IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 2, February 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002021

Email Spam Detection and Filtering Using Machine Learning

Tausif Mansuri [1], Neeraj Oswal [2], Taher Patanwala [3],Yogesh Sharma [4]

U.G. Student, Department of Computer, Sinhgad Academy of Engineering, Pune, Maharashtra, India [1]

U.G. Student, Department of Computer, Sinhgad Academy of Engineering, Pune, Maharashtra, India [2]

U.G. Student, Department of Computer, Sinhgad Academy of Engineering, Pune, Maharashtra, India [3]

U.G. Student, Department of Computer, Sinhgad Academy of Engineering, Pune, Maharashtra, India [4]

ABSTRACT: Many techniques have been proposed to combat the upsurge in spam. All the proposed techniques have the same target, trying to avoid the spam entering our inboxes. Spammers avoid the filter by different tricks and each of them needs to be analyzed to determine what facility the filters need to have for overcoming the tricks and not allowing spammers to full our inbox. Different tricks give rise to different techniques. This work surveys spam phenomena from all sides, containing definitions, spam tricks, anti spam techniques, data set, etc. Finally, we discus the data sets which researchers use in experimental evaluation of their articles to show the accuracy of their ideas. Machine learning methods of recent are being used to successfully detect and filter spam emails. We present a systematic review of some of the popular machine learning based email spam filtering approaches. Our review covers survey of the important concepts, attempts, efficiency, and the research trend in spam filtering. The preliminary discussion in the study background examines the applications of machine learning techniques to the email spam filtering process of the leading internet service providers (ISPs) like Gmail, Yahoo and Outlook emails spam filters.

KEYWORDS: Computer science, Analysis of algorithms ,Machine learning, Spam filtering, Deep learning ,Neural networks, Support vector machines, Naïve Bayes.

I. INTRODUCTION

For sharing of important and official information, email is used as a default medium of communication. It is used widely as it also provides the confidentiality of the data shared. But with the pros also comes the cons, as many people misuse this reliable and easy way of communication by sending unwanted and useless bulk messages for their own personal benefits. These unwanted emails affect the normal user to face the problems like flooding of the mail box with unwanted emails making it harder to look for the useful once, even sometimes one may skip through important and useful emails because of all these unwanted emails. So, this gives rise to a need of a strong email spam detector which can filter maximum amount of spam emails with a greater accuracy so that a genuine email does not get filtered as spam .In recent times, spam has become a huge problem on the internet. The person sending the spam messages is referred to as the spammer. Such a person gathers email addresses from different websites, chatrooms, and viruses. Spam prevents the user from making full and good use of time, storage capacity and network bandwidth. The huge volume of spam mails flowing through the computer networks have destructive effects on the memory space of email servers, communication bandwidth, CPU power and user time .It is responsible for over 77% of the whole global email traffic .

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002021

II. RELATED WORK

Sr. No.	Method	Title	Description	Advantages	Disadvantages
1	Naive Bayes	Enhancing the Naive Bayes Spam Filter through Intelligent Text Modification Detection	Bayesian Classifier Naive Bayes Classifier Multinomial Naive Bayes Algorithm	Naïve Bayes classifier can get highest percentage spam messages to block.	The classifier is trained under the assumption that the spam letter(s) is/are constant whereas over time their pattern will change
2	K Nearest Neighbour (kNN)	Email filteration using K Nearest Neighbour	Learning by analogy and distance based algorithm	Less work on training data Simple and easy to learn	Computationally expensive. Requires efficient storage technique
3	Support Vector Machine (SVM)	Email Filteration using Support Vector Machine	Non-Linear mapping	Highly Efficient and accurate classifier. Less prone to overfitting	Complex algorithm -difficult to understand. Training time is more.
4	Decision Trees	Decision Trees Induction Algorithm	Top down, recursive. divide and conquer based.	Can handle high dimensional data. It is simple and fast and have good accuracy.	Branches may contain outliers in the training data sets.

III. METHODOLOGY

The methodology that is used for the filtering method is machine learning techniques that divide by three phases.The methodology is used for the process of e-mail spam filtering based on Na¨ive Bayes algorithm

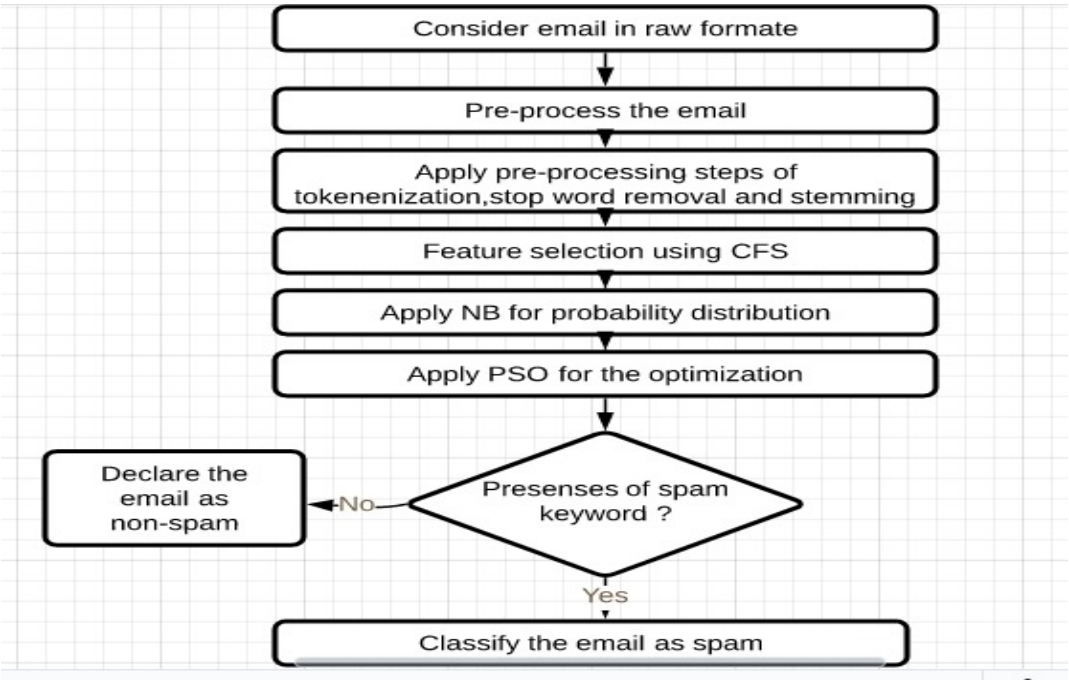


Fig. 1 Workflow of Methodology

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002021

Naïve bayes:
Naïve bayes are basically a family of simple "probabilistic classifiers" based on the bayes theorem with strong naïve independence assumptions between the features.These are used generally to achieven high accuracy levels.. Naive Bayes uses probabilistic models based on Bayes Theorem, which states that.

$$P(A|B) = P(B|A)P(A)/ P(B)$$

Where,
• P(A|B) is the posterior probability that document A is in a class given evidence B;
• P(B|A) is the conditional probability that evidence B is in document A;
• P(A) is the class prior probability that any document is in the certain class;
• P(B) is the predictor prior probability that evidence B is true.

Bayesian classifier:
For detection of spam emails, we apply Bayesian classifier .We have two classes : let S stand for spam and L stand for ham . P(A|B) is the a-priori probability, given a message B, what category A produced it, if a word is in B, :

$$P(A|B) = P(B|A)*P(A)/ P(|S*)P(S) + P(B|L)*P(L)$$

Where,
• P(A) is the overall probability that a message is spam;
• P(B|A) is the probability that the word appears in spam messages;
• P(L) is the probability that a given message is legitimate mail;
• P(B|L) is the probability that the word is in ham.

Multinomial naïve bayes:
We choose the Multinomial Naive optimization technique of the Naive Bayes classifier mainly for its multi-class predicting ability. Multinomial Naive Bayes is very accurate than the other optimization method. Multinomial Naive Bayes is the conditional probability of the evidence tk, or words, within a message d given a class of the message, spam or ham. It assumes that the message is a bag of tokens or words, such that the order of the tokens is irrelevant. Multinomial Naive Bayes essentially counts the relative occurrences of a particular token within the message to determine the conditional probability:

$$P(c|d) = P(c)\ 1 \leq k \geq nd\ P(tk|c)$$

Where,
• d is the document;
• c is the class, either spam or not spam;
• tk is the evidence, where t1 to tnd are tokens
.
Theoretically, the best class is determined by multiplying all the probabilities that each individual word is spam together as shown in the equation to get an overall probability, with probabilities closer to 1 being spam. However, there are instances where the spam word does not occur at all in a message

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002021

STEPS:

1. Consider a random email from the ling spam dataset for experimentation.

2. The considered email is in raw form. To perform the feature extraction/selection and classification procedure, initially email is needed to pre-process. Pre-processing involves the steps of tokenization, stemming and stop word removal.

2.1. Initially, tokenize the email into individual keywords. Tokenization split each individual word into different tokens.

2.2. Remove the stop words from the obtained tokens.

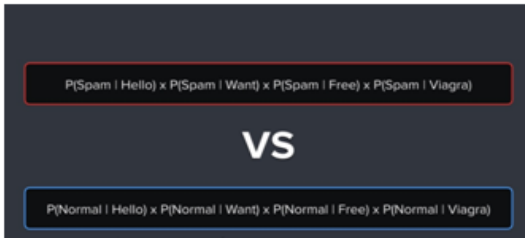
2.3. Perform stemming on the tokens obtained from the previous step. Stemming process reduces the size of the word to its root word. For stemming, a predefined list of possible words with their respective stem words is considered.

2.3.1. For stemming, a list of suffix words is stored into an array with their respective root words.

2.3.2. Consider check_token = availability in considered array of root word

2.3.3. If suffix of check_token = true, stem the word to its respective root word from the list of arrays.

2.3.4. Else, there is no need of stemming. Word is already in its root word format. Move to the next token



.Fig 2. Posterior Probability

3. Apply Correlation based feature selection approach to select the useful feature words from the pre-processed data. Correlation based feature selection method only selects the feature set which are most related to the particular class. If f is the feature set with k number of features and c is number of classes then CFS can be applied as mentioned in Equation 2.

Where, rcf is the average of feature-class correlation,rcf is the average of feature-class correlation, rff is the average of feature-feature correlation.

4. Calculate the probability distribution of the tokens along with selected features using NB approach. The formulation for the probability distribution is presented in Equation 3.

Where, f is any feature vector set (f1, f2, f3,….fn) and y are the class variables with m possible outcomes (y1, y2, y3,….yn). P(y|x) stands for posterior probability, P(x|y) is any particular class on which P(y|x) is dependent. P(x) is evidence depending on the known feature variables, P(y) is the prior probability.

5. Apply PSO approach to optimize the parameters of NB approach.

5.1. All the tokens are considered as the particles. Initially these particles randomly fly and search for the food sources in the form of the best feature match for tokens. Then search for the local and global solution.

5.2. The performance of each particle depends upon the similarity with the features that has to optimize.

5.3. Each particle flies over the n- dimensional outer search space and keep updating the following information: Xi – current position of particle Pi – the personal best position of particle x Vi – the current velocity of particle

5.4. The velocity updates in PSO can be calculated using the formula given below by Equation (4) Now, Vi is the new velocity. So, the position of the particle updates with the velocity as defined with Equation (5)

5.5. Update the positions for each particle and store the global best solutions.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002021

6. Based on the evaluated feature similarity using PSO, classification of tokens is declared as spam or non spam.

7. Further, final classification is performed by evaluating the probability of spam or non-spam tokens in a sentence.
   7.1. If the probability value of spam tokens is more, then email is considered as spam email.
   7.2. Else, email is considered as non-spam email.

8. Store the email as spam or non-spam and repeat the process for all the emails.

Fig 3. Spam Classification

IV. CONCLUSION

Through the Naive Bayes classifier, we were able to improve upon the baseline for spam filtering. Naive Bayes has a very fast processing speed and allows for a small training set, hence is suitable for real-time spam filtering. Previous research showed that most spam classifications failed when exposed to text modifications. By creating an addition to Naive Bayes, we were able to improve the multi-class prediction ability. We also found that our new addition helped improve ham classification due to the high recall and precision rates. We are able to classify mails as SPAM mails or genuine mails. Some mails might be considered into suspicious list. Na¨ive Bayes classifier also can get highest precision that give highest percentage spam message manage to block if the data_set collect from single-mail accounts. With so many people, enterprises and companies using mails as their daily part of their lifes , it is manually difficult to handle all possible mails as our projects deals with limited amount of corpus

REFERENCES

[1] “A bayesian approach to filtering junk e-mail,” https://robotics.stanford.edu/users/sahami/papers-dir/spam.pdf, Stanford University, accessed
[2] K. Netti and Y. Rahdika, “A novel method for minimizing loss of accuracy in naive bayes classifier,” in Proc. of IEEE International Conference on Computational Intelligence and Computing Research, 2015.
[3] “Text classification and nave bayes, the task of text classification,” https://web.stanford.edu/ jurafsky/slp3/slides/7 NB.pdf, Stanford University, 2011, accessed: 2017, 2018
[4] “Exim internet mailer,” https://www.exim.org, accessed: 2017, 2018.
[5] “Spamassassin: Welcome to spamassassin,” https://spamassassin.apache.org, accessed: 2017, 2018.
[6] csmining.org, “Ling-spam datasets - csmining group,” http://csmining.org/index.php/ling-spam-datasets.html, accessed: 2017, 2018
[7] S. Raschka, “Naive bayes and text classification i: introduction and theory,” https://arxiv.org/abs/1410.5329, Cornell University Library, 2014, 2015, 2017, accessed: 2017, 2018.
[8] N. NaveenKumar and A. Saritha, “Effective classfication of text,” in International Journal of Computer Trends and Technology, vol. 11, no. 1, 2014.
[9]Mathswork, Detector Performance Analysis Using ROC Curves–MATLAB&Simulink Example, Retrieved August 11, 2017 from, 2016,http://www.mathworks.com/help/phased/examples/detector-performance-analysis-using-roc-curves.html.
[10] G. He, Spam Detection, 1st ed. 2007.sharma, a. and jain, D. (2017). A survey on spam detection.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details