IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

EMAIL Spam Classification-LinearSVC approach using Machine Learning based technique

K.Ushma[1], S. Monisha Reddy[2], M.Rani[3], Sk.Adil Basha[4]

[1]UG Student, [2]Assistant Professor, [1,2]Department of Computer Science and Engeneering

[1,2]Kallam Haranadareddy Institute of Technology(UGC-Autonomous),Chowdavaram,

Andhra Pradesh, India

ABSTRACT

Email spam has become a major problem nowadays, with rapid growth of internet users, email spams also increasing. people are using them for illegal, phishing and fraud. Sending malicious link through spam emails which can harm system and can also seek into our system. Spam emails are irrelevant and unwanted messages typically sent to breach security. Spam fills inbox with number of ridiculous emails.it also degrades the internet speed to a great extent and steals useful information like our details on contact list. email spam is an operation to send messages in bulk by mail. These unsolicited mail emails contain phishing URLs, advertisements, industrial segments, and a huge variety of indiscriminate recipients. Thus, such content is usually a danger for the user, and many researches have taken area to come across such unsolicited mail content.

So, it is needed to identify those spams by using techniques of machine learning techniques and apply these on our datasets which can give better accurate results on classifying emails. The proposed study utilizes the knowledge of supervised machine learning algorithm such as LinearSVC to discover and categorize the email content as spam or ham. With the help of proposed system, the specified message can be stated as spam or ham.The proposed study utilizes the prevailing gadget gaining knowledge of algorithm LinearSVC to discover and categorize the email content. The LinearSVC representation of better than other models with a greatest rating of 98.74% precision.

Keywords: email spam classification, spam filtering, LinearSVC, Machine learning

1. INTRODUCTION

Now-a-days, large volume of undesired huge emails called as spam has turn out to be a big trouble facing by the on the online [2]. Character passing the junk mail information is termed to be hacker. Those characters assemble electronic mail inscription along with special webpage, meetings [3]. Unsolicited avoids consumer along with doing thorough and right way of using, storing capability and section transmission capacity. Big extended junk e-mails passing via pc websites having unfavourable outcomes at the reminiscence area of e-mail servers, verbal exchange bandwidth, CPU strength and user time [4]. The growth of these junk e-mails effects the annual base along with accountable above 77% for appeal to identify it more traumatic. [5] Likewise derived to limitless monetary dropping to more customers who were the sufferer of the network hackers also different dishonest applications of hackers who ships e-mails affecting from legitimate groups having the aim to influence separate to communicate touchy non-public records like passkeys, Banking details and credit card digits.

[6] Increasing use of email has created problems caused by unsolicited bulk email messages commonly referred as spam. spam emails are the emails that the receiver does not wish to receive [7]. there are many effects of spam: fills our inbox with no of ridiculous emails, degrades our internet

speed at great extent, steals useful information, alters your search result in system programs, a huge waste of time, identifying these spam contents is labourious task. these may contain links to phishing or malware hosting websites used to steal confidential information [8]. Email has become one of the most important forms of communication. In 2014, there are estimated to be 4.1 billion email accounts worldwide, and about 196 billion emails are sent each day worldwide [9]. Links in spam emails may lead to users to websites with malware or phishing schemes, which can access and disrupt the receiver's computer system. These sites can also gather sensitive information.

Therefore, an effective spam filtering technology is a significant contribution to the sustainability of the cyberspace and to our society.

[10] Considering the significance and high usage of emails, the number of spam emails has also increased rapidly. Spam emails are unwanted emails that contains various contents like advertisements, offers, malicious links, malware, trojan, etc. Spammers send junk mails with an intention of committing email fraud, thus it is important to filter spam emails from emails. The motivation of our project is to use spam detection by using machine learning techniques like Linear SVM.

[11] To solve this problem the different spam filtering techniques are used. These ML techniques have the capacity to learn and identify spam mails and phishing messages by analysing loads of such messages throughout a vast collection of computers.

[12]A spam filter is a program used to detect unsolicited, unwanted, and virus-infected emails and prevent those messages from getting to a user's inbox. Like other types of filtering programs, a spam filter looks for a specific criterion on which to base its judgment [13], though there are several email spam filtering methods in existence, We explained below one of the categories of spam filtering techniques that have been widely applied to overcome the problem of email spam.

Content Based Filtering Technique: Content based filtering is usually used to create automatic filtering rules and to classify emails using machine learning, such as Naïve Bayesian classification, Support Vector Machine, K Nearest Neighbor.It analyses words occurrence and distributions of words and phrases in the content of emails and then use generated rules to filter the incoming mails

2. LITERATURE SURVEY

[1] "A Machine Learning based Spam Detection Mechanism" written by Ashi Bansal, Nikhil Govil, Astha Varshney, Kunal Agarwal.

This arises the need for presenting prudent mechanism to hit upon or discover such junk e-mails so that time and reminiscence space of the gadget can be saved as much as a wonderful quantity. In this paper, we provided the same mechanism that could filter junk mail and non-junk e-mails. This proposed algorithm generates dictionary and features and trains them via system mastering for effective results.

[2] "Detection of Spam E-mails with Machine Learning Methods" written by İbrahim Alper Dogaru , Hamdallah Karamollaoglu, Murat Dorterler .

In this look at, it's far goal to examine the factors material message of mails write down in Turkish by the service of Naive Bayes Classifier and a model which is Space Vector from device getting to know methods, to discover if or not those mails are unwanted mail and dividing them. Both models are dominated to one-of-a-kind evaluation standards and their performances are as compared.

[3] "Email Spam Detection Using Machine Learning Algorithms" written by Nikhil Kumar, Nishant, Sanket Sonowal.

Discovering the unreal identity and mail accounts are lot clean for hackers, they act to be as a real man or woman of their unwanted emails, the hackers aims the peoples who has no knowledge about

these frauds. Hence, it is required to find the junk mails which can be fake, the task can discover the spam emails by utilize different strategies of the gadget Study, In this paper
,We'll talk over the system mastering methods, follow a majority of these algorithm on our facts sets and fine set of rules is chosen for the e-mail junk mail detection having exceptional precision and accuracy.

[4] "Optimizing semantic LSTM for spam detection" written by Manisha Sharma, Gauri Jain, Basant Agarwal.

In this, we have applied an existing location of method referred to as deep gaining knowledge of method. A unique structure referred to as Long Short-Term Memory (LSTM), a version of the Recursive Neural Network (RNN) termed to be as spam division. Before the usage of the LSTM for class undertaking, the textual content transferred to semantic phrase vectors with the service of word2vec, WordNet and Concept Net. Categories effects are differentiated with the standard classifiers like SVM, Random Forest, Naïve Bayes, okay-NN.

[5] "Spammer Detection and Fake User Identification on Social Networks" written by Ghana Ammad, Assad Abbas, Ikram Ud Din, , Mansour Zuair, Faiza Masood, Ahmad Almogren, Hasan Ali Khattak, Mohsen Guizani.

In this, we state and evaluate of strategies utilised for identifying hackers on Twitter [6]. Not only this, a harmful of the Twitter unwanted e-mail identification procedures are supplied those divides the policies primarily evolved on to locate: (i) fake information, (ii) junk e-mail primarily evolved on Uniform Resource Locator, (iii) fake customers. The offered strategies are also as compared based on various capabilities, which include consumer characteristics, information capabilities, graphs, design capabilities, and capabilities.

[7] "Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers" written by Shubhangi Suryawanshi, Pramod Patil,, Anurag Goswami.

The mail spam datasets from the UCI ML repository and the Kaggle websites is used to evaluate and test classifiers. Accuracy scores, F measures, and ROC are among the accuracy measures employed. This Classifier with a voting technique is better to utilize, according on the preliminary results. It has a less false +ve rate and a highest level of precision.

[8] "Machine learning for email spam filtering: review, approaches and open research problems" written by Emmanuel Gbenga Dada,, Haruna Chiroma, Shafi'i Muhammad Abdulhamid, , Opeyemi Emmanuel Ajibuwan, Joseph Stephen Basi, Shafi'i Muhammad Abdulhamid, Adebayo Olusola Adetunmbi. This paper is published on June 2019. Starting discussion in the learning background shows that how ML models are applied to the e-mail spam filtering processes of Major Internet Service providers like Yahoo. The complete e-mail junk filtering technique is considered, along with some innovations by various examiners to take action on junk utilizing ML ways.

[9]"Email Spam Filtering using Supervised Machine Learning Technique" written by VS.Karpagavalli ,V. .Christina , G.Suganya . Even though we are being inundated with spam emails on a daily basis. This is not due to filter aren't strength required; It's because of spammers' rapid learning of modern methods and junk filter inability to receive modifications. Here, the usage of ML which is supervised approaches to strain e-mail junk messages in our research. For learning the features of spam emails, largely usage of ML methods which are supervised such as C 4.5 Decision tree classifier Multilayer Perceptron, and Nave Bayes Classifier, Multilayer Perceptron are employed, and this technique is formed by learning about all junk and non-junk mails.

[10]"Contribution to the study of SMS spam filtering: new collection and results" written by Jose Maria G. Hidalgo Tiago A. Almeida, Akebo Yamakami. Messages are often short, content-based junk to strain may be affect a working hit. Here, presented the largest non-artificial-time, not-private,

which is not encoded message junk assemble, we are aware of. Furthermore, we compare the results of many well- known ML algorithms.

[11]“Online Active Learning Methods for Fast Label- Efficient Spam Filtering” written by D. Sculley. We investigate an online active learning scenario in which the filter is presented with a stream of messages that must be classified one by one. A label for a message can only be requested by the filter after it has been classified[12]. With limited label requests, the goal is to obtain good online classification performance. We use linear classifiers to evaluate multiple techniques to selective sampling in this context, drawing on the label efficient machine learning literature. We show that online active learning can drastically cut labeling and training expenses while retaining excellent levels of classification performance with minimal additional overhead

3. PROPOSED SYSTEM

[1]In this project, we have worked on identifying spam emails from bulk of emails[2]. To detect the spam emails, we have used LinearSVC for binary classification on the spam-ham dataset. We have used the architecture shown in the figure.

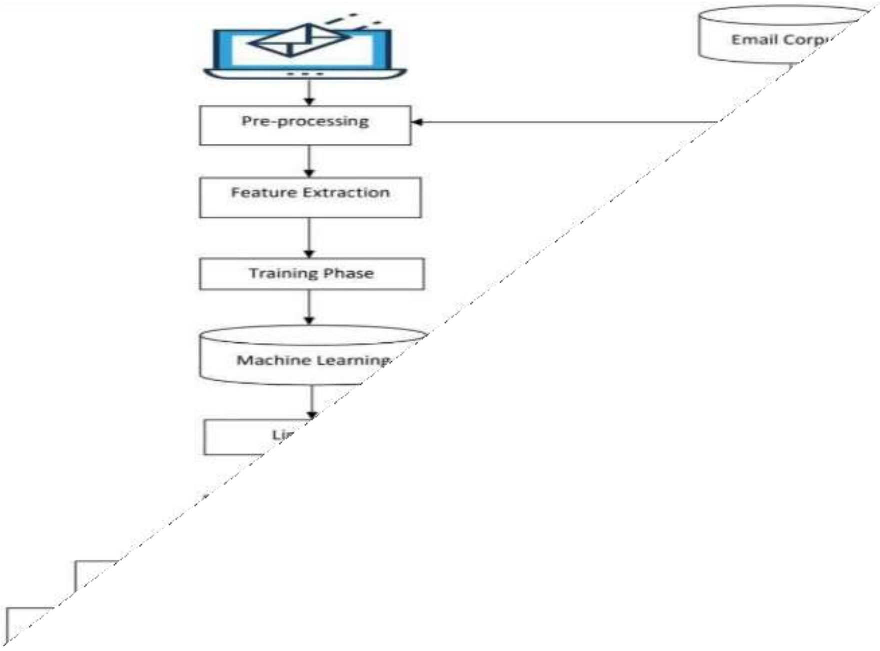
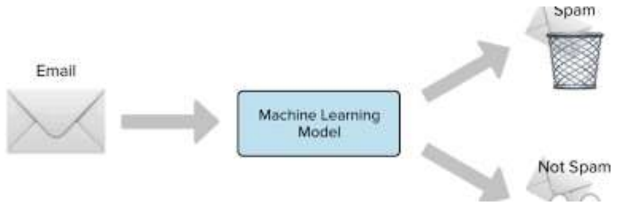


Fig1:Proposed system Architecture

[3] First, we loaded the dataset to pandas dataframe and then replaced the null values with the null string. The spam emails are labeled as '0' and the ham emails are labeled as '1' and the text data is separated.

[4] The data is splitted into the training and testing data and the text data is transformed into feature vectors using *TfidfVectorizer* so that it can be provided as input to the SVC model[5]. The Support Vector Machine SVMmodel is trained with training data and then prediction on tested data is done.



3.1. Advantages of using LinearSVC

1. There is no need for extra manual cost since all resources are available

2. It gives result with greater accuracy and with good performance.

3. This proposed system will optimize the data storage by blocking and deleting spam mails.
4. This proposed system will save the user's time and it destroys the risk of spam mails.
5. It is more convenient and easy to classify the mails from spam mails.
6. It makes better predictions than other existing systems

3.2. Modules

1. INPUT: This datasets contains samples of spam and ham which are used for providing input to training data and testing data.We have splitted the data into 2 steps in ratio 80:20(80% of data for training and 20% for testing).

2. PREPROCESSING: In this module, it is used to clean the unwanted data and make it free from errors , redundancy and gives desired content.so we need to preprocess the data.We have preprocesses the data by loading the dataset to the pandas dataframe and replacing null values as null strings.

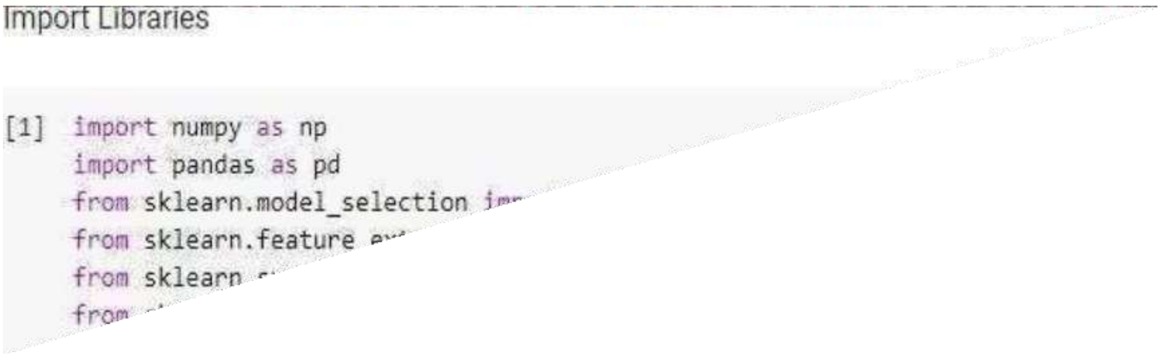
3. FEATURE EXTRACTION: It transforms the text data to feature vectors so that it can be used as input to SVM model using TfidfVectorizer.Then converts text into lowercase letters and label as '0'and '1',where 0 as spam mails and 1 as ham mails.

4. ENSEMBLE ALGORITHM: SVM is considered to be classification approach but it is used in both regression(continuous)and classification problems.It can easily handle multiple continuous and categorical variables.it constructs hyperplane in multidimensional space to separate different classes.It is used to minimize the error

3.3. Algorithm

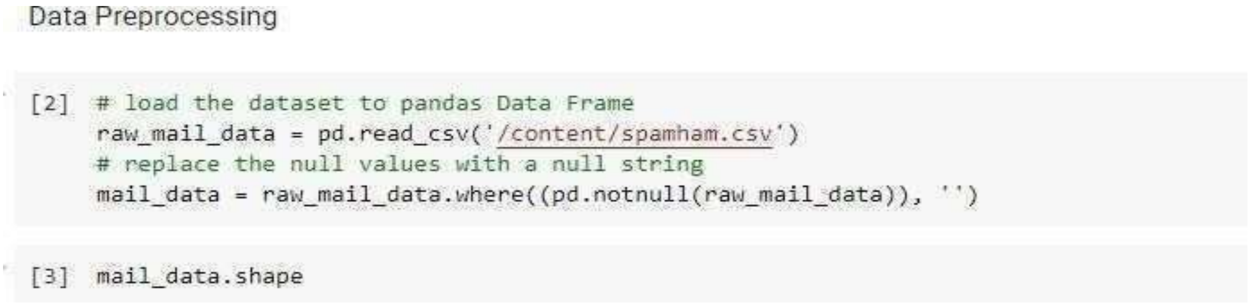
Step 1: Accumulate email dataset
Step 2: Preprocessing the data obtained
Step 3: Replacing null values to null strings
Step 4: Labeling the spam mails as '0' and ham mails as '1'
Step 5: Now split the data into two partitions i.e training data and testing data
X_train_Val, X_test_Val, Y_train_Val, Y_test_Val = train_test_split(X, Y, train_size=0.9, test_size=0.1, random_state=3)
Step 6: Transform the text data to feature vectors that can be used as input to the SVM model using TfidfVectorizer
feature_extraction=TfidfVectorizer(min_df=1, stop_words='english' , lowercase='True')
Step 7: Training the support vector machine model with trained data
model1 = LinearSVC() model1.fit(X_trainVal_features, Y_train_Val)
Step 8: Accuracy on test data
prediction_on_testVal_data = model1.predict(X_testVal_features) accuracyOn_test_data = accuracy_score(Y_test_Val, prediction_on_testVal_data)
After the model gets trained with the training data, we have evaluated model taking accuracy score as the metric.

4. RESULT SCREENS AND DISCUSSION



Import Libraries

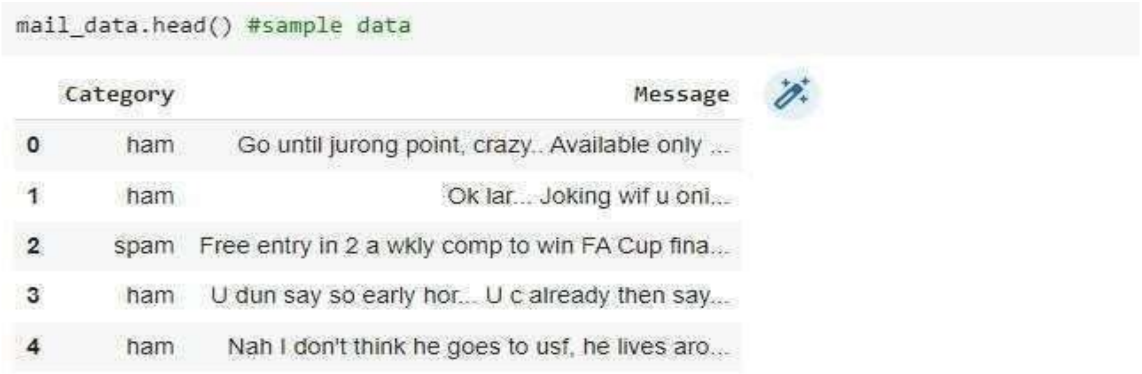
```
[1]  import numpy as np
     import pandas as pd
     from sklearn.model_selection imp
     from sklearn.feature_ex
     from sklearn.
     from
```



Data Preprocessing

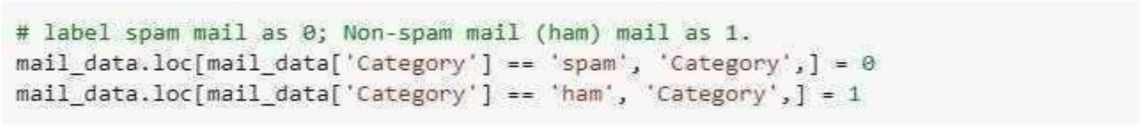
```
[2]  # load the dataset to pandas Data Frame
     raw_mail_data = pd.read_csv('/content/spamham.csv')
     # replace the null values with a null string
     mail_data = raw_mail_data.where((pd.notnull(raw_mail_data)), '')
```

```
[3]  mail_data.shape
```

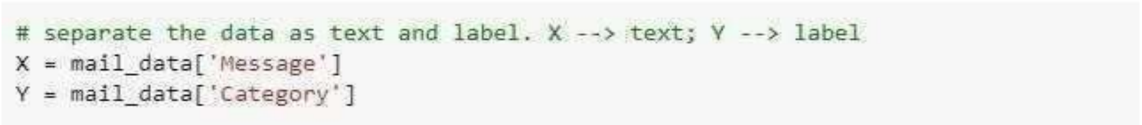


```
mail_data.head() #sample data
```

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

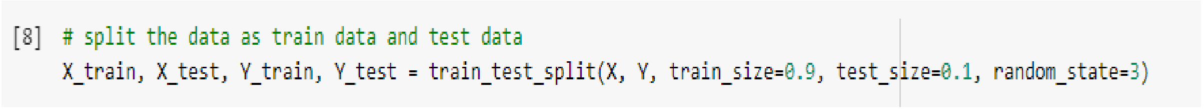


```
# label spam mail as 0; Non-spam mail (ham) mail as 1.
mail_data.loc[mail_data['Category'] == 'spam', 'Category',] = 0
mail_data.loc[mail_data['Category'] == 'ham', 'Category',] = 1
```



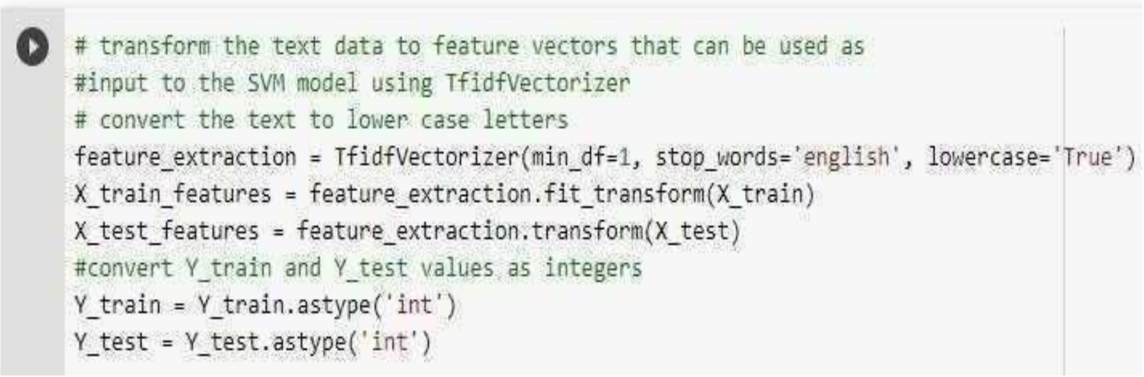
```
# separate the data as text and label. X --> text; Y --> label
X = mail_data['Message']
Y = mail_data['Category']
```

Train Test Split



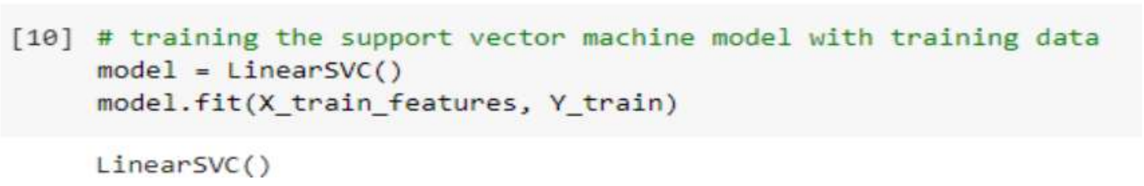
```
[8]  # split the data as train data and test data
     X_train, X_test, Y_train, Y_test = train_test_split(X, Y, train_size=0.9, test_size=0.1, random_state=3)
```

Feature Extraction



```
# transform the text data to feature vectors that can be used as
#input to the SVM model using TfidfVectorizer
# convert the text to lower case letters
feature_extraction = TfidfVectorizer(min_df=1, stop_words='english', lowercase='True')
X_train_features = feature_extraction.fit_transform(X_train)
X_test_features = feature_extraction.transform(X_test)
#convert Y_train and Y_test values as integers
Y_train = Y_train.astype('int')
Y_test = Y_test.astype('int')
```

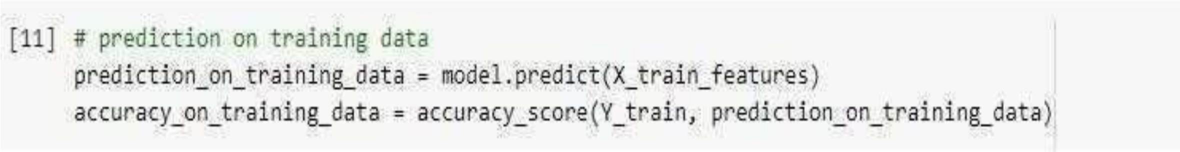
Training the model --> **Support Vector Machine**



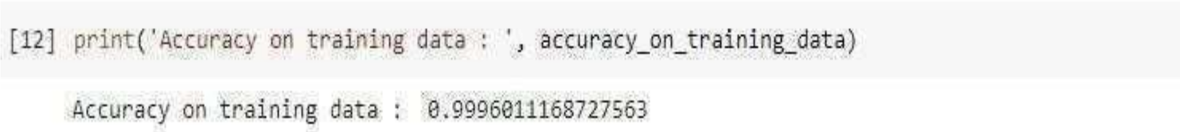
```
[10] # training the support vector machine model with training data
     model = LinearSVC()
     model.fit(X_train_features, Y_train)

     LinearSVC()
```

Evaluation of the model

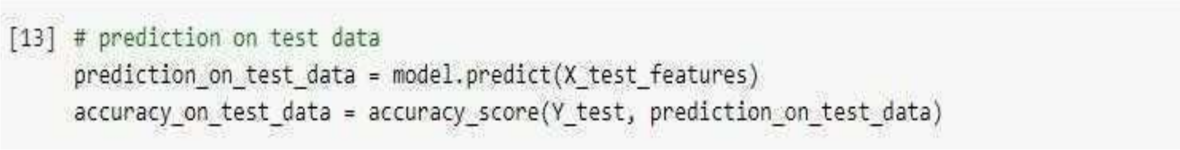


```
[11] # prediction on training data
     prediction_on_training_data = model.predict(X_train_features)
     accuracy_on_training_data = accuracy_score(Y_train, prediction_on_training_data)
```

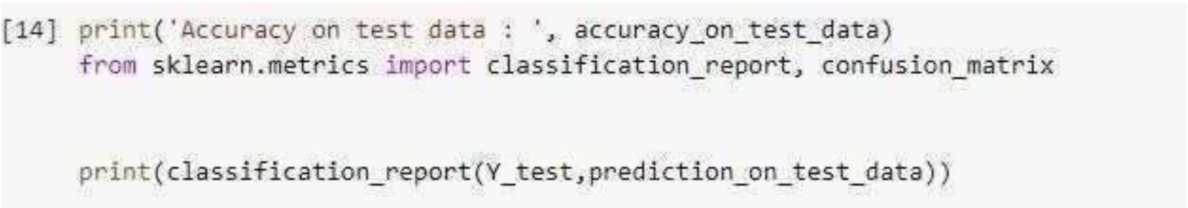


```
[12] print('Accuracy on training data : ', accuracy_on_training_data)

     Accuracy on training data :  0.9996011168727563
```

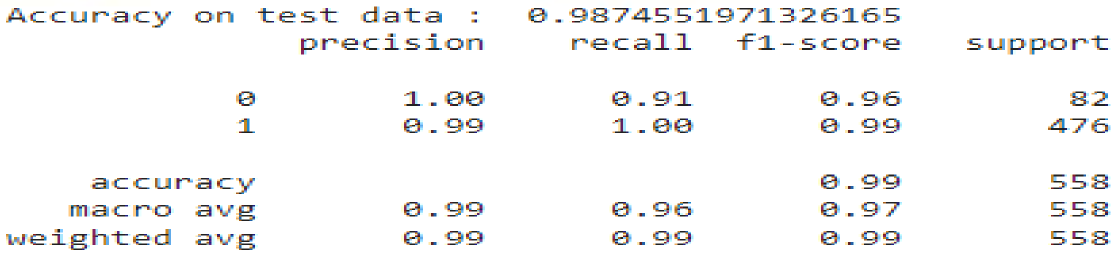


```
[13] # prediction on test data
     prediction_on_test_data = model.predict(X_test_features)
     accuracy_on_test_data = accuracy_score(Y_test, prediction_on_test_data)
```



```
[14] print('Accuracy on test data : ', accuracy_on_test_data)
     from sklearn.metrics import classification_report, confusion_matrix


     print(classification_report(Y_test,prediction_on_test_data))
```



```
Accuracy on test data :  0.9874551971326165
              precision    recall  f1-score   support

           0       1.00      0.91      0.96        82
           1       0.99      1.00      0.99       476

    accuracy                           0.99       558
   macro avg       0.99      0.96      0.97       558
weighted avg       0.99      0.99      0.99       558
```

4.1. Performance Evaluation Parameters:

The following evaluation parameters used

Precision:

Precision is the number of correct results divided by the number of all returned results. Precision = $tptp+fn$

Recall:

Recall is the number of correct results divided by the number of results that should has been returned. Recall = $tptp+fn$

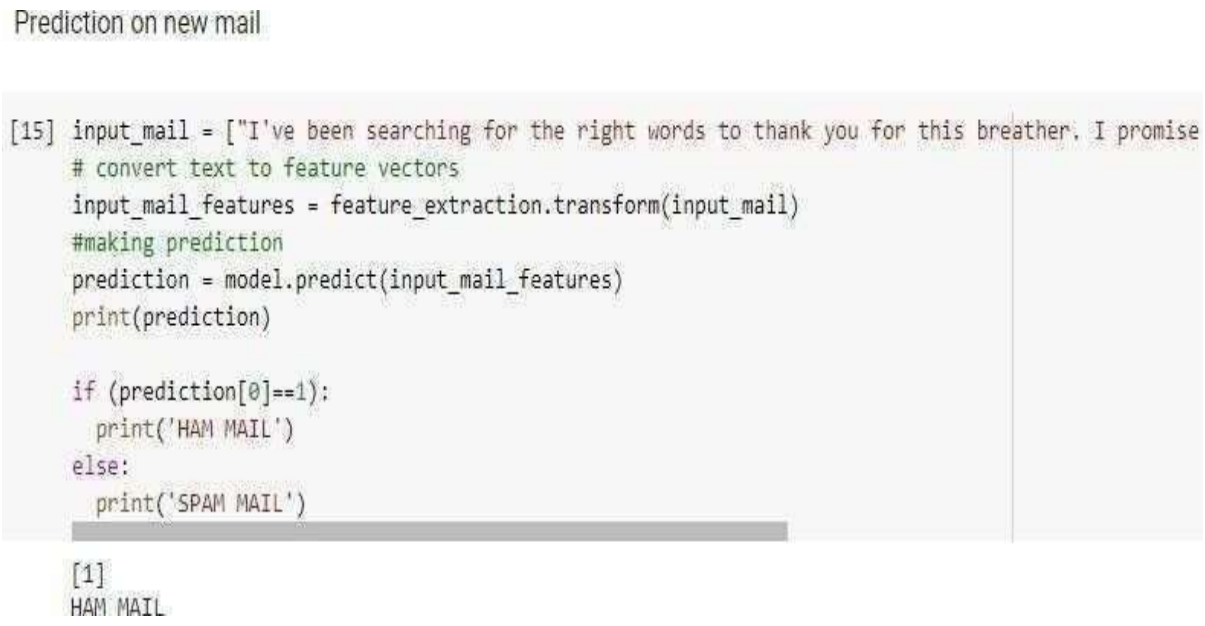
F1-Score:

A measure that combines precision and recall is the harmonic mean of precision and recall. F = 2 * $precision*recallprecision+recall$

Accuracy:

Accuracy = $tp+tntp+tn+fp+fn$

4.2. PREDICTION



Prediction on new mail

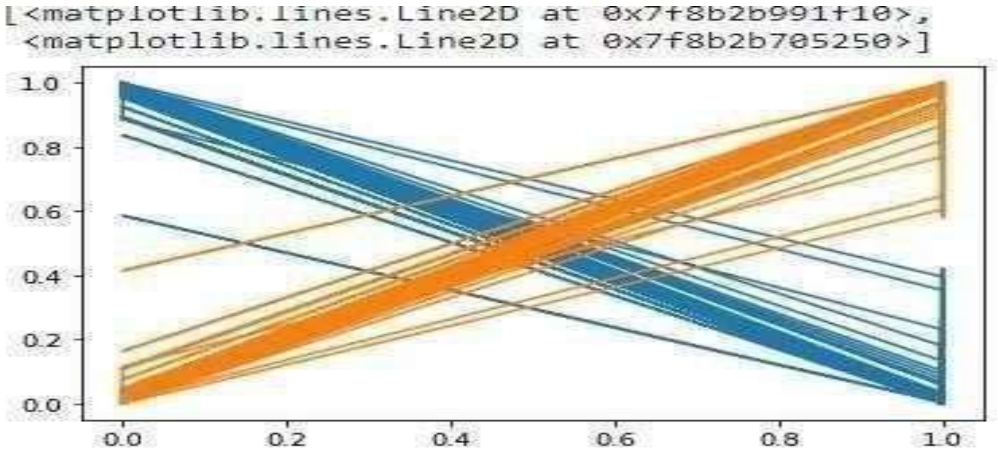
```
[15] input_mail = ["I've been searching for the right words to thank you for this breather. I promise
     # convert text to feature vectors
     input_mail_features = feature_extraction.transform(input_mail)
     #making prediction
     prediction = model.predict(input_mail_features)
     print(prediction)

     if (prediction[0]==1):
       print('HAM MAIL')
     else:
       print('SPAM MAIL')

     [1]
     HAM MAIL
```

The figure shows the accuracy that is obtained for the spam mail detection using LinearSVC. It also shows values for different performance metrics like precision, recall, f1-score and support.

GRAPH:The graph is generated with the predictions and probablities that have obtained from the X_test_features.



```
[<matplotlib.lines.Line2D at 0x7f8b2b991f10>,
 <matplotlib.lines.Line2D at 0x7f8b2b705250>]
```

6. CONCLUSION

Due to the increase in usage of emails, this work focuses on using self-regulating ways to detect spam emails. The proposed model by using LinearSVC algorithm exhibits accuracy rate of 98.74%. The accuracy rate of LinearSVC is higher than that of LSTM (98%) and also LSTM takes a little longer to train the model. The proposed model's accuracy and performance are vey much far better than that of the LSTM and also other existed algorithms which makes better prediction on categorizing the emails as spam or ham.

7. REFERENCES

1. Machine learning for email spam filtering by Stephen Joseph
https://www.academia.edu (2019)

2. detection of spam emails with machine learning methods by murat dorterla at gazi university
https://www.researchgate.net (oct 2018)

3. optimizing semantic lstm for spam detection by Gauri Jian,Basant Agarwal
https://www.semanticscholar.org (June 2019)

4. email spam filtering using supervised learning machine learning technique by V.Christina
https://wwwresearchgate.net (2010)

5. email spam fetection using ML algorithms by Nikhil kumar and Nishanth
https://www.researchgate.net (July 2020)

6. email spam detection:an empirical comparative study on ML algorithms by Anirag and Pramod Patil
https://www.researchgate.net (Dec 2019)

7. Performance analysis for different ML techniques on email filtering system by M.S.Osho , shafi, ismaila
https://www.irjet.net (2018)

8. Filtering of emails using naïve bayes and hash function by V.Sharma,S.Ajaz
https://www.irjet.net (2017)