and the TF - IDF module was developed. Review and comparison of some modern Naive Bayesian classifiers were performed based on their ability to classify a large number of text documents efficiently [8]. Significant results in text categorization performance with MNB Model were achieved which improved the performance of the datasets. Multinomial Naïve Bayes algorithm is a fast, easy to implement with almost modern text categorization algorithm. Their conclusion suggested that some changes can be made to the classifier for greater accuracy in the future work and would involve the use of artificial intelligence to improve the accuracy up to the best extent [8].

In this research, the authors develop a text classification framework for Hausa Sentiment Analysis based on MNB and Logistic Regression. Comparison is done to ascertain the best possible model for classifying Hausa text.

III. METHODOLOGY

This research used both approaches (MNB and LR) to combined Hausa text, which is the Lexicon-based and Machine learning for sentiment analysis on Twitter data. The algorithms were implemented for pre-processing data set for filtering as well as reducing the noise from the data set. The data set were trained using machine learning Multinomial Naïve Bayes and logistic regression algorithm for measuring the performance and accuracy of the trained data set. The diagram in Figure 1 shows an abstract view of the approach that combines the lexicon-based and machine learning for sentiment analysis.

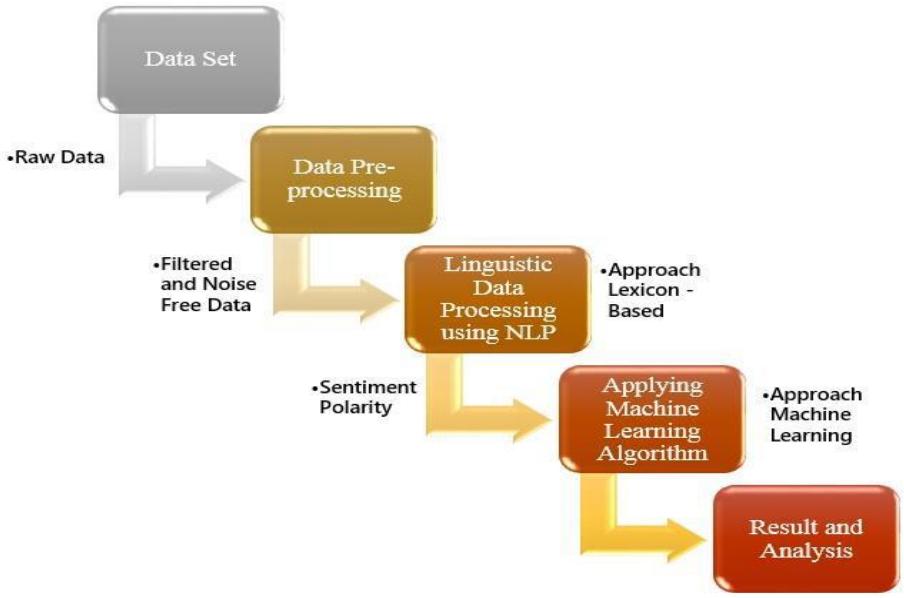


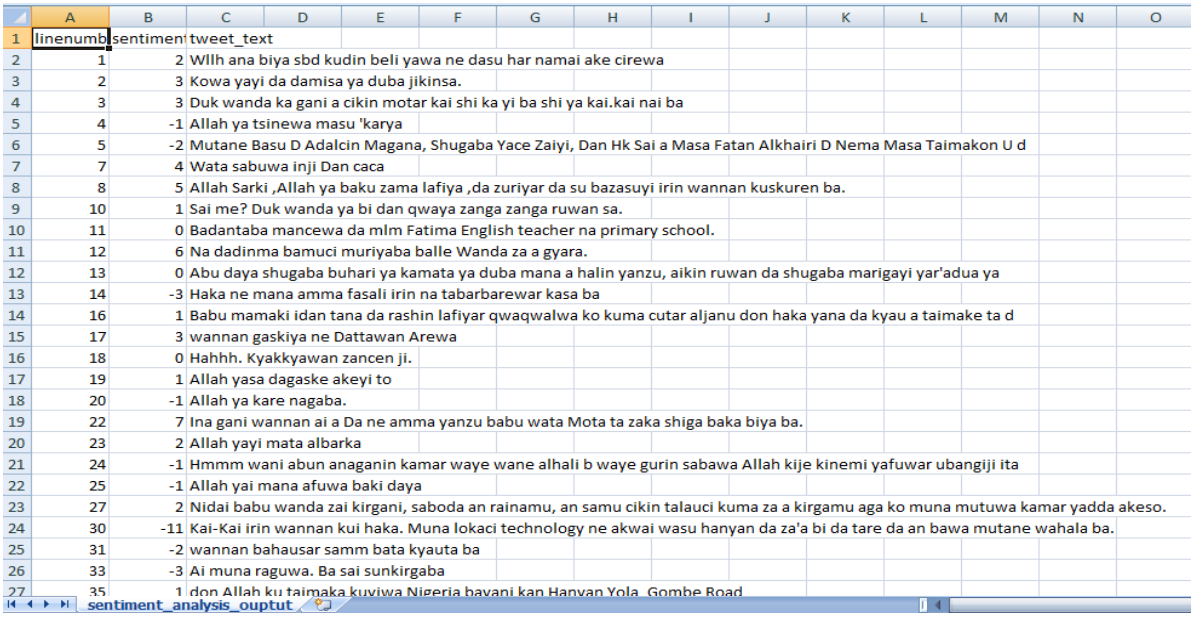
Fig 1: Overview on approach for Hybrid Sentiment Analysis

Dataset

In this dataset, more than 20,000 words were collected from SentiWord.Net. Each of the words were translated and manually annotated as positive and negative using the same value for each word on SentiWord.Net. The dataset was derived from BBC Hausa twitter handle. A total of 4790 tweets were collected automatically from @BBCHAUSA twitter handle using the Twitter application programming Interface (API) and they were annotated as positive or negative using R programming. The database created contains main tweet and comments of Hausa sentences in CSV file. The attributes that are in the dataset consist of comments, opinions, rating and review.

A	B	C	D	E	F	G
116	acaudal	karin	-0.25	negative	0	
117	acaudate	wadata	-0.25	negative	0	
118	acaulescent	karin	0.25	positive	1	
119	accelerative	hanzari	-0.125	negative	0	
120	accelerator factor	mai hanzari	-0.25	negative	0	
121	acceleratory	hanzari	-0.125	negative	0	
122	accent	lafazi	0.4375	positive	1	
123	accenting	karin magana	0.125	positive	1	
124	accentuation	accentuation	0.125	positive	1	
125	accept	karba	0.2	positive	1	
126	acceptably	karba	0.5	positive	1	
127	acceptation	yarda	0.375	positive	1	
128	accepted	karba	0.5	positive	1	
129	accepting	karba	0.25	positive	1	
130	acceptive	karbi	0.125	positive	1	
131	acceptor	mai karbar rna	-0.375	negative	0	
132	accessary	dama	0.375	positive	1	
133	accessible	samun dama	0.125	positive	1	
134	accessional	karin girma	0.125	positive	1	
135	accessory		0	negative	0	
136	accident	hadari	-0.4375	negative	0	
137	accidental	mai hadari	-0.125	negative	0	
138	accidental injury	rauni na hadari	-0.75	negative	0	
139	accidentally	bazata	0.125	positive	1	
140	acclaim	yabo	0.375	positive	1	

Fig 2: A sample of manually annotated words translated from SentiWord.Net



	A	B	C
1	linenumb	sentimen	tweet_text
2	1	2	Wllh ana biya sbd kudin beli yawa ne dasu har namai ake cirewa
3	2	3	Kowa yayi da damisa ya duba jikinsa.
4	3	3	Duk wanda ka gani a cikin motar kai shi ka yi ba shi ya kai.kai nai ba
5	4	-1	Allah ya tsinewa masu 'karya
6	5	-2	Mutane Basu D Adalcin Magana, Shugaba Yace Zaiyi, Dan Hk Sai a Masa Fatan Alkhairi D Nema Masa Taimakon U d
7	7	4	Wata sabuwa inji Dan caca
8	8	5	Allah Sarki ,Allah ya baku zama lafiya ,da zuriyar da su bazasuyi irin wannan kuskuren ba.
9	10	1	Sai me? Duk wanda ya bi dan qwaya zanga zanga ruwan sa.
10	11	0	Badantaba mancewa da mlm Fatima English teacher na primary school.
11	12	6	Na dadinma bamuci muriyaba balle Wanda za a gyara.
12	13	0	Abu daya shugaba buhari ya kamata ya duba mana a halin yanzu, aikin ruwan da shugaba marigayi yar'adua ya
13	14	-3	Haka ne mana amma fasali irin na tabarbarewar kasa ba
14	16	1	Babu mamaki idan tana da rashin lafiyar qwaqwalwa ko kuma cutar aljanu don haka yana da kyau a taimake ta d
15	17	3	wannan gaskiya ne Dattawan Arewa
16	18	0	Hahhh. Kyakkyawan zancen ji.
17	19	1	Allah yasa dagaske akeyi to
18	20	-1	Allah ya kare nagaba.
19	22	7	Ina gani wannan ai a Da ne amma yanzu babu wata Mota ta zaka shiga baka biya ba.
20	23	2	Allah yayi mata albarka
21	24	-1	Hmmm wani abun anaganin kamar waye wane alhali b waye gurin sabawa Allah kije kinemi yafuwar ubangiji ita
22	25	-1	Allah yai mana afuwa baki daya
23	27	2	Nidai babu wanda zai kirgani, saboda an rainamu, an samu cikin talauci kuma za a kirgamu aga ko muna mutuwa kamar yadda akeso.
24	30	-11	Kai-Kai irin wannan kui haka. Muna lokaci technology ne akwai wasu hanyan da za'a bi da tare da an bawa mutane wahala ba.
25	31	-2	wannan bahausar samm bata kyauta ba
26	33	-3	Ai muna raguwa. Ba sai sunkirgaba
27	35	1	don Allah ku taimaka kuyiwa Nigeria bayani kan Hanyan Yola, Gombe Road

sentiment_analysis_ouptut

Fig 3: Sentiment Analysis of tweet data using R.

Data Pre-processing

The second step includes the data pre-processing, this is where the extracted data is transformed to remove the inconsistencies and improve the quality of the data. First, the key attribute fields were identified and then tokenization was performed where a stream of text is split into smaller units called tokens and the NLTK tokenization package was then used for pre-processing. This was used to remove repeated letter, hash-tags, URLs, special symbols, references, emoticons, stop words etc. The basic step which includes removal of the URLs, whitespaces at the beginning, in between or at the end of the tweets, special characters like punctuation and repetition of characters, username, and Hash tags from the text data makes it more meaningful and helps to reduce noise as well as the size of the dataset to increase the performance for further data processing task. In this report, all extra white space was removed using the built in function strip (). Secondly, all the meaningless and unnecessary special characters from the tweets were eliminated. These characters include: "[ ] $ - + ( ) <># % *", and a few more. These characters do not have specific meaning neither do they explain if the characters are used for positivity or negativity.

Linguistic Data Processing Using NLP

Analysing data requires some basic steps; Firstly, the documents were prepared in a proper text format. The second step involves tokenization/feature extraction (which means dividing the data into different set of statement so that the computer understands very well). The third step is detection and negation which involves targeting the keyword in the data, if the word is found return "True" to verify else "False" for negated. The dependency parser analyzes the grammatical structure of the sentence if the value is "True". Co-reference parser analyzes the expression and it is the main object in NLP. The last step involves analyzing the result.

Applying Machine Learning

The nature of the learning "signal" or "feedback" available to a learning system depends on the Machine learning task. In this research the Multinomial Naïve Bayes and Logistic Regression classifier were employed.

Naïve Bayes is a probabilistic machine learning algorithm that can be used in a wide variety of classification task. It is a commonly used method for text classification due to its effective grading assumptions, quick and easy implementation. The classifiers of Naïve Bayes (NB) are a family of classifiers based on bayes' popular probability theorem. The Multivariate Bernoulli Naïve Bayes model (BNB) is done to classify documents and treat the absence of each word as a logical attribute as in one of the initial statistical models of language. It is well thought out but it only focuses on the appearance of words which make it a baseline for text classification. In BNB, when a word appears in the document, the value of the attribute equivalent to that word is written either as one, otherwise zero. The Multinomial Naïve Bayes (MNB) was proposed as an improved method of BNB. The MNB assumes that the document is a bag of words and takes word frequency and information into account

Principle of Multinomial Naive Bayes Classifier

Naïve Bayes multinomial classifier compute class probabilities for a given text for purpose of classification. If *C* is the set of classes and N is the size of a vocabulary, Naïve Bayes multinomial classifier classifies a text document to the class which has the highest-class membership probability Pr ($c|t_i$), which can be expanded by Bayes' rule,

$$\boldsymbol{Pr(c|ti) = \frac{Pr(c)\,Pr(ti|c)}{Pr(ti)}, c \in C} \qquad (1)$$

Where the prior probability, pr(c) is calculated as:

$$\boldsymbol{Pr(c) = \frac{numberofdocumentsofclassc}{totalnumberofdocuments}} \qquad (2)$$

The probability of a document given a class c is considered as a multinomial distribution:

$$\boldsymbol{Pr(ti|c)=(\sum_n f_{ni})!\prod_n \frac{Pr(Wn|c)fni}{fni}} \qquad (3)$$

Where f_{ni} = the count of word *n* in our test document t_i and $Pr(W_n|c)$ = probability of word *n* given class c, which is estimated from the training documents as:

$$\boldsymbol{Pr(Wn|c)= \frac{1+Frnc}{N+\sum_{x=1}^{N} Frxc}} \qquad (4)$$

Where Fr_{xc} = count of word *x* in all the training documents in class c and the normalization factor $Pr(t_i)$ for Eq. 1 can be computed as:

$$\boldsymbol{Pr(t_i)= \sum_{k=1}^{|C|} Pr(k)Pr(ti|k)} \qquad (5)$$

It is obvious that the terms and in Eq. 3are computationally expensive and neither of these depends on the class c. So, these terms can be omitted to rewrite the Eq. 3as:

$$\boldsymbol{Pr(t_i|c)= \alpha \prod_n Pr(Wn|c)fni,} \qquad (6)$$

where α is a constant. For implementing our sentiment analysis system, we have chosen the classifier "Multinomial Naïve Bayes", included in Python. To apply multinomial naïve Bayes classifier, we have considered each tweet as a document.

Logistic regression is a very popular method in machine learning and statistics for past centuries. It is a statistical method for describing relationships. Its recognizably used in various fields like; medicine, business, technology and more. Logistic Regression analysis focuses on the classification of individuals in different groups which establishes the relationship between a categorical variable and one or more independent variable. This relationship is used in machine learning to determine and predict the result of a categorical variable [11]. Generally, the two outcomes of the response variable are called 'success' and 'failure', 'yes or no' or 'true and false' represented by '1 and 0' respectively. Logistic regressions work with odds rather than proportions. The odds are simply the ratio of the proportions for the two possible outcomes. If *p* is the proportion for one outcome, then 1-*p* is the proportion for the second outcome.

Logistic Regression Model

The statistical model for logistic regression is;

$$\boldsymbol{log\left(\frac{p}{1-p}\right) - \beta_0 + \beta_1 x}$$

Where *p* is a binomial proportion and *x* is the explanatory variable. The parameters of the logistic regression model are b_0 and b_1.

IV. RESULTS / FINDINGS

In this section we discuss the results obtained through the machine learning technique used in this paper for training sentiment labeled dataset, the result obtained using each technique were both analyzed. The performance evaluation, accuracy and classification result obtained using multinomial Naïve Bayes and logistic regression supports the objective for processing linguistic data set using Natural language processing (NLP)

techniques and measures the accuracy of the sentiment labeled data set. In comparison of using MNB and Logistic Regression for measuring the accuracy for sentiment labeled dataset, Logistic Regression algorithm gives high accuracy in the data classification. Even though, Naïve Bayes gives better performance throughout for the data classification and trains the data set with the huge file size with greater speed.

Enhancement

Since there are words in several documents of both classes, they do not provide any relevant information. To overcome this problem, we have used two vectorizer; the term frequency inverse document frequency (TF-IDF) and Count Vectorizer. It takes into account the frequency and uniqueness of words. Figure 4, 5, 6 and 7 show the estimated predictive result of MNB classifier and figure 8,9,10 and 11 shows the estimated predictive result for Logistic Regression classifier. Similarly, Table 1 shows a comparative measurement on the basis of overall accuracy.

```
### Model Built ###

              precision    recall  f1-score   support

           0       0.77      0.87      0.82       306
           1       0.84      0.72      0.77       284

    accuracy                           0.80       590
   macro avg       0.80      0.80      0.80       590
weighted avg       0.80      0.80      0.80       590
```

Fig 4: Result of Analysis Using Multinomial Naïve Bayes with TF-IDF

```
### Model Built ###

              precision    recall  f1-score   support

           0       0.84      0.76      0.80       306
           1       0.77      0.84      0.80       284

    accuracy                           0.80       590
   macro avg       0.80      0.80      0.80       590
weighted avg       0.80      0.80      0.80       590
```

Fig 5: Result of Analysis Using Multinomial Naïve Bayes with CountVectorizer

The prediction of sentiment classes (positive and Negative) using Naïve Bayes classifier has accuracy rate of 80% each with TF-IDF and Count Vectorizer. The accuracy of a classifier on a given data set is the percentage of the data set tuples that are correctly classified by the classifier. The confusion matrix generated is shown in fig. 6 and 7 respectively. A confusion matrix is a useful tool for analyzing how well your classifier can identify tuples of the different class.

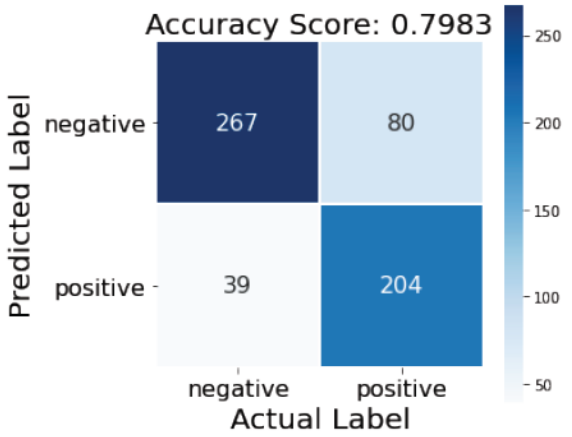


Fig 6: Confusion Matrix generated by Naïve Bayes with TF-IDF

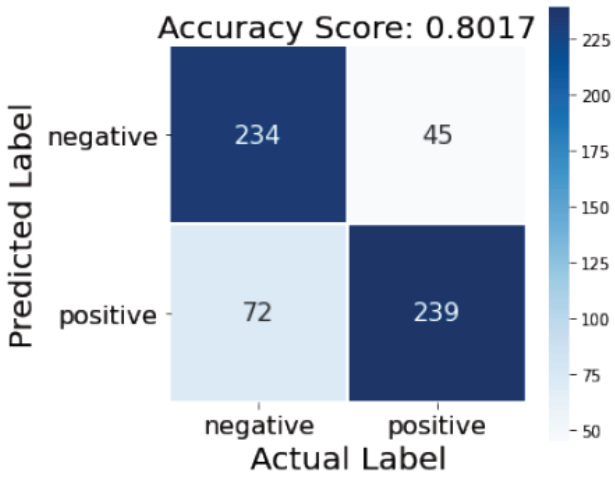


Fig 7: Confusion Matrix generated by Naïve Bayes with CountVectorizer

```
### Model Built ###

              precision    recall  f1-score   support

           0       0.83      0.93      0.88       306
           1       0.91      0.80      0.85       284

    accuracy                           0.86       590
   macro avg       0.87      0.86      0.86       590
weighted avg       0.87      0.86      0.86       590
```

Fig 8: Result of Analysis Using Multinomial Logistic Regression with countVectorizer

```
### Model Built ###
```

	precision	recall	f1-score	support
0	0.79	0.93	0.85	306
1	0.90	0.73	0.81	284
accuracy			0.83	590
macro avg	0.85	0.83	0.83	590
weighted avg	0.84	0.83	0.83	590

Fig 9: Result of Analysis Using Multinomial Logistic Regression with TF-IDF

The prediction of sentiment classes (positive and Negative) using Logistic Regression classifier has accuracy rate of 83% and 86% with TF-IDF and Count Vectorizer respectively. The confusion matrix generated is shown below:

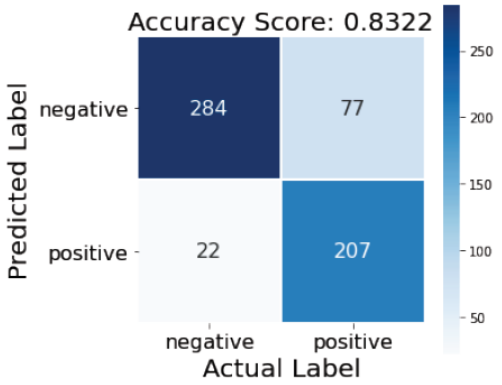


Fig 10: Confusion Matrix generated by Logistic Regression with TF-IDF

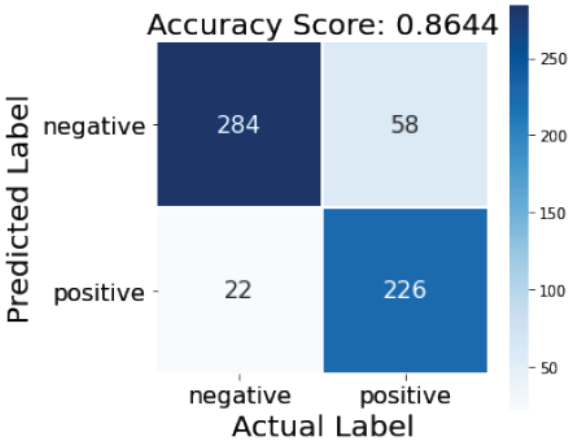


Fig 11: Confusion Matrix generated by Logistic Regression with count Vectorizer

Table 1: The summary of the performance of the models using both vectorizer.

	Parameters	Logistic Regression				Naïve Bayes			
		Count Vectorizer		TF-IDF		CountVectorizer		TF-IDF	
		0	1	0	1	0	1	0	1
1	Precision	0.83	0.91	0.79	0.90	0.84	0.77	0.77	0.84
2	Recall	0.93	0.80	0.93	0.73	0.76	0.84	0.87	0.72
3	Fi-Score	0.88	0.85	0.85	0.81	0.80	0.80	0.82	0.77
4	Supports	306	284	306	284	306	284	306	284
5	Accuracy	86%		83%		80%		80%	

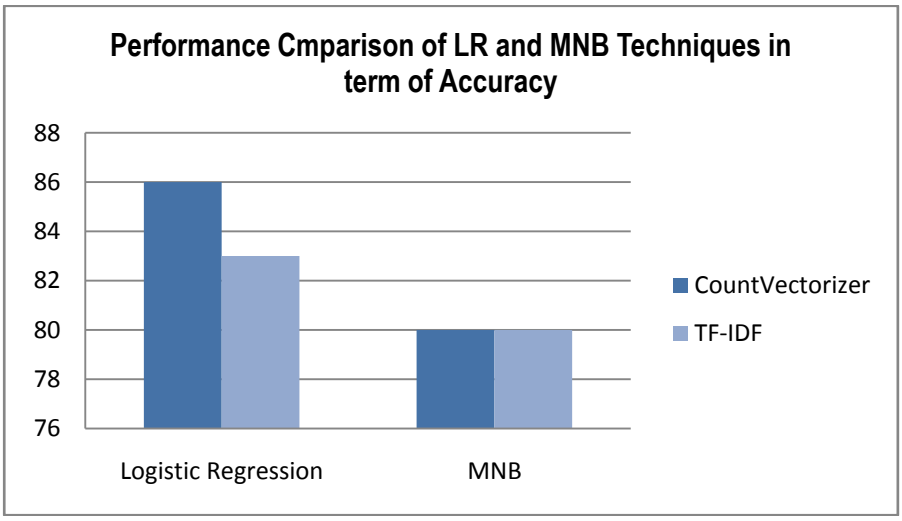


Fig 12: Performance Comparison of LR and MNB Techniques in term of Accuracy

V. CONCLUSION

In this research, we presented a text classification framework for Sentiment Analysis based on MNB and Logistic Regression classification algorithm, using the Count Vectorizer and TF – IDF method. A significant result has been achieved in text categorization performance with the help of the Model. The result clearly shows that the Logistic Regression algorithm stands ahead in comparison to Multinomial Naïve Bayes classification algorithm except for the time taken to train the data and the memory consumption during data classification. It also shows that the time for training the same data set using Logistic Regression takes longer time in comparison to Naïve Bayes. Therefore, it has been concluded that both classifiers used for analyzing sentiment data set are accurate, but logistic regression algorithm stands ahead in terms of the overall accuracy.

For the future work on sentiment analysis, it is necessary to work on abbreviations and acronyms, as most internet users tend to write comment, replies or opinions by shortening some words which may reduce the sentiment polarity.

REFERENCES

[1]. Aminu Tukur, Kabir Umar and Anas Saidu Muhammad. Tagging Part of Speech in Hausa Sentences. *15th International Conference on Electronics, Computer and Computation (ICECCO).* Abuja, Nigeria. 2019

[2]. Cabrera-Diego A. Luis, Bessis N and Korkontzelos. Classifying emotions in Stack Overflow and JIRA using a multi-label approach. Elsevier. 2020 195, 105633

https://doi.org/10.1016/j.knosys.2020.105633
[3]. Dalmia, A., Manish, G., Vasudeva, V. Twitter Sentiment Analysis The good, the bad and the neutral! IIIT-H at SemEval. 2015
[4]. Deepali A., Kin Fun, L., and Stephen, W. N. Consumers' sentiment analysis of popular phone brands and operating system preference using Twitter data: A feasibility study", IEEE. 2015.
[5]. Deng,L,.and Yu D. Deep Learning: Methods and Applications. 2014. http://research.microsoft.com/pubs/209355/DeepLearning-NowPublishing-Vol7-SIG-039.pdf
[6]. Friedman, C., Rindflesch, T. C., and Corn, M., (2013). Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. 2013 46(5): pp. 765–773.
[7]. Harshal, K., Kalyani, G. & Tanmay, S. A review on: Sentiment polarity analysis on Twitter data from different Events. *International Research Journal of Engineering and Technology (IRJET)*, 2018, 5(3): 1479.
[8]. Muhammad A. S., Muktar M. Aliyu and Sani I. Zimit. Towards the Development of Hausa Language Corpus. *International Journal of Scientific & Engineering Research*, 2019 10(10): pp1598 - 1604.
[9]. Oxford Online Dictionary. Accessed, 2:35 AM, 10th, January 2021:https://www.lexico.com/definition/sentiment_analysis
[10]. Ravikumar, P. Sentiment Analysis on Twitter Data Using Machine Learning. 2017.
[11]. Sucky, R. N. A complete Logistic Regression Algorithm from Scratch. *Towards Datascience*. 2020
[12]. Trstenjak B., Mikac S., and Donko D., (2014) "KNN with TF-IDF based framework for text categorization," Procedia Eng., 2014, 69 pp. 1356–1364.