Email Spamming Filter: A Systematic Review

MD. MOHAIMANUL HAQUE, 17-33833-1 , C, CSE,American International University Bangladesh
MD.ASIFUL ISLAM, 17-33100-1, C, CSE, American International University Bangladesh
MAHMUD ZAMAN BALI, 17-33159-1, C, CSE, American International University Bangladesh

E-mail is the cheaper and fast method of transmission. E-mail is used in both personal and expert levels of life. Various types of e-mail are lies within social websites. The spam is one of them. Spam is the unwanted messages on marketing or other unhealthy on the internet site which is nothing but wastes the time and resources. Through this study, the aim is to distinguish between harm emails and spam emails by making an effective and sensitive classification model that gives good integrity with a low false-positive rate. In this study, the following comparison will be made between the most popular machine learning classiers such as Naïve Bayes, Support Vector Machine (SVM), J48 (decision tree), K nearest neighbor(KNN), etc.

Additional Key Words and Phrases: data-sets,K-NN,SVM,NB,Bagged,machine learning,AI,
data mining,spam detector,email filtering,j48,c4.5,bagged approach,hybrid machine

ACM Reference Format:
Md. Mohaimanul Haque, Md.Asiful Islam, and Mahmud Zaman Bali . 2018. Email Spamming Filter: A Systematic Review . In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In this modern world, is the time of computers, one of the well-organized and easier type of communication is the email. Reading an email is becoming a periodic habit of many people. This is an efficient, fast and cheaper means of communication. Email containing unwanted emails irritates the user and occupies the half of the bandwidth of the inbox. These mails are identified as spam. The problems of spam emails are a grim issue. E-mail spam refers to sending dissimilar, incorrect and spontaneous email messages to numerous users. Spammers are normally technically skilled persons. So, most of the companies hired spammers for sending spam. A third party is hired to avoid any legal action on the company itself. Spamming activity can cost attractively to a company if done properly. In this case spam classification, various spam filtering methods are used. The function of a spam filter is to identify spam email and prevents it from going to the mailbox. With the help of filters, the adverse impact of spam email is mitigated and operates like a predictable and certified tool to eliminate unwanted emails. In the research, various spam filtering methods such as Naïve Bayes, Support Vector Machine (SVM), J48 (decision tree), K nearest neighbor (KNN), etc. and they will be evaluated in different measurements. We use those approaches to find the spam emails for that any important information for the user cannot a loss by others.

Manuscript submitted to ACM

2 RESEARCH METHODOLOGY

To conduct this literature review we followed the directive that was advised by Lavallée, Robillard and Mirsalari [7]. We took an iterative procedure to perform this review. As this paper will perform the review within the known methods of email spam. So, It was concluded that it would be a great method.

2.1 Research Objective

The main of this study is to systematically review email spam classifier systems such as Naive Bayes (NB) classifier, K nearest neighbor(KNN), Support vector machine(SVM) and comparing them on different measures to get an image of their comparison. thus based on the study and review, a machine should be proposed with an object to enhance the outcome.

2.2 Research Questions

Category	Research Questions	Main Motivation
Target	To minimizing the existing and possible upcoming threats, how much effective Filtering Emails can be?	to identify the necessity of classifying emails apart from several kinds of spams. Such as Blatant Blocking, Bulk Email Filter, Division Filters, Null Sender Disposition, Null Sender Header Tag Validation.
approach	What are the similarities, uniqueness and differences between work principles of the existing systems?	to achieve an image of the work procedures, actions and implementation similarities and differences among the present systems.
approach	What are the Data-mining based systems used classifying spam-emails?	to learn the participation of data-mining based systems via behaviour based, decision-tree.
Outcome	Are these systems effective enough to achieve 100 per cent accuracy in filtering Spams from Emails?	To learn the accuracy achieved by the existing systems individually.
Outcome	What is the hold-backs for and improvement areas for the existing Email filtering systems?	to acknowledge and consider the lacking of the studied systems and discuss improvement possibilities.
Target group	Is it possible to combine or merge several existing methods to acquire a better the outcome for filtering Emails?	to explore the possibilities of customizing or resolving a method supplementing other method's perks for better accuracy and effects.

2.3 Article Selection

To help with article selection, we first searched through some prestigious digital Library and collected paper that is relevant to our research. From that collection, we manually selected some paper which is strongly related to email spam detection. We confirmed the relationship by reading the title, abstract, introduction and conclusion of the paper that we thought were relevant. Then we read them and performed this literature review.

2.3.1 ***Keywords and Search String.*** Here are the keywords that are the main focus of this paper.

Keywords: Email, Spam, Spam Filter, Data Mining, SVM, Naive Bayes, K-NN, Machine Learning, c4.5, Literature Review, J48, hybrid method.

Here is the search string that helped with the collection of research paper that we wanted to do a review on.

Search String: (email or electronic mail),(spam or junk-mail) and (data mining) and (classification, categorisation, taxonomy) and (machine-learning or artificial intelligence) and (algorithms or formula).

2.3.2 ***Digital Libraries to Search***. We have taken help from this online digital library to search for all the relevant paper that we need to research this topic.

a) Microsoft Academic Search

b) ACM

c) ScienceDirect

d) IEEE

e) Google Scholar

f) Springer

2.3.3 ***keyword search and Manual Selection***. **Keyword search**:Email spam algorithm, spam detection, classing or grouping of algorithms.

Manual selection:

Automated search is great for finding paper fast, but to find exactly the relevant paper we had to do it manually. We have selected 15 papers that were related to our research topic. We read the title, abstract, conclusion to find out which one of them is heavily relevant to our area of the research topic. After finding them we discarded the papers that we thought were not strongly connected to our chosen subject.

1. Detection, analysis and investigation of spam.
2. Using a large amount of data to classify email spam.
3. Limitations and capability of Statistical method based on Spam Email Filters.
4. Spam Mail Detection through Data Mining – A Performance Analysis
5. Spam campaign detection, analysis, and investigation
6. Identifying spammers on social networks.

2.3.4 ***Final set of Articles***. This is the papers that were finally selected after discarding manually selected papers.

1. Machine-learning techniques for junk-mail classification.
2. An inspection of Machine Learning Classifiers for Spam Detection.
3. Spam Email Detection using Data Mining – A Comparative Performance Study.
4. Application of Improved KNN algorithm for Email Spam Detection.
5. Junk Electronic-mail Identification Technique using Different Decision Tree Classifiers through Data Mining procedure-A Performance Analysis.
6. Machine Learning-based application of Spam E-Mail Detection.
7. Machine learning for email spam filtering: review, approaches and open research problems.
8. E-mail Spam Filtering using Machine Learning: Methods and Trends.
9. Filtering Electronic-mail Spam: A Review of Procedure and Current Algorithms.

3 DISCUSSION

3.1 To minimizing the existing and possible upcoming threats, how much effective Filtering Emails can be?

In this contemporary era, spam -email can pose a great loss to the constant e-communication system. Using existing methods which are mostly adapting with situations it is possible to analyze about 97% of the emails [2]. As the systems are evolving due to its adaptive domains it is expected to keep a conventional rate of identifying and filtering emails.

3.2 How Machine Learning and Data-mining based systems contributes solutions to Spam-Email detection constraints?

Almost all the methods and engines available to filtering emails basically followed the domains of machine learning and Data-mining. Filtering systems taken for this study which are Naïve Bayes Classifier, K-NN , J48 decision tree and Support vector machine all are under the domains of machine learning and Data Mining.

Naïve Bayes Classifier:

Bayesian Classifier, proposed in 1998, Mainly detects the dependent events and the probability of an event occurring in the future based on previous occurrences of that event [2]. In an Email spam detection or filtering system, the technique is to predict future emails by predicting the words from the previous emails as it only requires a little substance of training data to evaluate parameters for analysis. If the summation of word possibilities exceeds a certain limit, Bayesian classifier classifies the email to either category which is SPAM or HAM. The basic Bayesian theorem formula defined as,

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

Modifying this formula aiming for the filtering based on tokens, the statistic mostly interested for a token T,stands for (spam rating) [2] calculated as follows:

$$S[T] = C_spam(T)/(C_spam(T) + C_ham(T))$$

Here, CSpam(T) & CHam(T) represents count of SPAM or HAM consisting token T.

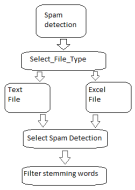

Fig. 1. KNN spam detection architecture;

KNN Classifier:

KNN stands for K Nearest Neighbor is a simple non parametric method used for classification and regression containing 2 stages, Training Filtering [8].

Training: Training data gets to be stored.

Filtering: For filtering, indexing method is used to reduce the time of comparison which leads to a sample with a

complexity of O (message_sample_size). This method also refers to a memory-based filtering method [2].

SVM:

SVM known as Support Vector Machine works by taking concept of "statistical learning theory and structural optimization principal" [8]. Due to its ability to deal with high dimensional data with the help of kernel function. SVM modelling algorithm figures an optimal hyperplane to separate two classes with the maximal margin[5].

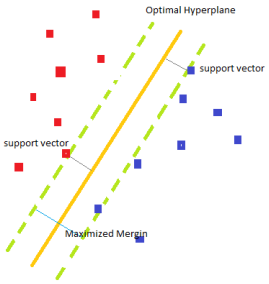

Fig. 2. SVM;

Assume a training set,

X= x, y where x Rn and Y +1, -1, indicates the unique class for i-th training sample where it takes +1 for SPAM and -1 as HAM [11]where output is measured by following,

$$Y = wx - b$$

Here, y=final classifier output;

x=feature vector;

w=normal vector comparable values consisting in x;

b=bias parameter, decided by the training process;

the separation between classes followed by,

minimize:

$$\sum_{i=1}^{n} ai * y\ i = 0$$

subject to:

$$Y(wx - b) \geq \forall i$$

J48 Decision Tree:

Being an open-source implementation of C4.5, J48 works by analyzing data from the nodes of which are used to assess the consequence of existing data based the concept of entropy. It evolves into a multi-class classifier based on the decision trees of the training data [4].

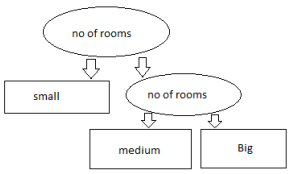

Fig. 3. Decision tree Example;

Here, at each node of the tree, an attributed is chosen to further split the samples into subsets where every leaf node represents a classification.

3.3 What are the similarities, holdbacks uniqueness and differences between work principles of the existing systems?

The basic factors followed to establish given characteristics and a comparison between the studied systems are,

- Dimensions of featured location
- Speed
- Accuracy
- Complexity

The Naive Bayes classifier, also known as avid learning classifier is comparatively faster than most other classifiers such as KNN, J48 excluding SVM. From a theoretical point of view, comparing SVM with the other methods studied is relatively complex. Where Naïve Bayes deals with anticipation [9], Support Vector Engine is geometric in nature. The main reason to use an SVM instead, if the problem appears to be linearly not detachable or the platform being a higher-dimensional space [8]. On the other hand, the problem being linear but highly denser and noisy with very big size K nearest neighbor is a robust choice.

Proceeding into specifics, K-NN classifier having local heuristics is a supervised indolent classifier. Decision boundaries achieved using the method is far more complex than any other decision trees. but in the case of Email filtering, it does a better job as it directly focuses on finding the correlation between observations. but the accuracy of K-NN decreases as the complexity of scope expands [1]. J48, on the other hand, prophesies a class for a provided input vector. KNN mostly used for clustering, a j48 decision tree for classification (in case of Email filtering, both for classification).

Performance Comparison:

In terms of spam recall, Precision and Accuracy performance selecting the top 55 features (dataset of 1000) [10] comparison is summarized in this study. (computed by Weka)

Algorithm	Spam Recall (%)	Spam Precision (%)	Accuracy (%)
Naïve Bayes	98.43	95.56	97.20
Support-Vector Machine	95.32	94.00	92.70
K nearest neighbor	97.92	87.00	96.29
J48 Decision Tree	96.12	96.31	95.80

From the dataset we can see that KNN has the lowest precision percentage being nearly independent of the value

of K. based on the data set NB seems to be the fastest method. Both J48 and SVM also have a good number under circumstances.

Fig. 4. Demonstration Of spam recall, precision, accuracy of the given methods;

3.4 Is it possible to combine or merge several existing methods to acquire a better outcome for filtering Emails?

It is possible to improve a system resolving its flaws. It is possible to combine several methods to achieve better performance. The implementation of hybrid approaches is well known in email filtering and other systems. Hybrid bagged method: From the systems studied, a bagged system can be manifested with a combination of J48 and Naive Bayes classification. Bagging approach also known as the bootstrap aggregating approach considers the combinations of the multiple repeated sets of the same dataset and decreases the variance.[6]

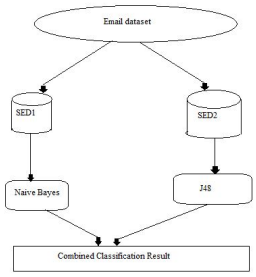

Fig. 5. workPrincipal baggage;

For this approach, the email dataset is randomly divided into two separate sample datasets, SED1 SED2 are considered to train as an individual classifier[9]. thus, the overall result will be the average of the included.

3.5 Are these systems effective enough to achieve 100% accuracy in filtering Spams from Emails?

Email spam detection systems are constantly evolving to be better at detecting spam, but there is still an error that is appearing that needs improvement. For example some email are not spam but the system is deciding to put them in spam folder. This is what counts as an error in the detection system. Let us assume that we get to a point where the email system is near 100% accuracy at detecting spam email. At that time there could be a new technique that has been developed to send spam more effectively and avoid the state of the art email detection system. So, 100% accuracy is not a plausible idea, but to go near that accuracy is possible. All of the algorithms that are being analyzed in this paper are

considerably near the 100% mark and that means that the existing algorithms are on the right track for keeping spam sending services at a distance.

3.6 What are the improvement rooms for the existing approaches used to classify Emails?

The email spam filter system that exists today is very sophisticated. But, there is always room for improvement. The way to achieve this is to fine-tune existing algorithms with an iterative approach. Clean data that are easily available would cause a significant improvement in the performance of the existing algorithm. So, creating a pipeline that gives access to a huge amount of clean data upon which this statistical and machine learning algorithms can work on would be immensely helpful for the further development and advancement of these existing algorithms. Another prospect that can be improved greatly is having access to computers that are powerful enough to process large chunks of data so that the algorithms can "learn" new pattern from spam email when it is already working within a mail service system, that way it can detect new spam when state of the art spam mailing system is sending new types of spam mail.

4 FUTURE RESEARCH DIRECTIONS

From the authors' findings, it was assumed, Email detection and classifying systems is a well-researched topic.

4.1 KNN:

As it is seen that the increase of complexity of scope causes a decrease in the precision accuracy of K nearest neighbor classifier[1]. So, further, analyze should be done to resolve this problem by merging with other available classifiers or improving the working flowchart.

4.2 SVM:

SVM being one of the most efficient machines available, it is hard to implement as it is complex to understand. so study should be done to reduce its dependable to a level for a broad implementation of the technique.

4.3 J48 decision tree:

Being one of the most simple and efficient techniques available, it is an unstable system , which points to a little change to the system causes a massive change to the overall arrangement [11]. Further studies should be done to strengthen its inherent flexibility.

5 VALIDITY THREAT

The integrity of data and theory that are written in this paper, to survey them perfectly and extensively is a manual task. We have taken a couple of measures to ensure they are accurately presented here.

Researcher bias:

Sometimes the researcher has an opinion about the topic that they can believe is true but, that is not the case. So peer reviewer must ensure that the data and theory has integrity and not biased in any way to prove their point without a sound scientific base.

Comprehensiveness:

Steps must be taken to ensure that all the proof and theory are were researched and presented exhaustively.No topics were left out that were required to discuss this topic in-depth.

Data Falsification:

Researchers should be honest about the data that they are collecting and not alter them in any way to fit their agenda. All data and theory should be collected or discussed in a manner that was scientifically examined.

6 CONCLUSION

In this research, some of the prominent Email-detecting systems are reviewed. Descriptions and calculations are presented and the comparison on different factor was presented. Comparing accuracy values measured on different factors it is found, the Naive Bayes and SVM techniques have a very accomplishing performance amongst the other techniques. more analysis should be done to raise the execution of the Naïve Bayes either through a hybrid system or else by deciding the background interdependence argument within the naïve Bayes classifier. Thus, the last hybrid System appears to be the most efficient to detect SPAM and filtering emails comparing with the other systems studied. Therefore, it is found that tree-like classifiers operate adequately in spam mail apprehension while the multidimensional platforms are compromised [3].

REFERENCES

[1] Haryana India Ambala. [n.d.]. Implementation of Improved KNN algorithm for Email Spam Detection. ([n. d.]).

[2] WA Awad and SM ELseuofi. 2011. Machine learning methods for spam e-mail classification. *International Journal of Computer Science & Information Technology (IJCSIT)* 3, 1 (2011), 173–184.

[3] Alexy Bhowmick and Shyamanta M Hazarika. 2016. Machine learning for E-mail spam filtering: review, techniques and trends. *arXiv preprint arXiv:1606.01042* (2016).

[4] Sarit Chakraborty and Bikromadittya Mondal. 2012. Spam mail filtering technique using different decision tree classifiers through data mining approach-a comparative performance analysis. *International Journal of Computer Applications* 47, 16 (2012).

[5] Emmanuel Gbenga Dada, Joseph Stephen Bassi, Haruna Chiroma, Adebayo Olusola Adetunmbi, Opeyemi Emmanuel Ajibuwa, et al. 2019. Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon* 5, 6 (2019), e015802.

[6] ONUR GÖKER. 2018. SPAM FILTERING USING BIG DATA AND DEEP LEARNING. (2018).

[7] Mathieu Lavallee, Pierre-N Robillard, and Reza Mirsalari. 2013. Performing systematic literature reviews with novices: An iterative approach. *IEEE Transactions on Education* 57, 3 (2013), 175–181.

[8] Megha Rathi and Vikas Pareek. 2013. Spam mail detection through data mining-A comparative performance analysis. *International Journal of Modern Education and Computer Science* 5, 12 (2013), 31.

[9] Priti Sharma and Uma Bhardwaj. [n.d.]. Machine Learning based Spam E-Mail Detection. ([n. d.]).

[10] Tarek Sobh and Khaled Elleithy. 2010. *Innovations in Computing Sciences and Software Engineering.* Springer Science Business Media.

[11] Shrawan Kumar Trivedi. 2016. A study of machine learning classifiers for spam detection. In *2016 4th international symposium on computational and business intelligence (ISCBI).* IEEE, 176–180.

A CONTRIBUTION RECORD

Details of each group member contribution according to the following tables are given.

A.1 Paper Assessment

following table with the required information is populated.

A.2 Paper writing contribution

Populate the following table with the required information.

Student id & name	Paper No frm Ref	Paper Title
17-33159-1 Mahmud Zaman Bali	1,5,9,11	Implementation of Improved KNN algorithm for Email Spam Detection,Machine learning for email spam filtering: review, approaches and open research problem,A Study of Machine Learning Classifiers for Spam Detection ,Performing Systematic Literature Reviews With Novices: An Iterative Approach
17-33100-1 Md.Asiful Islam	4,9,3	Spam Mail Filtering Technique using Different Decision Tree Classifiers through Data Mining Approach-A comperative performance Analysis,Machine Learning based Spam E-Mail Detection,Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends
17-33833-1 Md. Mohaimanul Haque	2,6,8,10	Innovations in Computing Sciences and Software Engineering,Spam Mail Detection through Data Mining – A Comparative Performance Analysis,SPAM FILTERING USING BIG DATA AND DEEP LEARNING,MACHINE LEARNING METHODS FOR SPAM E-MAIL CLASSIFICATION

Table 1. Paper collected and read by the group member

Student id & name	Section No	Section Title
17-33159-1 Mahmud Zaman Bali	2.1, 2.3.1, 2.3.3, 2.3.4, 3.5, 4.3	Introduction, Article selection, Discussion, Future Research Direction, Validity thread
17-33100-1 Md.Asiful Islam	1, 2.3.4, 3.6, 4.2, 5, 6	Introduction, Article selection, Discussion, Future Research Direction, Validity thread, Conclusion
17-33833-1 Md.Mohaimanul Haque	2.2, 2.3.2, 3.1, 3.2, 3.3, 3.4, 4.1	Introduction, Article selection, Discussion, Future Research Direction, Conclusion

Table 2. Section(s) Written in the paper by the group member