EMAIL SPAM DETECTION

Project report for Data Mining, Dr Xiaofei Zhang instructing

Srividya Polneni
Computer Science
University of Memphis
U00848973
spolneni@memphis.edu

Sindhuja Yerramalla
Computer Science
University of Memphis
U00839259
syrrmlla@memphis.edu

1. ABSTRACT :

Recently unsolicited commercial/bulk e-mail also known as spam became a big trouble over the internet. Spam is waste of time, storage space and communication bandwidth. The problem of spam e-mail has been increasing for years. Automatic e-mail filtering seems to be the most effective method for countering spam now. Only several years ago most of the spam could be reliably dealt with by blocking e-mails coming from certain addresses or filtering out messages with certain subject lines. Spammers began to use several tricky methods to overcome the filtering methods, like random sender addresses and/or appending random characters to the beginning or the end of the message subject line. Machine learning techniques nowadays are used to automatically filter spam e-mails at a very successful rate. The machine learning field is a subfield of the broad field of artificial intelligence, this aims to make machines able to learn like humans. Learning here means understanding, observing, and representing information about some statistical phenomenon. First, data collection and representation are mostly problem specific (i.e., e-mail messages), second, e-mail feature selection and feature reduction attempt to reduce the dimensionality (i.e., the number of features). Finally, the e-mail classification phase of the process finds the actual mapping between the training set and testing set. The Machine Learning approach includes lots of algorithms that can be used in e-mail filtering like Naïve Bayes, K-nearneighbourhbor, Support Vector Machine, and classifiers. In conclusion, we try to summarize the performance results of the few machine learning methods in terms of spam precision and accuracy.

2. PROBLEM STATEMENT:

Different spam filtering formulas are used by Gmail, Outlook.com, and Yahoo Mail to deliver solely valid emails to their users and strain illegitimate messages. Conversely, these filters additionally typically mistakenly block authentic messages. it's been according to that concerning twenty p.c of authorization based mostly emails sometimes fail to urge to the inbox of the expected recipient. the e-mail suppliers have designed varied mechanisms to be used in email anti-spam filters to curtail the risks posed by phishing, email-borne malware, and ransomware to email users. The mechanisms area unit will not decide the danger level of every incoming email. samples of such mechanisms embody satisfactory spam limits, sender policy frameworks, whitelists and blacklists, and recipient verification tools. These mechanisms may be utilized by single or multiple users. once the satisfactory spam thresholds are simply too low it will cause a lot of spam to evade the spam filter and get into the users' inboxes. in the meantime, having an awfully high threshold will cause some vital emails to be isolated unless the administrator redirects them.

CHALLENGING PART:

To detect spam emails from several irrelevant emails.

3. EXISTING SYSTEM :

Email spam is one of the unsolved problems of today's Internet, annoying individual users and bringing financial damage to companies. Among the approaches developed to stop spam emails, filtering is a popular and important one. Common uses for email filters include organizing incoming emails and computer viruses and removal of spam. As spammers periodically find new ways to bypass spam filters and distribute spam messages, researchers need to stay at the forefront of this problem to help reduce the number of spam messages. Currently, spam emails occupy more than 70% of all email traffic. The negative effect spam has on companies is greatly related to financial aspects and the productivity of employees in the workplace. In this paper, we are proposing a new approach to classifying spam emails using ML techniques.

4. PROPOSED SYSTEM :

Most of the business and general communication is done through email because of its cost-effectiveness. This efficiency leads emails exposed to various attacks including spamming. Nowadays spam email is the foremost concern for email users. These spams are used for sending fake proposals, advertisements, and harmful content in the form of executable files to attack user systems or link to malicious websites resulting in the unessential consumption of network bandwidth. This paper elucidates the

different Machine Learning Techniques such as K-Nearest Neighbor, Naive Bayes, and Support Vector Machine algorithms for filtering spam emails using email datasets. However, here the comparison of different spam email classification techniques is presented and summarizes the overall scenario regarding the accuracy rate of different existing approaches.

5. DATASETS

In this system, the email dataset is shown in Figure.1 which is contain ham and spam messages with two columns 'Class' and 'Email imported from the UCI repository. This dataset contains 5574 messages and among them, 4827 were ham messages, and 747 were spam messages. The dataset is stored in CSV (Comma Separated Value) file format where each row represents a single message.

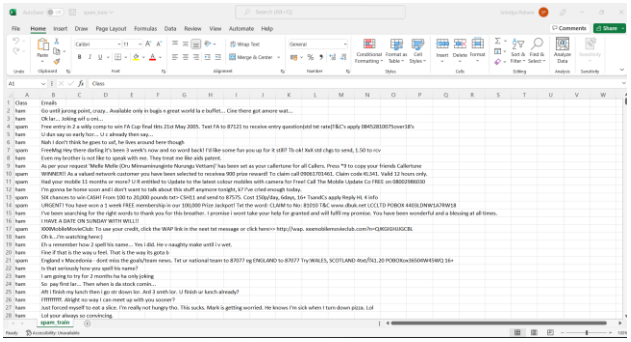


Fig1: SCREENSHOT OF DATASET USED

6. SOLUTION

We have applied the Navies base theorem and K-nearest Neighbor, Support vector Machine techniques. These are the Machine learning techniques.

Naive Bayes:

The naïve Bayes machine learning algorithms are useful for the categorization of documents and email spam filtering and this algorithm is working based on Bayes' rule.

$$P(A \mid B) = P(B \mid A)\, P(A)\, P(B) \quad (1)$$

Wherein:
A is a class

B is a message

P (A) is a class probability

P (B) is the probability of a message

P (B|A) is the conditional probability of the class for the given message B

P (A|B) is the conditional probability that message B belongs to class A.

In our system for implementing the naïve Bayes algorithm we use a python library whose name was sklearn.naive_bayes.multinomialnb class. This algorithm has a function like fit() which is building the training model whose inputs are independent and dependent values of the dataset and uses predict () function which takes input as a test message then it can predict the spam email detection as spam or ham.

K-Nearest Neighbour:

The K-nearest neighbour classifier is the simplest algorithm for the prediction of any dataset with help of Euclidean distance. Here K means the number of neighbours, so if we take k=1 then it gives the very nearest neighbour as output. It is working on a voting-based system which means while prediction takes the nearest neighbour's votes which output gets more votes that become the predicted output value.

$$D = \sum k\, i = 1\, (xi - yi)\, 2 \quad (2)$$

Wherein:

D is the distance between x and y

K is the nearest neighbour

x and y are independent attribute values

In our system for implementing the K-nearest neighbour algorithm, we use a python library whose name was sklearn.neighbors.KNeighborsClassifier class. This algorithm has functions like fit () which is building the training model whose inputs are independent and dependent values of the dataset and use predict () function which takes input as a test message then it can predict the spam email detection as spam or ham.

Support Vector Machine:

The support vector machine is a supervised machine learning algorithm that is used for the classification of instances. It can separate the data linearly and for non-linear data, it can use kernel functions. The spam email classifies the two classes with help of a hyperplane which has the largest margin to separate the dataset into classes. The margin between the two classes represents the longest distance between the closest data points of those classes which are called support vectors.

$$w.u+b \geq 1 \quad (3)$$

$$w.u+b \leq -1 \quad (4)$$

$$y\,(w.u+b)-1 \geq 0 \quad (5)$$

Wherein

b is a constant distance.

w, u are vectors.

y is the output is 1 for positive samples and -1 for negative samples.

In our system for implementing support vector machine algorithm, we use python library which named was sklearn.svm.SVR class. This algorithm has functions like fit () which is building the training model whose inputs are independent and dependent values of the dataset and using predict () function which takes input as a test message then it can predict the spam email detection as spam or ham.

Pre-processing:

In this system first, we need to gather the training dataset which is in CSV file format so that we need to read the file data with help of panda's library for conversation to list array as well as the input message which is given by using that one should be appended to list array because the machine cannot understand the file format data. The finishing of converting process then can remove the stop words and stemming words from training and testing data which is reduced irrelevant data.

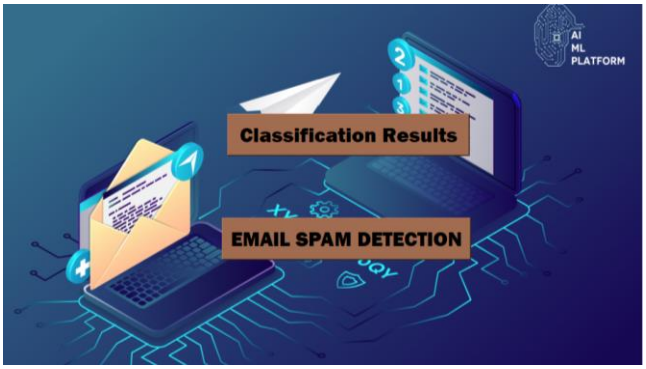
Features Extraction:

In this system, we are using an email dataset that is in text format so that for comparison word to word it takes a long time for spam email detection which is not recommended. To overcome this after finishing the pre-processing stage we need to convert the text (messages) to numeric data. So, in this system, we are implementing the TfidfVectorizer python library for generating numerical data from text data which is available in list format. Here we need to apply TfidfVectorizer for both training and testing with n-gram weights which is useful for splitting keywords. The TfidfVectorizer class contains the fit_transform() function which takes input as a number of messages in a list format and generates features like attributes with help of the get_feature_names () function. Finally, these features become independent values and the Class column becomes a dependent value and applies any machine learning classifier then predict whether the given message is spam or ham.

7. RESULTS/ OUTPUT SCREENS:

When we run the main.py file it would popup the main screen, Now we should enter the Username and password where the Username is admin, Password is admin

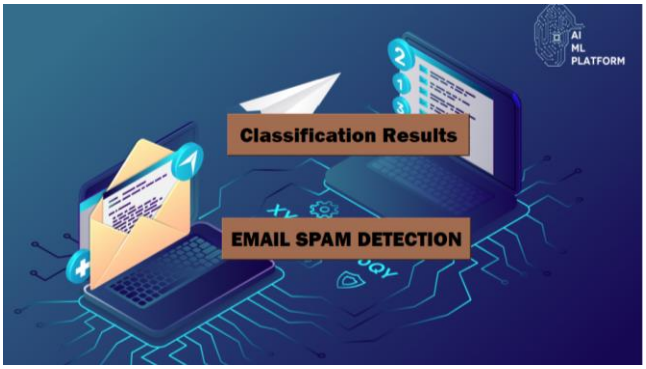
After entering the details it would show the below screen. You can click on Email Spam detection.



To check whether a mail is spam or Ham, enter the email subject in the given text box. The result below it shows that it is a SPAM

Based on the above results we can say that the Email subject entered is a HAM.

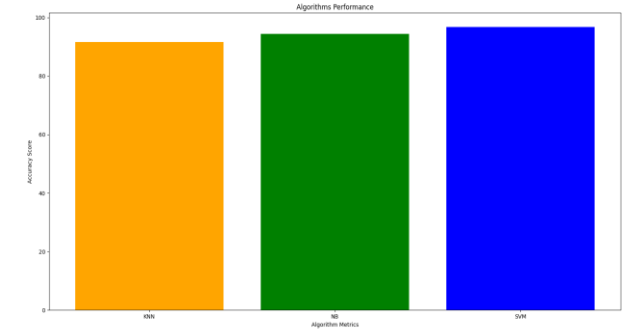
Also when click on the Classification results



We have applied techniques like Accuracy, precision and recall values for all three algorithms. We can clearly observe from the below results that Algorithm performance is better in the SVM classification technique.

Results

	Algorithm	Accuracy	Precision	Recall	F1_Score
1	KNN	91.63179916317992	63.703703703703695	47.175141242937855	51.343823979947075
2	NB	94.38135086670651	64.61739699149771	53.672316384180796	57.53936122357175
3	SVM	96.77226539151226	64.42606859556011	60.14813283726057	62.0542543588769



8. DELIVERABLES:

Feel free to clone and experiment with my GitHub repo. I am leaving the link below. After cloning into the repository, run all the files and you will be redirected to a window mentioned above in the results and screen shorts.

Here is the link:

https://github.com/PolneniSrividya/EMAIL_SPAM/tree/main/Detect_SpamMail/Detect_SpamMail

9. CONCLUSION:

In this study, we have a tendency to review machine learning approaches and their application to the sector of spam filtering. A review of the state-of-the-art algorithms applied for the classification of messages as either spam or ham is provided. The makes attempt created by totally different researchers to find the matter of spam through the utilization of machine learning classifiers was mentioned. The evolution of spam messages over the years to evade filters was examined. the essential design of email spam filters and therefore the processes concerned with filtering spam emails were looked into. The paper surveyed a number of them in publicly accessible datasets and performance metrics that may be wont to live the effectiveness of any spam filter. The challenges of the machine learning algorithms in expeditiously handling the menace of spam were found and comparative studies of the machine learning techniques accessible in literature were done. we have a tendency to additionally disclosed some open analysis issues related to spam filters. In general, the figure and volume of literature we have a tendency to review show that vital progress is created and can still be created in this field. Having mentioned the open issues in spam filtering, more analysis to reinforce the effectiveness of spam filters got to be done. this may create the event of spam filters to still be an energetic analysis field for academician and business practitioners researching machine learning techniques for effective spam filtering. Our hope is that analysis students can use this paper as a springboard for doing qualitative analysis in spam filtering mistreatment machine learning, deep learning and deep adversarial learning algorithms. The overall accuracy of the results achieved is 99.9% accuracy on training data and 98.2% on testing data with less false positive rate. It shows that classifiers give better training data and less compared to testing data. It has also been further observed that the proposed system has the least per cent error and hence can be deemed the most accurate method. The future enhancement will be to extend this design to take into account more attributes that could classify the emails using images and also include different datasets in training the algorithm into producing more accurate results.

10. ACKNOWLEDGMENTS:

We thank Dr Xiaofei Zhang for teaching us a lot of new techniques in data mining. It helped us to understand the concepts and apply them in a practical way.

11. REFERENCES:

[1]. M. Awad, M. Foqaha Email spam classification using hybrid approach of RBF neural network and particle swarm optimizationInt. J. Netw. Secur. Appl., 8 (4) (2016)

[2]. D.M. Fonseca, O.H. Fazzion, E. Cunha, I. Las-Casas, P.D. Guedes, W. Meira, M. Chaves

[3]. Measuring characterizing, and avoiding spam traffic costs IEEE Int. Comp., 99 (2016)

[4]. Visited on May 15, 2017 Kaspersky Lab Spam Report (2017) 5. 2012 https://www.securelist.com/en/analysis/204792230/Spam_Report_April_2012