International Journal of Advanced Trends in Computer Applications
www.ijatca.com

Email Spam Detection using Extended KNN algorithm

[1]Ritu Saini, [2]Er. Geetanjli Chawla
[1]Research Scholar
Department of Computer Science Engineering
Galaxy Global Group of Institutions, India
[2]Assistant Professor
Department of Computer Science Engineering
Galaxy Global Group of Institutions, India
[1]ritusaini71@gmail.com, [2]geetanjli@galaxyglobaledu.com

Abstract: *E-mail is the cheaper and fast way of communication. E-mail is used in both personal and professional levels of life. Various types of e-mail are lies on social websites. The spam is one of them. Spam is the undesired messages on the internet site which is nothing but wastes the time and resources. Spam messages are sent by the spammer for marketing, promotion, spreading the virus. Various detection and filtering approaches are used to manage the spam. One is the most useful and simple approach is KNN algorithm which is content based approach. In this paper, the authors are trying to improve KNN algorithm which can be later used for better Spam Email Detection.*

Keywords: *Email, Spam, KNN, Websites.*

I. INTRODUCTION

E-mail is the most essential part or application of the internet. Without e-mail, internet is nothing. E-mail is the written communication and convenient way of communication. E-mail is the way to exchange the information of lifestyle and work. E-mail messages are sent to a single user and broadcasted to group as well. E-mail are used in every field like academic institutions, medical, finance, marketing, banking etc. Most useful mailing servers of e-mail is g-mail and yahoo. E-mails has many types like compose mails, sent mails, outbox mails, trash mails, spam mails, important mails etc. From all of these, the most challenging mail is spam mail. User may receives the hundreds spam mails every day with new content. Spam in emails are the most difficult and vast problem in the emails. Spam e-mails are the unwanted, unsolicited emails that are not intended for specific receiver that are sent for the marketing purposes, scam, hoaxes etc. The purpose of email spam is advertising, promotion, and spreading backdoors or malicious programs. Currently, Phishing is also considered as one of the main goals of spammers when employing email spams. Spam mails are used for spreading virus or malicious code, for fraud in banking. So it can cause serious problem for internet users such as loading traffic on the network, wasting searching time of user and energy of the user, and wastage of network bandwidth etc. Spam e-mails do not waste the resources but also pose serious security problems like stealing the personal information of user, spammer may sent fraudulent notifications. Today user receives the more spam mails than non-spam mails.

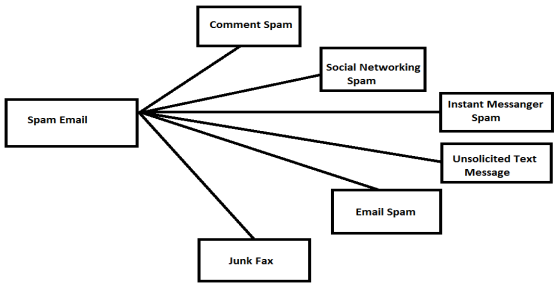

Fig. 1: Spam Emails

According to "Elements of Computer Security 2010", 97% e-mails were classified as spam. The cost to managing the spam e-mail is high rather than the cost of sending spam e-mails. To fighting with spam e-mail various filtering and detecting techniques are used. Spam filtering and detection is the process of detecting unsolicited and unwanted email and prevents those messages from getting to a user's inbox. In general, email classification techniques can be classified

into two categories: rule-based approaches and content-based approaches. In the content based approach the main focus is on classifying the email as spam or as ham, based on the data that is present in the body or the content of the mail. Content base approaches are Chi-Square, Adaboost classifier, Bayesian Classifiers, KNN classifier. In the rule based approach applied on the header part of the email of the mail. Whitelist and blacklist are the main approaches of rule based classifier for fast filtering without considering the content.

II. LITERATURE SURVEY

Wenjuan Li et al [1] in their paper many algorithms have been applied to email classification, while they identify those larger studies should be conducted to explore the practical performance of SML in different environments. Authors perform an empirical study with three different environments and over 1,000 participants regarding this issue. It is found that decision tree and SVMs are acceptable by most users in real environments and that environmental factors would greatly affect the performance of SML classifiers.

Michal Prilepok et al. [2] performed two settings of the LCBoD algorithm. The first setup was without community closure. The second setup was with community closure. For the spam detection filter, they used all feature vectors from selected communities. In the other experiment, they computed the average vector for every selected community. They used only communities that consisted of spam or non-spam emails.

Sunil B. Rathod et al. [3] have emphasized Bayesian approach for classifying Spam and legitimate mails using supervised learning across features extracted. Applying the Bayesian classifier, they experimentally demonstrated that spam mails can be detected with an accuracy of more than 96.46% with respect to real world gmail data sets.

Wanli Ma et al. [4] in their study the feasibility of negative selection in detecting spam emails without using any prior knowledge of any spam emails. They use TREC07 corpus for experiments. The outcomes, under the assumption of no prior knowledge about spam emails, are very encouraging.

Xiao Li et al.[5] considered the requirements of improving the efficiency in processing e-mail data for a specific purpose, this paper Bayesian classification approach is used to filter e-mails based on feature analysis combined with text classification. It fully utilizes and improves the basal technology of traditional spam filtering, describes the e-mail filtering outline, and tests in English data set. Experimental results show that this method is reasonable and effective.

Xianchao Zhang et al. [6] studied promoting and spam campaigns in Twitter and proposed a detection framework and proposed a URL-driven estimation method to measure the similarity between two accounts and linked the accounts with high similarity then they integrated a graph-based approach to extract dense sub graphs as candidate campaigns. Finally they proposed multiple features to distinguish promoting and spam campaigns from the legitimate ones based on SVM algorithm.

Soghra M.Gargari et al. [7] in their paper they introduce a novel framework for spam detection task in social bookmarking systems. They propose a set of new features to improve the accuracy of spammer detection. Experiments show that features demonstrate a high discriminative power. A performance evaluation of proposed method over different spammer detection methods indicate that the proposed framework yields an improvement of the prediction accuracy.

Hesham Altwaijry et al. [8] developed an intrusion detection system using Bayesian probability. The system developed is a naive Bayesian classifier that is used to identify possible intrusions. The system is trained a priori using a subset of the KDD dataset. The trained classifier is then tested using a larger subset of KDD dataset. The Bayesian classifier was able to detect intrusion with a superior detection rate.

Izzat Alsmadi et al. [9] in their paper presented a large set of personal emails is used for the purpose of folder and subject classifications. Algorithms are developed to perform clustering and classification for this large text collection. Classification based on NGram is shown to be the best for such large text collection especially as text is Bi-language (i.e. with English and Arabic content).

Sarwat Nizamani et al. [10] presented fraudulent email detection method and various classification algorithms including SVM, NB, J48 and CMM.
They achieved the accuracy of fraudulent email detection as high as 96%. The research study also concludes that for the fraudulent email detection task, choice of efficient features affects the accuracy of the task.

Loredana Firte et al. [11] In their paper proposed a new approach for a spam detection filter. Messages are classified with the kNN algorithm based on a set of features extracted from the email's properties and content.

Nitin Jindal et al.[12] In their paper studied review spam and spam detection and they take the data from the Amazon.com. Three main types of spam were identified. Results showed that the logistic regression model is highly effective.

Snehal Dixit et al. [13]. In their work surveyed existing techniques and algorithms created for Review centric and Reviewer spam detection. To draw a general picture of the review spam detection, first provide proposed work in each paper and also presented a brief overview of evaluation method used to determine accuracy. They also provided a comparative study about different spam detection techniques depending upon their accuracy.

Ommera Jan et al. [14] proposed that filtered mails are further filtered to measure the misclassification using different data mining techniques. The results show that the decision tree is the best classifier. It is easy to interpret and explain the executives. In comparison to random forests are time efficient. Decision tree requires relatively less effort from users for data preparation.

Carmona-Cejudo et al. [15] paper is related to real time email classification and introduced GNUsmail open source for email folder classification. The application is developed to parse emails from different email clients and perform some data mining analysis using WIKA data mining tool. In email folder classification is also based on the time of email messages.

Bekkerman et al.[16] The paper used Enron and SRI email datasets for the case study. Some new classification methods such as: MaxEnt were evaluated in the paper. The major decision to make in all email classification papers is what features to select. Features can be related to email title, from or to addresses or can be related to the content; words, sequence of words, etc. Natural language processing activities such as parsing and stemming are then involved to parse email contents and eliminate any words that may not be relevant for the classification process.

III. KNN ALGORITHM

KNN is the K Nearest Neighbor classification algorithm. It is simple and most useful method for classification. It is non-parametric method used for classification and regression. KNN algorithm has two stages one is training and other is filtering.

Stage 1 Training- In training stage the training messages are stored.

Stage 2 Filtering- In filtering stage for a given message x, its k nearest neighbours among the messages in the training set are determined. If there are more spam's among these neighbours, then classify given message as spam. Otherwise classify it as ham.

Steps

1. Determine parameter K= number of nearest neighbours.
2. Calculate the distance between the query-instance and all the training samples.
3. Sort the distance and determine nearest neighbours based on the K-th minimum distance
4. Gather the category Y of the nearest neighbours.
5. Use simple majority of the category of nearest neighbours as the prediction value of the query instance.

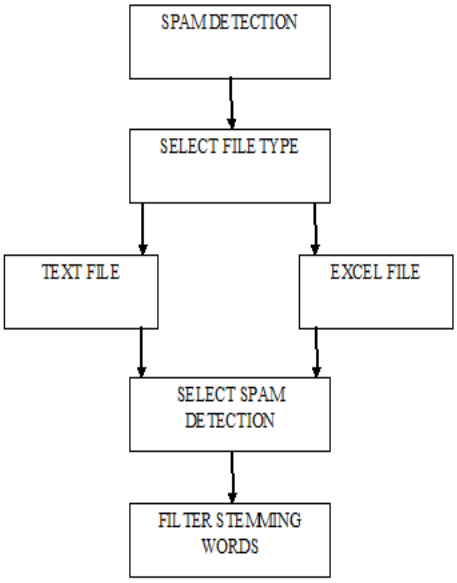

Fig. 2: Spam Detection System Architecture

IV. PROBLEM FORMULATION

In this work, the authors are trying to extend the KNN algorithm in terms of parameters like accuracy, Fmeasure, Precision, Gmean and Recall which can be later collectively used to provide better results in terms of Email Spam Detection. In KNN algorithm, whenever training phase is performed, it recalculates its internal centers and also recalculates its values. So KNN is able to perform recalculations and able to calculate closest match. But in this proposed work, the authors are trying to improve KNN algorithm where calculation is done only once due to which I-KNN will turn to be time efficient algorithm. Moreover one time calculation in proposed I-KNN will improve accuracy too.

CONCLUSION

In this paper, the authors gave a new idea of extending KNN algorithm for E-mail Spam Detection. The idea was to improve KNN algorithm by eliminating recalculations of internal centers and values so as to improve in terms of parameters like accuracy, precision and fmeasure. Later the authors gave an idea that this improved KNN algorithm can be used for filtering emails for better Spam Detection.

REFERENCES

[1] Wenjuan Li and Weizhi Meng, "An Empirical Study on Email Classification Using Supervised Machine Learning," IEEE, pp. 7438-7443, 2015.
[2] Michal Prilepok and Milos Kudelka, "Spam Detection Based on Nearest Community Classifier," IEEE, pp. 353-359, 2015.
[3] Sunil B. Rathod and Tareek M. Pattewar, "Content Based Spam Detection in Email using," IEEE, pp. 1257-1261, 2015.
[4] Wanli Ma, Dat Tran, and Dharmendra Sharma, "A Novel Spam Email Detection System Based on Negative Selection," IEEE, pp. 987-992, 2009.
[5] Xiao Li, Junyong Luo, and Meijuan Yin, "E-mail Filtering Based on Analysis of Structural," IEEE, 2010.
[6] Xianchao Zhang, Shaoping Zhu, and Wenxin Liang,"Detecting Spam and Promoting Campaigns in the Twitter Social Network," IEEE, pp. 1194-1199, 2012.
[7] Soghra M. Gargari and Sule Gunduz oguducu, "A Novel Framework For Spammer Detection In," IEEE, pp. 827-834, 2012.
[8] Hesham Altwaijry and Saeed Algarny, "Bayesian based intrusion detection system," elsevier, pp. 1-6, 2012.
[9] Izzat i Alsmadi and Ikdam Alhami, "Clustering and classification of email contents," Elsevier , pp. 1-11, 2015.
[10] Sarwat Nizamani, Nasrullah Memon, Mathies Glasdam, and Dong Duong Nguyen, "Detection of fraudulent emails by employing," Elsevier, pp. 169-174, 2014.
[11] Loredana Firte, Camelia Lemnaru, and Rodica Potolea, "Spam Detection Filter using KNN Algorithm and Resampling," IEEE, pp. 27-33, 2010.
[12] Nitin Jindal and Bing Liu, "Analyzing and Detecting Review Spam," IEEE, pp. 547-552, 2007.
[13] SNEHAL DIXIT and A.J. AGRAWAL, "SURVEY ON REVIEW SPAM DETECTION," vol. 4, no. 2, pp. 68-72, 2013.
[14] Ommera jan ,heena khana "An analysis of misclassification error detection in mails using data mining techniques"MAY 2015.
[15] Carmona-Cejudo, Jose´ M., Baena-Garcı´a, Manuel, Morales Bueno, Rafael, Gama, Joa˜ o, Bifet, Albert, 2011. Using GNUsmail to compare data stream mining methods for on-line email classification.J. Mach. Learn. Res. Proc. Track 17, 12–18.
[16] Bekkerman, Andrew McCallum, Gary Huang, 2004. Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora.