DETECTING SPAM EMAIL WITH MACHINE LEARNING OPTIMIZED WITH BIO INSPIRE META HEURISTIC ALGORITHM

A.Durga Devi[1], Manchem, Ramya[2]

[1] **Assistant Professor MCA, DEPT,** Dantuluri Narayana Raju College **, Bhimavaram, Andharapradesh**

Email id:- adurgadevi760@gmail.com

[2]**PG Student of MCA,** Dantuluri Narayana Raju College **, Bhimavaram, Andharapradesh**

Email id:- ramya.m44446@gmail.com

ABSTRACT

Spam emails have been a chronic issue in computer security. They are very costly economically and extremely dangerous for computers and networks. Despite of the emergence of social networks and other Internet based information exchange venues, dependence on email communication has increased over the years and this dependence has resulted in an urgent need to improve spam filters. Although many spam filters have been created to help prevent these spam emails from entering a user's inbox, there is a lack or research focusing on text modifications. Currently, Naive Bayes is one of the most popular methods of spam classification because of its simplicity and efficiency. Naive Bayes is also very accurate; however, it is unable to correctly classify emails when they contain leetspeak or diacritics. Thus, in this proposes, we implemented a novel algorithm for enhancing the accuracy of the Naive Bayes Spam Filter so that it can detect text modifications and correctly classify the email as spam or ham. Our Python algorithm combines semantic based, keyword based, and machine learning algorithms to increase the accuracy of Naive Bayes compared to Spam assassin by over two hundred percent.

1 INTRODUCTION

Spam refers to an email aimed manipulating an individual to whom it is aimed at or just randomly flooding the inbox. It is also called as junk mail and it floods Internet clients Inboxes. Today spam emails are of a variety of types ranging from ads to business promoting to doubtful products to some objectionable services. Therefore it is difficult to identify and classify an email as spam or non-spam.

Usenet also called as User Network is an email service that distributes group talks or emails aimed at a particular group of people associated with a certain service or product and are mostly informative but do crowd up the inbox of the user. The data that goes over the Internet is called Netnews an accumulation of these data that is aimed at providing message about a specific topic is called a newsgroup. People that read such news from these newsgroups are the prime target of Spammers. Spammers use these news groups for the promotion of certain unrelated ads or unrelated posts. Usenet spam robs clients of the utility of the newsgroups by promoting other unrelated posts.

2.LITERATURESURVEYANDRELATEDWORK

Email Spam Detection using integrated approach of Naïve Bayes and Particle Swarm Optimization[1]

Naïve Bayes algorithm is a Bayes theorem based statistical machine learning based approach having properties of strong independence, probability distribution and ability to handle large datasets. In NB, probability distribution is evaluated from the frequency distribution of dataset.Particle Swarm Optimization (PSO) is swarm intelligence based concept derived in 1995 by Eberhart and KennedyPSO work on the property of stochastic distribution and initially find the local search solution, then individual particle share their solution and global solution is obtained.NB having probability distribution property determines the possible class for the email content from the spam class or non-spam class on the basis of keywords present in the email textual data.PSO is used to further optimize the parameters of NB approach to improve the accuracy, search space and classification process.

Email Spam Classification by Support Vector Machine[2]

This paper uses Support Vector Mechanism algorithm to identify spam emails. Descriptions as provided on Spam Assassin website for the dataset used in this paper. SVM is also considered as an important kernel methods, which is one of the most important areas in machine learning concepts. Smart Traffic Control System with Application of Image Processing Techniques In this work they have also compared Linear and Gaussian as two of the very popular kernel and employed them for the problem of email spam detection The two models have been proposed, trained and tested using popular and often used standard database.

Intelligent Model for Classification of SPAM and HAM[3]

In this paper they have used machine learning and non machine learning approaches. Machine learning approaches like support vector mechanism, neural network etc. Non machine learning approaches like strong key word searching and whitelisting and blacklisting of words. The sets so formed are further used as training set and the classification set.

The process is to use the first set as training set and the remaining N-1 sets as the sets to be classified. In the next iteration the second set is used as the training set and the remaining sets are sets to be classified. The process is repeated until all the sets are used as training sets. The emails are classified based on the spam percentage each mail gains.

3 PROPOSED WORK AND ALGORITHM

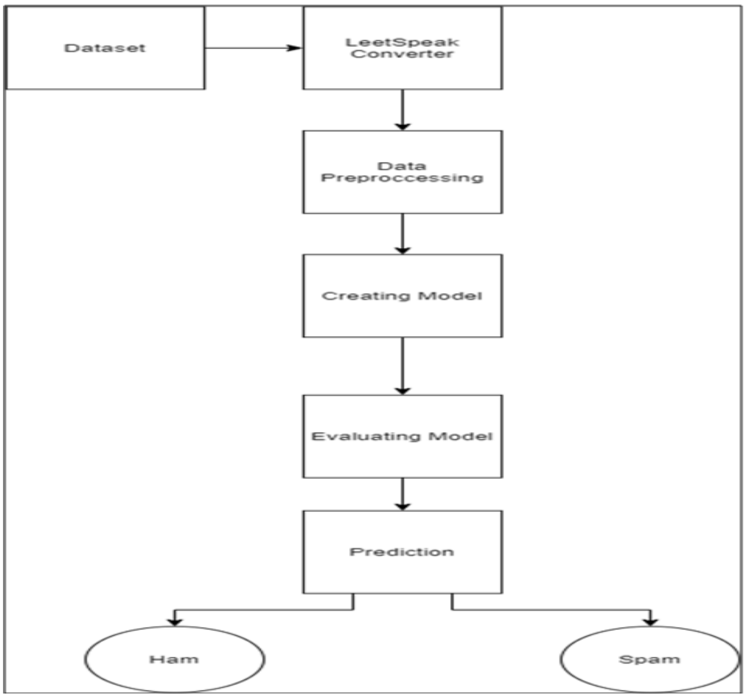

Fig-1: System Architecture

The above figure represents the Proposed System of the project. We first take the dataset then it is passed through the leet speak converter which converts all the leetspeak characters into normal text. Then that data set is set for data pre processing where in which all the unwanted and stop words are removed in order to minimize the dataset as much as possible. Then a model is created to test the data which is evaluated. Finally the dataset is fed to this model which predicts the output whether the given text is spam or ham.

4 METHODOLOGIES

1.Load Dataset:
Load data set using pandas read_csv() method. Here we will read the excel sheet data and store into a variable.

2.Split Data Set:
Split the data set to two types. One is train data test and another one is test data set.here we will remove missing values from the dataset.

3.Train data set:
Train data set will train our data set using fit method. 80% of data from dataset we use for training the algorithm.

4.Test data set:
Test data set will test the data set using algorithm. 20% of data from dataset we use for testing the algorithm.

5.Predict data set:
Predict() method will predict the results. In this step we will predict the ranking of the google play store app.

5.RESULTSANDDISCUSSION SCREENSHOTS

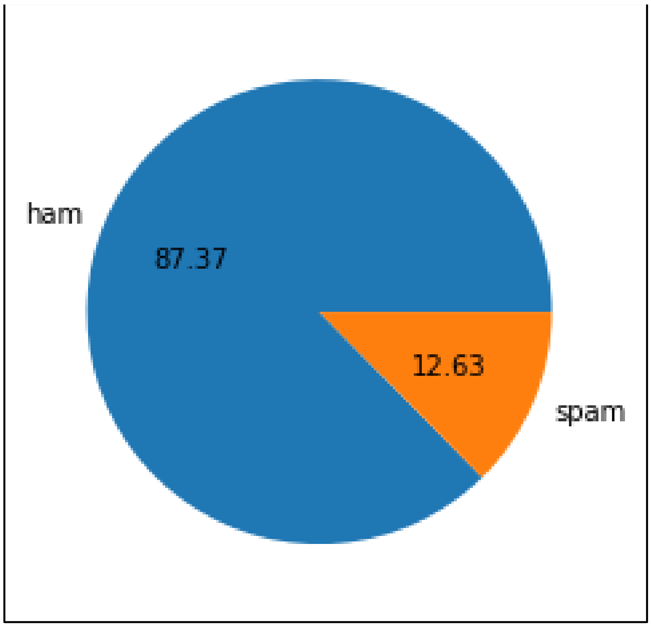

Fig 2:- Pie-chat representation of spam and ha mails

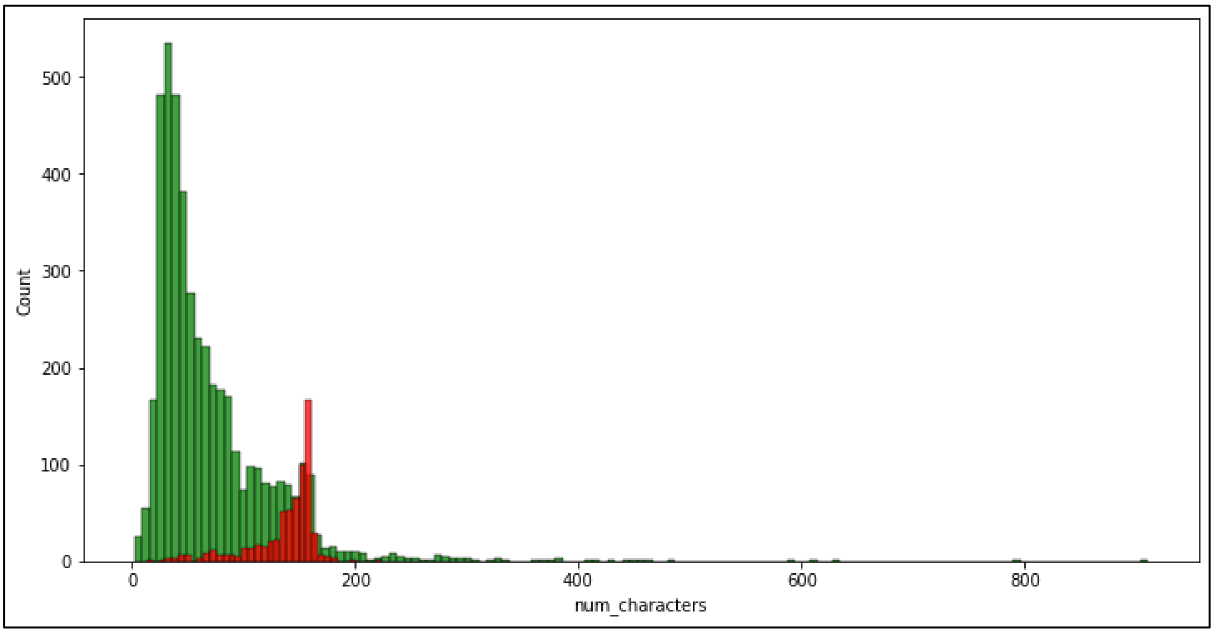

Fig 3:- Graph-2

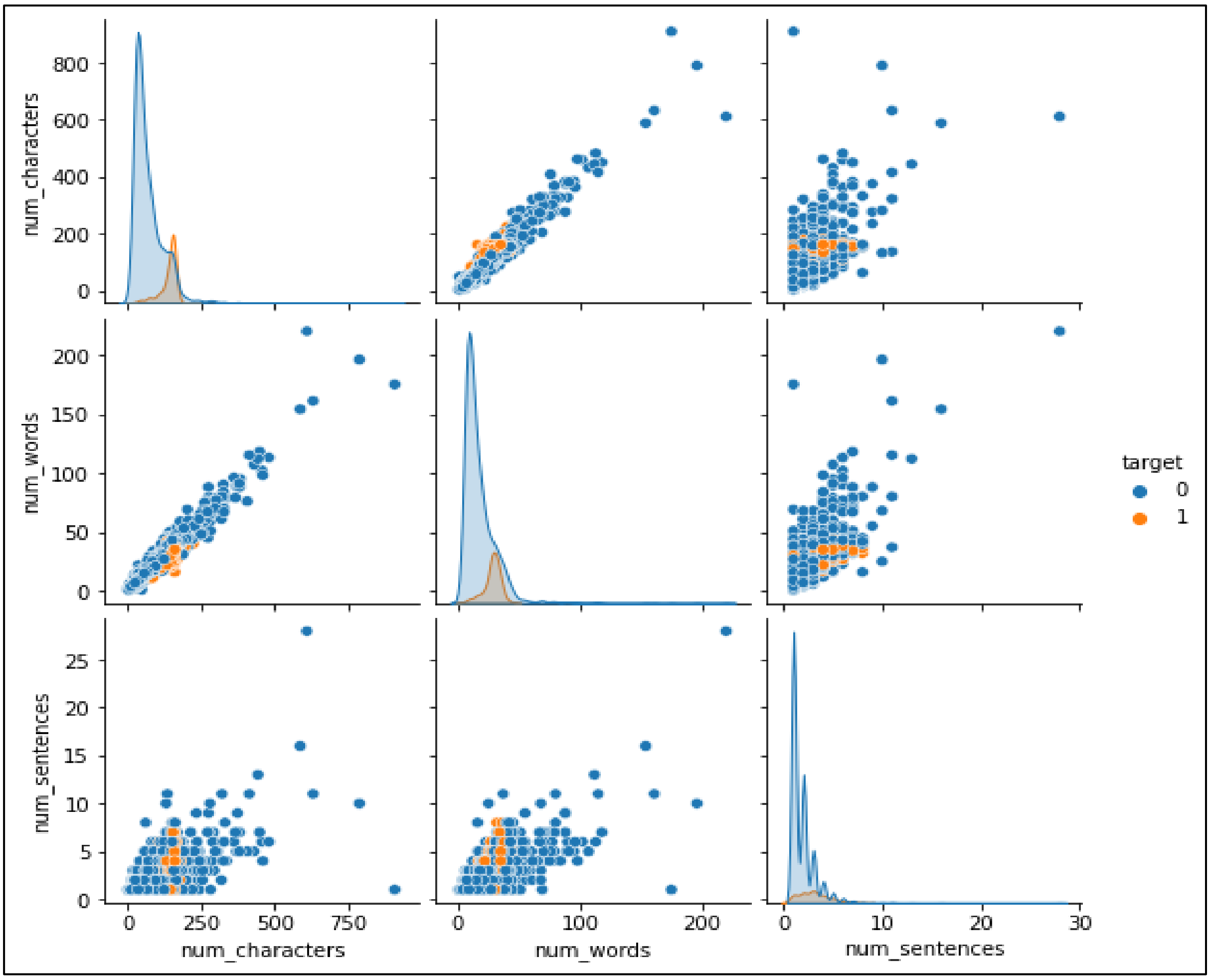

Fig 4 :- Sub plots

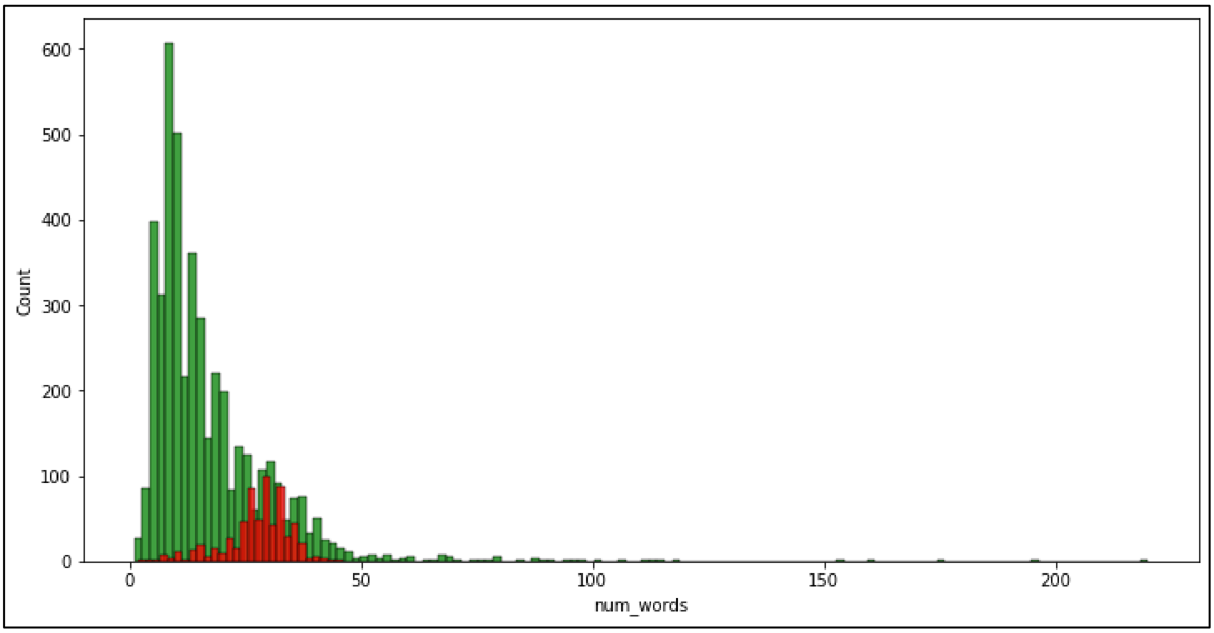

Fig 5 :- Graph-4

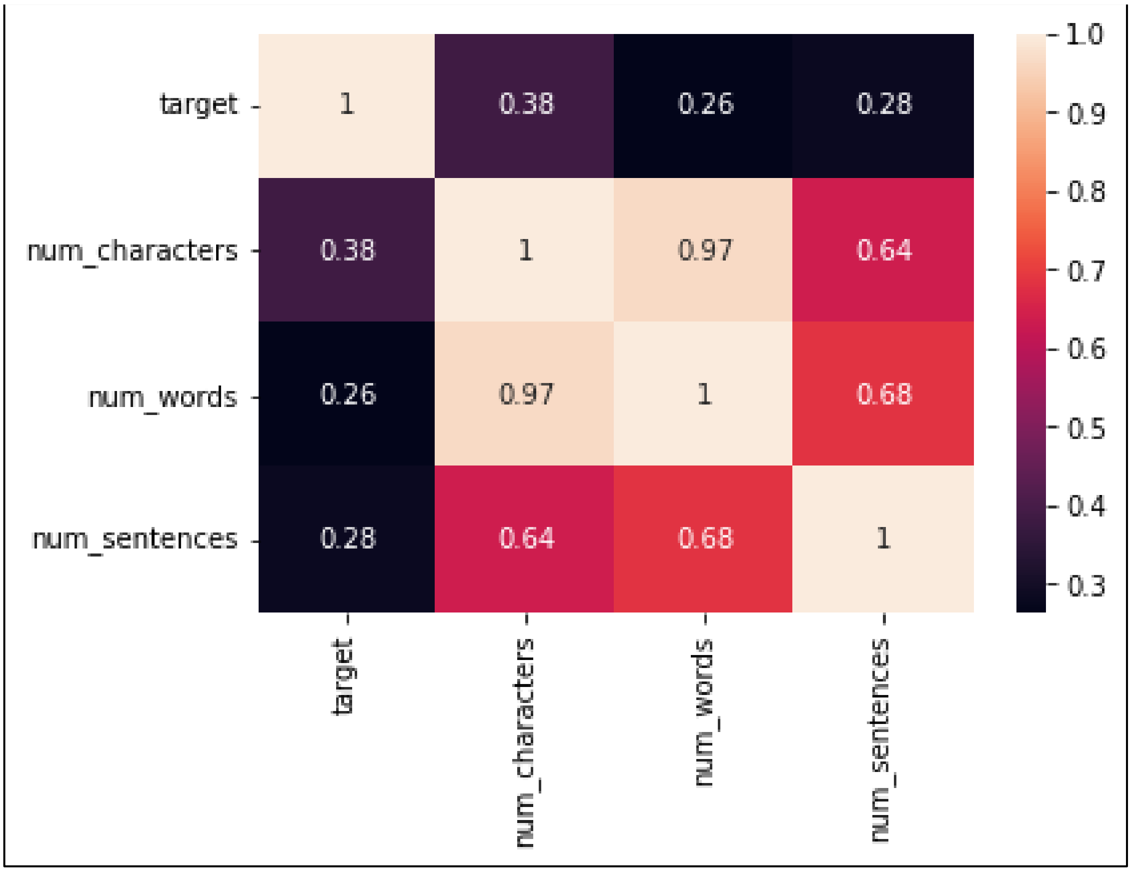

Fig 6 :- Box plot

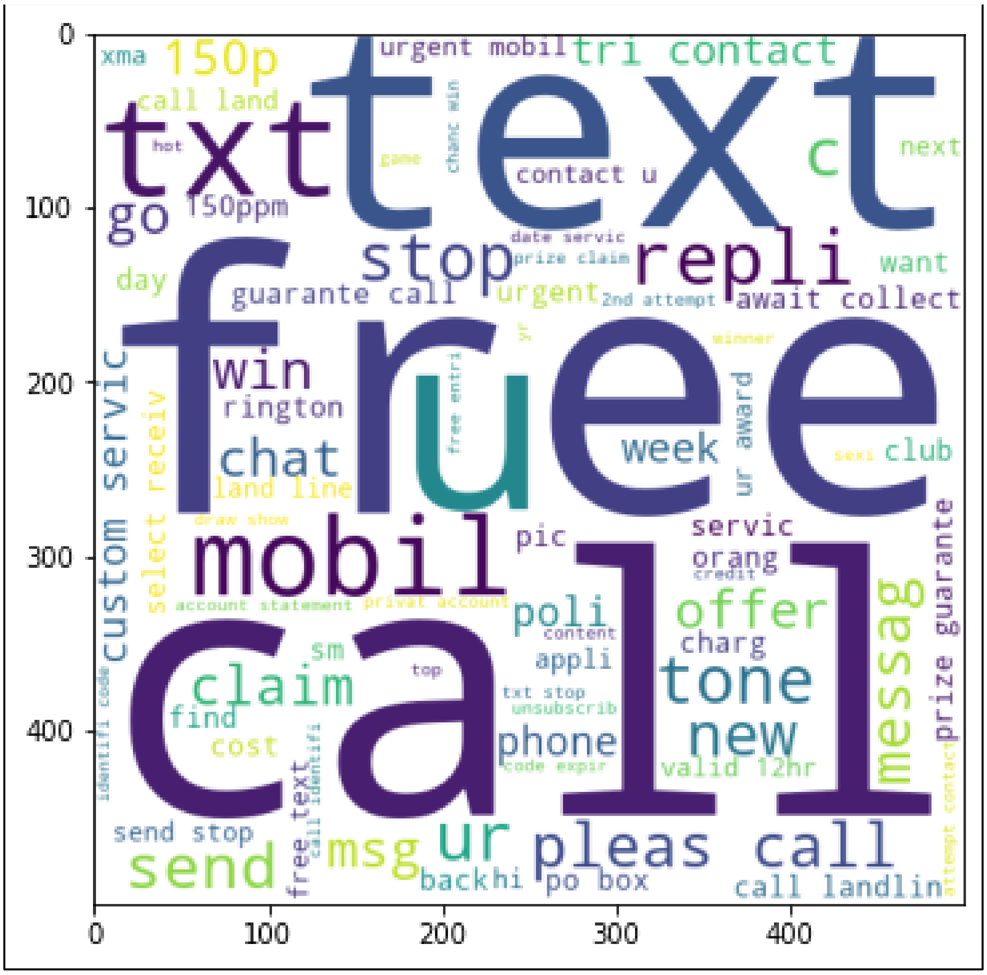
0
100
200
300
400
0
100
200
300
400
text
free
call
txt
mobil
u
repli
stop
xma
150p
urgent mobil
tri contact
call land
hot
game
chanc win
contact u
c
next
go
150ppm
date servic
prize claim
want
day
guarante call
urgent
2nd attempt
await collect
yr
winner
win
rington
free entri
ur award
custom servic
select receiv
chat
week
sexi
club
land line
draw show
pic
servic
orang
credit
account statement
privat account
poli
offer
content
messag
prize guarante
sm
appli
charg
claim
top
tone
txt stop
unsubscrib
identifi code
find
cost
call identifi
phone
new
code expir
valid 12hr
attempt contact
send stop
free text
msg
ur
pleas call
send
back
hi
po box
call landlin

Fig 7 :- Graph-6

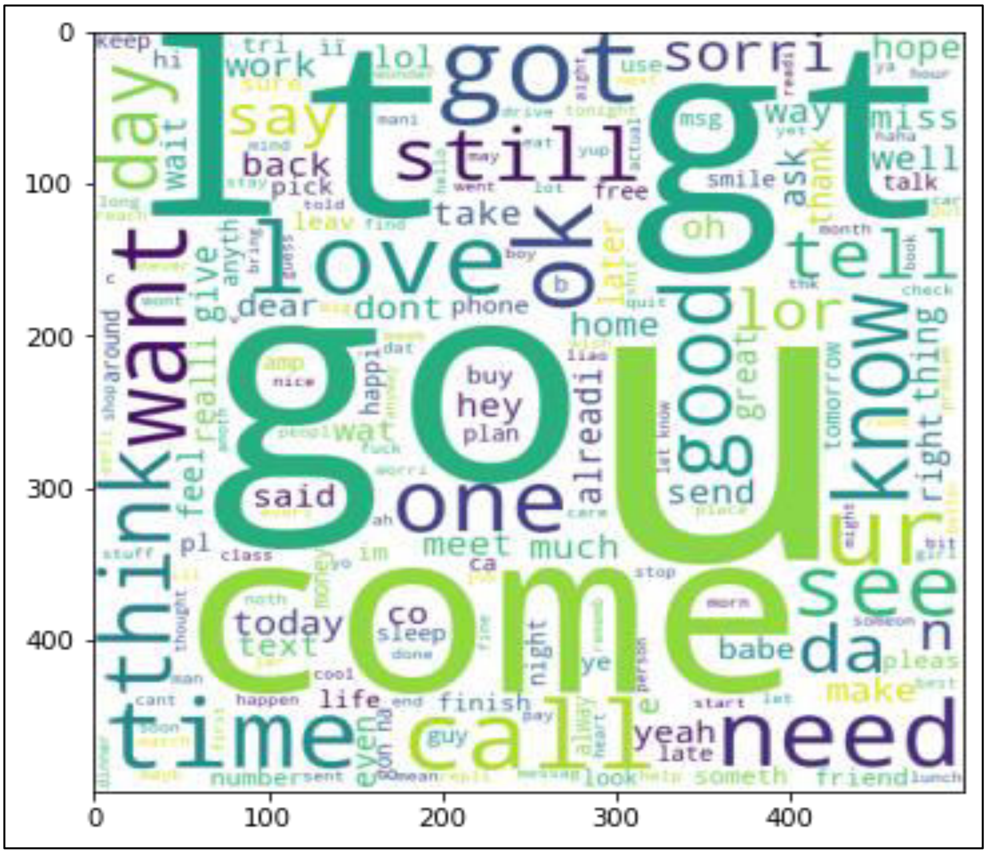

Fig 8 :- Graph-7

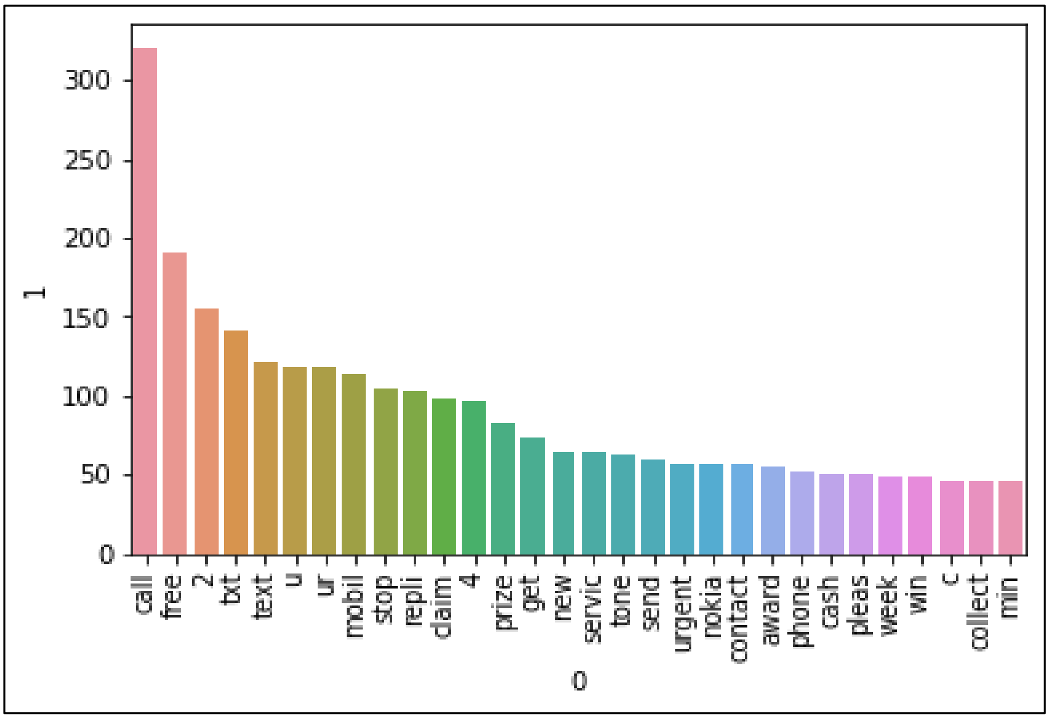

Fig-9 :- Word count graph

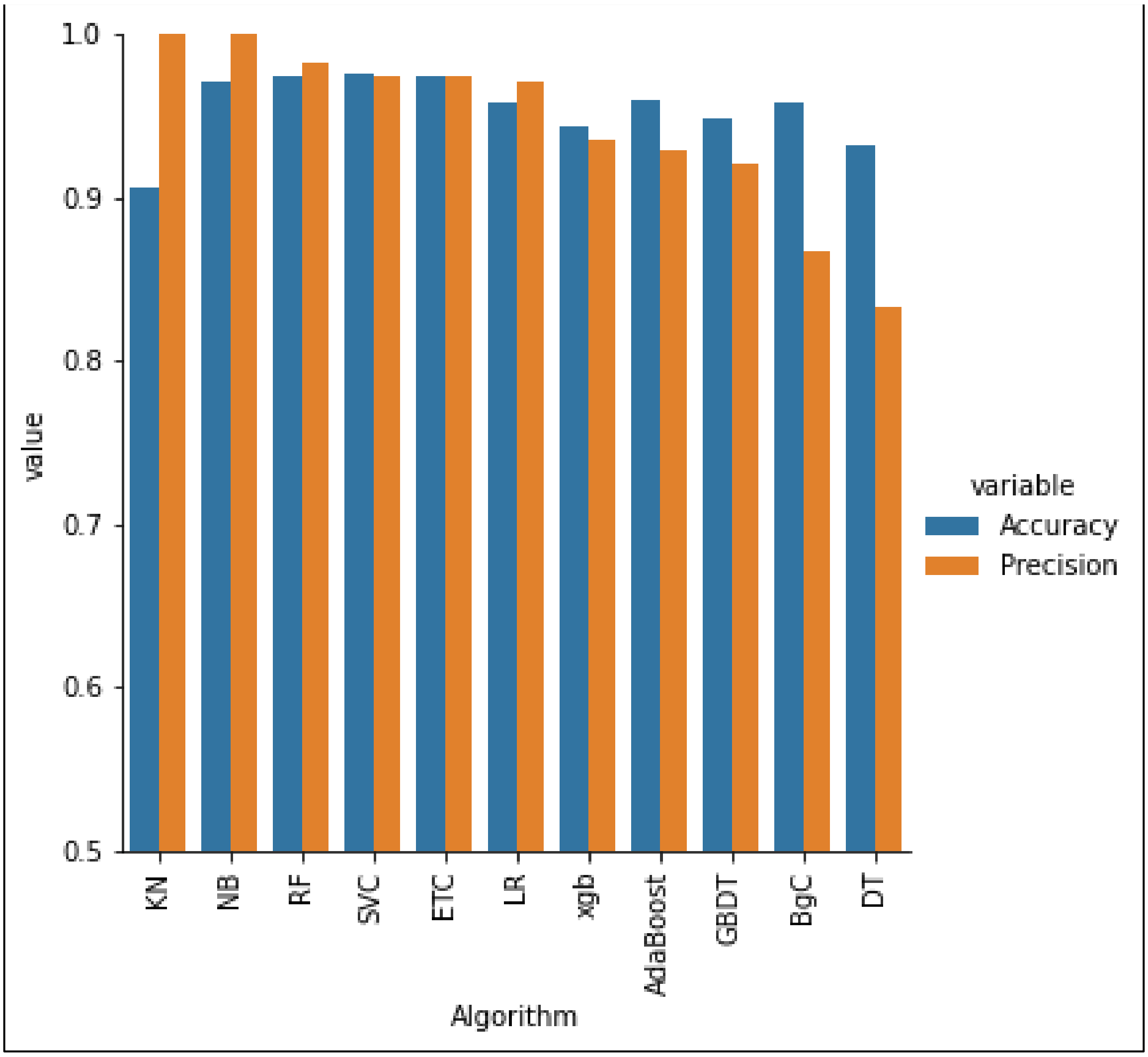

Fig-10:- Algorithm comparison graph

6.CONCLUSION

We proposed a novel algorithm for enhancing the accuracy of the Naive Bayes Spam Filter. The algorithm was implemented as an enhancement for Naive Bayes Classifier and also tested with logistic regression model. Naive Bayes has a very fast processing speed and allows for a small training set, hence is suitable for real-time spam filtering. We are also using Intelligent Text Modification method to identify messages containing leetspeak and diacritic. We are able to classify email as spam or ham. By creating an addition to Naive Bayes Classifier. We also found that our new addition helped improve ham classification due to the high recall and precision rates. We demonstrated that our algorithm consistently reduced the amount of spam emails misclassified as ham email.

7.REFERENCES

1. Email Spam Detection using integrated approach of Naïve Bayes and Particle Swarm Optimization

2. Email Spam Classification by Support Vector Machine

3. Intelligent Model for Classification of SPAM and HAM

4. Team, Radicati. "Email Statistics Report, 2015-2019. The Radicati Group." (2015).

5. Androutsopoulos I., J. Koutsias, K. V. Chandrinos, G. Paliouras,and C. D. Spyropoulos, "An evaluation of naive bayesian antispam filtering", In: 11th European Conference on Machine Learning, pp.9-17, Barcelona, Spain, 2000.

6. GitHub, Inc, "Spam Assassin," 21 April 2016. [Online]. Available: in the given link below is : https://github.com/dmitrynogin/SpamAssassin.git. [Accessed 20 August 2017].