Volume: 04 Issue: 06 | June -2020 ISSN: 2582-3930

Email Spam Detection using NLTK

Narasingh Pratap, Mr Amit Kumar, Vipin Kumar

Department of Information Technology

Rajkiya Engineering College Ambedkar Nagar Uttar Pradesh

Abstract..

Most of the person in world use electronic mail for communication between two party and it is a very useful platform for many type businesses. Large amount of unwanted emailreceive into user's email inbox on adaily basis. Due to large amount of unwanted email user's memory consume more storage .Spam putspressure on the IT infrastructure of any organizations and consume more cost in lost efficiency. So it increases need of spam detection for received email . We use NLTK filtering algorithm using machine learning tool to classify the spam and ham in the given data sets.

Keywords—

Spam email detection, machine learning, feature selection,NLTK, ham, spam.

Introduction.

In our daily life people communicated to each other using email. which contains billions of spam [unwanted or unwelcome email] and in harm the storage of memory , it also consume the high cost .the email which are spam it would be virus, controversial content ,adult content etc.

Although ,Spam mail have low importance than ham messages ,spam message can stop the network traffic due to low bandwidth. And it can harm user's system storage. Spam affect the production of any organization .in the table we have show the email spam condition in the world in daily life.

Ham is known as the useful messages which user want to send and receive.

Ham is the welcome message and it does not harm the system storage of the user. Ham also known as the bussiness message of the user by using ham message user can take decisions according there need and which action will be comfortable to them .

There are various technology which will be used to detect the spam message in email message but in the current scenario most the people uses machine learning technology to detect the spam message from the given email.

Table 1.

Email accounted as spam	45% of all email
Spam email sent by users daily	14.4 billion(approx.)
Spam cost to non corporate sector email users	$255 million
Average loss per employee on yearly basis	$1934
Predicted global spam cost by 2012	$198 billion
Estimated Spam mcrease by 2007	63%(approx.)
Spam received in 1,000 employee company yearly	2.1 million
Spam email received by a person on daily basis	8

There are many techniques which can be used to filter the email spam like WEKA,MATLAB,and Machine learning but in machine learning technology ,it has decision tree, SVM, naïve bayes, NLTK etc .

In this paper we will use NLTK(Natural Language tool kit) using machine learning to filter the spam from the given data set.

NLTK.

NLTK stand for Natural Language Tool Kit. Which is uses in machine learning algorithm to provide the relation between System and human language data.

It is very easy to use between python libraries and connect easily. NLTK is most used technology for Natural Language Processing Technique .NTLK provide all necessary resources to import the data in the Machine Learning algorithm. NLTK first load the data set in the given tool then it convert it into machine understandable format . After converting these messages it will detect spam messages easily from the given email messages .NLTK convert all character of messages into lower case order to find spam easily.

Data Set .

In this data set total email messages are 5558 and 4812 (approx. 80%) messages are ham and 747 (approx. 20%) messages are spam. We use all these data to evaluate the performance of the NLTK technique using machine learning algorithm. To perform the operations we separate these data into training and testing data sets. In machine learning algorithm all the operation will be perform using training and testing data set. After separating this data we can evaluate confusion matrix to know the accuracy of the algorithm performing technique. After knowing the accuracy of the given technique we draw the relationship between ham and spam message received by any user in the world .table 2 shows the features of given data set ie, ham,spam.

Table 2. features of data sets

Data Set	**Spam**	**Ham**	**Total**
Original Data Set	**747**	**4812**	**5559**

Training Data Set	**225**	**900**	**1125**
Testing Data Set	**75**	**300**	**375**

To detect the spam messages from the given data set using NLTK technique first it remove the repeated words and also remove stopwords like comma, punctuation mark, dot etc. after removing all these words from the datasets we draw confusion matrix to evaluate the accuracy of data set .

Table 3 shows the confusion matrix functionality which take false positive ,false negative and true positive ,true negative value of given dataset . to check the Accuracy of given dataset using Natural Language Tool Kit technique in machine learning we take data from the confusion matrix.

Table 3.

Confusion Matrix	**True Positive**	**True Negative**
False Positive	**T P**	**T N**
False Negative	**F P**	**F N**

Confusion _Matrix:

[[0 0 0 ... 0 0 0]
[0 0 0 ... 0 0 0]
[0 0 1 ... 0 0 0]
...
[0 0 0 ... 1 0 0]
[0 0 0 ... 0 0 0]
[0 0 0 ... 0 0 1]]

After using the NLTK using machine learning technique we get the confusion matrix which is given above .

Working of NLTK using Machine Learning.

Figure 1, shows various steps which are used in this technique to evaluate the performance the datasets .first it load the example and analyse the data set to perform the operations using machine learning aalgorithm.

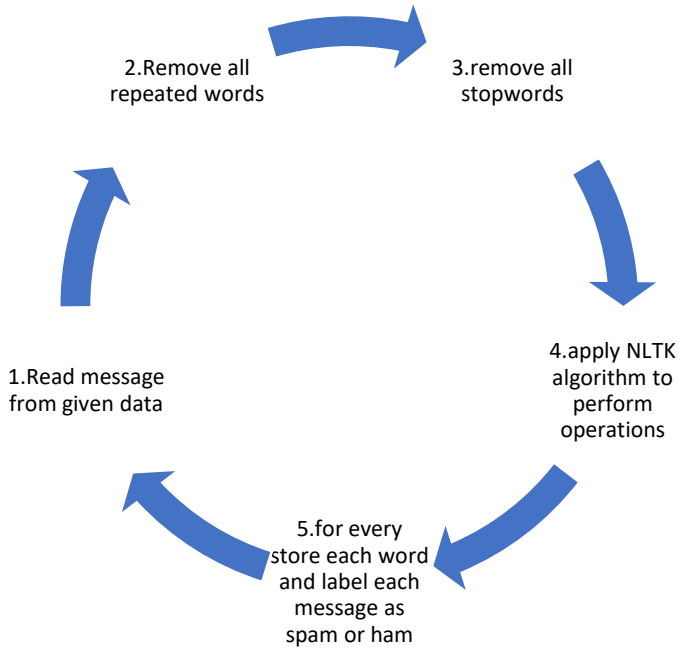

Figure-1 (Representation of steps which are used in NLTK)

Accuracy= TP + TN /(TP + TN + FP + FN)

Accuracy = 0.6453789071284012

After getting Accuracy of the given data set we saw it's evaluation performance of Natural Language Tool Kit using Machine Learning applications. Machine learning Technology is one of the most usable tool to perform the spam detection operation in current world.
We can draw the relationship between ham and spam messages received by a user. Figure-2 shows the relationship between spam message and ham message.

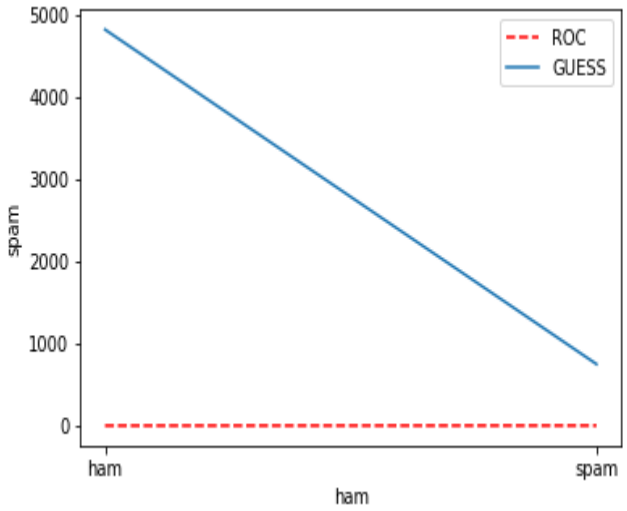

Figure-2(Relationship between ham and spam messages received by user)

Conclusion:

In this paper we saw the performance of NLTK using Machine Learning for the given dataset. Using NLTK we can detect spam messages and ham messages which are presents in the given data set. And also we can say that number ham email more than number spam messages which are received by a user in daily life.

ACKNOWLEDGMENT.

We want to thank all the references and for useful content and information which improve the quality of this paper.The author is appreciative to our Supervisor for availing essentialservices which helped for successful conclusion of work.

References:

[1]"Email Spam filtering using BPNN classification algorithm"**https://ieeexplore.ieee.org/document/7877720/metrics#metrics**[Online last accessed 26-November -2019]

[2]"A proposed data science approach for email spam classification using machine learning techniques" https://ieeexplore.ieee.org/document/8260935/metrics#metrics[last accessed 26-November-2019]

[3] '' download data set to perform operations''https://www.kaggle.com/uciml/sms-spam-collection-dataset[last accessed 20-November-2019]

[4]"to know the total number spam received by user"https://www.statista.com/statistics/420391/spam . -email-traffic-share/[last accessed 1-june-2020]

[5]"to know About the NLTK "https://www.nltk.org/[last accessed 1-june-2020]