Third International Conference on "Materials, Computing and Communication Technologies"
International Journal of Scientific Research in Computer Science,
Engineering and Information Technology | ISSN : 2456-3307 (www.ijsrcseit.com)

A Machine Learning Approach for Identification of Spam Content in Email

Winit Anandpwar

Assistant Professor, Dr. Ambedkar College, RTM Nagpur University, Nagpur, Maharashtra, India

Article Info
Volume 8, Issue 7
Page Number : 229-235

Publication Issue :
May-June-2022

Article History
Accepted: 01 June 2022
Published: 20 June 2022

ABSTRACT

The internet has become the most important component of our lives, and it is used for everything. One of the most important applications of this is the exchange of information from one person to another. The increase in internet usage has resulted in an exponential surge of spam in the internet world. E-mail is the most often used online communication tool. The emails contain some unsolicited messages labelled as spam, which causes problems for consumers and necessitates the usage of dependable anti-spam filters. Many methods for detecting spam in email have been investigated. Spam consists of text and graphics that can affect the system. Spam senders grossly abuse email by broadcasting unsolicited facts. As a result, spam is one of the most common issues that an internet user must deal with. This paper proposes two classification methods for spam email detection: k-nearest neighbor's algorithm (KNN) and support-vector machines (SVM). During this process, the dataset is divided into many sets and fed into each algorithm. The findings of three studies are compared in terms of precision, recall, accuracy, f-measure, true negative rate, false positive rate, and false negative rate.

Keywords : Spam Filtering, Classification, KNN, SVM.

I. INTRODUCTION

Email system is a standout amongst the best and regularly utilized sources of correspondence. The reason of the prevalence of email system lies in its financially savvy and quicker correspondence nature. Unfortunately, email system is getting compromised by spam messages. Spam messages are the excluded messages sent by some unknown users also called spammers with the intention of profiting. The email users invest the greater part of their valuable time in arranging these spam mails. Numerous copies of same message are sent commonly which influence an organization financially as well as bothers the getting users. Spam messages are barging in the client's messages as well as creating vast measure of undesirable information and consequently influencing the system's ability and utilization. In this

paper, a Spam Mail Detection (SMD) system is proposed which will arrange email information into spam and ham messages. The procedure of spam sifting centers around three primary dimensions: the email address, subject and substance of the message.

Spammers are generally technically skilled persons that are hired by companies for sending spam. A third party is hired to prevent any legal action on the company itself. Spamming activity can cost attractively to a company, if done right.

E-mails are quick and cheap method for data sharing and correspondence in this day. Perusing inbox E-mails turns into the regular habit for the peoples. Email containing undesirable content irritates the user and possesses the half of the transfer speed of the inbox. These Emails are recognized as spam. The issues of spam mails are a horrid issue. Email spam alludes to sending different, erroneous and unconstrained email messages to various clients. The motivation behind these sends is attention, headway and dissipating indirect accesses or pernicious programs. The time spends by individuals in perusing and erasing the spam mail is waste. A spam mail can't just irritating yet in addition hazardous to beneficiaries. Tapping on connections contained in spam messages may send client to phishing and malware.

A spam mail cannot only be annoying but also dangerous to recipients. Clicking on links contained in spam emails may send user to phishing and malware. Machine learning approach has been widely studied and there are lots of algorithms can be used in e-mail filtering. They include Naïve Bayes, support vector machines, Neural Networks, K-nearest neighbor, Rough sets and the artificial immune system.

Naïve bayes classifier is based on Bayes theorem with an assumption of strong independence. The classifier is a probability based classifier which computes the class probabilities of the given instances. The probability set is calculated by computing the combinational and frequency values of the data set. The class probability which is nearest to the rear end will be picked by the classifier. The Naïve Bayes classifier is a multiclass classifier and works efficiently with supervised learning approach.

The section I explains the Introduction of pam filtering detection using classification method. Section II presents the literature review of existing systems and Section III present proposed system implementation details Section IV presents experimental analysis, results and discussion of proposed system. Section V concludes our proposed system. While at the end list of references paper are presented.

Literature Review

In the literature work, various spam detection techniques are introduced. In linguistic approach natural language processing technique is used to identify similarity among multiple reviews. Feng et al. [3] uses n-gram and their composition. Some studies [2][4] Language modeling also include study for features between multiple reviews like capital words in statements. Lai et al. [5] proposes the probabilistic language modeling technique to find similarity between multiple reviews.

This technique is based on metadata analysis of a review. Metadata includes user behavior and review behavior analysis. Feng et al. [6] proposes a technique that studies metadata of review based on distribution of user rating on different products. 36 different behavior analysis techniques are proposed by Jindal et.al [7] with supervised learning mechanism. [11]

Indicates behavioral features show spammers` identity better than linguistic ones. Fi. Al [12] proposes machine learning method to identify spam reviews. Paper [13] investigates syntactic stylometry for deception detection

Network based algorithms can be applied for spam detection. In this techniques heterogeneous network is established between reviews and users. Fei et al. in [8] proposed a network based Loopy Belief Propagation (LBP) algorithm to find burstiness in reviews to find spam reviews. Li et al. in [10] proposes a technique to analyze a review from multiple users from same IP address. For this heterogeneous network is established between users, reviews and user IPs.

The study of all categories is done independently. Netsapm[1] is the technique proposed by Saeedreza Shehnepoor, Mostafa Salehi, Reza Farahbakhsh, and Noel Crespi. In this technique simultaneous study of Behavioral (RB) Based, Linguistic (RL) Based and graph based approach is proposed. EuijinChoo, Ting Yu , and Min Chi [9] detects the spammer groups in review systems. This is done using sentiment analysis on user interactions and graph theory. It analyses user relationship graph and annotating the graph by sentiment analysis and then pruning is done. According to the studies in literature, a common platform is required that make the study of spam reviews and relationship among various spam detection techniques along with the spammer community identification.

Adem Tekerek, Omer Faruk Bay aimed to detect spam e-mails using BC, RT, and SVM which are some of the machine learning methods. Although there are many e-mail spam filtering studies, due to the existence of spammers and adoption of new techniques, email spam filtering becomes a challenging problem to the researchers. Generally, Performance of proposed model was calculated using training set and observed that RT classifier outperforms other classifiers.

Md. Al Mehedi Hasan1, Mohammed Nasser, Biprodip Pal, Shamim Ahmad [14] developed two models for intrusion detection system using Support Vector Machine (SVM) and Random Forest (RF). The performances of these two approaches have been observed on the basis of their accuracy, false negative rate and precision. The results obtained indicate that the ability of the SVM classification produces more accurate results than RF and it takes less time to train the classifier than SVM.

For improving the accuracy of spam detection, author present an improved Filtering technique [15] which is based on the Improved Digest algorithm and DBSCAN clustering algorithm.

Siddu.Pacingill. Algur et.al, in [16] proposed a system in which link and content spam detection are used to detect the web pages as spam. System also classifies the web page as spam based on threshold set by statistical method.

R.Malarvizhi et al.in [17] an overview for spam filtering, and the ways of evaluation and comparison of different filtering methods is present in the paper. Fisher Robinson Inverse chi square, Ad boosts classifier, Bayesian classifiers are discussed. Bayessian method is used to create the spam filter.

Author says opposite of "spam", e-mail which one wants, is called "ham"[18], usually when referring to a message's automated analysis (such as Bayesian filtering). Machine learning techniques now days are utilized to automatically filter the spam e-mail in a very successful and efficient way. Author consider some of the machine learning methods such as Naïve

Bayes, Artificial Neural Networks, Artificial Immune System Classifier methods, and fuzzy logic.

system architecture

System Architecture

Following Fig. 1 Shows the proposed system architecture. In our proposed system, email is used as input. e-mail classification task can be viewed as a two dimensional matrix, whose axes are the messages and the features. E-mail classification tasks are often divided into several sub-tasks. First, Data collection and representation are mostly problem specific (i.e. e-mail messages), second, e-mail feature selection and feature reduction attempt to reduce the dimensionality (i.e. the number of features) for the remaining steps of the task. Finally, the e-mail classification phase of the process finds the actual mapping between training set and testing set.

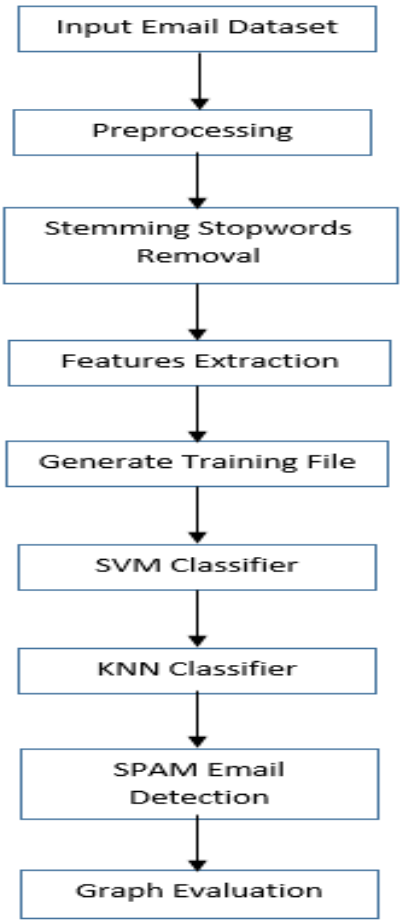

Fig 1. System Architecture

Algorithm

1. K-nearest neighbor classifier method

The k-nearest neighbor (K-NN) classifier is considered an example-based classifier, that means that the training documents are used for comparison rather than an explicit category representation, such as the category profiles used by other classifiers. As such, there is no real training phase. When a new document needs to be categorized, the k most similar documents (neighbors) are found and if a large enough proportion of them have been assigned to a certain category, the new document is also assigned to this category, otherwise not. Additionally, finding the nearest neighbors can be quickened using traditional indexing methods. To decide whether a message is spam or ham, we look at the class of the messages that are closest to it. The comparison between the vectors is a real time process. This is the idea of the k nearest neighbor algorithm:

Stage1. Training
Store the training messages.

Stage2. Filtering

Given a message x, determine its k nearest neighbours among the messages in the training set. If there are more spams among these neighbours, classify given message as spam. Otherwise classify it as ham.

The use here of an indexing method in order to reduce the time of comparisons which leads to an update of the sample with a complexity O(m), where m is the sample size. As all of the training examples are stored in memory, this technique is also referred to as a memory-based classifier [6]. Another problem

of the presented algorithm is that there seems to be no parameter that we could tune to reduce the number of false positives. This problem is easily solved by changing the classification rule to the following l/k-rule:

If l or more messages among the k nearest neighbors of x are spam, classify x as spam, otherwise classify it as legitimate mail.

The k nearest neighbor rule has found wide use in general classification tasks. It is also one of the few universally consistent classification rules.

2. SVM

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships, the SVM modeling algorithm finds an optimal hyperplane with the maximal margin to separate two classes.

Input: sample x to classify training set T.
T = {(x1,y1),(x2,y2),……(xn,yn)};
number of nearest neighbours k.

Output: decision yp ∈ {-1,1}
Find k sample (xi,yi) with minimal values of
K(xi,xi) – 2 * K(xi,x)
Train an SVM model on the k selected samples
Classify x using this model, get the result yp
Return yp

RESULT AND DISCUSSIONS

A. Experimental Setup

All the experimental cases are implemented in Java in congestion with Netbeans tools and MySql as backend, algorithms and strategies, and the competing classification approach along with various feature extraction technique, and run in environment with System having configuration of Intel Core i5-6200U, 2.30 GHz Windows 10 (64 bit) machine with 8GB of RAM

B. Result

Fig. 2 shows the performance Analysis Graph. We summarize the performance result of the two machine learning methods in term of spam recall, precision and accuracy. In term of accuracy we can find that the SVM method is the most accurate while the k nearest neighbour give us lower percentage, while in term of spam precision we can find that the SVM method has the highest precision among the two algorithms while KNN has better recall percentage compare to SVM. In Following graph x-axis show different classification algorithms while y-axis show percentage.

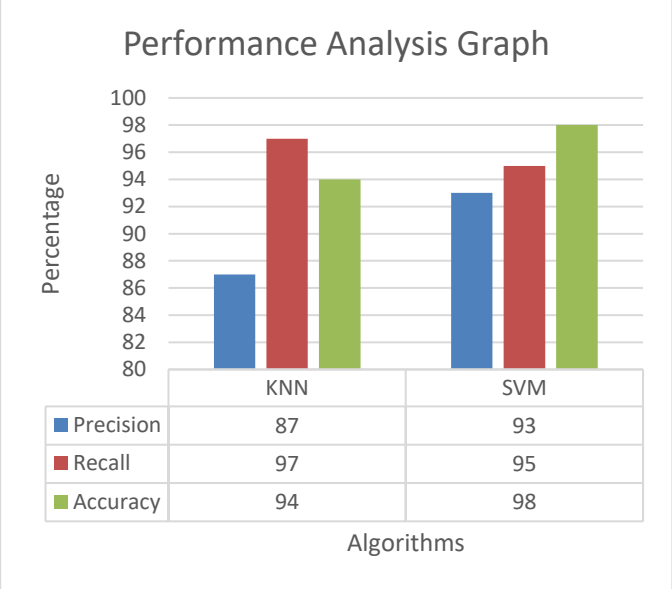

Fig 2. Performance Analysis Graph

Conclusion

The proposed system is designed for detection of spam filtering using classification algorithms like KNN, SVM. Numerous clustering algorithms are used to

detect the spam. The previous methods have few limitations of having less accuracy or precision. This problem would be solved by using the SVM algorithm. The result obtained by using this algorithm may be compared with the NB. Comparison result shows that SVM is better than KNN. Because it requires less time to execute.

REFERENCES

[1] Saeedreza Shehnepoor, Mostafa Salehi*, Reza Farahbakhsh, Noel Crespi NetSpam:a Network-based Spam Detection Framework for Reviews in Online Social Media IEEE Transactions on Information Forensics and Security 2017.

[2] J. Donfro, A whopping 20 percent of yelp reviews are fake.

[3] M. Ott, C. Cardie, and J. T. Hancock. Estimating the prevalence of deception in online review communities. In ACM WWW, 2012.

[4] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination.In ACL, 2011.

[5] Ch. Xu and J. Zhang. Combating product review spam campaigns via multiple heterogeneous pairwise features. SIAM International Conference on Data Mining, 2014.

[6] N. Jindal and B. Liu. Opinion spam and analysis. In WSDM, 2008.

[7] F. Li, M. Huang, Y. Yang, and X. Zhu. Learning to identify review spam. Proceedings of the 22nd International Joint Conference on Artificial Intelligence; IJCAI, 2011.

[8] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Exploiting burstiness in reviews for review spammer detection. In ICWSM, 2013.

[9] Choo E., Yu T., Chi M. (2015) Detecting Opinion Spammer Groups Through Community Discovery and Sentiment Analysis. In: Samarati P. (eds) Data and Applications Security and Privacy XXIX. DBSec 2015. Lecture Notes in Computer Science, vol 9149. Springer, Cham.

[10] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao. Spotting fake reviews via collective PU learning. In ICDM, 2014.

[11] N. Jindal, B. Liu, and E.-P. Lim. Finding unusual review patterns using unexpected rules. In ACM CIKM, 2012.

[12] F. Li, M. Huang, Y. Yang, and X. Zhu. Learning to identify review spam. Proceedings of the 22nd International Joint Conference on Artificial Intelligence; IJCAI, 2011.

[13] S. Feng, R. Banerjee and Y. Choi. Syntactic stylometry for deception detection. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers; ACL, 2012.

[14] Md. Al Mehedi Hasan1, Mohammed Nasser, Biprodip Pal, Shamim Ahmad" Support Vector Machine and Random Forest Modeling for Intrusion Detection System (IDS)", Journal of Intelligent Learning Systems and Applications, 2014, 6, 45-52

[14] Adem Tekerek, Omer Faruk Bay "Spam E-Mail Detection Based On Machine Learning", Conference Paper · April 2018

[15] Alaa H. Ahmed and Mohammed Mikki, "Improved Spam Detection using DBSCAN and Advanced Digest Algorithm", International Journal of Computer Applications, Vol. 6 May 2013.

[16] Siddu p. Algur and Neha tarannum Pendari, "Hybrid Spamicity Approach to web Spam Detection", IEEEC onference on Pattern Recognition, Informatics and Medical Engineering, March 2012

[17] R. Malarvizhi and K. Saraswathi, "Content–Based Spam Filtering and Detection Algorithms-An Efficient Analysis and Comparison", International Journal of Engineering Trends and Technology, Vol.4,Issue 9,Septmber 2013

[18] Mehdi Samiei yeganeh, Li Bin, G.Praveen Babu "A Model for Fuzzy Logic Based Machine Learning Approach for Spam Filtering" Volume 4, Issue 5(Sep.-Oct. 2012), PP 07-10.