INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

IJARIIT

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 2)
Available online at: www.ijariit.com

Hybrid spam detection using machine learning

Diksha S. Jawale
jawalediksha27@gmail.com
SSBT's College of Engineering and Technology,
Jalgaon, Maharashtra

Ashwini G. Mahajan
ashwinimahajan2807@gmail.com
SSBT's College of Engineering and Technology,
Jalgaon, Maharashtra

Kalyani R. Shinkar
jawaleduhita27@gmail.com
SSBT's College of Engineering and Technology,
Jalgaon, Maharashtra

Vaishnavi V. Katdare
jawalediksha@gmail.com
SSBT's College of Engineering and Technology,
Jalgaon, Maharashtra

ABSTRACT

Social networks are recognized as popular communication channel but in this, there is one of the problems is spam messages. Spam messages can contain malware in the form of the executable file and the link to the malicious websites or the links which do not exist. Most of the existing machine learning solutions are based on either Support Vector Machine or Naive Bayes but the existing solutions either slow or inaccurate in solving spam filtering problem. Support Vector Machine-based spam filter has great advantages on high precision and recall rate and Naive Bayes based spam filter give faster classification speed and require small training sets. By taking the advantages of both, we propose hybrid spam filtering algorithm which has more accuracy than separately implemented NB and SVM.

Keywords: *Naive Bayes, Support vector machine, SVM trimming technique, Spam filtering technique, Spam, Ham.*

1. INTRODUCTION

Online social networks are recognized as popular communication channels, it becomes a convenient method to communicate with each other because of its popularization, low cost, and fast delivery of a message. Messages and emails are still most popular forms and main sources of the communication through the internet in our daily life. Along with the growth of communication on the internet, one of the long-last problem in messages is the existence of spam messages, there has been a dramatic growth in spam in recent years. Spamming is the abuse of electronic messaging systems to send unsolicited bulk messages or to promote products or services, which are almost universally undesired bulk data received by massive recipients. Spam can originate from any location where internet access is available. Spam data may contain malware, in the form of scripts or executable files, or contain disguised links that lead to phishing websites or the files attached to the email that contains viruses or many of emails are containing the URL of any website which may exist but malicious or which may not exist. Because of this many peoples face the problems, so it's necessary to know which email is spam and which email is regular. Efficiently and accurately detect and filter spams is a critical and important research issue. There are many spam filters are available but some of them having the accuracy problem and some are slow in finding spam emails. Spam filtering is the processing of data to organize it according to specified criteria. Most often this refers to the automatic processing of incoming messages, and to outgoing emails as well as those being received. Some email filters are able to edit messages during processing some of the filters automatically detect that the email is spam but while detecting the spam email automatically, this type of spam filters allowed to change the contents of the messages. The mechanism that only necessary email is automatically taken out of a large amount of email. Because the content of the mail is basically described by the text it can be said that task of classifying email into spam email and other email is text Classification task. Therefore, various text classification algorithms can be applied for the mail classification task. There are many solutions to spam filtering, e.g. the black-list and white-list filtering techniques, decision tree based approaches, and machine learning based methods. Among various solution, machine learning based ones are receiving more and more attention, due to its high accuracy in detecting spams having different algorithms as algorithms as support vector machine, naive Bayes, k-means, decision tree, neural network. Most of the existing machine learning based solutions are based on either the support vector machine

(SVM) or Naive Bayes (NB) methods. Typically, an SVM based spam filtering solution has great advantages of great accuracy on high precision and recall rate but it works slowly and requires a large dataset, while an NB based solution has faster classification speed but accuracy provided by naïve Bayes is less than the support vector machine and requires a smaller training set, but these existing solutions are either too slow or inaccurate in solving the spam filtering problem. Because of this, the some of the spam emails cannot detect due to low accuracy and sometimes it detects slowly. So it is necessary to develop an innovative system which will have the great accuracy as well as fastest classification of emails or messages.

2. RELATED WORK

The existing work undergoes an implementation on detection of malicious URL, malware, and bulk data in spam messages and mails. Weimiao Feng, Jianguo Sun, Liguo Zhang, Cuiling Cao and Qing Yang, had shown that the need for effective spam filters increases [1]. They discussed spam filtering methods and their correlated problems by using Bayesian algorithms and Support Vector Machine (SVM) and implemented a hybrid spam [2] filtering process which has advantages of both of the algorithms. Recent spam filters are discussed in this paper for determining spam messages which utilize. Semantic analysis information. They simplify both the algorithms as Support Vector Machine and also the Naive Bayes algorithm, separately implement the system of both and also an innovative spam detection system and then give the accuracy performance of both which shows that the Naive Bayes is giving the fastest classification of the data but it has less accuracy and In SVM, it has greater accuracy higher precision and recall rate but require more time. The SVM-NB algorithm having the advantages of both require less time and higher precision and accuracy. The SNM-NB has a better performance than the separate SVM, NB algorithm. In their training phase, it having 96.70% in SVM, 98.60% in NB, 99.14% in SVM-NB. And testing phase gives the result as in 92.49% SVM, 98.53% in NB, and 98.77% in SVM-NB accuracy performance. Sunil B. Rathod and Tareek M. Pattewar, has given a spam filtering method by using Bayesian classifier [3] to improve the accuracy by reduced feature sets and considered phish Tank dataset, the work was restricted to URL in Email only. A Bayesian classifier is statistical classifier works on independence computation of probability. They have considered the content of Email with features of a domain and shown that accuracy can be increased. In their training phase, it had 96.46%and testing phase gives the result as in 95% accuracy in NB. Ayahiko Niimi, Hirofumi Inomata, Masaki Miyamoto and Osamu Konishi, has explored two main semantic methods: [4] Bayesian algorithms and Support Vector Machine (SVM). Recent spam filters are discussed in this paper for determining spam messages which utilize semantic analysis information. They simplify both the algorithms as Support vector machine and also the naive Bayes algorithm, separately implement the system of both and them give the accuracy performance of both which shows that the Naive Bayes is given the fastest classification of the data but it has little on inaccuracy and In SVM, it has greater accuracy higher precision and recall rate. SVM based algorithms usually offer high precision and recall rate but NB based solutions have faster classification speed and require a fewer number of training samples. In summary, none of the existing solutions can quickly and accurately classify spam emails. In their training phase it having 97.59% in SVM, 98.66% in NB and testing phase gives the result as in 90.15% SVM,98.33% in NB accuracy performance. Priyanka Chhabra, Rajesh Wadhvani, and Sanyam Shukla had presented spam detection based on an application [5] to Anti-Spam Email using a newly improved Support vector ma- chine based Email filter. They have used vector weights for representing word frequency and adopted attribute selection based on word entropy and deduce its corresponding formula. It is proved that their filter improves total [6] performances apparently. In their training phase, it had 97.56%and testing phase gives the result as in 90.28% accuracy in SVM.

3. SPAM FILTERING USING NB-SVM

Most of the existing machine learning based solutions are based on either the support vector machine (SVM) or Naïve Bayes (NB) methods. Typically, an SVM based spam filtering solution has great advantages on high precision and recall rate, while an NB based solution has faster classification speed and requires a smaller training set. Overall, existing solutions are either too slow or inaccurate in solving the spam filtering problem. Another technical challenge in spam detection lies in the dependence of the features extracted from spam emails, which may cause inaccurate classifications in NB based algorithms.

3.1 Overview of Spam Filtering

A spam filter is used to detect spam emails using machine learning algorithm. Spam filter fig [1] classifies the data into spam emails and ham emails. Spam filters filter the messages with the malware, malicious data, phishing mails which contain a link to the malicious website which may not exist otherwise it contains viruses. These spam filters are either develop in a supervised algorithm or unsupervised algorithms. Mostly supervised algorithms are used in spam filters, that are SVM (Support Vector Machine), NB (Naive Bayes), K-means, decision tree algorithms. We use SVM and NB to implement a hybrid spam filter.

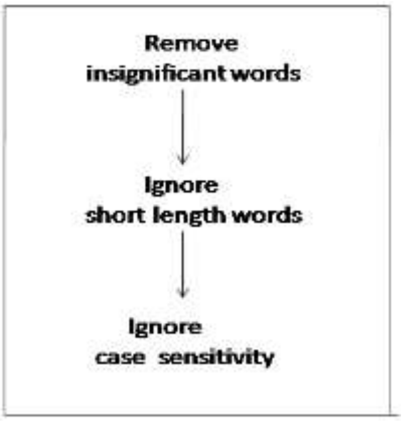

Fig. 1. Filtration Process

3.2 Naive Bayes

Naive Bayes (NB) is an algorithm to classify the data set expressed by the probability into two classes. In the spam filter of NB, emails are classified as spam and regular email by NB algorithm. Navie Bayes algorithms is a probability-based algorithm which requires small dataset to train the system. These algorithms have fastest classification speed but inaccurate due to precision and recall rate. Naive Bayes algorithms fig [2] first filter the dataset and then separate the message into tokens. For each token, it calculates the spam probability.

For each word (A), the probability of email i.e. spam probability is expressed as follows:

$$P\ (A) = \frac{a^{spam}/_{s}}{w\left(a^{ham}/_{h} + a^{spam}/_{s}\right)}$$

Where,
P(A): Is a probability for word.
a^{spam}: Appearance time of word in spam mail.
a^{ham}: Appearance time of word in ham mail.
s: Number of spam mail.
h: Number of ham mail.
w: weight i.e. tfidf, which is calculated by taking the probability of spam to ham.
After calculating spam probability, it calculates the composite probability for the message where, P(A_1) is a spam probability.

$$P(A_1A_2 \dots A_n) = \frac{P(A_1)*P(A_2)*\dots P(A_n)}{PA_1*PA_2*\dots*PA_n+(1-P(A_1))\dots(1-P(A_n))}$$

3.2.1 Naive Bayes algorithm

Process for classifying whether an email is a spam or ham.

1) Filter the data
2) Calculate spam probability.
3) Calculate Composite probability.
4) If the composite probability is higher than 0.5, then the email is classified as spam.
5) If the composite probability is smaller than 0.5, then the email is classified as ham.

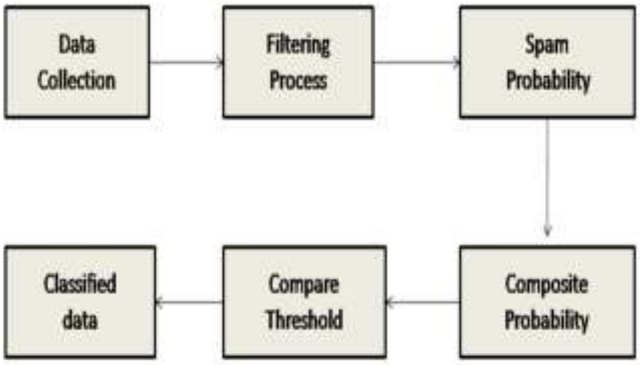

Fig. 2. NB Architecture

3.3 Support Vector Machine

Support vector machine (SVM) is an algorithm to classify the dataset express by the feature vector into two classes. In the spam filter by SVM, email is classified in spam email and regular email by using support vector machine algorithms. SVM uses data expressed by the vector to classify the data. SVM have great accuracy due to high precision and recall rate, but it has slow classification speed and requires large dataset to train the system. In SVM filter it requires a labeled dataset which is label as spam and ham. This dataset is the f filter and then all the messages are separated into a number of tokens. Token code is allocated to each token. For each word calculate the appearance frequency. Along with the feature make a feature vector with token code and appearance frequency.

(X1; Y1), (X2; Y2).... (Xn; Yn) where,

Xi is a vector with a numeric value as the number of times token occurs in the message.
Yi (+1, -1)
which define two classes, +1 = Spam, -1 = Ham
Along with the feature vector, SVM constructs a hyper plane by plotting a vector point's fig [3]. The hyper plane is a line which is closer to more vector points and classifies the data.
Where,
Y = +1 is a class +1 having label spam.
Y = -1 is a class +1 having label ham.
From this two classes, data is classified as spam and ham email fig [4].

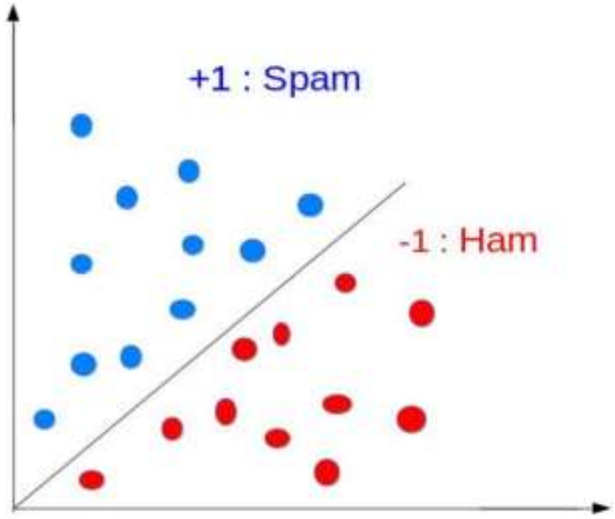

Fig. 3. Hyper plane

3.3.1 Support vector machine algorithm

Process for classifying whether an email is a spam or ham.

1) Filter the data
2) Separate all messages into tokens
3) Make a vector with the token code and its appearance frequency.
4) Construct the hyper plane:- class +1 for spam class -1 for ham
5)

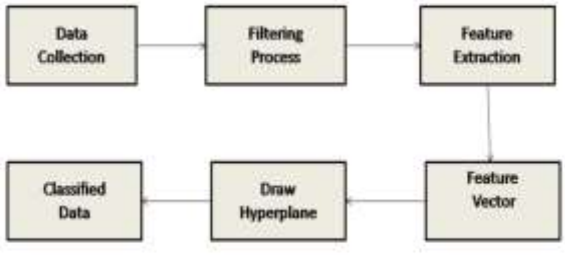

Fig. 4. SVM Architecture

3.4. Implementation of NB-SVM

Spam filters are used to classify the email into spam and ham. We NB-SVM algorithm to detect spam emails. NBSVM is a hybrid spam filtering algorithm which requires the advantages of both NB and SVM. Naive Bayes (NB) algorithm is having fast classification and also requires small dataset but it has low accuracy performance, because naive Bayes has major limitation of assumption of independence of the features extracted from training samples and support vector machine (SVM) algorithm having highest accuracy performance due to their high precision rate and recall rate because SVM algorithm is capable to find out a perfect hyper plane which divides training samples into two categories, it maximizes the distance between boundaries of two categories and not allowed the mixing of two boundaries. Due to such an independence assumption in NB, recall rate and accuracy affected and to remove the problem we apply an SVM trimming technique which eliminates the samples that are classified into the wrong categories by NB algorithm. SVM works on the vector, that is independent on the feature vector and improves the accuracy and NB increases the speed, because of this we implement an innovative hybrid NB-SVM algorithm to increase its performance. In the NB-SVM algorithm, the dataset is divided a training set and the testing set. Training data is first processed by NB algorithm in which it calculates the probability for each word and message in the dataset and compares with a threshold which classifies the data. NB processed data is going to SVM to improve accuracy by calculating the feature vector, it draws the hyper plane along with this vectors and classifies the data fig [5].

3.4.1Algorithm:

Divide the labeled dataset into training data and testing data.

Training phase

Input: 80% of data

1) Filtration:

a) Remove insignificant words.
b) Remove words having a length more than 3.
c) Remove case sensitivity that converts all letter in lower case.

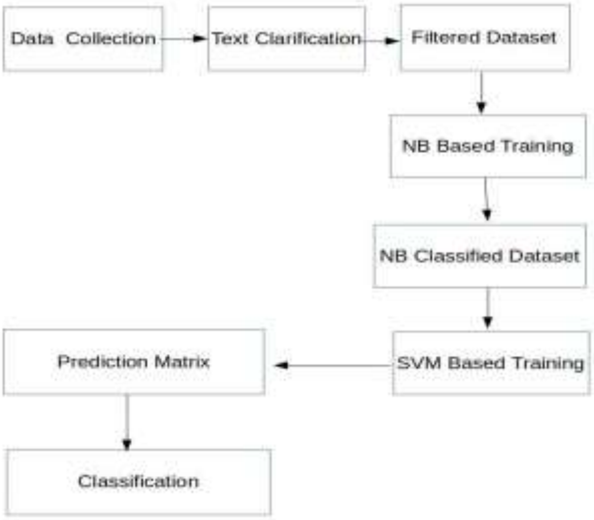

Fig. 5. NB-SVM Architecture

2) NB based training:

Input: filter dataset

a) Calculate the spam probability for each word in the message.
b) Then calculate the composite probability for each message
c) Compare composite probability with a neutral value 0.5.
d) If the composite probability is less than 0.5 then: the message is ham. Otherwise, spam.

3) SVM based training:

Input: NB classified dataset.

a) Separate each word into tokens.
b) Each token is given a token code.
c) Make vector with token code and appearances time of the word in a message i.e. (xi,yi).......(xn,yn) where xi is a vector yi +1, -1 y= +1, define class 1 having label spam. y= -1, define class -1 having label ham.
d) Construct the hyper plane along with the vector and plot the vector point and draw the hyper plane line as more point should closer to the line.
e) Classify data as spam or ham.

Testing Phase Input: 20% of the dataset

1) Filtration

a) Remove insignificant words.
b) Remove words having a length more than 3.
c) Remove case sensitivity that converts all letter in lower case.

2) NB-SVM testing

Test the data as per NB-SVM training

3.5. Performance Evaluation

The performance evaluation in spam filtering makes the use of related indexes in text classification. It can decide whether text classification is proper or not that is accuracy and speed. The speed is decided by the complexity of the arithmetic evaluation and the accuracy is getting evaluated by the information retrieval evaluation. The two common indexes as follows: Recall Ratio and Precision Ratio of information retrieval in the spam filtering techniques Recall Ratio is the ratio of the number of spam that has been filtered to the number of emails that should be filtered. The formula for Recall Ratio is:

$$\text{Recall} = \frac{\text{Number of filter spam emails}}{\text{Number of emails that should be filtered}}$$

Precision Ratio is the ratio of the number of spam that has been filtered to the number of emails that have been filtered. The formula for Precision Ratio is:

$$\text{Precision} = \frac{\text{Number of filtered spam emails}}{\text{Number of emails that have been filtered}}$$

4. RESULT ANALYSIS

In spam detection system, we use two algorithm NB and SVM. NB having fast classification speed and SVM with the highest accuracy. Using NB we got 96.65% accuracy in the training phase and 95.78% in the testing phase. By using SVM we got 99.43% accuracy in the training phase and 97.13% in the testing phase. We combine these two algorithms for getting highest accuracy with fastest classification speed and required small dataset. Along with this combine NB-SVM algorithm, we got 99.44% accuracy in the training phase and 97.57% in the testing phase. Which shows that their combination has more accuracy than by separately implementing them.

Table-1: Result

NB		SVM	
Training	Testing	Training	Testing
96.65%	95.78%	99.43%	97.13%
NB-SVM			
Training		Testing	
99.44%		97.57%	

5. CONCLUSION

Existing spam filters are either developed on SVM or on NB, but it has some drawbacks as SVM has great accuracy but slow classification speed and require more dataset and NB has fast classification speed but having low accuracy and requires small dataset, thus we implemented combine NBSVM hybrid spam filter which having more accuracy than both separately implemented NB and SVM.

4. REFERENCES

[1] L. Z. C. C. Weimiao Feng, Jianguo Sun and Q. Yang, A support vector machine based naive Bayes algorithm for spam filtering, 2016.
[2] K. M. Dr. Deve ndra K. Tayal, Amita Jain, Development of the anti-spam technique using modified k-means and naive Bayes algorithm, 2016.
[3] T. M. P. Sunil B. Rathod, Content-based spam detection in email using bayesian classifier, 2015.
[4] M. M. Ayahiko Niimi, Hirofumi Inomata and O. Konishi, Evaluation of bayesian spam filter and svm spam filter, 2004.
[5] S. S. Priyanka Chhabra, Rajesh Wadhvani, Spam filtering using support vector ma- chine, 2010.
[6] S. Wang and C. D. Manning, Baselines and bigrams: Simple, good sentiment and topic classification.