An Effective Spam and ham word Classification Using Naïve Bayes Classifier

Heena Tamboli*, Sambhaji Sarode**

Dept. of Computer Science and Engineering, MIT School of Engineering, MIT-ADT University, Pune Maharashtra, India
E-mail address: - *heenatamboli1991@gmail.com; **sambhajisarode@mituniversity.edu.in

DOI: 10.29322/IJSRP.11.07.2021.p11510
http://dx.doi.org/10.29322/IJSRP.11.07.2021.p11510

Abstract- : Email is a popular way for individuals on the Internet to communicate. The Number of emails received and sent has increased due to huge growth of internet users. Emails are used for many purposes like in business, colleges, marketing, for job searching, for chatting purpose, for internships and many more. It is used for personal or in business for sharing information. Spam emails are unwanted emails or junk emails. Spam emails are become a big problem on Internet and it may causes problem in computer security. So they are very risky for networks and computers. As the use increased of emails there need to secure computers as well as networks from spam emails. So we implement this model by using Naïve Bayes classification Algorithm for classify spam and ham emails. In which we filter emails into ham and spam emails and find its performance by calculating its accuracy.

Index Terms- Email Classification, Naïve Bayes Classification, Spam, Email, Spam Filter, Spam Detection, Filtering Web Application.

I. INTRODUCTION

Email is most popular way of electronic communication which sent messages from one system to another system. Emails uses are rapidly increased because now a day's emails are used for many purposes like in companies, marketing, searching for jobs, colleges , for communication etc. Spam emails are junk emails that we does not want. Spam emails are something sent by persons which we do not want. Spam emails are risky for computers and networks. And another drawback of spam emails are they wasting the time of users , if many unwanted emails are received , user may reads all emails then it was wastage of time and some time user forgot to read ham emails due to lots of spam emails. Sometimes spam emails contains virus if we open that email it will harm our system. So there is need to filter emails and protect our mailbox from spam emails.

Spammer is one who sent us spam messages. Spammer find our email addresses from different sites, chat rooms and from viruses. So if we receives lots of spam emails then it effects on our memory space, bandwidth, speed, time etc. It irritates user if lots of unwanted emails receives in email box. It waste lots of users time in reading spam emails and deleting spam emails from inbox. So it is better if we have spam filter. Spam filter is used to detect spam emails and protect our inbox from receiving spam emails. Some Spam filters arrange the incoming emails and remove spam emails and computer infections. Classification of emails/ messages are based on different criteria such as subject, content in emails, URL, address of sender etc.

We detect spam emails in this research by differentiating spam and ham terms from a dataset. We use Naive Bayes classifier algorithm for separating spam and ham word. . Which method is the most effective for classifying two objects in a dataset. We also find its accuracy. It gives 93% accuracy. This method is applicable for any two classifier for their classification. In this paper we see how to use binary classifier for particular text processing application.

II. LITERATURE REVIEW

M.Rubin Julis et al[1] proposed system to detection of spam in sms using machine learning. This system divided into some iterations. Every iteration has four phases: Inception, elaboration, construction and transition. In iteration it identify the idea of work. In elaboration

it design the architectural part. In construction implementation of code is done. And in transition validation of the developed part of system is done. They uses various algorithms such as logistic regression algorithm in that they use logistic function for calculate relation between the categorical dependant variable and independent variable. For training and testing data they uses Knearest neighbours algorithms. They also uses different classifiers that are Naïve Bayes Classifier , decision tree classifiers, support vector machines and compares all. It checks accuracy for all and support vector machines gives 98% accuracy which is good as compares to others.

In this paper authers Emmanuel Gbenga Dada et al[2] described the various spam detection filtering techniques that are commonly used to detect spam emails and overcome the problem of spam emails. The various techniques explained that are Case Base Spam Filtering Method, Content Based Filtering Technique, Heuristic or Rule Based Spam Filtering Technique, Previous Likeness Based Spam Filtering, Adaptive Spam Filtering Technique. Content Based Filtering Technique used to classify emails and for create automatic filtering rules. Case Base Spam Filter is used to train dataset and testing data of incoming emails for filter spam and non-spam emails. Heuristic or Rule Based Spam Filtering Technique is used to filter spam emails using already created rules and regular expressions. Previous Likeness Based Spam Filtering Technique uses k-nearest neighbor to filter spam emails. Adaptive Spam Filtering Technique detect spam and filter by grouping them into different classes. In this paper they discuss some open research problems related to spam filters.

Priyanka Sao et al[3] proposed email spam classification using Naïve Bayes classifier. They used the Lingspam dataset for classification of spam and non-spam emails .For extraction of features they used feature extraction techniques. Features are extracted for accurate result. When data is large in amount that time feature extraction technique is used. They also used word-count algorithm for extracting words. In this paper they conclude that Naive Bayes algorithm gives low error rate as compare to vector support machine so Naïve Bayes algorithm is more efficient for classify spam emails.

Aditya Gupta et al[4] proposed spam filter using Naive Bayesian Technique. For identify spam emails they used supervised learning method. It identifies spam and non spam emails after receiving massages. Spam filter is used to find unwanted emails and prevents messages from reaching users inbox. Different python libraries are used that are NLTK, WordCloud, Panda, Matplotlib for filtering emails and finds frequently used keywords. For processing NLTK libraries used ,for visualization WordCloud and Matplotlib used and for loading data they used Pandas. They used dataset from Kaggle contains 5572 test cases of ham and spam messages. Data is split into trained and test dataset for testing the model.

In this paper authors Linda Huang et al[5] proposed Naive Bayes Spam Filter through intelligent text modification detection. For increase the accuracy of Naive Bayes classifier they implemented Novel algorithm so it can detect spam emails more correctly and classify it into spam and ham more effectively. They also discover the correlation between spam score and length of email. They code new additional server for improving accuracy of Spam Server Spamassassin. Because of improving ham classification it gives high precision rate and high recall.

Satyam Sagar et al[6] proposed spam classification filter using Naïve Bayes classifier which is developed as a web application for classification of emails into spam and ham. They use Python's Micro Flask Framework for developing web application in which input is new incoming emails and it predict output as spam and non spam emails. Their system contains main two parts first one is train the classifier and another is deploy the model. In train classifier it contains dataset of spam and ham emails and it generate classification model. In second part it deploy the model on server.

In this paper authors Sebastian Romy Gomes et al[7] focuses on comparatively study on classification of Hidden Markov Model and Naïve Bayes classifier. In this paper they studied both classifier techniques to investigate that which technique is best. For this they collect databases and check accuracy, recall, f-metrics, precision for both techniques and also check for different combinations of lemmatizing, removal of stopwords ,stemming techniques to see which combination gives good result. HMM technique gives best accuracy that's why they use HMM classifier for classifying database. Their structure of research is made in such a way that it will work to distribute and classify massages in any number of categories.

Nurul Fitriah Rusland et al[8] studied on Naïve Bayes algorithm for filtering spam emails in different datasets.To measure the performance of Naïve Bayes algorithm they used two datasets which are SPAMDATA and SPAMBASE and measured its performance. Performance measured is based on their recall, accuracy,F-measure and precision. For filtering of spam emails they used WEKA tool. SPAMBASE dataset contains data from single email account and SPAMDATA contains data from multiple email accounts. After analysis of Naives Bayes algorithm on both datasets it is observed that Naives Bayes algorithms performance is good when data comes from single email accounts as compare to data comes from different email accounts.So Naïve Bayes algorithm gives best performance when SPAMBASE dataset used.

The author Nikhil V Mathew et al[9]] proposed a system to study efficiency of N-gram technique in Naïve Bayes spam classification. They study on Naïve Bayes algorithm on multiple datasets for filtering of emails. In this paper by using N-gram technique they make feature set which is used as training set in Naïve Bayes classifier and checks its efficiency for various range of grams. They check

efficiency for three n-grams techniques: 3-grams, 4-grams, 5-grams and analysis that 4-grams feature set gives more efficiency than others because it has more accuracy and low false positive rate.

Wanqing You et al[10] developed web service enabled spam filtering with Naïve Bayes classification. In this they develop spam filter in which we gives email as input and it finds how much that email is spam. This system is developed by Resteasy technology has 3 phases for training emails. They use Naïve Bayes theorm to find spamacity of emails. They developed web service as filter in which user upload emails and this filter predicted spamacity of that emails. For process on large amount of email data they integrated Hadoop Map. In [11-20], data aggregation approaches are discussed to improvise the overall performance of the system. These techniques give the insight to enhance the data preprocessing approaches.

III. PROPOSED WORK

Firure1 shows the system architecture of our proposed work.

A. Dataset

Dataset contains spam and ham emails. Enron dataset is used for spam and ham classification.

B. Pre-processing

In this we remove all punctuation marks. Then we do word tokenization for every word. Tokenization is a preprocessing type in which large text split into small words.

C. Feature selection and extraction

We make wordlist of every tokenization words and take count of that words.

D. Training

Wordlist assemble into x and y pattern. In x contains actual word list and in y contains its labels. Prediction of every word is done that is the word is spam or ham. Gaussian noise removed by smoothing. First we train data for first 100 list data and take true results from them and then next 100 list data goes for prediction.

E. Testing

We compare predicted data and ground truth data. Ground truth value is 0 or 1. 0 for ham and 1 for spam. Predicted result is also 0 or 1.If input value is 1 and predicted result is also 1then the result value is 1 means that word is spam word. If input value is 1 and predicted result 0 then the word is ham word.

F. Classification

We classify data into spam and ham data .And take count of that data.

G. Performance

We measure the accuracy by following formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{FN}}{(\text{TP} + \text{FN} + \text{TN} + \text{FP})}$$

Where
TP- True Positive (input spam and detection spam).
TN- True Negative (input spam and detection ham).
FN- False Negative (input ham and detection ham).
FP- False Positive (input ham and detection spam).

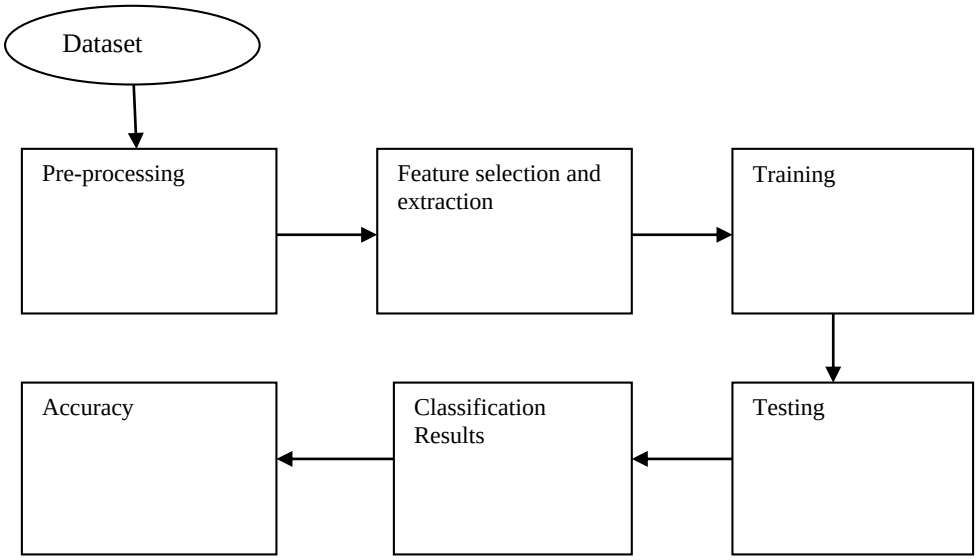


Figure1:System architecture of proposed work

IV. METHODOLOGY

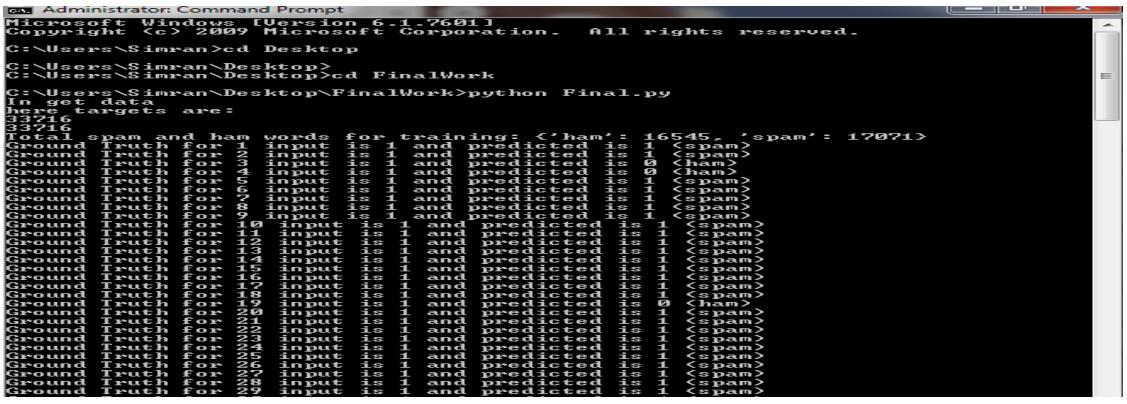
We use Naïve Bayes classifier method for filtering emails. This machine learning methodology helps us to classify spam and ham emails from given dataset.

Naive Bayes classifier:

The Naive Bayes algorithm is machine learning algorithm reached from Bayes theorm used for filtering and classifying data. Naïve Bayes theorm gives quick and fast prediction. Naïve Bayes theorm used in text classification. It predicts on the basis of objects probability.
This is based on occurance of words in emails or datasets which is called bag of words. It measures probabilities by computing the frequency and combination of values in given datasets. Naïve Bayes classifier is used to classification of any two objects in dataset.

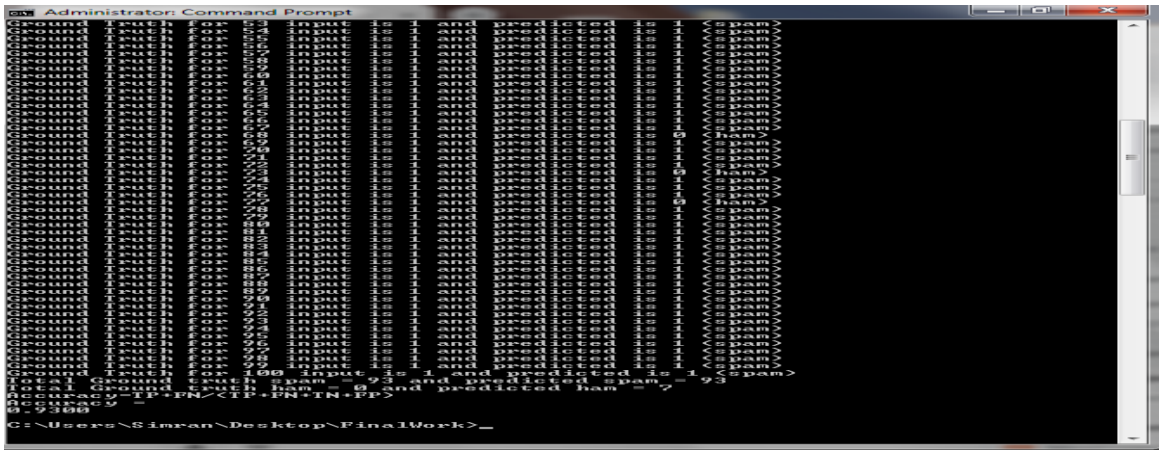
$$P\left(\frac{x}{y}\right) = \frac{P\left(\frac{y}{x}\right).P(x)}{P(y)}$$

Where x and y are two cases and $P(y) \neq 0$.
P(x) and P(y) are probabilities of x and y which are not dependant on each other.
P(x|y): is the conditional probability of case x when y is true.
P(y|x): is the conditional probability of case y when x is true.



It shows result total ground truth spam 93 and predicted spam 93. Total ground truth ham 0 and predicted 7.

Then calculates its accuracy. It gives 93% accuracy.



V. RESULT AND DISCUSSION

In this project we discuss about classification of emails to identify the spam and ham mails. For this purpose we are using Naïve Bayesian Classifier. Which is best technique to classify two objects in datasets.

In this first it count of spam and ham words in dataset and print as target. In our dataset target are 33716. Then it shows separate count of spam and ham words for training that is in our data set spam words are 16545 and ham words are 17071.Then it find Ground Truth value for each word that is that word is ham or spam.

VI. CONCLUSION

Now a days Spam is a huge problem in world. The spam messages are the messages which user don't need but that are receiving daily. Spam emails are message of anything it may any advertisement or may be any URL , or any king of virus. Naive Bayes classfier is very good technique for filtering spam emails. It is used to classify any two objects in datasets. Performance of Naïve Bayes algorithm is based on datasets that used. We can use this method on any datasets for classification of any two objects. We filter email spam from dataset and calculate accuracy ,it gives 93% accuracy.

REFERENCES

[1] M.Rubin Julis, S.Alagesan. (2020). Spam Detection In Sms Using Machine Learning Through Text Mining. *International Journal Of Scintific & Technology Research Volume9, Issue 02, February* 2020 ISSN 2277-8616

[2] Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), e01802.

[3] Sao, P., & Prashanthi, K. (2015). E-mail Spam Classification Using Naïve Bayesian Classifier. *International Journal of Advanced Research in Computer Engineering &* Technology (IJARCET), 4(6), 2792-2797.

[4] Aditya Gupta, Khatri Mrunal Mohan, Sushila Shidnal.(2018,Jane) Spam Filter using Naïve Bayesian Technique. ISSN (e): 2250 – 3005 , Volume, 08 ,Issue, 6,Jun – 2018. International Journal of Computational Engineering Research (IJCER)

[5] Linda Huang, Julia Jia, Emma Ingram, Wuxu Peng(2018) Enhancing the Naive Bayes Spam Filter through Intelligent Text Modification Detection. *2018 17th* IEEE *International Conference On Trust, Security And Privacy In Computing And Communications/ 12th*

[6] Satyam Sagar, Piyush Kumar Shukla, Raju Baraskar.(2019,October) An Effective Spam Classification Filter As A Web Application Using Naïve Bayes Classifier. *International Journal Of Scientific & Technology Research Volume 8, ISSUE 10, OCTOBER 2019* ISSN *2277-8616*

[7] E Gomes, S. R., Saroar, S. G., Mosfaiul, M., Telot, A., Khan, B. N., Chakrabarty, A., & Mostakim, M. (2017, September). A comparative approach to email classification using Naive Bayes classifier and hidden Markov model. In 2017 4th International Conference on Advances in Electrical Engineering (ICAEE) (pp. 482-487). IEEE.

[8] Rusland, N. F., Wahid, N., Kasim, S., & Hafit, H. (2017, August). Analysis of Naïve Bayes algorithm for email spam filtering across multiple datasets. In *Proceedings of the IOP Conference Series: Materials Science and Engineering*.

[9] Mathew, N. V., & Bai, V. R. (2016, October). Analyzing the effectiveness of N-gram technique based Feature Set in a Naive Bayesian Spam Filter. In 2016 International Conference on Emerging Technological Trends (ICETT) (pp. 1-5). IEEE.

[10] Wanqing You, Kai Qian, Dan Lo, Prabir Bhattacharya, Minzhe Guo(2015). Web Service-enabled Spam Filtering with NaïvBayes Classification. 2015 IEEE *First International Conference on Big Data Computing Service and Applications*

[11] Sarode, Sambhaji, and Jagdish Bakal. "PFPS: Priority-first packet scheduler for IEEE 802.15. 4 heterogeneous wireless sensor networks." International Journal of Communication Networks and Information Security 9, no. 2 (2017): 253.

[12] Bang, Raghav, Manish Patel, Vasu Garg, Vishal Kasa, Jyoti Malhotra, and Sambhaji Sarode. "Redefining smartness in township with Internet of Things & Artificial Intelligence: Dholera city." In E3S Web of Conferences, vol. 170, p. 06001. EDP Sciences, 2020.

[13] Rathod, Akshada, Prachi Ayare, Ramchandra Bobhate, Rajneeshkaur Sachdeo, Sambhaji Sarode, and Jyoti Malhotra. "IoE-enabled smart embedded system: An innovative way of learning." In Information and Communication Technology for Sustainable Development, pp. 659-668. Springer, Singapore, 2020.

[14] Sarode, Prachi, and R. Nandhini. "Intelligent query-based data aggregation model and optimized query ordering for efficient wireless sensor network." Wireless Personal Communications 100, no. 4 (2018): 1405-1425.

[15] Sarode, Prachi, T. R. Reshmi, and Venkatasubbu Pattabiraman. "Combination of Fitness-Mated Lion Algorithm with Neural Network for Optimal Query Ordering Data Aggregation Model in WSN." Wireless Personal Communications 116, no. 1 (2021): 513-538.

[16] Malhotra, Jyoti, Prachi Sarode, and Aparna Kamble. "A review of various techniques and approaches of data deduplication." International Journal of Engineering Practices 1, no. 1 (2012): 1-8.

[17] Sarode, Sambhaji, Jagdish Bakal, and L. G. Malik. "Reliable and Prioritized Data Transmission Protocol for Wireless Sensor Networks." In Proceedings of the International Congress on Information and Communication Technology, pp. 535-544. Springer, Singapore, 2016.

[18] Sarode, Sambhaji S., and Jagdish W. Bakal. "A data transmission protocol for wireless sensor networks: A priority approach." Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 10, no. 3 (2018): 65-73.

[19] Sarode, Sambhaji, and Jagdish Bakal. "Performance analysis of beacon enabled prioritized CSMA/CA for IEEE sensor networks." International Journal of Applied Engineering Research 12, no. 8 (2017): 1622-1627.

[20] Sarode, Prachi, and R. Nandhini. "APDA: Adaptive pruning & data aggregation algorithms for query based wireless sensor networks." In 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), pp. 219-224. IEEE, 2016.

[21] Deshpande, Vivek, Prachi Sarode, and Sambhaji Sarode. "EDCAM-early detection congestion avoidance mechanism." *International Journal of Computer Application* 7, no. 2: 11-14.

AUTHORS

First Author – Heena Tamboli, pursuing M.Tech degree in Computer Science Engineering, will be Post- Graduation in the year 2021, from MIT-ADT University in Pune., heenatamboli1991@gmail.com

Second Author – Sarode received BE and ME in information Technology from the SPPU university Pune in 2005 and 2010, respectively and Ph.D. in Computer Science & Technology from RTM Nagpur University India in 2019 under the supervision of Prof. Dr. J. W. Bakal. Currently, he is working as an associate professor with the Department of Computer Science & Engineering, School of Engineering, MIT ADT University Pune. sambhajisarode@mituniversity.edu.in