A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms

Mansoor RAZA and Nathali Dilshani Jayasinghe
School of Computing and Mathematics
Charles Sturt University, Study Centre
Melbourne VIC 3000, Australia
maraza@studygroup.com
nathalidj@gmail.com

Muhana Magboul Ali Muslam
Department of Information Technology
College of Computer and Information Sciences
AI Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh 11432, Saudi Arabia
mmmuslam@imamu.edu.sa

2021 International Conference on Information Networking (ICOIN) | 978-1-7281-9101-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICOIN50884.2021.9334020

***Abstract*—Email is the most used source of official communication method for business purposes. The usage of the email continuously increases despite of other methods of communications. Automated management of emails is important in the today's context as the volume of emails grows day by day. Out of the total emails, more than 55 percent is identified as spam. This shows that these spams consume email user time and resources generating no useful output. The spammers use developed and creative methods in order to fulfil their criminal activities using spam emails, Therefore, it is vital to understand different spam email classification techniques and their mechanism. This paper mainly focuses on the spam classification approached using machine learning algorithms. Furthermore, this study provides a comprehensive analysis and review of research done on different machine learning techniques and email features used in different Machine Learning approaches. Also provides future research directions and the challenges in the spam classification field that can be useful for future researchers.**

***Index Terms*—Spam Detection, Spam Classification, Spam Filter, E-mail, Supervised Learning, Machine Learning Algorithms, Email Classification, Spam Email Detection, Email Categorization, Email Feature Set Analysis, Spam Detection Using Machine Learning Algorithms.**

I. Introduction

Email has become the most extensively used official communication mechanism for most of the internet users. In the past few years, the increased number of email usage has emerged and escalated the problems caused by spam emails. Spam or junk email is referred to the act of distributing bulk unsolicited messages [3]. Whereas the emails which are meaningful with opposite nature are called 'Ham'. An average email user receives about 40-50 emails per day. Spammers earn around 3.5 USD million from spam every year making financial losses to both personal and institutional front [4]. Because of that, the users spend a significant amount of working time on these emails. According to [1] spam accounts more than 50 percent of email server traffic, transmitting massive volume of unwanted and unsolicited bulk emails. They consume user resources to no-useful output, reducing the productivity. The spam which are propagated by spammers have the objective of marketing purposes to unfold malicious criminal activities such as identity theft, financial disruptions, stealing sensitive information and reputational damage. Because of these reasons, email management and classification of spam emails are a vital necessity for organizations in order to increase their productivity and reduce the financial losses.

A. Relevant Spam Statistics

In this following subsection, we will emphasis on some of the global statistical information on spam vs. financial impact. Some country specific metrices for Australia are also discussed in the analysis. As per [5] there are a little over 4 billion email accounts that are actively in use in 2020 and this number is projected to grow up to 4.48 billion by the year 2024. This means nearly half of the world population are actively using emails at the year 2020. From this, spam accounts for 57.26 percent of total email traffic for the year 2019 [5]. This shows that, out of the total emails going around the world, more than half of it accounts for unwanted, unsolicited spam emails. As for year 2019 FBI recently reported that a global financial loss of $12.5Billion has incurred for the businesses due to business email and business email account compromise as a result of spam and phishing [6]. These financial losses incurred to the businesses are expected to skyrocket in the upcoming years as the growth of the email usage is increasing day by day. As for the Australian context the following are the statistics for digital scams for the previous years [7] (Australian Competition and consumer Commission, 2019. This statistic shows that, these digital scams are a global issue and Australia also have a significant problem because of that. Apart from that, data demonstrates that the trends are heading upwards for both losses and the number of cases for each year. Investment scams offers promising business opportunities which are fraudulent in exchange for monetary investments. Dating scams focuses on victimizing individuals who are looking for romantic partners using internet. Emails are the number one used method by the scammers when for delivering malware and other fraudulent scams. The solutions that are used in the current context are most of the time lagging because of the innovativeness and problem-solving skills of the spammers. Because of these reasons, it is important to understand and develop systems which can detect and classify spam emails from legitimate ham emails.

TABLE I
TOP THREE DIGITAL SCAM LOSSES(AU$) AND FREQUENCY (NUMBER OF CASES) IN AUSTRALIA FROM 2017 – 2020. [1]

Year	Total Loss	Digital Scam Amount	Digital Scam Frequency
2017	$90801407	Investment Scams $31,326,476 Dating and Romance $20,530,578 Other business & employment $ 5,270,948	Phishing 26,386 Identity Theft 15,703 False billing 13,455
2018	$107001471	Investment Scams $38 846 635 Dating and Romance $24 648 024 False Billing $ 5 512 502	Phishing 24,291 Threats to Life 19,455 Identity Theft 12,800
2019	$142898217	Investment Scams $61,813,801 Dating and Romance $28,606,215 False Billing $ 10,110,753	Phishing 25,170 Threats to Life 13,375 Identity Theft 11,373
2020	$52971358	Investment Scams $20,650,486 Dating and Romance $14,708,686 False Billing $ 4,378,559	Phishing 10,689 Threats to Life 4 255 Identity Theft 4,237

II. LITERATURE REVIEW

As we are reviewing spam detection systems which uses Machine Learning (ML) algorithms, it is important to review on the history of ML in the field and the different algorithms that are used in the current context to classify spam. Researchers have pointed out that, content and the operational mechanisms of the spam emails changes over the time. Therefore, the techniques that are working now may not be useful soon. This phenomenon is identified as the conceptual drift [8]. Machine Learning is the engineering approach formulated to enable computational instruments to act without being programmed explicitly. This approach is a huge boon to detect and tackle spam issue because of the ML system's ability to evolve itself over the time minimizing the concept drift. In the following section, we will discuss on number of ML techniques, approaches and algorithms and their associated benefits with Supervised, Unsupervised and Semi Supervised Machine Learning Algorithm Approaches.

A. Supervised Machine Learning Algorithms

Supervised machine learning algorithms learns from a set of pre-labelled data, with the possible outputs for the corresponding spends have been already been given [9]. This algorithm learns gradually using the labelled data provided and eventually builds up its own probabilistic mapping system to use for new inputs. This technique has two different subtypes called, Regression and classification [9]. This technique is mostly used to generate the outputs which are in categorical nature. In this context; Spam and Ham as the two categories.

B. Unsupervised Machine Learning Algorithms

As the name describes, in this technique there are no labelled data or explicit instructions to pre trained the designed model. Therefore, these systems are not provided with a training. In this algorithm the analysis is carried out based on the dataset and feature out the common characteristics, structures and features in a group. Then rearrange the output data in different based structure or the pattern [10]. The output data can be organized in different types such as clustering, anomaly detection, association and autoencoders

C. Semi-supervised Machine Learning Algorithms

In this approach the system is trained with both labelled and unlabelled data in the testing phase and the system analysis are carried out using both techniques. The main objective of this approach is to achieve better accuracy and precision than the traditional supervised and unsupervised approaches. There are two different types of output presentations; Semi supervised clustering and semi supervised classification in this approach. All the research papers that have been selected are categorized using the above approaches to carry out the analysis in an effective manner. The categorization details and its analysis are presented in the analysis and findings section. In the following section we will be focusing on the different machine learning algorithms that have been used in the reviewed studies. These have been analyzed after categorizing them under the above discussed machine learning algorithm approach.

III. SUPERVISED LEARNING BASED MACHINE LEARNING ALGORITHMS

In this following section ML algorithms, which are used in supervised based nature approach have been discussed. In a system one or several ML algorithms have been used to achieve the expected performance measurements.

A. Artificial Neutral Network (ANN)

[11] Has used a system using ANN approach to classify spam and ham emails. This developed model is based on thirteen pre labelled-fixed email features which are associated with spam emails. ANN is built using artificial neutrons, Hence the name come from. The number of artificial neutrons that are been used in the system can be varied and depend on the requirements of the system. These neurons are connected to different layers such as Input layer, Hidden layers and output layers. ANN systems 'learns' through a process named, 'Back-Propagation'. The produced new output of the network is compared and matched with the ideal match that should have been produced. The variation is taken into account and adjust the weights between the neutron connections with many iterations [4] .

B. Naïve Based Machine Learning Algorithm

This is one of the commonly used supervised machine learning algorithm. This has been developed using the Bayes' rule which tries to derive the probability of an event occurrence based on even related prior knowledge and conditions [12]. This approach is highly scalable, fast and easy to implement into a system. Naïve Based algorithm treats the features as independent from each other. This has been used in the system developed by [13] to provide the solution to the problem independence of random variables with 23 different classification rules. This system uses Decision tree along with Naïve Based to generate the expected outcome. The main drawback of this algorithm is this can be only used if the input features are 'completely independent on each other'. In the practical scenario, this is not always possible.

C. Support Vector Machine

Support Vector Machine Support Vector Machine (SVM) is another well established and most frequent used Machine learning classification algorithm which was proposed by [14]. Some of the systems have used only SVM as their system classification algorithm while some researchers have used combination of algorithms including SVM. [15] Has used a system with SVM and Weighted SVM. The weights are reflecting the importance of different analysis categories; 'classes'. As per the researchers, the advanced weighted SVM algorithm has higher performance metrices. In the SVM algorithm a hyperplane is created generating different classes to analyze various features derived from the dataset. SVM can be adopted into any number of vector dimensions. In the 2D dimension the approach would be a line. In the 3D dimension it would be a hyperplane.

D. Decision tree (DT)

Decision tree machine learning algorithm is another algorithm that have been used more commonly in the reviewed supervised learning approach studies. The reasons to use this more often are this is an algorithm that can be used easily, easier explanations and visualizations. This can be used with both large and small data sets. Has the ability to handle both numerical data and the categorical data in the system [4]. In the developed system done by [16], they have used DT along with other algorithms in their system. DT has been used in the tier three stage with binomial categorization of spam and ham emails. The model could classify the spam in real time, for this feature DT has provides significant insights as it has simple computational mechanism which is required for efficient real time computational requirements.

IV. Unsupervised Learning Based Machine Learning Algorithms

In this section we are focusing on the unsupervised machine learning algorithms that have been used in the reviewed systems. The adoption of unsupervised machine learning algorithms is low compared to supervised machine learning algorithms. There are two algorithms used by the researchers.

A. K-nearest Neighbour machine learning algorithm (KNN)

This algorithm is effective to use when there is noise in the input dataset. This can be used to generate both classifications and regression outputs for the developed system. The main drawback of this algorithm is it is highly sensitive for the outliers in the data set. Apart from that, computational cost for this algorithm is comparatively higher with regard to other machine learning algorithms [4]. This may be the main reason that this has not been adopted more commonly in the reviewed studies.

B. K- means Clustering machine learning algorithm

This algorithm has straightforward implementation mechanism and the computational cost is comparatively lower than KNN ML algorithm. These are the reasons for this algorithm to be one of the commonly used unsupervised machine learning algorithm in spam classification field [4]. In the K means clustering the data mining process initiates with the first group which is selected randomly. There is a randomly selected centroid for each cluster to begin the process. Repetitive calculations are carried out starting from that centroid to generate the optimized position.

V. Analysis and Findings

In this section we will focus on some key insights and the findings derived from the critically analyzed studies. We will start with a basic overview of the testing approaches used and the different machine learning algorithms used by each of the reviewed paper. Based on the information in the table 2, we can obtain information on the nature of the approaches of the studies and the distribution of the different machine learning algorithms used to classify spam and ham.

A. High adoption of Supervised Based Approach

The figure 01: The pie chart demonstrates the distribution of the different approaches in the reviewed systems. High adoption of the supervised learning technique can be seen in the distribution with 54 percent of the selected sample. The next most adopted framework is unsupervised approach. As the figure 01 suggest majority of the researchers have adopted Supervised learning technique as their first choice. This clearly signifies that there is high degree of opportunities to expand and the availability for the research in the field of semi-supervised and unsupervised Machine learning Approach.

B. Consistent and Higher accuracy in Supervised Based Approach

The main objective of majority of the researchers are based on increasing a higher accuracy of spam detection from the developed systems. As we can see in the figure 2, The scatter plot shows that the accuracy of the systems using supervised learning model has a distribution of accuracy in higher level with tight close range with the minimum variation. This shows that, the outcomes are consistent with higher accuracy (average accuracy is above 90percent for supervised learning model).

TABLE II
SUMMARY OF MACHINE LEARNING ALGORITHMS AND RESULTS OBTAINED

Techniques	Model Types			Primary Algorithms Used													Result	
	Supervised	Semi-Supervised	Unsupervised	K-means	Naïve Based	Decision Tree/C4.5	Support Vector Machine	Maximum Entropy	Ann	Swarm Optimization	Multilayer Perceptron	Key word-based	Gradient Boosting	Weighted SVM	KPCM	KFCM	Accuracy%	P F Score F R
T1		✓			✓	✓	✓	✓									88	R,F
T2	✓					✓			✓								75	
T3	✓					✓											91	
T4	✓				✓	✓											80	F,R
T5	✓					✓	✓										71	P
T6			✓	✓		✓	✓										98	P
T7	✓				✓							✓						
T8		✓		✓		✓	✓			✓							61	
T9		✓					✓										82	
T10			✓	✓	✓			✓					✓				88	
T11			✓	✓	✓												93	P
T12	✓						✓							✓	✓	✓	93	F
Sum	7	3	2	4	6	5	7	1	1	1	1	1	1	1	1	1		

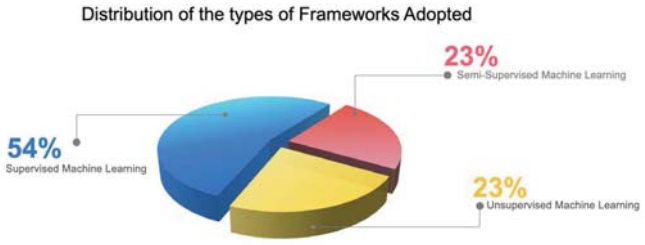

Fig. 1. Distribution of types of framework adopted.

From finding A and B, we can conclude that there is a huge opportunity to develop systems using semi-supervised and unsupervised frameworks for future researchers. Apart from that, the performance evaluation bar for the existing systems for these approaches are very low compared to supervised framework, meaning there is a higher flexibility in achieving higher percentage outcomes.

C. Algorithm Preference.

The bar chart on the figure 3 illustrates the distribution of the Machine Learning algorithms used in the reviewed papers. Majority of the developed system have used SVM machine learning algorithm. Some of the studies have used, their own advanced algorithms that have been derived from the main ML algorithm. These have been included in the original ML format for above analysis purpose. These can be further analyzed in future studies in this section.

D. Using Single Algorithm Vs. Multi Algorithm Framework

As we can see in the table 2, most of the systems (83 percent) have used a combination of different ML algorithms for their systems in-order to achieve higher results from their study. Out of twelve studies only two systems have used single ML algorithm as their approach. All the other studies have used two or more algorithms combined to achieve higher and better results from their studies.

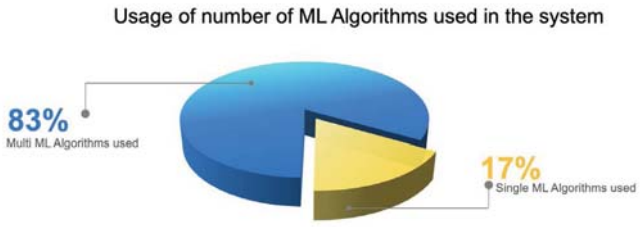

Fig. 2. Usage of number of ML algorithms in system.

E. Email Features Analysed using ML algorithms

The table 3 demonstrates the different email features that are analysed in the reviewed articles. As we can see majority of the systems that are developed are focused on analysing for spam using the Body of the email and the BoW. BoW (bag of word) is the approach of analysing the email using key words, phrases and texts in the email. From this we can identify

that, there is a higher opportunity to develop spam detection systems analysing the email features such as attachments in the email, structure of the email and spam hyperlinks added in the emails. These research areas have been approached only by a few current researchers leaving these as the higher opportunity areas to explore research areas for the future studies.

F. Analysing the Email Content

Almost all the studies that have used in this paper have developed their systems based on analysing the content such as checking on keywords or phrases that are pre identified to be included in spam emails. This approach is reasonable and logical for a certain state for current context. However, as we have explained in the previous sector, the spammers are using new techniques day by day. Therefore, the spamming techniques are evolving. Hence, this traditional approach would not be adequate in the future to detect spam as spammers will find creative and efficient spamming techniques to go under this detecting method. Table 4 provides a summary of the machine learning algorithms.

VI. CHALLENGES AND FUTURE RESEARCH DIRECTIONS

In this section, we highlight the several research challenges and open issues that have been identified during the review process. In this regard, future research work that is yet to be focused on to enhance the performance of the spam email classification in different application areas and features are presented in below section.

A. Real Time Spam Classification

Majority of the researchers that are reviewed are based on datasets which are not included with real time environment features and elements. Therefore, these studies are unable to classify spam emails in real time. Only one of the studies that are in the review database could detect spam in real time. As we are living in a fast pacing environment, it is essential to work in real time. This is one of the major untapped direction that needs future studies to be focused on. The online stream and analysing of emails for spam is complicated and more advanced compared to the existing studies, therefore there will be higher potential research challenges that needs to be addressed when developing systems catering to these requirements.

B. Dynamic Feature Updates

Another research area that needs to be focused on the future studies is to addition and removal of spam detection features in the system without rebuilding or restructuring the entire developed system. As the techniques used by spammers are changing and developing day by day, it is essential to have a detection system which can be updated easily with minimum time are resource consumption. Most of the reviewed systems have not mentioned whether the systems can be updated easily, which means these areas is not much focused on during their studies.

C. Reduction of System Process time

Focus on reduction of the process and classification time for a spam classification system using advanced hardware rather than using traditional hardware system is another area which needs to be focused on in the future studies. Real time classifications and user centric email classifications are the next generation of spam classification which requires higher processing time. Therefore, to reduce classification time suitable developed hardware technologies should be adopted for the systems. This area should be focused on future researchers in order to cater the future spam classification requirements.

TABLE III
EMAIL FEATURES ANALYZED BY ML ALGORITHMS

Article	Email features analyzed by ML Algorithms							
	Header	Body	BoW	Structure	Hyper Links	Attachments	Term Frequency	Other
A1		✓	✓					
A2		✓	✓	✓				
A3	✓							✓
A4	✓	✓	✓					
A5							✓	
A6					✓		✓	
A7		✓	✓					
A8		✓	✓					
A9				✓	✓			
A10		✓	✓		✓	✓	✓	
A11	✓	✓	✓					
A12			✓	✓				
Sum	3	7	8	3	3	1	3	1

D. Emphasis on different email features

From the review we have found out that, majority of the designed systems are focused on detecting spam emails using the analysis on BoW or the body of the email. But spams can come in different formats such as hyperlinks, images and attachments. Efficient methods to detect spam for these spam approaches should be more focused on in the future studies.

E. Focus Semi Supervised and Unsupervised approach

From the review conducted, we have understood that most of the Machine learning approaches are based on supervised learning techniques. Therefore, there is an untapped and higher opportunities to develop spam email classification systems are available using the other two approaches; semi and unsupervised approaches.

F. Reducing the false positive rate

More researchers should be carried out to deliver the outcome based on reduced false positive rate. False positive is marking the legitimate emails as spam and risk of losing the important emails during the process. A good designed system should have a better false positive rate. This is another area

the future studies should be focused on apart from focusing on achieving the higher accuracy rate from their systems.

VII. CONCLUSION

After the comprehensive analysis on the selected research studies, we have identified several research findings and observation. These have been detailed discussed in the prior sections with adequate explanations. In this section, we will be more focused on main findings and the conclusions of the study. High adoption rate for supervised machine learning approach can be seen throughout the review. This approach is used mainly because it generates higher accuracy results with less variation giving high consistency for this approach. Apart from that, we have found out that certain algorithms such as Naïve Based and SVM have high demand compared to other Machine Learning Algorithms. The multi algorithm used systems are more common in use to cater better outcome rather than using single algorithm. Researchers have more focused on email features such as BoW and Body text creating future research opportunities to develop systems to detect spam on other email features.

REFERENCES

[1] "Global spam volume as percentage of total e-mail traffic from January 2014 to September 2019, by month." https://www.statista.com/statistics/420391/spam-email-traffic-share/.

[2] T. Ouyang, S. Ray, M. Allman, and M. Rabinovich, "A large-scale empirical analysis of email spam detection through network characteristics in a stand-alone enterprise," Elsevier, vol. 2015, pp. 101–102.

[3] O. Saad, A. Darwish, and R. Faraj, "A survey of machine learning techniques for Spam filtering," IJCSNS Int. J. Comput. Sci. Netw. Secur.

[4] K. Asif, A. Sami, S. Bharindhan, and K. Krishan, "A Comprehensive Survey for Intelligent Spam Email Detection," IEEEXplore, 2019.

[5] "Number of e-mail users worldwide from 2017 to 2024." [Online]. Available: https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/.

[6] M. Guntrip, "https://www.proofpoint.com/us/corporate-blog/post/fbi-reports-125-billion-global-financial-losses-due-business-email-compromise." [Online]. Available: https://www.proofpoint.com/us/corporate-blog/post/fbi-reports-125-billion-global-financial-losses-due-business-email-compromise.

[7] "Australian Competition and consumer Commission," Scam Stat., [Online]. Available: https://www.scamwatch.gov.au/scam-statistics?scamid=all & date=2018.

[8] K. Jackowski, B. Krawczyk, and M. Woźniak, "Application of adaptive splitting and selection classifier to the spam filtering problem," Cybern. Syst. An Int. J.

[9] Sathya and A. Abraham, "Comparison of supervised and unsupervised learning algorithms for pattern classification," ResearchGate.

[10] F. Qian, Y. C. H. Abhinav Pathak, Z. M. Mao, and Y. Xie, "A case for unsupervised-learning-based spam filtering," Univ. Minnesota J., 2010.

[11] Y. Alamlahi and A. Muthana, An Email Modelling Approach for Neural Network Spam Filtering to Improve Score-based Anti-spam Systems. Modern Education and Computer Science Press, 2018.

[12] L. Melian and A. Nursikuwagus, "Prediction student eligibility in vocation school with Naïve-Byes decision algorithm," 2018.

[13] A. S. Aski and N. K. Sourati, "Proposed efficient algorithm to filter spam using machine learning techniques," Elsevier, vol. 2016, pp. 145–149.

[14] K. Pawar and M. Patil, "Pattern classification under attack on spam filtering," IEEExplore, 2015.

[15] A. K. Rajan, V, and A K. "V, V., & Rajan, "An Improved Spam Detection Method With Weighted Support Vector Machine," IEEE Explor. ." IEEExplore.

[16] H. Kaur and A. Sharma, "Improved Email Spam Classification Method Using Integrated Particle Swarm Optimization and Decision Tree," IEEE Xplore, vol. 2016, pp. 516–521.

TABLE IV
SUMMARY OF THE MACHINE LEARNING ALGORITHMS

Article	Machine Learning Algoruthm	Results and Summary
A1	Naive Bayes (NB) SVM Decision Tree (DT) Maximum Entropy (MaxEnt)	Hybrid ensemble learning approach to detect review spam Active semi supervised learning approach Dealing with duplicates precisely
A2	DT with binary classification	Fixed features improves ANN classification Classifies spam emails from Arabic and English languages
A3	DT with binary classification	Recognize the spam features more accurately Detects a pattrn of repetitve keywords in spam Classifications of spam based structure such as Cc/Bcc, domain and header
A4	C4.5 DT classifier Multilayer perception Naive bayes classifier	Filter spam from valid emails with low error rates and high efficiency using a multilayer perception model Demonstrates higher effeciency than NB classifier and J48 with a low rate of false positive
A5	SVM C4.5 Random forest	Identifies spam using feature representation preserving class separability and lower dimensional space Identify spam when feature size is small with a good generalization irrespective of the data source
A6	SVM NB and active learning K-Nearest Neighbours (KNN) classifiers	Reduces the consuming time of email classification while guaranteeing the accuracy Proposed method perfoms better than other methods on the two corpuses by using F1 measurement
A7	NB Classifer ML Algorithm with combination of semantic-based, key-word base and machine learning in Python	Effectively increase the accuracy of Navie bayes and reduce false positives of detecting spam Detects text modifications and correctly classify the email as spam or ham Spam detection with an error rate of 38 % and an accuracy of 62 % Discovery of relationship of exponential regression between email length and spam score
A8	Integrated particle swarm optimization based on decision tree algorithm with unsupervised filtering SVM K means	Sam and non-spam mails are classified with 98.32 % accuracy for the experimental
A9	Support Vector Machine (SVM)	Evaluates the performance of non linear SVM based classifiers with two different kernel functions Compare the training and testing accuracy of the kernels and find out which kernel better Decision tree classifier requires higher the memory System identifies the spam emails from its contents Spam can be blocked by user and ham can be retained by the user
A10	SVM NB and active learning K-nearest neighbours (KNN) classifiers DT classifiers SVM linear kernel Gradient boosting (GB)	Out of the analysed algorithms, SVM for both text and file classification has the highest spam classification accuracy
A11	Modified K-Means NB classification	Improved classification accuracy 96 % precision in detection Decreased the number of iteration step
A12	SVM linear kernel Weighted SVM KPCM (Kernel based Probabilistic C-Means) KFCM (Kernel based	Evaluate the impact of spam detection using SVM, WSVM with KPCM and WSVM with KFCM.UC