Hierarchical Email Spam Filtering

Ismail M. Khater
Department of Electrical and Computer Engineering
Birzeit University
Ramallah, Palestine

Omar M. Al-Jarrah
Amman Arab University
Jordan Street Mubis
Amman 11953, Jordan

Basheer Al-Duwairi
Department of Network Engineering and Security
Jordan University of Science and Technology, Irbid, Jordan
Email: basheer [AT] just.edu.jo

Abstract—**Email spam continues to be a major problem in the Internet. There have been great research efforts to combat email spam. However, a major problem in most email spam filters is that they may result in filtering some legitimate emails. Such a problem could be prohibitively expensive in practice especially if the misclassified email is of a great importance to the recipient. To address this problem, we propose a hierarchical email spam filtering system that is composed of two main phases. Email messages entering the first phase are classified into ham or spam using header-based email spam filtering. Email messages that are classified as spam are further inspected in phase two using content-based email spam filtering. In this context, we identify the combination of machine learning algorithms that would provide the best performance when used in the two phases. We evaluate the proposed work through a combination of theoretical analysis and experimental studies based on publicly available datasets. Our studies show that the proposed Hierarchical Email Spam Filter (HESF) achieves a precision of 99.99% and 100% in some case studies with very low false positives.**

Keywords: two phases email spam filtering; header-based spam filtering; content-based spam filtering; image spam filtering; image texture analysis

I. INTRODUCTION

Email spam, defined as unsolicited bulk email, continues to be a major problem in the Internet. With the spread of malware combined with the power of botnets, spammers are now able to launch large scale spam campaigns covering wide range of topics (e.g., pharmaceutical products, adult content, etc.) causing measure traffic increase and leading to enormous economical loss. Recent studies such as [1] and [2] revealed that spam traffic constitute more than 89% of Internet traffic. According to Symantec [3], in March 2011 the global Spam rate was 79.3%. According to the same report, spam accounted for approximately 52 billion emails per day at the beginning of March and decreased to 33 billion emails per day at the end of March. The cost of managing spam is huge compared with cost of sending spam which is negligible, this cost include the waste of network resources and network storage, the traffic and the congestion over the network, in addition to the waste in employees productivity. It was estimated that an employee spends 10 minutes a day on average sorting through unsolicited messages [4]. Other studies [5], [6], [7] reported that spam costs billions of dollars.

Spammers are increasingly employing sophisticated methods to spread their spam emails. Also, they employ advanced techniques to evade spam detection. A typical spam campaign involves using thousands of spam agents to send spam to a targeted list of recipients. In such campaigns, standard spam templates are used as the base of all email messages. However, each spam agent substitutes different set of attributes to obtain messages that do not look similar. Moreover, spammers are increasingly adopting image-based spam wherein the body of the spam email is converted to an image which renders text-based and statistical spam filters useless. While header-based email spam filtering is considered to be one of the main approaches to combat email spam, content-based email spam filtering is another approach that is equally important, especially when spammers intelligently craft their spam emails with header attributes that are indistinguishable from that of legitimate emails rendering header-based approach less efficient.

Generally, content-based email spam filtering approach involves digging into the content of email messages searching for certain signatures or specific patterns. Spammers are continuously adopting new techniques to evade detection. Image spam is one of these techniques that have gained a lot of popularity among spammers and that is being increasingly used in recent years. This type of spam began to appear in late 2005 and reached a peak of over 50% of spam emails from 2006 to 2007 [38]. In April, 2009 the amount of image spam was about 15-22% of all spam [39]. In this technique, spammers launch their campaigns through images attached to their emails instead of text based spam.

With the widespread of viruses, worms, malware, and botnets, email spam detection has always been a challenging problem. While there are enormous research efforts that have been made to increase the accuracy of email spam detection, a major problem of most email spam filters is that they may result in filtering some legitimate emails. Such problem could be prohibitively expensive in practice especially if a misclassified email is of a great importance to the recipient. Therefore, there has always been a great concern not only regarding the rate of misclassified spam emails (i.e., false negatives) but also regarding the rate of misclassified ham emails (i.e., false positives). To address this issue, we propose a hierarchical email spam filtering system, called HESF, which consists of two phases. HESF applies header based filtering on

incoming email message (Phase I). Email messages that are classified as spam are further processed by content-based filter (Phase II) to reduce the false positives rate. In this context, this paper extends the work presented in [49] and [50] by combining header-based filtering and content-based filtering in such a way that we achieve the best of both worlds. HESF is evaluated theoretically based on the results obtained in [49] and [50]. Our studies show that the proposed HESF system achieves excellent performance for different scenarios.

The rest of this paper is organized as follows: Section II discusses related work. Section III presents the proposed HESF system. Section IV presents performance evaluation of the proposed HESF system. Finally, Section V concludes the paper.

II. RELATED WORK

Email spam filtering represents a major approach to combat spam. The goal of email spam filtering is to classify email messages into ham or spam. Typically, email spam filtering involves inspecting message content, header or both. In all cases, it is necessary to apply some technique (e.g., data mining, machine learning, pattern recognition, etc.) to distinguish ham from spam. Generally, combating email spam techniques can be categorized into three main categories as follows [12]: pre-send methods which focuses mainly on blocking supply lines of spam (e.g. [10]), post-send methods which deals with filtering email spam after being sent (e.g., [11], [12], [13]), and new protocols which are based on modifying the email transfer process itself to avoid most of the waste of resources caused by spam, such as network traffic and workload on receiving server (e.g., [9], [11], [12]).

Machine Learning-based email spam filtering represents a major approach of post-send techniques. In this approach, a machine learning-based classifier is applied to certain features extracted from the email message in order to classify it as ham or spam. The machine learning-based spam filters may be further classified into two main categories [13], namely "Non-content-based (Header-based) spam filtering", "Content-based spam filtering", and "Combining multiple classifiers for email spam filtering". In the following subsections, we discuss the previous work done in each category and point out how it differs from the work presented in this paper.

A. Header-based Email Spam Filtering

An email message typically consists of header and body. The header is a necessary component of any email message. The Simple Mail Transfer Protocol (SMTP) [33] defines a set of fields to be contained in the email message header to achieve successful delivery of email messages and to provide important information for the recipient. These fields include: email history, email date, time, sender of the email, receiver(s) of the email, email ID, email subject, etc. Header-based email spam filtering represents an efficient and lightweight approach to achieve filtering of spam messages by inspecting email message header information. Typically, a machine learning classifier is applied on features extracted from email header information to distinguish ham from spam. For example, Sheu [15] categorized emails into four categories based on the title: sexual, finance and job-hunting, marketing and advertising, and total category. Then he classified them according to the attributes from email message header. He proposed a new filtering method based on categorized Decision Tree (DT), namely, applying the Decision Tree technique for each of the categories based on attributes (features) extracted from the email header. The extracted features are from the sender field, email's title, sending date, and the email's size. Sheu applied his filter on a Chinese emails and obtained accuracy, precision, and recall of 96.5%, 96.67%, 96.3%, respectively.

Wu [16] proposed a rule-based processing that identifies and digitizes the spamming behaviors observed from the headers and syslogs of emails by comparing the most frequent header fields of these emails with their syslog at the server. Wu noticed the differences in the header filed of the sent email from what is recorded in the syslog, and he utilized that spamming behavior as features for describing emails. A rule-based processing and back-propagation neural networks were applied on the extracted features. He achieved an accuracy of 99.6% with ham misclassification of 0.63%. YE et al. in [17] proposed a spam discrimination model based on SVM to sort out emails according to the features of email headers. The extracted features from email header fields are the return-path, received, message-id, from, to, date and x-mailer; they used the SVM classifier to achieve a recall ratio of 96.9%, a precision ratio of 99.28%, and an accuracy ratio of 98.1%. Wang in [18] presented a statistical analysis of the header session message of junk and normal emails and the possibility of utilizing these messages to perform spam filtering. A statistical analysis was performed on the contents of 10,024 junk emails collected from a spam archive database. The results demonstrated that up to 92.5% of junk emails are filtered out when utilizing mail user agent, message-id, sender and receiver addresses as features.

Recently, Hu et al. [14] proposed an intelligent hybrid spam-filtering framework to detect spam by analyzing only email headers. This framework is suitable for extremely large email servers because of its scalability and efficiency. Their filter can be deployed alone or in conjunction with other filters. The extracted features from the email header are the originator field, destination field, x-mailer field, sender server IP address, and email subject. Five popular classifiers were applied on the extracted features: Random Forest (RF), C4.5 Decision Tree (DT), Nave Bayes (NB), Bayesian Network (BN), and Support Vector Machine (SVM). The best performance was obtained by the RF classifier with accuracy, precision, recall, and F-measure of 96.7%, 92.99%, 92.99%, 93.3%, respectively.

B. Content-based Email Spam Filtering

Content-based techniques inspect the body of an email searching for specific keyword(s) that are typically used by spammers or associated by certain spam campaign. Other techniques use pattern recognition to detect spam that follows certain behavior or pattern. Email body itself may be text, image, or both. Also, attachments are possible. Therefore, content-based filtering techniques usually deal with all these content types. Generally, Image based spam filtering techniques can be categorized into:

- *OCR-based Techniques*: The philosophy of OCR-based techniques is based on extracting the text embedded into attached images, then the same approaches used in spam filters to analyze emails' body text is used [20], which are keyword detection and text categorization techniques. The power of OCR-based techniques is determined by the OCR system itself. OCR errors is considered as one of the drawbacks of this kind of filters, especially when spammers obscure the content of the image by adding noise, dots, changing the background colors and rotating images, which affects the efficiency of OCR text extraction. This fact has led to other techniques based on low-level image features [21] and a combination of OCR with low-level image features [22], [23], [24]).
- *Techniques based on low-level Image Features*: In these techniques, image classification is based on a set of low-level features extracted from images. The classification process depends on the chosen features. For example, Wu et. al., [25] proposed a classification technique based on the presence of text features such as number of text regions, fraction of images with detected text regions, and the text area. L. Qiao et. al., [26] used corner and edge detection to characterize text area, and the color variance, the number of colors contained in the image, and the prevalent color coverage to characterize graphic properties of spam images. Low-level features such as color, shape and texture are used by [27], based on the fact that spam images often contain clearer and sharper objects than ham images. A different approach based on image metadata was proposed in [28], [29]. Image metadata and information include image width, height, aspect ratio, image area, image compression, image file extension and file size. Another technique is the near-duplicate detection technique. Spam images are often generated from a common template, and randomized to evade signature-based filters. Besides, the spam images are sent in batches to many users. Thus, images generated from the same template are visually similar (near-duplicate), these images can be recognized by a comparison with a known spam images stored in a database [30].

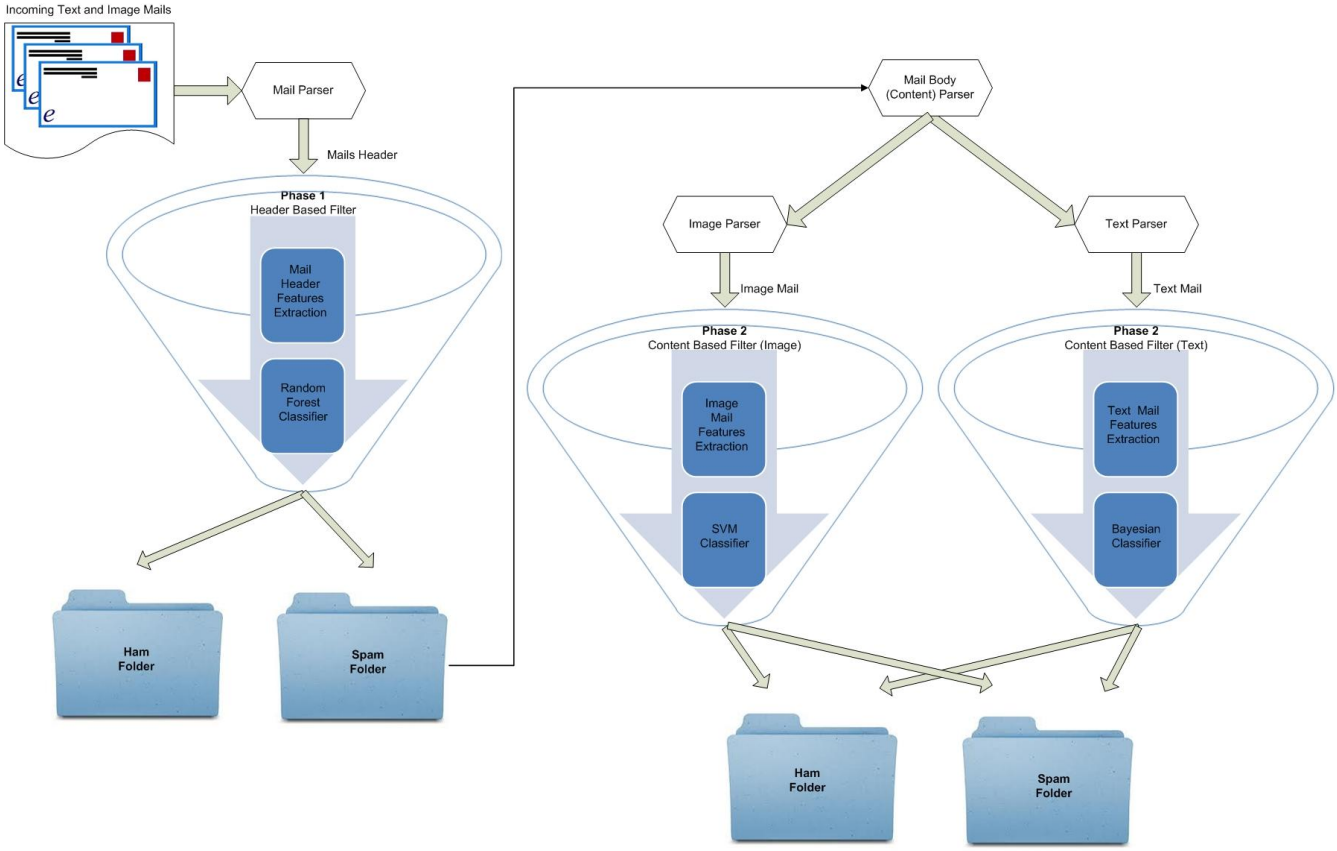

Figure 1. Hierarchical email spam filtering HESF approach

C. Combining Multiple Classifiers for Email Spam Filtering

There have been some research efforts (e.g., [31], [45], [46]) to enhance the accuracy of email spam filtering by

combining multiple classifiers. The main idea in these approaches is to apply multiple classifiers to incoming email messages in parallel, meaning that each classifier processes the email independently from other classifiers. After that, a decision is made about the legitimacy of the email by combining the results of the different classifiers. Majority voting is usually adopted in such systems. The main problems of these approaches are:

- The same email is applied to all the filters of the system at the same time.
- The system has to wait for the result of each filter separately [45], that is, the overall system must wait the slowest filter to obtain the final result.
- The system requires all filters to be available at the same time [45].
- Applying majority voting can result in Gray List (GL) (i.e., is a set of emails that is not classified as spam or ham. Or emails that voting does not come with final decision as it is purely spam or ham emails).

III. HIERARCHICAL EMAIL SPAM FILTERING SYSTEM (HESF)

Figure 1 depicts the proposed Hierarchical email spam filtering system. The system consists of two main phases. Phase I implements header-based email spam filtering, while Phase II implements content based email spam filtering. All incoming email messages go through Phase I to be classified into ham or spam using header-based email spam filtering. ***Only email messages that are classified as spam are further inspected in phase II using content-based email spam filtering***. This means that only fraction of the total incoming email messages is subject to Phase II filtering. This is particularly important to limit the problem of misclassifying some legitimate emails (i.e., minimizing false positive rate). In Phase II, we deal with two main content types:

- *Text-based emails*: For this type, we apply statistical filter to determine whether an email is ham or spam.
- *Image-based emails*: For this type, we apply image-based filter to determine whether an email is ham or spam.

It is to be noted that email messages entering Phase II, after being initially classified as spam in Phase I, are deeply inspected by only one of the content-based filters of Phase II depending on the email content type (i.e., text or image). We believe, that using the best classifier at each phase (i.e., header-based and content-based) and combining them in this fashion would result in a highly efficient and accurate email spam filter.

In Phase I, we perform header-based email spam filtering where certain email header features are extracted and provided as input to several machine learning algorithms. We refer the reader to our previous work [49] for detailed description. It is important to mention that the selection of email header features is based on analyzing large publicly available datasets to determine the most distinctive features. It is also important to point out that we include most of the mandatory and optional email header fields in order to fill any gab or missing information that is required for email classification. The process of building a feature vector of an email starts by preprocessing of email messages to convert them into a standard format as described in RFC 2822. After that, we extract the header of the email to select the required features and build the feature vector which summarizes all the needed information from an email. This feature vector is then used to build the feature space for all emails that are needed for the classification phase.

Phase II of the proposed HESF system includes two types of content-based email spam filtering: text-based email spam filters and image-based email spam filters. Filtering of text-based email spam was studied extensively in the literature and the research community agrees that statistical filters are the most efficient for this type email spam. For image-based email spam, we adopt the approach proposed on our previous work [50] which focused on selecting image-texture features for image spam filtering. Generally, textures are complex visual patterns composed of entities that have characteristic brightness, color, slope, size, etc. The main reason for choosing image texture features for image spam filtering is the fact that non-computer generated images have a different quality of texture as compared to textures in computer generated images. In our work, we use the following features which are considered to be among the most important features for texture analysis as pointed out in [39], [40], [41]:

- *Image Histogram*: is a graphical representation of the tonal distribution in digital images (i.e., for each tonal value, it plots the number of pixels).
- *Image Gradient*: is a directional change in the intensity or color in an image.
- *Run-Length Matrix* (*RLM*): the run-length matrix $p(i, j)$ is the number of runs with pixels of gray level i and run length j [43]. Various texture features can be derived from RLM.
- *Co-Occurrence Matrix* (*COM*): is a matrix that is defined over an image to be the distribution of co-occurring values at a given offset.
- *Autoregressive Model* (*AR*): assumes a local interaction between pixels of the image in that the intensity is a weighted sum of neighboring pixel intensities.
- *Wavelet Transform*: in digital image processing, a Discrete Wavelet Transform (DWT) is used. DWT is any wavelet transform for which the wavelets are discretely sampled. It captures both frequency and location information (location in time), and considered as a key advantage over Fourier transform.

IV. PERFORMANCE EVALUATION OF HESF SYSTEM

Evaluating the proposed hierarchical email spam filtering system requires a representative dataset that includes email messages with full header information and image content. Unfortunately, to the best of our knowledge there are no publicly available datasets with the required information. Therefore, instead of using real datasets, the proposed system was theoretically evaluated based on the results obtained in [49] and [50] for header-based and image-based email spam filters, respectively. Subsection IV-B summarizes the main results obtained in [49] and [50]. Subsections IV-C and IV-D present the results obtained for two theoretical scenarios to evaluate the proposed HESF system. In both scenarios, we assume a dataset with complete email information (i.e., header and body).

A. Performance Metrics

We use the following standard performance metrics to evaluate machine learning classifiers used in Phases I and II: accuracy, precision, recall, F-measure. These metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F - measure = \frac{2Precision.Recall}{Precision + Recall} \quad (4)$$

Where FP, FN, TP, and TN are defined as follows:

- *False Positive* (*FP*): The number of misclassified legitimate emails.
- *False Negative* (*FN*): The number of misclassified spam emails.
- *True Positive* (*TP*): The number of spam messages that are correctly classified.
- *True Negative* (*TN*): The number of legitimate emails that are correctly classified.

Precision is the percentage of correct prediction (for spam email), while spam Recall examines the probability of true positive examples being retrieved (completeness of the retrieval process), which means that there is no relation between precision and recall. On the other hand, F-measure combines these two metrics in one equation which can be interpreted as a weighted average of precision and recall. In addition, we use Receiver Operating Characteristics (ROC) curves which are commonly used to evaluate machine learning-based systems. These curves are basically two-dimensional graphs where TP rate is plotted on y-axis and FP rate is plotted on x-axis. Therefore, ROC curves depicting the tradeoffs between benefits TP and costs FP [37]. A common method to compare between classifiers is to calculate the Area Under ROC Curve (AUC).

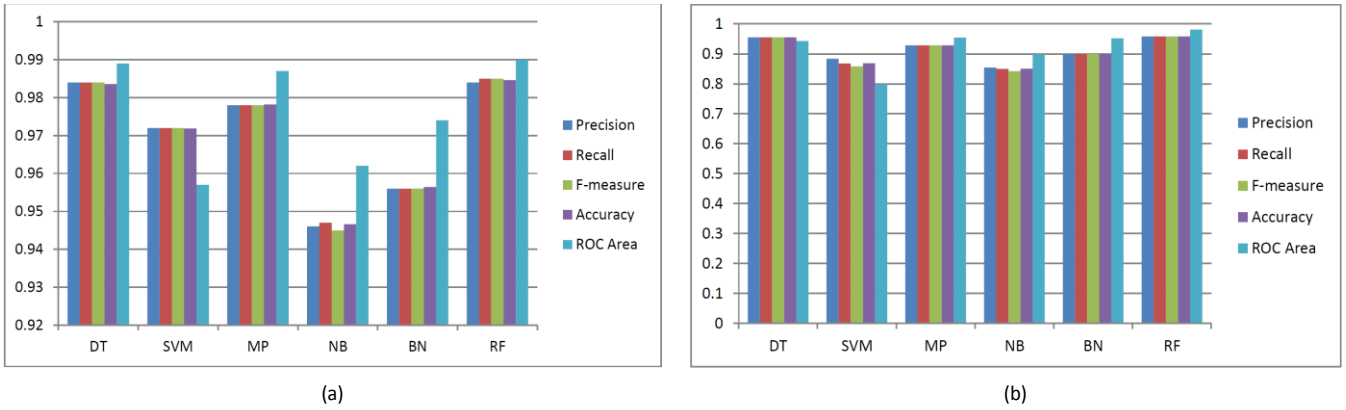

Figure 2. Header-based email spam filtering. The Performance of different machine learning classifiers applied on (a) CEAS2008 dataset and (b) CSDMC2010 dataset in terms of accuracy, precession, recall, F-measure, and ROC area

B. Summary of the results obtained in [49] and [50]

For the sake of completeness, this subsection presents a summary of the results obtained for header-based email spam filtering [49] and for image-based spam filtering [50]. Table I summarizes datasets used to evaluate each filter. Figures 2 and 3 depict the results obtained for each of the following machine

learning classifiers: C4.5 Decision Tree (DT), Support Vector Machine (SVM), Multilayer Perception (MP), Nave Bays (NB), Bayesian Network (BN), and Random Forest (RF).

TABLE I. SUMMARY OF THE DATASETS USED IN [49] AND [50]

Filter type	Dataset used	Ham	Spam
Header-Based	CEAS2008 [11]	6523	26180
	CSDMC2010 [36]	2949	1378
Image-Based	Dredze [29]	1770	3209
	Image Spam Hunter [44]	810	926

Figure 2-a depicts the performance of the different classifiers in terms of accuracy, precision, recall, F-measure and the area under ROC for CEAS2008 dataset. It can be seen that RF classifier outperform all the other classifiers with an average accuracy, precision, recall, F-Measure, ROC area of 98.5%, 98.4%, 98.5%, 98.5%, and 99%, respectively. In order to confirm the results obtained using CEAS2008 dataset, the experiments were repeated using another recent dataset (however, with smaller size). Figure 2-b depicts the performance of the different classifiers using CSDMC2010 dataset in terms of accuracy, precision, recall, F-measure and the area under ROC. It can be seen that RF classifier outperform all the other classifiers with an average accuracy, precision, recall, F-Measure, ROC area of 95.8%, 95.8%, 95.8%, 95.8% and 98.1%, respectively. It is to be noted that all classifiers achieved comparable performance this time indicating that the performance of some classifiers depends on the dataset used for testing and training.

Figure 3-a depicts the performance of the classifiers applied to the features extracted from ISH dataset. We also show the performance of SVM for different values the parameter γ (the radial basis kernel of the SVM classifier). It can be seen that both the RF classifier and the SVM classifier (with $\gamma = 0.1$) performs very well. RF classifier achieved precision, recall, F-measure, accuracy, and ROC Area of 98.1%, 98.1%, 98.1%, 98.1%, and 99.5% respectively. While the same metrics for SVM classifier (with $\gamma = 0.1$) were as follows: 98.6%, 98.6%, 98.6%, 98.56%, and 98.6%. It is also obvious that as we increase the value of γ, the overall performance of SVM classifier decreases, but with a very low false positive. This means that the value of γ could be adjusted to obtain the increase or decrease FP while maintaining a good performance for this classifier.

Figure 3-b depicts the performance of the classifiers applied to the features extracted from Dredze dataset. We also show the performance of SVM for different values the parameter γ (the radial basis kernel of the SVM classifier). It can be seen that both the RF classifier outperforms all other classifiers with precision, recall, F-measure, accuracy, and ROC Area of 98.6%, 98.6%, 98.6%, 98.55%, and 99.4% respectively. It is to be noted that the performance of SVM was very close to that of RF classifier, and it did not vary much for different values of γ.

C. HESF-Theoretical Scenario I

In this scenario, our discussion is based on a theoretical dataset of 1000 emails, all of these emails are assumed to be image emails with complete header information. We further assume that the dataset is a balanced dataset meaning that it contains 500 spam emails and 500 legitimate emails. Assuming that a dataset with this specification is used as input for the proposed hierarchical email spam filtering system, our objective is to evaluate the performance of the system in terms of precision, recall, F-measure, FP rate, FN rate, and accuracy. As mentioned in Subsection IV-B, the RF classifier has the best performance among all other classifiers for the header-based email spam filtering for both data sets. Based on that, we decide to use this classifier for header-based filtering of Phase I. We need to expect the confusion matrix of the assumed dataset. Then, we can compute the performance of Phase I in the context of proposed hierarchical email spam filtering system based on the following Equations.

$$FP\ Rate = \frac{FP}{FP + TN} \tag{5}$$

From Equation 5, we can compute FP and we can find TN based on the fact that FP + TN = the actual number of ham emails. Similarly, FN and TP can be calculated as follows:

$$FN\ Rate = \frac{FN}{FN + TP} \tag{6}$$

TP can be found based on the fact that TP+FN= the actual number of spam emails.

Phase I: To evaluate the performance of the classifier used in Phase I, FP, FN, TP, and TN can be found as follows:

FP = FP Rate × (FP + TN) = 0.046 × 500 = 23 ➔ FP + TN = 500 ➔ TN = 477.

FN = FN Rate × (FN + TP) = 0.008 × 500 = 4 ➔ FN + TP = 500 ➔ TP = 496.

The confusion matrix is shown in Table II. Based on this confusion matrix, the performance metrics for Phase I would be as follows: Precision = 95%, Recall = 99.2%, Accuracy = 97.3% and F-measure = 97.1% (all of these metrics are for spam).

Phase II: In Phase II, we need to filter the predicted spam emails from Phase I, because these emails may contain misclassified emails. The predicted number of spam emails = TP + FP = 496 + 23 = 519 will be subject to further analysis by Phase II. Based on the results obtained in [50], we have two candidate classifiers for this phase; the RF classifier and the SVM classifier. Therefore, it is necessary to evaluate the performance of Phase II for both classifiers and compare them to decide which one is the best. The performance evaluation of the RF classifier presented here is based on the results obtained for this classifier using the Dredze dataset experiment. The confusion matrix of this classifier obtained after using the 519

spam emails The confusion matrix of this classifier obtained after using the 519 spam emails resulting from phase I is shown in Table III.

TABLE II. THE CONFUSION MATRIX FOR PHASE I- FIRST SCENARIO

Prediction	Actual	
	Spam	***Ham***
Spam	496	4
Ham	23	477

TABLE III. THE CONFUSION MATRIX FOR RF CLASSIFIER IN PHASE II- FIRST SCENARIO

Prediction	Actual	
	Spam	***Ham***
Spam	493.024	2.976
Ham	0.437	22.563

TABLE IV. THE CONFUSION MATRIX FOR SVM ($\Gamma = 0.3$) CLASSIFIER IN PHASE II- FIRST SCENARIO

Prediction	Actual	
	Spam	***Ham***
Spam	484.096	11.904
Ham	0.161	22.839

The following are the resulting values of the performance metrics for the hierarchical email spam filter after using RF classifier in phase II: Overall Precision = 99.91%, Overall Recall = 99.4%, Overall Accuracy = 99.3%, Overall F-measure = 99.7%. Based on these results, it can be seen that the proposed hierarchical email spam filtering system improves the overall performance in terms of precision, recall, accuracy and F-measure as compared with the results from one phase. The performance evaluation of the SVM classifier is based on the results obtained in [50] for this classifier using the Dredze dataset experiment with ($\gamma = 0.3$). The confusion matrix of this classifier obtained after using the 519 spam emails resulting from Phase I is shown in Table IV. Which would result in the following values of the performance metrics for the hierarchical email spam filter after using SVM classifier in Phase II: Overall Precision = 99.97%, Overall Recall = 99.6%, Overall Accuracy = 97.7%, and Overall F-measure = 99.8%.

Similar to the case of RF classifier, it can be seen that the proposed hierarchical email spam filtering system improves the overall performance in terms of precision, recall, accuracy and F-measure as compared with the results from one phase. However, it can be seen that using SVM in Phase II would result in less accuracy as compared to the case of using RF classifier. This is due to the relatively high false negative of the SVM. At the same time, it is important to highlight that the false positive rate is less than that when using RF classifier.

TABLE V. THE CONFUSION MATRIX FOR SVM ($\Gamma = 0.2$) CLASSIFIER IN PHASE II- FIRST SCENARIO

Prediction	Actual	
	Spam	***Ham***
Spam	461.28	34.72
Ham	0	23

Now we present the performance evaluation of the SVM classifier based on the results obtained in Figure 3 for this classifier using the ISH dataset experiment with ($\gamma = 0.2$). Using SVM ($\gamma = 0.2$) classifier results when applied on ISH dataset experiment. The confusion matrix of this classifier obtained after using the 519 spam emails resulting from Phase I is shown in Table V. Which would result in the following values of the performance metrics for the hierarchical email spam filter after using SVM classifier in Phase II: Overall Precision = 100%, Overall Recall = 93%, Overall Accuracy = 93.3%, and Overall F-measure = 96.4%.

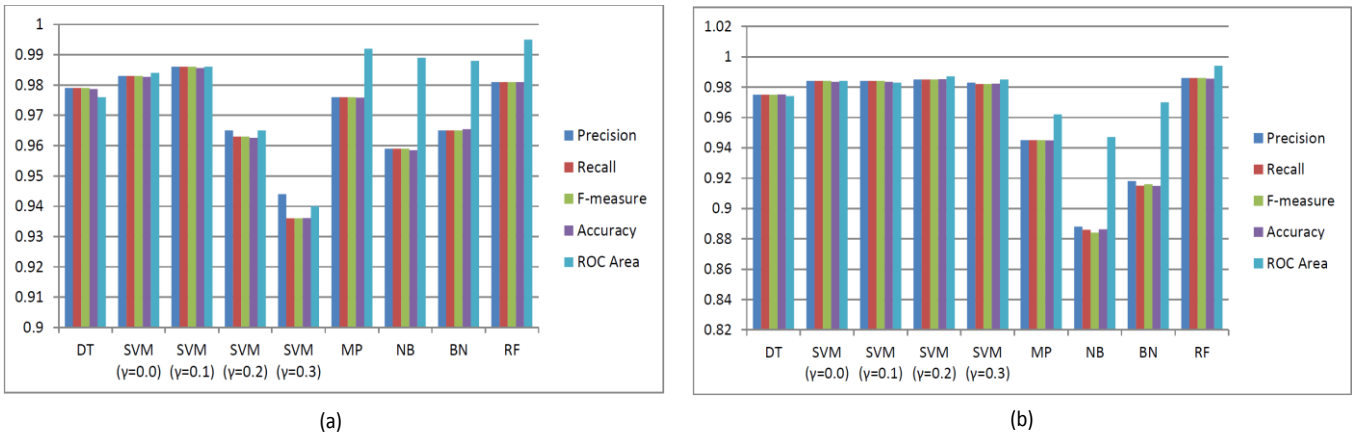

Figure 3. Image-based email spam filtering. The Performance of different machine learning classifiers applied on (a) ISH dataset and (b) Dredze dataset in terms of accuracy, precession, recall, F-measure, and ROC area

From the above results, we can see the effect of the hierarchy on the overall evaluation measures, the overall spam precision is 100%. However, there is a slight decrease in recall, accuracy, and F-measure. This is due to the tradeoffs between FN and FP, because we designed a filter with FP equal zero, this will increase the value of FN. When using SVM in Phase II, we can see that the accuracy decreased compared with the RF classifier. This is due to the relatively high false negative of

the SVM. At the same time, the false positive is less than that when using RF classifier and that is the main point.

D. HESF-Theoretical Scenario II

In this scenario, we study the performance of the proposed hierarchical email spam filter using a theoretical dataset that is assumed to have text and image spam emails with complete header information for these emails. However, a dataset with this specification is not publicly available. Therefore, we used the CEAS2008 dataset which has 32703 emails divided into 26180 spam emails and 6523 ham emails. However, we assume that this contains mixed spam emails of text and images with header information. We further assume that image spam represents 22% of all email spam in the dataset. This percentage is the same percentage of image spam that was reported recently in the literature [38]. Assuming that a dataset with this specification is used as input for the proposed hierarchical email spam filtering system, our objective is then to evaluate the performance of the system in terms of precision, recall, F-measure, FP rate, FN rate, and accuracy.

Phase I: We decided to use the RF classifier in Phase I because this classifier was the best among all other classifiers based on the results obtained in [49] with precision, recall, accuracy and F-measure as follows: Precision = 98.9%, Recall = 99.2%, Accuracy = 98.5%, and F-measure = 99%.

TABLE VI. THE CONFUSION MATRIX FOR RF CLASSIFIER IN PHASE II APPLIED ON IMAGE EMAILS- SECOND SCENARIO

Prediction	Actual	
	Spam	***Ham***
Spam	5680.71	34.29
Ham	1.254	64.746

TABLE VII. THE CONFUSION MATRIX FOR SPAMASSASSIN FILTER IN PHASE II APPLIED ON TEXT EMAILS- SECOND SCENARIO

Prediction	Actual	
	Spam	***Ham***
Spam	19956.1	303.9
Ham	0.1638	233.8362

TABLE VIII. THE PROCESS OF EXTRACTING FEATURES OF THE IMAGE ATTACHED TO AN EMAIL

Prediction	Actual	
	Spam	***Ham***
Spam	25636.81	338.19
Ham	1.4178	298.5822

Phase II: In this phase, we want to filter the predicted spam emails from phase I, because the emails may contain misclassified emails. The confusion matrix values are computed using the Equations (5) and (6) as well. The predicted number of spam emails = TP + FP = 25975 + 300 = 26275 will be subject to further analysis in this phase. Since we assume that 22% of email spam is an image spam, then the number of image spam emails that will be considered by the image-based filter of Phase II = 0.22 × (300 + 25975) = [66 (TP) + 5715 (FP)] = 5781, and the remaining 20494 spam emails will be considered for text-based email spam filter of Phase II. For text-based spam filtering, our discussion is based on the results obtained by Cormack et. al., [47] where SpamAssassin [48] filter was applied on a dataset of 49086 email messages, consisting of 9038 ham and 40048 spam emails. The results reported in [47] were as follows: FP Rate = 0.07%, FN Rate = 1.5%, Precision = 99.98%, Recall = 98.49%, Accuracy = 98.76% and F-measure = 99.23%.

It is to be mentioned that the RF classifier is used for image-based spam filtering of Phase II because it is the classifier of best performance as discussed in [50]. To evaluate the performance of the proposed hierarchical email spam filter, it is needed to build the confusion matrix for each of the image spam filter and text spam filter. The confusion matrix of RF classifier when applied on 5781 image spam emails coming from Phase I is shown in Table VI. The confusion matrix of SpamAssassin filter when applied on 20494 text spam emails coming from Phase I is shown in Table VII.

To calculate the performance metrics for the hierarchical email spam filtering system, we add the confusion matrix in Table VI to the confusion matrix in Table VII. The resulting confusion matrix Table VIII of the Phase II is used to compute the overall performance.

Note that from the confusion matrix of Phase II, FP + TN = 1.4178 + 298.5822 = 300 which is equal to the number of ham emails coming from Phase I, and TP + FN = 25636.81 + 338.19 = 25975 which is equal to the total number of spam emails from phase I. The overall performance of the proposed hierarchical email spam filter is as follows: Overall Precision = 99.99%, Overall Recall = 98.7%, Overall Accuracy = 99.85% and Overall F-measure = 99.34%. Based on these results, it can be seen the effect of the hierarchy on the overall evaluation measures, the precision, accuracy and F-measure are improved strongly compared with the results from one phase only, we notice that the recall decreased slightly. This is due to an increase in FN, but the FP is decreased greatly.

V. CONCLUSION

In this work, we have proposed a Hierarchical Email Spam Filtering system (HESF). The proposed system performs spam filtering in two phases. In Phase I, a header-based that uses features extracted from the mandatory and optional header fields is applied to all incoming email messages. Therefore, achieving fast classification of these emails into ham and spam. In Phase II, we apply content-based filtering mechanisms to confirm that initially classified spam is indeed spam. Non confirmed spam is moved to the ham folder. In this phase, we apply text-based filters for text content and image-based filter for image content. To this end, we study several machine learning-based classifiers and compare their performance in filtering email spam based on email header information in the first phase, and based on the email body in the second phase. These classifiers are: C4.5 Decision Tree (DT), Support Vector Machine (SVM), Multilayer Perception (MP), Nave Bays (NB), Bayesian Network (BN), and Random Forest (RF). We evaluate the proposed work through a combination of

theoretical analysis and experimental studies based on publicly available datasets. Our studies show that:

- The RF classifier outperform all the other classifiers in Phase I with an average accuracy, precision, recall, F-Measure, ROC area of 98.5%, 98.4%, 98.5%, 98.5%, and 99%, respectively.
- RF classifier and SVM classifier outperform all other classifiers in Phase II. RF achieved precision, recall, F-measure, accuracy, and ROC Area of 98.1%, 98.1%, 98.1%, 98.1%, and 99.5% respectively. While the same metrics for SVM classifier (with $\gamma = 0.1$) were as follows: 98.6%, 98.6%, 98.6%, 98.56%, and 98.6%.
- The performance of the SVM classifier is affected by the value of its radial basis kernel (i.e., γ parameter). This means that the value of could be adjusted to obtain the increase or decrease FP while maintaining a good performance for this classifier.
- The overall performance of the system is greatly increased when using hierarchical approach, Hierarchical Email Spam Filter (HESF) achieves a precision of 99.99% and 100% in some case studies with very low false positives.

Our future work will focus on using larger and more recent datasets to validate our results in the two phases. Also, we plan to investigate using other features such as server log information, DNS information, and other optional header fields such as ”Return-path” field for the Phase I and color info, image header info for the Phase II. Also, an additional phase(s) can be added to the hierarchy for further filtrating, if needed.

REFERENCES

[1] C. Kreibichy, et al., “Spamcraft: An Inside Look At Spam Campaign Orchestration,” Proceedings of the Second USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET ’09), Boston, Massachusetts, April 2009.

[2] M. Intelligence, “MessageLabs Intelligence: 2010 Annual Security Report,” 2010.

[3] Symantec. March 2011 Intelligence Report. Available at: http://www.symantec.com/about/news/release/article.jsppri d=20110329 01

[4] S. Hinde, “Spam, scams, chains, hoaxes and other junk mail,” Computers & Security, vol. 21, pp. 592 - 606, 2002.

[5] A. R. B. Blog. October, 2010, The Dangers of SPAM. Available: http://www.anthonyricigliano.info/the-dangers-of-spam/

[6] A. C. Solutions. January 7, 2011 Statistics and Facts About Spam. Available: http://www.acsl.ca/2011/01/07/statistics-and-facts-about-spam/

[7] H. R. Cournane A, “An analysis of the tools used for the generation and prevention of spam,” Computers & Security, vol. 23, pp. 154-66, 2004.

[8] R. JENNINGS. JANUARY 28, 2009, Cost of Spam is Flattening - Our 2009 Predictions. Available at: http://ferris.com/2009/01/28/cost-of-spam-is-flattening-our-2009-predictions/

[9] E. Blanzieri and A. Bryl, “A survey of learning-based techniques of email spam filtering,” Artif. Intell. Rev., vol. 29, pp. 63-92, 2008.

[10] G. Cormack and T. Lynam, “Spam corpus creation for TREC,” in Proceedings of Second Conference on Email and Anti-Spam CEAS, 2005.

[11] H. Stern, “A Survey of Modern Spam Tools,” CEAS’08, 2008.

[12] A. G. K. Janecek, et al., “Multi-Level Reputation-Based Greylisting,” in Availability, Reliability and Security, 2008. ARES 08. Third International Conference on, 2008, pp. 10-17.

[13] A. Banit, et al., “Controlling spam Emails at the routers,” in Communications, 2005. ICC 2005. 2005 IEEE International Conference on, 2005, pp. 1588-1592 Vol. 3.

[14] Y. Hu, et al., “A scalable intelligent non-content-based spam-filtering framework.,” Expert Syst. Appl., vol. 37, pp. 8557-8565, 2010.

[15] J.-J. Sheu, “An Efficient Two-phase Spam Filtering Method Based on E-mails Categorization “ International Journal of Network Security, vol. 9, pp. 34-43, July 2009.

[16] C.-H. Wu, “Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks,” Expert Systems with Applications, vol. 36, pp. 4321-4330, April, 2009

[17] M. Ye, et al., “A Spam Discrimination Based on Mail Header Feature and SVM,” presented at the Wireless Communications, Networking and Mobile Computing, 2008. WiCOM ’08. 4th International Conference on Dalian Oct. 2008.

[18] C.-C. Wang and S.-Y. Chena, “Using header session messages to anti-spamming,” Computers & Security, vol. 26, pp. 381-390, January 2007.

[19] V. P. Pedram Hayati, “Evaluation of spam detection and prevention frameworks for email and image spam: a state of art,” presented at the Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services, Linz, Austria, 2008.

[20] I. P. Giorgio Fumera, Fabio Roli “Spam Filtering Based On The Analysis Of Text Information Embedded Into Images,” J. Mach. Learn. Res., vol. 7, pp. 2699-2720, 2006.

[21] B. B. G. F. I. P. F. Roli, “Image Spam Filtering by Content Obscuring Detection,” presented at the Fourth Conference on Email and Anti-Spam (CEAS 2007), Mountain View, California, 2007.

[22] G. F. I. P. F. R. B. Biggio, “Image spam filtering using textual and visual information,” presented at the MIT Spam Conference 2007, Cambridge, MA, USA 2007.

[23] P. Klangpraphant and P. Bhattarakosol, “PIMSI: A Partial Image Spam Inspector,” in Future Information Technology (FutureTech), 2010 5th International Conference on, 2010, pp. 1-6.

[24] F. Gargiulo and C. Sansone, “Combining visual and textual features for filtering spam emails,” in Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, 2008, pp. 1-4.

[25] W. Ching-Tung, et al., “Using visual features for anti-spam filtering,” in Image Processing, 2005. ICIP 2005. IEEE International Conference on, 2005, pp. III-509-12.

[26] L. Qiao, et al., “Efficient Modeling of Spam Images,” in Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on, 2010, pp. 663-666.

[27] B. Mehta, et al., “Detecting image spam using visual features and near duplicate detection,” presented at the Proceeding of the 17th international conference on World Wide Web, Beijing, China, 2008.

[28] S. Krasser, et al., “Identifying Image Spam based on Header and File Properties using C4.5 Decision Trees and Support Vector Machine Learning,” in Information Assurance and Security Workshop, 2007. IAW ’07. IEEE SMC, 2007, pp. 255-261

[29] R. G. Mark Dredze, Ari Elias-Bachrach., “Learning Fast Classifiers for Image Spam.,” presented at the in Proc. CEAS 2007, Mountain View, California, August 2-3, 2007.

[30] W. J. Zhe Wang, Qin Lv, Moses Charikar, Kai Li., “Filtering Image Spam with Near-Duplicate Detection,” presented at the In Proceedings of the Fourth Conference on Email and AntiSpam, CEAS’2007, 2007.

[31] R. Islam and Z. Wanlei, “Email Categorization Using Multi-stage Classification Technique,” in Parallel and Distributed Computing, Applications and Technologies, 2007. PDCAT ’07. Eighth International Conference on, 2007, pp. 51-58.

[32] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. “The WEKA Data Mining Software: An Update. SIGKDD Explorations”, 2009.

[33] P. R. Network Working Group, Editor. (April 2001, Request for Comments RFC 2822 Available: http://tools.ietf.org/html/rfc2822.html

[34] CEAS 2008 Live Spam Challenge Laboratory corpus. Available from: http://plg1.uwaterloo.ca/cgi-bin/cgiwrap/gvcormac/fooceas.

[35] R. Beverly and K. Sollins, "Exploiting Transport-Level Characteristics of Spam," presented at the CEAS, Mountain View, CA, August 2008.

[36] C. GROUP. (2010, Spam email datasets, CSDMC2010 SPAM corpus. Available: http://csmining.org/index.php/spam-email-datasets- .html

[37] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters - Special issue: ROC analysis in pattern recognition, vol. 27, pp. 861-874, June 2006

[38] I. X. F. Report. (May, 2009, Image spam - reborn and trying to rejuvinate YOUR health! Available: http://blogs.iss.net/archive/image-spam-rebirth.html

[39] M. S. Andrzej Materka "Texture Analysis Methods - A Review," Institute of Electronics, Technical University of Lodz, Brussels 1998.

[40] P. S. Andrzej Materka "MaZda User's Manual " Instytut Elektroniki Politechnika Lodzka, Lodz, Poland1998-2005.

[41] P. M. Szczypinski, et al., "Mazda - a software for texture analysis," in Information Technology Convergence, 2007. ISITC 2007. International Symposium on, 2007, pp. 245-249.

[42] T. Xiaoou, "Texture information in run-length matrices," Image Processing, IEEE Transactions on, vol. 7, pp. 1602-1609, 1998.

[43] R. M. Haralick, "Statistical and structural approaches to texture," Proceedings of the IEEE, vol. 67, pp. 786-804, 1979.

[44] G. Yan, et al., "Image spam hunter," in Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, 2008, pp. 1765-1768.

[45] R. Bhuleskar, et al., "Hybrid Spam E-mail Filtering," in Computational Intelligence, Communication Systems and Networks, 2009. CICSYN '09. First International Conference on, 2009, pp. 302-307.

[46] M. R. Islam, et al., "MVGL Analyser for Multi-classifier Based Spam Filtering System," in Computer and Information Science, 2009. ICIS 2009. Eighth IEEE/ACIS International Conference on, 2009, pp. 394-399.

[47] G. Cormack and T. Lynam, "Online supervised spam filter evaluation," ACM Trans. Inf. Syst., vol. 25, 2007.

[48] SpamAssassin, "The apache spamassassin project. http://spamassassin.apache.org/," ed, 2005.

[49] O. Al-Jarrah, I. Khater, B. Al-Duwairi, "Identifying Potentially Useful Email Header Features for Email Spam Filtering," The Sixth International Conference on Digital Society, ICDS 2012, January 30 - February 4, 2012 - Valencia, Spain.

[50] B. Al-Duwairi, I. Khater, O. Al-Jarrah, "Detecting Image Spam Using Texture Features," International Journal for Information Security Research (IJISR), Volume 2, Issues 3/4, September/December 2012, pp 344-353.