SPAM MESSAGE CLASSIFICATION

[1] Mrs. Saranya N, [2] Ms. Tharanisri A, [3] Ms. Ranjitha B
[1]Assistant Professor, [2]UG scholar, [3]UG scholar
[1,2,3]Department of Computer Science and Engineering
[1,2,3]Sri Shakthi Institute of Engineering and Technology, Coimbatore, India

Abstract : In this world of technology spam messages and emails are a huge problem. Spam emails are the unsolicited messages sent bulk by email. They are mainly commercial but may contain malwares too. Already, we have spam classification in Gmail application but still some spam messages are moving to inbox. Our objective is to classifying spam e-mails from inboxes. Two classifiers are applied on one benchmark dataset to evaluate which classifier gives better result. The comparison has been done among distinctive machine learning classifiers (such as Bayesian, SVM (Support vector machine) the specified classifiers are tried and assessed on measurements (such as exactness, accuracy etc). Results of the classification algorithms are compared with the spam dataset. The experimental results approve that the spam mails can be classified correctly, with accuracy reaching up to 98.5% using Naive and SVM algorithm, which produces low false positive rate.

***IndexTerms* - Accuracy, Ham, Naive Baye's, Prediction, Spam filtering, Spam, SVM.**

I. INTRODUCTION

Today, emails are used for communication purpose by several users. Emails are broadly classified as spam mails and non-spam mails. First we will try to explain what is spam mails and non-spam mails and how affects on email user. Spam is defined as bad emails and unwanted emails sent with the aim of spreading viruses, for fraud in business and inflicting damage to email users. Non-spam mails are nothing but our regular emails that is beneficial for email users. In spam classification, various spam classification methods are used. The function of a spam message classification is to identify spam email and avoid it from going to the mailbox. With the help of filters, the impact of spam email is prevented. It operates like a predictable and reliable tool to eliminate unwanted emails. Though there exists a risk of misclassification. Now it's becoming difficult for a user to distinguish between emails by reading the subject or the email content, thereby it increases the necessity of spam filter. Rarely filters also makes mistake but it minimizes the error. Therefore, an effective spam filtering technology is significant to our society.

II. RELATED WORKS

Ajay Sharma, Anil Suryawanshi [1], focuses on to make a RBF NN technique and then compared it with SVM based on two parameters i.e. precision and accuracy. It provides high precision and accuracy using this efficient spam filtering technique. Spearman's connection coefficient is a factual measure of the quality of monotonic relationship between matched information. Then KNN algorithm is used with Spearman Correlation. Spearman coefficient of correlation is employed as distance live in KNN classification technique.

Deepak Agarwal [2] focuses on a popular machine learning algorithm SVM with different parameters using different kernel-functions. It is evaluated to get best accuracy. Different kernel functions are implemented for spam filtering. The author has used four types of kernels: linear kernel, polynomial kernel, RBF and sigmoid kernel. After this the accuracy is calculable for all the kernels in any respect the combinations of train files and test files. Various Machine learning strategies area unit being employed to classify spammer's emails from legitimate emails.

Harshal D.et al. [3] article proposes a spam discovering system to detect text as well as image based mostly spam victimization ANN formula. In this system, pre-processing of email text before execution the algorithms is employed to create them predict higher. Using this system High level, low level, and combination of each the features of image in a very spam mail may be expected.

In [5] Social Filter system that enables nodes with no email classification functionality to query the network on whether a host is a spammer.

In [6] author, proposes a technique to classify text and images.

Reena S. et al.,[7] article focuses on to create a RBF NN technique then compared it with SVM supported 2 parameters i.e. exactitude and accuracy. This can be Associate in Nursing economical spam filtering technique which provides high exactitude and accuracy. Here the author has planned the RBF technique. It is the neural network technique that uses hidden neurons to method the input and to present the output. In this technique the RBF has collected the spam words and shaped the spam word lexicon. These words square measure used for coaching and testing. The Liebenberg algorithmic program is employed during this technique. Results obtained from the RBF square measure compared with the SVM.

In [8] Ren Wang (IEEE CCECE/CCGEI, Ottawa, May 2006)" On Some Feature Selection Strategies for Spam Filter Design "concluded that use of optimization techniques as feature selection strategies reduce the dimension of email likewise as improve the performance of the classification filter.

In [9] the author proposed a system which automatically classify spam messages and non-spam messages.

In [10] author, uses a method which drastically reduces the consumption of internet bandwidth by spam.

In [11] author, uses a technique based on machine learning and content feature.

In [12] author incorporating keyword-based filtering to document classification for email spamming" contributed to “the analysis of email filtering to melt the hard clustering decision and also achieved the result of value analysis of ham to spam is healthier than spam to ham.

SOAP exploits the social relationship among email correspondents to detect the spam adaptively and automatically [14].

III. PROPOSED WORK

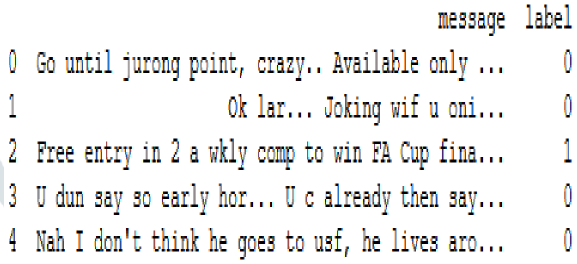
Our proposed methodology consists of four main stages. They are

3.1 Import libraries

The first stage is importing libraries, in which we import all the necessary libraries like tensor flow, scikit learn, pandas, numpy, matplotlib etc ., Here we use NTLK for processing the messages, Word Cloud & matplotlib for visualization and pandas for loading data, Numpy for generating random probabilities for train-test split.

3.2 Loading datasets

We don't need the columns ‘Unnamed: 2’, ‘Unnamed: 3’ and ‘Unnamed: 4’, so we remove them. We change the column ‘v1’ as ‘label’ and ‘v2’ as ‘message’. ‘Ham’ is replaced by zero and ‘spam’ is replaced by one within the ‘label’ column. Finally we obtain the following,

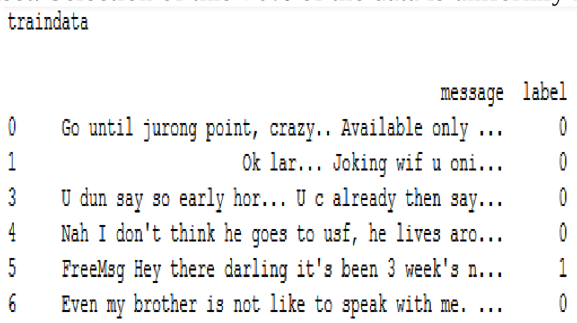


```
                                             message  label
0  Go until jurong point, crazy.. Available only ...      0
1                      Ok lar... Joking wif u oni...      0
2  Free entry in 2 a wkly comp to win FA Cup fina...      1
3  U dun say so early hor... U c already then say...    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

Fig.1Dropping the unnamed columns

3.3 Train – Test split

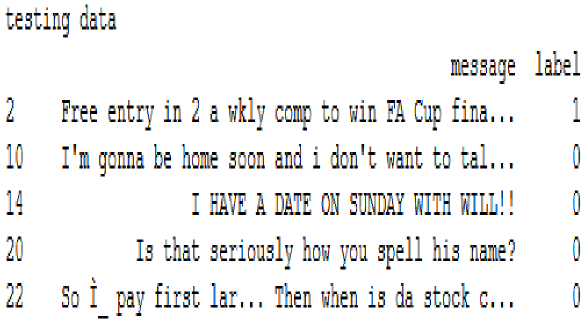
To check our model we should always split the information into train dataset and test dataset. We shall use the train dataset t0 train the model so it'll be checked on the test dataset. We shall use 70% of the dataset as train dataset and the rest as test dataset. Selection of this 70% of the data is uniformly random.



```
traindata

                                             message  label
0  Go until jurong point, crazy.. Available only ...      0
1                      Ok lar... Joking wif u oni...      0
3  U dun say so early hor... U c already then say...      0
4  Nah I don't think he goes to usf, he lives aro...      0
5  FreeMsg Hey there darling it's been 3 week's n...      1
6  Even my brother is not like to speak with me. ...      0
```

Fig.2 Training Dataset



```
testing data
                                              message  label
2     Free entry in 2 a wkly comp to win FA Cup fina...      1
10    I'm gonna be home soon and i don't want to tal...      0
14                     I HAVE A DATE ON SUNDAY WITH WILL!!      0
20                Is that seriously how you spell his name?      0
22    So Ì_ pay first lar... Then when is da stock c...      0
```

Fig.3 Testing Dataset

3.3.1 Visualizing data

Now, we use word cloud library for identifying the most repeated words in the spam and ham messages.

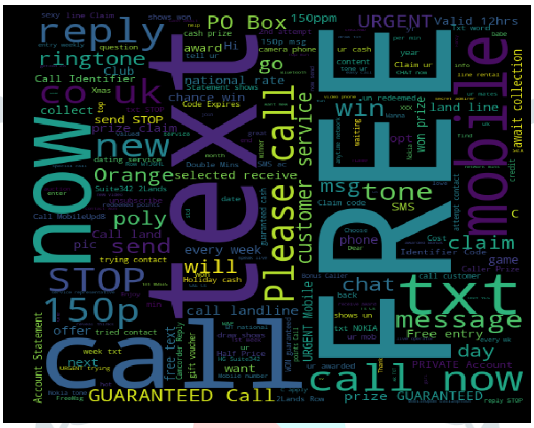


Fig.4 Word Cloud

3.3.2 Train the model

Here, we implement two techniques: Bag of words and TF-IDF.

3.3.2.1 Bag of Words

In Bag of words show we discover the 'term recurrence', for example number of events of each word in the dataset.

$$P(w)=\frac{\text{Total no of occurrences of w in dataset}}{\text{Total no of words in dataset}}$$

3.3.2.2 TF-IDF

TF-IDF stands for Term Frequency-Inverse Document Frequency.

$$IDF(w)=\log\frac{\text{Total no of messages}}{\text{Total no of messages containing w}}$$

For instance, there are two messages in the dataset. 'hi world' and 'hi foo bar'. TF ('hello') is 2. IDF ('hello') is log(2/2). In the event that a word happens a great deal, it implies that the word gives less data.

3.3.3 Preprocessing

Before beginning with preparing we should preprocess the messages. Above all else, we will make all the character lowercase. This is on the grounds that 'congrats' and 'CONGRATS' mean the equivalent and we would prefer not to regard them as two distinctive words. Then we tokenize each message in the dataset. Tokenization is the undertaking of part up a message into pieces and discarding the accentuation characters. The words like 'go', 'goes', 'going' demonstrate a similar movement. We supplant every one of these words by a solitary word 'go'. This is called stemming. We are utilizing Porter Stemmer, which is a well known stemming algorithm. We then proceed onward to evacuate the stop words. Stop words are those words which happen incredibly as often as possible in any content. For instance words like 'the', 'an', 'an', 'is', 'to' and so on. These words don't give us any data about the substance of the content. In this way it ought not make any difference on the off chance that we evacuate these words for the content.

Discretionary: You can likewise utilize n-grams to improve the exactness. Starting at now, we just managed single word. Be that as it may, when two words are as one the importance absolutely changes. For instance, 'great' and 'not great' are inverse in importance. Assume a content contains 'not great', it is smarter to consider 'not great' as one token as opposed to 'not' and 'great'. In this manner, here and there exactness is improved when we split the content into tokens of (at least two) words than just word.

3.3.4 Analysis of text

We need to discover the frequencies of words in the spam and non-spam messages.

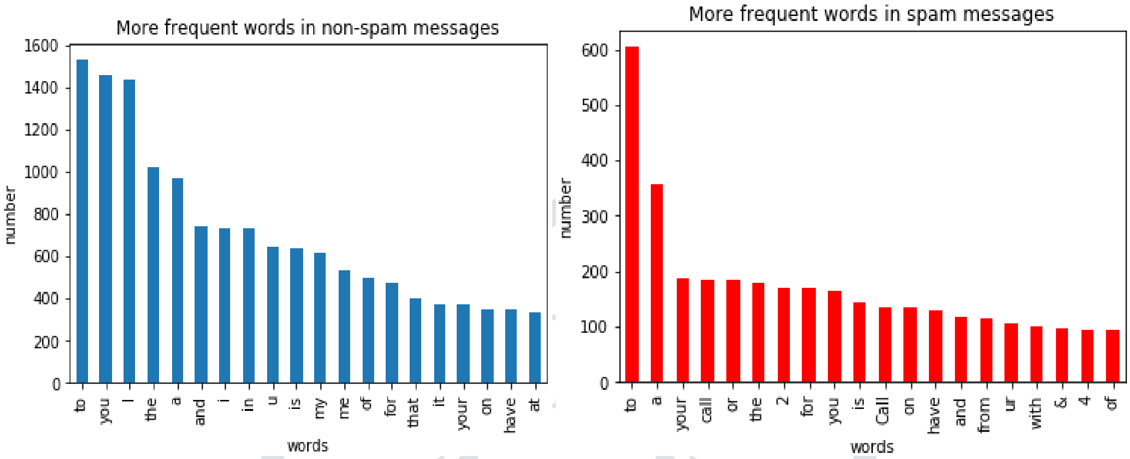


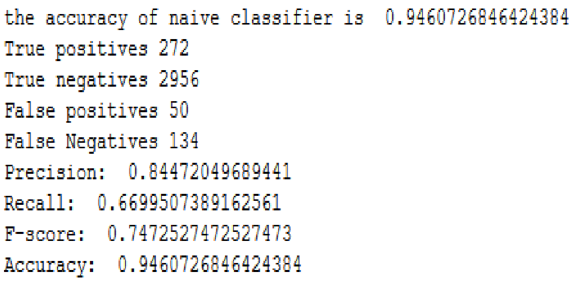
Fig.5 Analysis of spam and non-spam words

3.4 Classification

We prepared two models here to be specific Naive Baye's classifier and Support Vector Machines (SVM). Naive Bayes classifier is a regular and exceptionally famous strategy for record grouping issue. It is an administered probabilistic classifier dependent on Bayes hypothesis accepting autonomy between each pair of highlights.SVM is directed paired classifiers which are exceptionally successful when you have higher number of highlights. The objective of SVM is to isolate some subset of preparing information from rest called the help vectors (limit of isolating hyper-plane). The choice capacity of SVM show that predicts the class of the test information depends on help vectors and makes utilization of a piece trap. When the classifiers are prepared, we can check the execution of the models on test-set. We extricate word include vector for each mail in test-set and foresee its class (ham or spam) with the prepared NB classifier and SVM show.

3.4.1 Naive baye's classifier

For grouping a given message, first we preprocess it. For each word w in the handled informed we discover a result of P(w|spam). In the event that w does not exist in the train dataset we take TF(w) as 0 and discover P(w|spam) utilizing above equation. We duplicate this item with P(spam). The resultant item is the P(spam|message). Essentially, we discover P(ham|message). Whichever likelihood among these two is more prominent; the relating tag (spam or ham) is allocated to the info message. Note than we are not separating by P(w) as given in the equation. This is on the grounds that both the numbers will be partitioned by that and it would not influence the correlation between the two. Then, we assess the exactness, review and accuracy of the model with the test set.

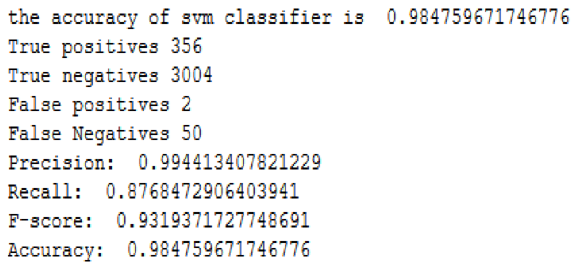


```
the accuracy of naive classifier is  0.9460726846424384
True positives 272
True negatives 2956
False positives 50
False Negatives 134
Precision:  0.84472049689441
Recall:  0.6699507389162561
F-score:  0.7472527472527473
Accuracy:  0.9460726846424384
```

Fig.6 Accuracy for Naive baye's classifier

3.4.2 SVM

We will similarly apply the help vector machine and demonstrate with the gaussian bit. We train distinctive models changing the regularization parameter C. We assess the exactness, review and accuracy of the model with the test set.

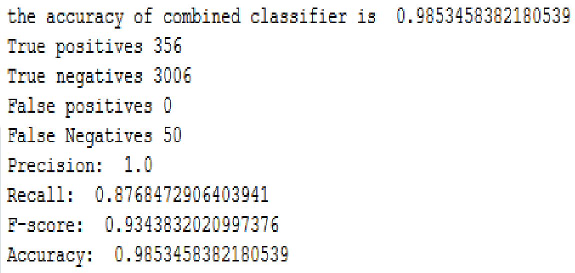


```
the accuracy of svm classifier is  0.984759671746776
True positives 356
True negatives 3004
False positives 2
False Negatives 50
Precision:  0.994413407821229
Recall:  0.8768472906403941
F-score:  0.9319371727748691
Accuracy:  0.984759671746776
```

Fig.7 Accuracy for SVM

3.4.3 Naive and SVM implementation

By combining Naive and SVM algorithm, we get the following accuracy, precision, recall, false positive rate etc.,



```
the accuracy of combined classifier is  0.9853458382180539
True positives 356
True negatives 3006
False positives 0
False Negatives 50
Precision:  1.0
Recall:  0.8768472906403941
F-score:  0.9343832020997376
Accuracy:  0.9853458382180539
```

Fig.8 Accuracy for hybrid classifier

Evaluation metrics	Naive Baye's Classifier	SVM classifier	Hybrid classifier
TP	272	356	356
TN	2956	3004	3004
FP	50	2	0
FN	134	50	50
Precision	0.84	0.99	1.0
Recall	0.67	0.88	0.88
F- Score	0.75	0.93	0.93
Accuracy (%)	94.6	98.5	98.5

Fig.9 Performance Evaluation

Therefore, by combining the following two classifiers, we reduced the false positive rate to zero and get the accuracy of 98.5%.

IV. CONCLUSION

It is important that spam mails do not reach the inbox of the users as this reduces efficiency of operations. But more importantly it is necessary that no ham mail goes to the spam folder as those lead serious problems to the user. In the proposed system, we discussed the requirements of improving the accuracy, precision parameters of data mining classification technique like Naive Bayesian and SVM and it is found that combination of SVM and Naive baye's is the best classifier of this study.

V. FUTURE WORK

In the future, various other algorithms are implemented to classification method to achieve better performance. Then, additionally try and collect additional datasets from the real world.

REFERENCES

[1] Ajay Sharma, Anil Suryawanshi ”A Novel Method for Detecting Spam Email using KNN Classification with Spearman Correlation as Distance Measure” ,2016.

[2] Deepak agarwal, Rahul Kumar”Spam filtering using SVM with Different Kernel Function”, 2016.

[3] Harshal Deshmukh, Chetan Nandeshwar, Sagar Wanjari, Pankaj Bhardwaj, Devendra Ramtekkar, Rajesh Nasare “Spam Mail Detection Using Artificial Neural Network”.

[4] https://www.kaggle.com/astandrik/simple-spam-filter-using-naive-bayes.

[5] Michael sirinivasan, Kyungbaek Kim, Xiaowei Yang” Social Filter: Introducing Social Trust to Collaborative Spam Mitigation”,2011.

[6] Rahul Bansod, R.S.Mangrulkar, V.G.Bhujade ” Spam Classification using Artificial Neural Network with Weight Measures”, 2016.

[7] Reena Sharma, Gurjot Kaur” Email Spam Detection using SVM and RBF”, 7 April 2016.

[8] RenWangI, Amr M. Youssef, Ahmed K. Elhakee “On Some Feature Selection Strategies for Spam Filter Design” 1-4244-0038-4 2006, IEEE CCECE/CCGEI Ottawa, May 2006.

[9] Savita Teli, Santoshkumar Biradar “Effective Email Classification for Spam and Non-Spam”, VOL.4, Issue 6,JUNE 2014.

[10] Sufian Hameed, Xiaoming Fu,Pan Hui,Nishanth Shastry ”LENS: Leveraging Social Networking and trust to Prevent Spam Transmission”, 2014.

[11] Suganya T, Hemlatha T” Spam Filtering in Online Social Networks using Machine Learning Technique”,Vol 2-Issue 1,Jan 2014.

[12] TAK-LAM WONG , KAI-ON CHOW, FRANZWONG“ Incorporating keyword-based filtering to document classification for email spamming” Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007.

[13] Wikipedia, “E-mail spam”. http://en.wikipedia.org/wiki/E-mail_spam

[14] Ze Li, Haiying Shen “SOAP: A Social Network Aided Personalized and Effective Spam Filter to wash Your E-mail Box”, 2011.