IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.8632

Email Spam Detection using Machine Learning Techniques

Rajesh Kumar J[1], Sudarshan P[2], Mahalakshmi G[3]

[1] PG – Master of Computer Application, CEG Anna University, Chennai, Tamil Nadu

[2] PG - Master of Computer Application, CEG Anna University, Chennai, Tamil Nadu

[3]Teaching Fellow, Master of Computer Application, CEG Anna University, Chennai, Tamil Nadu

Abstract: Email spam is one of the biggest threats to today's Internet. To dealwith this threat, many anti-spam filters have been developed. One big challengefor these filters is to predict the labels of emails in a personalized mailbox. These spam messages can lead to loss of private data as well. Modern day researchershave used some stylistic features of text messages to classify them to be hamor spam. E-mail spam detection can be greatly influenced by the presence of known words, phrases, abbreviations and idioms. This paper aims to compare different classifying techniques on different datasets collected from previous research works, and evaluate them on the basis of their accuracy, recall, and precision. The comparison hasbeen performed between traditional machine learning techniques. Most of the time such emails are commercial. But many times, such emails may contain some phishing links that have malware. This arises the need for proposing prudent mechanism to detect or identify such spam emails so that time and memory space of the system can be saved up to a great extent. In this paper, we presented the NLP mechanism which can filter spam and non-spam emails and also categorize into different spam mails. Our proposed algorithm generatesdictionary and features and trains them through machine learning for effective results.

Keywords: Naive Bayes, Support Vector Machine, Natural Language Processing, analysis.

I. INTRODUCTION

Consider a case in which someone over the internet is sending bulk emails regarding the promotion of their products for purchasing it, or someone sending a link to click or activate to win some lucrative prizes; such emails are generally considered as negative marketing strategies or fraud activities. As a receiver, you are helpless in this scenario. These unwanted emails may consume a lot of memory of your system also waste your precise time. It is also observed that one can be distributed by receiving such bogus emails again and again. So, there is a need for some mechanism that can reduce or even provide some sort of panacea to from these spam emails. Keep this situation in mind, in this paper; we are presenting a machine learning-based spam detection mechanism that uses a dataset of approximately 6000 valid and invalid collection of emails. Our proposed model will first make a dictionary that remove helping verbs form the contents of the email. Now our proposed algorithm will run to check whether the entered email address is spam or not. By applying this mechanism, a user can work efficiently as comparatively fewer spam emails will be received. This mechanism also saves the time and memory of the system.

II. EXPERIMENTAL METHODS OR METHODOLOGY

Out[2]:

	labels	text
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...
...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will ?_ b going to esplanade fr home?
5569	ham	Pity, * was in mood for that. So...any other s...
5570	ham	The guy did some bitching but I acted like i'd...
5571	ham	Rofl. Its true to its name

5572 rows × 2 columns

Figure 1: Email spam dataset extracted

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.8632

2.1 Machine Learning Algorithms

Machine learning is used to predict the system output easily without doing any programs. Basically, it builds the algorithm to receive the input then perform some analysis to generate the output. The expected output is going to be most accurate for the respective system.

2.1.1 Naive Bayes Classifier

Naive Bayes could be a reasonably classifier which uses the Bayes Theorem. It predicts membership probabilities for every class like the probability that given record or datum belongs to a specific class. The category with the very best probability is taken into account because the possibly class.

2.1.2 Support Vector Machine

Support Vector Machine may be a supervised machine learning algorithm which might be used for classification. Support Vectors are simply the co-ordinates of individual observation. It performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors that outline the hyperplane are the support vectors. Hyperplane separates the vectors (cases) into two non-overlapping classes. Perfect separation might not be possible.

2.2 Natural Language Processing

Natural language processing (NLP) is a field of artificial intelligence in which computers analyse, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.

III. RESULTS AND DISCUSSION

3.1 Dataset Collection

The figure 2 is referred to as the sample dataset in the csv file. Dataset Name: Spam Mails Dataset (enron-1 folder of spam dataset). This dataset is collected from Kaggle global public database repository. Used enron1 folder. Dataset Link: https://www2.aueb.gr/users/ion/data/enron-spam/

	A	B
1	labels	text
2	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
3	ham	Ok lar... Joking wif u oni...
4	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
5	ham	U dun say so early hor... U c already then say...
6	ham	Nah I don't think he goes to usf, he lives around here though
7	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, ?1.50 to rcv
8	ham	Even my brother is not like to speak with me. They treat me like aids patent.
9	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
10	spam	WINNER!! As a valued network customer you have been selected to receivea ?900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
11	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
12	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
13	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
14	spam	URGENT! You have won a 1 week FREE membership in our ?100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
15	ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.
16	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
17	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
18	ham	Oh k...i'm watching here:)
19	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
20	ham	Fine if that?s the way u feel. That?s the way its gota b
21	spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/?1.20 POBOXox36504W45WQ 16+
22	ham	Is that seriously how you spell his name?
23	ham	I?? going to try for 2 months ha ha only joking
24	ham	So ?_ pay first lar... Then when is da stock comin...
25	ham	Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?

Figure 2: Datasets (in CSV) Which are Collected Using Kaggle

3.2 Evaluation Indicators

To evaluate the results of the two algorithms we use four of the most popular measures: Accuracy, Precision, Recall, and F1 score. These four metrics are explained in the following:

1. **Accuracy Rate** is the most popular measure and also very easy to understand because is a simple ratio between the number of instances correctly predicted to the total number of instances used in the observation, in other words, accuracy gives the percentage of correctly predicted instances.
2. **Precision** is a measure that provides for each class the ratio between correctly positive predicted instances and total of positive instances predicted.

$$\mathbf{P = T\ P\ T\ P+FP}$$

3. **Recall** is a measure that provides for each class the ratio between the true positive instances predicted and the sum of true positives and false negatives in the observation.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.8632

R = T P T P+FN

4. **Fl score** is the weighted average of Precision and Recall and it is considered perfect when it is 1.0 and the worst possible value is 0.0, so a good F1 score means that we have low false positives and low false negatives.

F = 2PxR P+R

3.3 Classification Analysis

The Figure 3 shows the 5572 unique email messages that is splitted into the legitimate mail and spam email using the confusion matrix heat map.

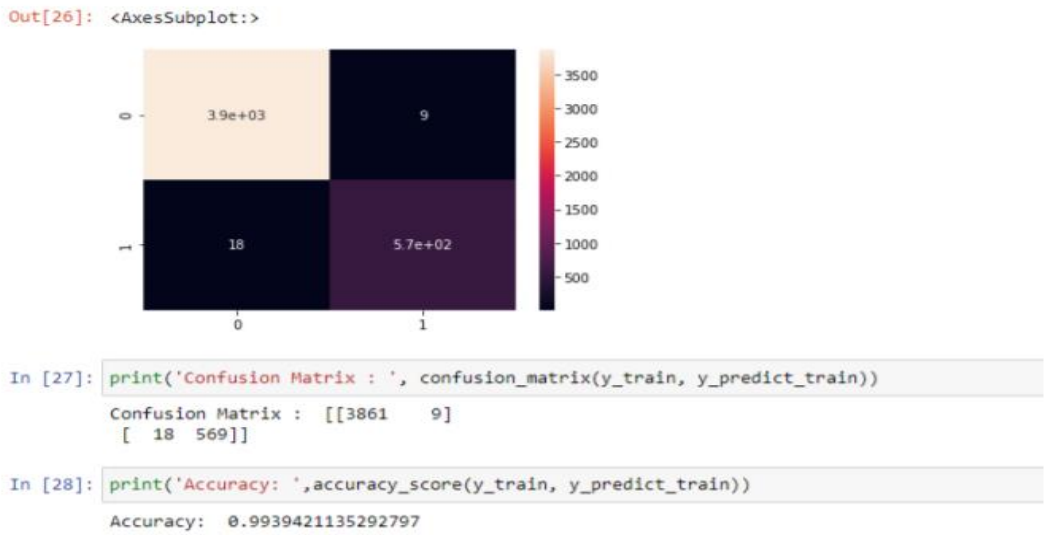

Figure 3: Legitimate and Spam mail

3.4 Naïve Bayes and SVM Results

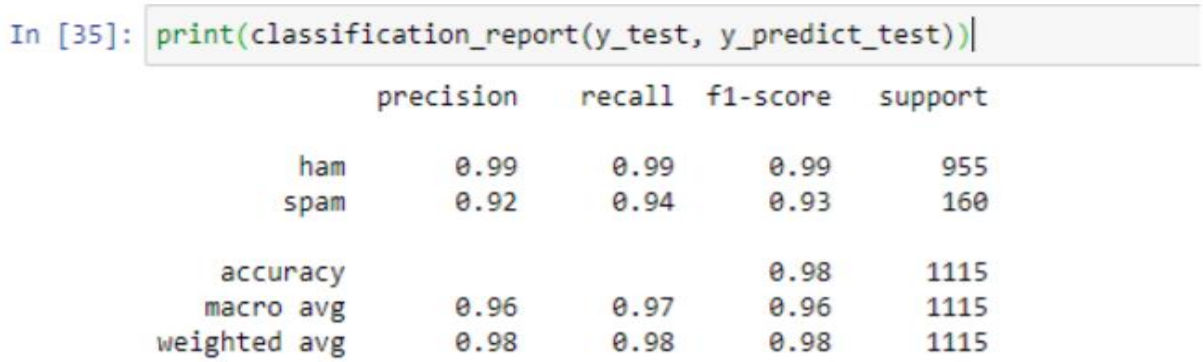

Figure 4: Naive Bayes Results

The Figure 4 are the details on most informative features after the classifier is executed on train data using Naive Bayes.

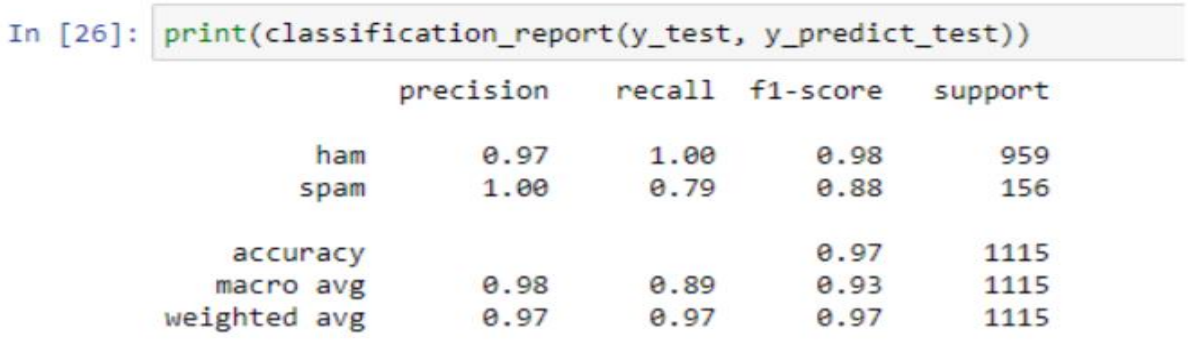

Figure 5: SVM Results

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.8632

The Figure 5 are the details on most informative features after the classifier is executed on train data.

3.5 Comparison and Evaluation Results

The Figure 6 shows that the accuracy rate of the twitter sentiment analysis based on Bayesian network can reach 94.09 percent in the top 5000 results, and the recall rate can reach 89.35 percent, and the F- measure can reach 90.45 percent, which shows that the proposed algorithm in this paper is still relatively satisfactory results.

Parameters	Accuracy	Precision	Recall	F-Measure
Naive Bayes	98	92	94	93
SVM	97	100	79	88

Figure 6: Evaluation Results of Two Algorithms

The above Figure 6 is a comparison with the similarity algorithm based on the SVM, the average accuracy rate is improved by 75.09 percent, the average recall rate is increased by 82.17 percent, and the average F-measure value is increased by 81.21 percent.

3.6 Result Analysis

Regardless of the evaluation index of the algorithm, the algorithm proposed is higher than the other algorithms. In this case, with the number of detected spam emails increasing, the accuracy decreases, and the recall rate and F measurement increases. This is mainly due to the number's increase of detected text, it increased the scope of the query, so resulting in increased recall rate. With continuous expansion of the inspection range, the error will increase and accuracy rate will decrease.

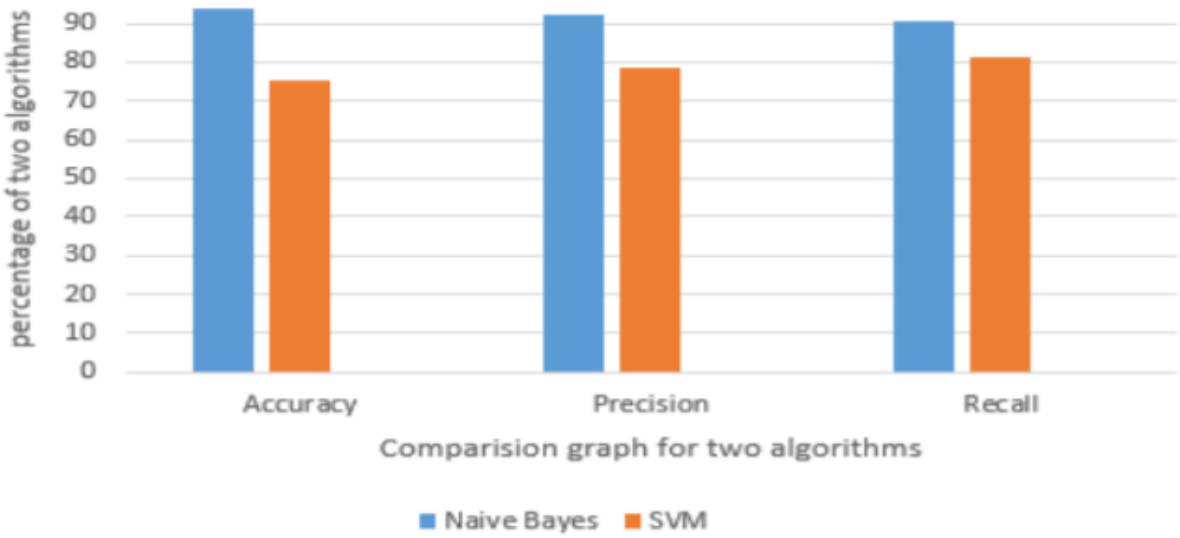

Figure 6: Graph for Comparison of Two Algorithms

It can be seen from Figure 5.9 that the accuracy rate of the similarity algorithm based on Naïve Bayes algorithm is obviously higher than that SVM network.

IV. CONCLUSION

This Project provides a work flow to understand and detect the legitimate and spam emails . To classify email spam dataset into five predefined categories to get in-depth knowledge of their learning experiences. It shows that the results derived from the Naive Bayes classifiers are much better than that of the SVM for text classification. Mining messages for understanding the email spam texts in python. It helps in understand the machine to understand human phrases and conversations.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.8632

REFERENCES

[1]. Sefat E Rahman and Shofi Ullah. Email spam detection using bidirectional long short-term memory with convolutional neural network. In 2020 IEEE Region 10 Symposium (TENSYMP), pages 1307–1311. IEEE, 2020.
[2]. Milivoje Popovac, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, & Andras Anderla. Convolutional neural network based sms spam detection. In 2018 26th Telecommunication Forum(TELFOR), page1–4. IEEE, 2018
[3]. Nikhil Govil, Kunal Agarwal, Ashi Bansal, and Astha Varshney. A machine learning based spam detection mechanism. In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pages 954–957. IEEE, 2020.
[4]. Aakash Atul Alurkar, Sourabh Bharat Ranade, Shreeya Vijay Joshi, Siddhesh Sanjay Ranade, Piyush A Sonewar, Parikshit N Mahalle, and Arvind V Deshpande. A proposed data science approach for email spam classification using machine learning techniques. In 2017 Internet of Things Business Models, Users, and Networks, pages 1–5. IEEE, 2017.
[5]. Mehul Gupta, Aditya Bakliwal, Shubhangi Agarwal, and Pulkit Mehndiratta. A comparative study of spam sms detection using machine learning classifiers. In 2018 Eleventh International Conference on Contemporary Computing (IC3), pages 1–7. IEEE, 2018.
[6]. N. Kumar, S. Sonowal, and Nishant. Email spam detection using machine learning algorithms. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pages 108–113, 2020.
[7]. K. Agarwal and T. Kumar. Email spam detection using integrated approach of naïve bayes and particle swarm optimization. In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), pages 685–690, 2018.
[8]. P. Liu and T. Moh. Content based spam e-mail filtering. In 2016 International Conference on Collaboration Technologies and Systems (CTS), pages 218–224, 2016.
[9]. M. K. Chae, A. Alsadoon, P. W. C. Prasad, and A. Elchouemi. Spam filtering email classification (sfecm) using gain and graph mining algorithm. In 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), pages 1–7, 2017. 51
[10]. S. Suryawanshi, A. Goswami, and P. Patil. Email spam detection: An empirical comparative study of different ml and ensemble classifiers. In 2019 IEEE 9th International Conference on Advanced Computing (IACC), pages 69–74, 2019.
[11]. E. E. Eryilmaz, D. Ozkan S¸ ahin, and E. Kılıc¸. Machine learning based spam e-mail detection system for turkish. In 2020 5th International Conference on Computer Science and Engineering (UBMK), pages 7–12, 2020.
[12]. W. Hijawi, H. Faris, J. Alqatawna, A. M. Al-Zoubi, and I. Aljarah. Improving email spam detection using content-based feature engineering approach. In 2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), pages 1–6, 2017.1.