Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 11077 – 11082

Research Article

Junk Mail Analysis And Blocking

Jauffar Sadiq A.R[1] , Jai Sharma[2]

Abstract

Email is one of the important tools used to send and receive the information. Through email the internet users can send and receive images, text files, audio and video images. Comparing other communication medium email is the simple and easiest way with low cost. Spam is the one type of the email message. Spam is the unwanted message. From 2004 onwards this unwanted messages are spread through internet. The advertisers send about their product and services through internet to multiple groups of peoples. The social problems are also sending to this type of spam mails. Most of the internet users are affected by this type of spam messages. Sometimes this spam messages contains malwares. If the user opens this malware spam message, the device entirely crashed. Hackers also try to collect the confidential information from the users through spam mail. To avoid such kind of problem various filter concepts are used. This article proposed a Naïve Bayes concept to filter the unwanted messages. Compared with other classifiers Naïve Bayes approach provides better result.

Keywords- Spam Detection, Naïve Bayes, Classifier, Accuracy, SVM classifier.

I. INTRODUCTION

In this modern world, to share the thoughts and some official information people prefers to use mails. They use emails mostly for their official purposes. E mails, electronic mail are an method of exchanging messages between people using electronic device with low cost. Now a days the users inbox are mostly filled with unwanted message like spam message which users does not like. The increased popularity of emails spam in both text and messages and images require a real time protection mechanism for the media flow. Advertisement by lazy advertising peoples is sent to the e mail users which are called spam messages. Spam messages totally crash the hard disk of the devices. Spam messages prevent email from well known content even if the spam message sometimes without their knowledge they also delete the needful message. Some types of spam messages are phishing spam and lottery spasms. Phishing spam is the message which has fake message which conveys that the users have won some lakhs and cores of money which is fake. And they usually ask the users to register their account number. Thos spam message are send to hack the user's personal information.

II LITERATURE SURVEY

P. U. Anitha et al., says that Email spam is one of the electronic type spams. Nowadays this is one of the important issues of internet users. Mostly spam mails are sent for business purpose. Sometimes malware type of mails is also send through the mail. The main aim of this research article is to classify the spam and ham mails using data mining classifier concept. Ham mail means it is a valid email and it also contains legally accepted messages. Spam mail consists of only unwanted messages. The users are also not interested to open spam type mails. In this study the authors classify the mails using modified Naïve Bayes data mining classifier concept. This proposed method generates better result in terms of accuracy compared with SVM concept and Naïve Bayes technique [1].

Nurul Fitriah Rusland et al., explained about spam mails. It is one of the major issues on internet. These types of mails send to the various people at the same time. The main purpose of the spam mails are send the advertisement message or unwanted messages to the various peoples at the same time. Various filtering concepts are used to filter the unwanted messages. In this paper the authors used Na¨ıve Bayes technique for email filtering. Two datasets are used for this filtering purpose. This algorithm is implemented by using WEKA tool. The datasets performance value can be measured using the terms of recall, accuracy and precision value. The experiment result shows that the Na¨ıve Bayes classifier provides better result compared with other type of data mining classifier. [2]

Priyanka Sao et al., explained that email spam message is the major problem of every individual people. Spam mail consists of advertisement messages or various types of malwares are received the client inbox without any kind of information. Spam filtering concepts are used to save the mailbox from unwanted mails. Na¨ıve Bayes concept is very simple and very effective classifier for spam mail classification. This research works the authors using Lingspam dataset to classify the mails. After receiving the mails the feature extraction technique is applied to remove unwanted features. The result shows that Na¨ıve Bayes classifier generates better result in terms of accuracy compared with other classification techniques [3].

G.Vijayasekaran et al., discussed about the nature of unwanted messages in client mailbox. Spam mail is the important issue for each and every individual people. Spam mail contains unwanted messages and viruses are transfer through internet to the various users at a same time. To overcome this problem various filtering methods are used. Here Na¨ıve Bayes concept is used for spam mail classification. Compare to other classification methods this Na¨ıve Bayes concept is very simple and provides better result. In this research work real time data set can be used for filtering spam mails. Features are extracted by using the bucked base concept [4].

Enaitz Ezpeleta et al., in this current internet world short message are sent instantly. Sometimes internet users threaten by illegal message. This type of advertisement message or illegal contents on the email is called as spam mail. It spoils the user's privacy. Here the authors filtered unwanted spam messages using the concept of sentiment analysis. In this method compute the division value of each message and consolidate all the values. Using this calculated values create a new data set [5].

Muhammad Hassan Arif · et al., explained about the two challenges of sending small messages to others. The first challenge is analysis the sentiments of public and political people. The second important challenge is detect the spam messages from the social media network. This article identifies the performance of the LCS (Learning Classifier Systems) using machine learning concept. This concept is

used to analyze social media messages and cinema reviews, and mail, spam datasets. In this study paper existing LCS concept us extended by using encoding method to handle the vector values. The output of this proposed system shows the better result compared with the existing LCS techniques [6].

Bin Ning says that Naïve Bayes is easiest and simple method to classify the spam email. In this model is constructed by using multi-classification technique and multi-two-classification concept via preprocessing and feature extraction method. The output of the Naïve Bayes performance compared with random forest technique and SVM approach. Naïve Bayes aconcept provides the better performance compared with other classifiers based upon multi-two-classification [7].

Johan Hovold et al., says that naive Bayes classifier is the one of the base email classifier. Various machine learning algorithms are used to filter the unwanted emails. But Naïve Bayes concept provides better result [8].

S.Jancy Sickory Daisy et al., discussed about the importance of the email messages. The major benefits of the email system are easy to send the data to multiple users with fewer amounts. Apart from this benefits email contains certain disadvantages also. Spam mail or junk mails contains unwanted messages. The lazy advertisers send the details about their products and services to the various groups of the people within the second. Sometimes the viruses are also sending through the mail. Here the authors proposed a new hybrid spam filter technique using Naïve Bayes and technique of Markov Random Field. These methods are used to detect the spam messages accurately. This method performance is evaluated in terms of consuming of time and it accuracy level [9].

Ishtiaq Ahmed et al., proposed a new a hybrid model for SMS spam filter. This hybrid model is constructed based upon Naïve Bayes classifier technique and Apriori data mining algorithm. This above mentioned model is executed based on fully logic. Naïve Bayes is the one of the easiest and important classification concept used to filter spam messages in email. Here UCI Data Repository is used for test the performance of the proposed work. This work generates the accuracy level 98.7%. But the tradition method provided 97.4% accuracy level only [10].

III PROPOSED METHOD

Spam is the unwanted email messages. Through spam messages the hackers are try to get the personal information from the internet users. Most of the advertisers also send the particulars about the products through spam messages to various peoples at the spam time. Illegal messages are also send to the using with the help of spam mails. Most of the users waste their time for deleting spam mails. Various filtering and data mining concepts are used to filter unwanted messages. The one of the way to detect the spam mail by using CC mail address. The important classification methods are SVM, Naïve Bayes, random forest etc. In this research article Naïve Bayes classifier concept is used to filter unwanted messages. Naïve Bayes classifier is the simplest and easiest method to compare with other classifiers.

With the help of Naïve Bayes approach each mail are tested by using important attributes. Based on NB concept , mails are divided into separate words 1 2 , ,..., w w wn and its features are mentioned as F . The probability value of emails is equal to the probability value of receiving the list of separated words.

$$P(F) = P(w_1, w_2, ..., w_n)$$ ----------(1)

From the above mentioned equation (1) Naive Bayes assumption becomes as follows

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i)$$ ----------(2)

Naïve Bayes concept is the important method is statistical analysis used for spam filtering. Naïve Bayes concept calculate the probability value of spam or not spam mail.

To compute the probability value of the email is spam or non-spam by using Na¨ıve Bayes Bayes theorem as mentioned below

$$P(spam \mid word) = \frac{\text{P(spam)}.\text{P(word|spam)}}{\text{P(spam)}.\text{P(word|spam)} + \text{P(non}-\text{spam)}.\text{P(word|non}-\text{spam)}}$$ ----------(3)

P(spamword) represents the probability of word in e-mail spam

P(spam) represents the probability value of the message in the spam mail

P(wordspam) denotes the probability value of the specific word available spam content

P(non − spam) means the probability word is not spam.

(v) P(wordnon − spam) means probability value of specific word in non-spam content. The following diagram 1 shows the block diagram of proposed system.

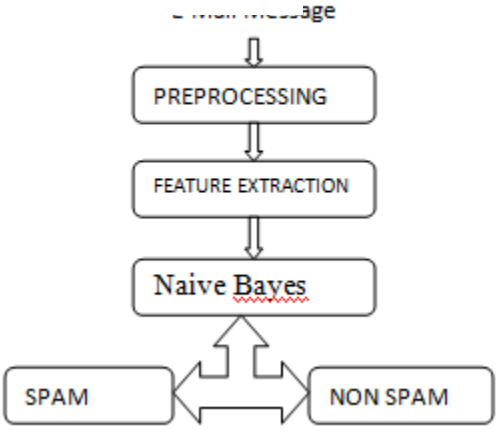

Figure 1 Block Diagram of Proposed System

To attain the goal of this research the task is conducted in three stages.

In Pre-processing stage the noisy data is removed from the original data. Feature extraction stage is used to extract the specified attributes from the preprocessed data. Finally in stage three the Na¨ıve Bayes Classifier concept is applied to classify the data.

IV RESULTS AND DISCUSSIONS

The unwanted emails are called as spam mail. The various advertisers are used the spam mails for marketing their products. The main benefit of the mail communication is easily sending the large messages to the various group of peoples within a second. Various machine learning algorithms are used to filter the unwanted mails. This proposed system is constricted by using Na¨ıve Bayes classifier concept. The performance of the classifier is measured in the following terms.

Evaluation Measure	Evaluation Function
Accuracy	$Acc = \frac{TN+TP}{TP+FN+FP+TN}$
Recall	$r = \frac{TP}{TP+FN}$
Precision	$P = \frac{TP}{TP+FP}$
F-measure	$F = \frac{2pr}{p+r}$

Accuracy means the percentage of properly detected spam and not spam content.
Recall represents the percentage of spam content to be block
Precision means correct content in spam mail
F-measure is the weighted average value of the of precision value and recall value
FP- number of miscalculate non spam mail content
FN- number of miscalculate spam emails
TP- spam message numbers are correctly categorized as spam message
TN: non-spam e-mail numbers correctly state as non-spam mail

V CONCLUSION

E-mail is one of the important communication medium. Using emails the internet users easily communicate with peoples with minimum amounts. Spam represented as the unwanted messages filled on the user's mailbox. Based upon the various survey recently most of the online users inbox consists large amount spam mails. To overcome this issue various filters are used. Data mining classifier concepts are used to classify spam and non spam messages. Compared with other data mining concepts Na¨ıve Bayes generates better result in terms of accuracy value. The entire process consists of three stages like pre processing, feature selection and classifier.

REFERENCES

[1] P. U. Anitha, C. V. Guru Rao, & Suresh Babu(2017), "Email Spam Classification using Neighbor Probability based Naïve Bayes Algorithm", 7th International Conference on Communication Systems and Network Technologies, pp. 350-355.

[2] Nurul Fitriah Rusland, Norfaradilla Wahid, Shahreen Kasim & Hanayanti Hafit(2017), "Analysis of Na¨ıve Bayes Algorithm for Email Spam Filtering across Multiple Datasets", International Research and Innovation Summit (IRIS2017), pp. 1-9.

[3] Priyanka Sao & Kare Prashanthi(2015), "E-mail Spam Classification Using Naïve Bayesian Classifier", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) , Vol. 4, No. 6, ISSN: 2278 – 1323, pp. 2792-2796.

[4] G.Vijayasekaran & S.Rosi(2018)," Spam And Email Detection In Big Data Platform Using Naives Bayesian Classifier ", International Journal of Computer Science and Mobile Computing, ISSN 2320–088X, Vol. 7, Issue. 4, pp.53 – 58.

[5] Ezpeleta , Urko Zurutuza & Jos´e Mar´ıa G´omez Hidalgo(2016), "Short Messages Spam Filtering Using Sentiment Analysis".

[6] Muhammad Hassan Arif · Jianxin Li · Muhammad Iqbal & Kaixu Liu(2018), "Sentiment Analysis and Spam Detection in Short Informal Text using Learning Classifier Systems", pp. 1-12.

[7] Bin Ning, Wu Junwei, Hu Feng(2019), " Spam Message Classification Based on the Naïve Bayes Classification Algorithm", IAENG International Journal of Computer Science, 46:1, IJCS_46_1_05.

[8] Johan Hovold (2006), "Naive Bayes spam filtering using word-position-based attributes and length-sensitive classification thresholds" Proceedings of the 15th NODALIDA conference, pp. 78–87.

[9] S.Jancy Sickory Daisy & A.Rijuvana Begum(2019), "Hybrid Spam Filtration Method using Machine Learning Techniques ", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075,Vol.8, No. 9,pp. 1818-1821.

[10] Ishtiaq Ahmed, Donghai Guan & Tae Choong Chung(2014)," 2014SMS Classification Based on Naïve Bayes Classifier and Apriori Algorithm Frequent Itemset ", International Journal of Machine Learning and Computing, Vol. 4, No. 2, pp.183-187.