Project Report:
Performance Evaluation of Intelligent Spam Classifiers with Split Local-Concentration-Based Feature Extraction Approach

Course ECE 9601A

Vasundhara Sharma
M.E.Sc. Student
Department of Electrical and Computer Engineering
University of Western Ontario
London, ON, Canada
vvasundh@uwo.ca

Abstract --- With the ever-growing demand for the Internet services and e-commerce, e-mail stands as the prime mode of communication. However, this led to the increased circulation of spam mails. Spam impacts productivity as well as exposes users to various other security threats. Applying spam filters is a good pro-active and preventive measure. Exploring artificial intelligence techniques for anti-spam, in this Project, we analyze the work of Zhu et al. [1] which proposes local concentration (LC)-based feature extraction approach for anti-spam which extracts position correlated information from the messages which is further used to contruct feature vector. Taking inspiration from their work, we propose a extended concept of Split LC-based feature extraction approach. Feature vector are constructed utilizing the above feature extraction model with variable-length sliding window. In addition, we evaluate the performance of machine learning (ML) classifiers with the Split LC-based features vector as input. To leverage the analysis, we conduct several experiments using the benchmark PU2 mail corpora with ten-fold cross validation techniques using support vector machines (SVM) and artificial neural networks (ANN). The efficacy of the proposed model is compared with the existing feature extraction approaches including LC-based approach [1]. It is observed that Split LC-based feature extraction approach peforms better than other prevalent approaches with higher prediction accuracy and is computationally efficient, scalable and robust. Further, we show that ANN classifier out performs SVM consistently with higher prediction accuracy. We demostrate and recommend the use of ANN classifier with Split LC-based feature extraction approach for effective spam filtering

Keywords— Artificial Immune system (AIS), feature extraction, bag-of-words (BoW), Spam Filtering, local concentration (LC), support vector machines (SVM), artificial neural netowrks (ANN)

I. INTRODUCTION

With the advances in Internet technologies, electronic mails are the most economical and fastest mode of communication on professional as well as personal front. However, with the increasing use of Internet and with the growing size of e-mail users, the problem of receiving unsolicited commercial e-mail (UCE) or unsolicited bulk e-mail (UBE), commonly referred as Spam, has seen tremendous growth as well. The social networking channels and e-advertisers often make use of spam to send identical e-mails to numerous recipients in bulk. Spam mails involve productivity costs as well as occupy storage spaces and network bandwidth. The problem of receiving spam can be dated back to early 1990's with increasing trend over the years. Moreover, in recent years, the magnitude of this problem has meandered into serious security threats when spam got associated with phishing tools for stealing sensitive information from the e-mail users. Spammers collect the e-mail addresses from social networking sites, web groups, customer visited websites, customer databases and exploit social media for infecting computers with malware and UBE. With the advent of cloud computing technologies and the growing trend for bringing own devices to workplace is posing serious threats to business through the compromised systems. Mobile malware attacks, like mobile botnets, has tremendously increased with the growing number of mobile devices and smartphone users.

To combat the various problems and threats posed by spammers, there is a need for filtering spam mails effectively as well as efficiently owing to the large mail databases. Spam filtering basically refers to the identification and separation of the spam mails from the legitimate mails. Spam filters are the

grave requirement for not just the government and business establishments but also the individual e-mail users. Most of the existing solutions for spam filtering involve filtering rule sets based on few selected words or phrases and the list of trusted sites. However, this methodology involves frequent manual intervention in order to update the information database. Thus, automatic detection of spam remains a challenging and open area of research interest. Some of the prevalent machine learning (ML) approaches for spam filtering is been discussed in Section II of this paper.

In this project, we propose a split local concentration (LC)-based feature extraction approach for anti-spam which is derived from the work of Zhu *et al.* [1]. Zhu *et al.* propose a LC-based feature extraction approach for anti-spam which is inspired from the functioning of the biological immune system (BIS). Split LC-based feature extraction approach transforms selective portions of mails into corresponding LC feature based on the position correlated information extraction for spam and legitimate mails. These feature vectors are used for training the ML classifiers. The trained ML classifiers then independently classify the spam mails from the fresh set of unlabeled mails. Further, we present a comparative analysis on the performance of two machine learning (ML) classifiers; Artificial Neural Networks (ANN) and Support Vector Machine (SVM) for spam detection using Split LC-based feature extraction approach.

The remainder of the project report is organized as follows. In Section II, we present the detailed literature survey covering the related research works on artificial intelligence (AI) based ML approaches for spam filtering. Section III and IV present the project methodology and the Split LC-based feature extraction approaches used in the project. Simulation details and the Experiment setup is presented in Section V and VI respectively. This is followed by detailed result analysis in Section VII. Section VIII provides the conclusion and related future works.

II. LITERATURE REVIEW

Spam emails are of major concern for Internet communication which incur productivity costs, exposure to security threats and annoying user experience. Several research works try to develop optimal solutions for filtering malicious traffic at the backbone network level. However, since spam mail may use unicast email id (addressing individual valid users), it may not get filtered successfully at the router level. Spam filtering is the next best corrective measure which can be deployed for automatically filtering and removal of spam from the user inbox. Several researchers have tried different approaches and methodologies for effective spam filtering. Some of the related works are discussed as follows:

Paulo et al. [21] present a novel distributed data mining approach for spam filtering. This method involves combining the content-based filtering (CBF) and collaborative filtering (CF) to extract distinct features for personalized spam filtering without compromising on the privacy and security.

Another technique for designing spam filters is to create blacklists based on the IP addresses of the spammers or compromised hosts. This can be fine-tuned to include user ratings and periodic review and update of the spam database. Some of these strategies are summed up in [25] which discusses the statistical data compression models for spam filtering.

Use of Artificial Intelligence (AI) and Machine Learning (ML) techniques for spam filtering is demonstrated in the work of Yang et al. [22]. They demonstrate the effectiveness of email classifiers based on feed forward back propagation neural networks and Bayesian filters. Results from this paper qualify feed forward back propagation neural network with higher predictive capabilities over the Bayesian filters. Moreover, Bayesian filters can get by-passed by smart spammers who resort to using sophisticated and convincing words in the mail to sound genuine. Shen et al. [23] propose a social network aided personalized and effective spam filter (SOAP). In SOAP framework, information is collected from the social network link which form a distributed overlay network. SOAP utilizes the social relationships, interests, and trust management to extract spams adaptively. Shen et al. verify the SOAP model with simulations utilizing facebook trace data. SOAP seems to perform better than the Bayesian filters.

Recently, spammers have started embedding images in spam mails to combat the various spam filters which work on the feature extraction from the text mails. These image spams successfully escape from the textual spam filters and make their way to user inbox. Image spam control is the most recent area of research. Ketari et al. [24] discuss the various image spam filtering strategies under study. Some of the intelligent techniques for filtering image spam, involve converting the pixels of images into a gradient histogram which is normalized and pre-processed to generate feature vector. Active learning with sparsely induced similarity method for evaluating image spam performs well with high prediction rate [26].

None of the spam filtering techniques can guarantee the complete solution owing to the varying magnitude of the problem. However, The Survey of the Internet Security Threats of 2013 based on the recent report released by Symantec Corp. [27] presents slight declining trend in spam rate between the years 2010 and 2012. Tore-downs of some of the massive spam botnets on Internet can be attributed for this achievement. However, spam continues to be a problem with 69% persistent grip overall. Fig. 1. Shows the estimated global spam rate for three consecutive years as per the Symantec threat report.

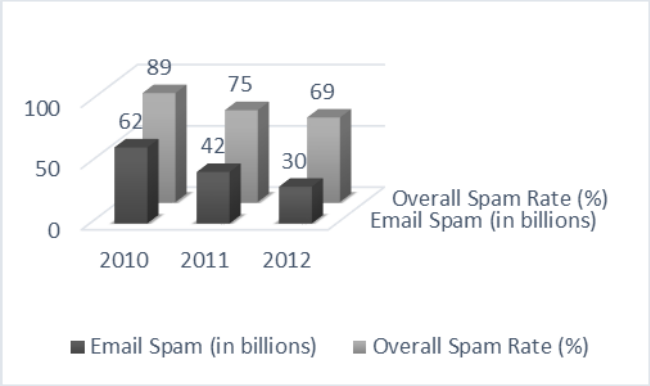

Fig. 1: Summary of the global spam rate in three consecutive years (data based on Symantec Corp. Internet Security Threat Report 2013::Vol 18)

We also extract cost related information for successful spam/phishing attack from the Symantec survey report which is represented in Fig. 2. This data high-lights the importance and requirement for effective spam filtering.

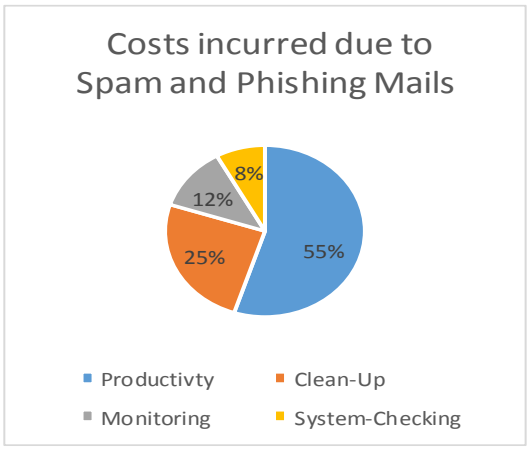

Fig. 2: Graph representing the Costs incurred in case of unsuccessful spam or phishing filtering (data based on Symantec Corp. Internet Security Threat Report 2013::Vol 18)

III. PROJECT METHODOLOGY AND SPLIT LC-BASED MODEL

The importance of designing effective spam filtering is evaluated in this project work. In addition to the LC-based approach followed by Zhu et al. [1], we propose split LC-based approach for segregating the local concentration of terms in spam and legitimate mails for effective spam filtering. SVM and ANN classifiers are trained with the LC-based feature vectors and the better classifier is derived based on the results of the experiment. The overall design of the proposed model is shown in Fig. 3.

A. Background

Zhu et al. designed LC model for spam filtering which is inspired from the BIS. BIS is an adaptive model in the humans which acts as a protection from the attacks of pathogens. It protects the system by applying the capability of distinguishing between "self-cells" and "foreign-cells". Lymphocytes in the blood produce antibodies, which keep circulating in the body and kill pathogens found nearby. BIS has two types of response mechanisms for the pathogens. Primary response develops when the pathogen appears in the system for the first time and it is tackled with the slow production of antibodies. The occurrence of this pathogen is stored in the body in the form of long-lived B memory cells which act much faster on the successive recurrence of the same pathogens in the body. Inspired from the functioning of BIS, AIS was developed in early 1990's [9]. Zhu et al. apply this concept for spam detection. Detector sets (antibodies) are developed which can discriminate spam from the legitimate mails based on term selection methods. These detector sets (DS) are used for feature extraction and construction of LC-based feature vectors.

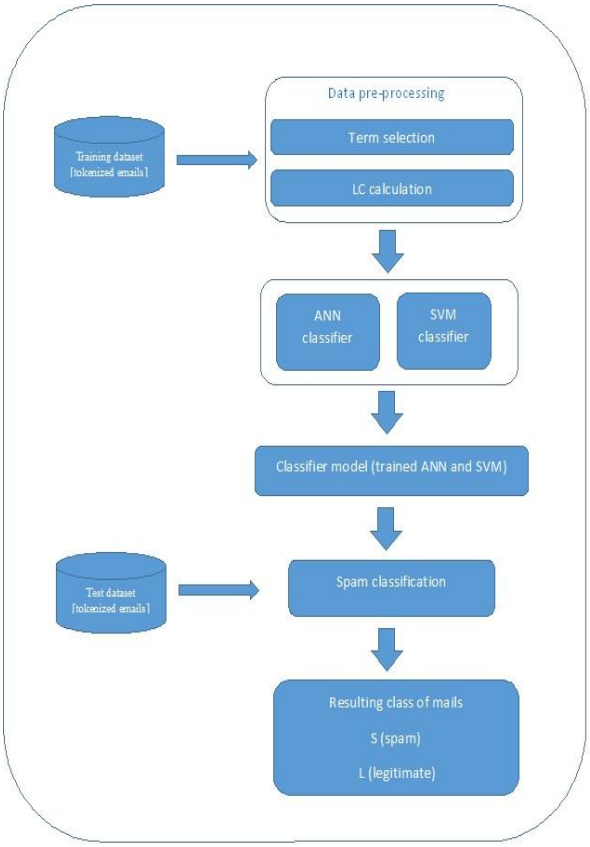

Fig. 3: Design of the proposed split LC-based spam classifier system

B. Structure of Split LC Model

To implement more effective spam filtering using LC-based model, we propose a generic system design which is represented in Fig. 3. Tokenized encrypted mails (message categorized in individual terms based on white spaces and delimiters) are taken as the input for pre-processing which involves term selection and LC calculations.

1) **Term selection:** Term selection methods are used to remove the less important terms from the datasets. This helps to reduce the computational complexity of large datasets which involve huge number of terms. There are various term selection methods like Information Gain (IG), Term Frequency Variance (TFV) and Document Frequency (DF). We choose to use DF in our work. DF calculates the number of documents in which a certain term appears and the terms whose DF falls below a predefined threshold values are removed from the set of terms. DF is a simple yet highly effective term selection method which eliminates the rare terms which are not equally useful in comparison with high frequency terms. Moreover, the choice of selecting DF in our project lies on the major factor that the DF has linear

computational complexity and hence can be effectively put to use for larger number of training messages.

2) **LC calculation:** Some of the prevalent feature extraction approaches used for spam filtering are Bag-of-Words (BoW) [13], Sparse Binary Polynomial Hashing (SBPH) [15]. We propose a novel concept of feature extraction approach of Split LC-based model which is derived from [1]. Instead of calculating tendency threshold of terms as considered in [1], we split the terms in spam and legitimate classes at the term selection stage itself and generate the detector sets (antibodies) for spams and legitimate mails. Further, we apply variable length sliding window over the messages to calculate the spam and legitimate densities (called as genes) which is used to construct the LC-based feature vector. The details of the feature extraction approach and the proposed algorithm is presented in section IV.
3) **Classification:** The tokenized messages from the training data set are pre-processed using term selection and LC calculation to generate feature vectors. These feature vectors are used as inputs for training the ML classifiers. Supervised learning is deployed for training the classifiers. The resulting classifier model is applied to unlabelled messages to predict the classification as spam and legitimate. Apart from generating LC-based feature vectors for spam detection, we also focus on choosing the better classifier which fits well in the spam filtering scenario based on classification accuracy and prediction consistency. We consider ANN and SVM for the performance analysis and present the details of the obtained results in section VII.

IV. SPLIT LC-BASED FEATURE EXTRACTION APPROACH

A. Motivation

Deriving the right set of features is of prime importance for Spam filtering. BoW is the most prevalent approach currently been used but it suffers from the problem of high-dimensionality of feature vectors. This affects the overall computational performance of the classifiers. Also, spam and legitimate emails may just vary on some portion of the mail content and hence extracting position related information is vital. Zhu et al. [1] propose LC approach to address these issues. They derive position related feature vectors (antibodies) for detecting and removing spam which draws inspiration from BIS. Taking inspiration from [1], we derive Split LC-based approach which differentiates the messages into spam and legitimate and then calculates the different detector sets based on the document frequency and likelihood of the terms to be present in spam or legitimate. This is used for deriving the feature vectors for training the classifiers.

B. Novel approach for DS Generation

The steps involved in DS generation from the tokenized emails are explained in Algorithm 1. The terms are selected using DF method in this project. Term selection method ensures that only the important terms (based on the number of occurrences across all message documents) are filtered out and the remaining terms are discarded. Term selection is important for reducing computational complexity of the LC calculation step. It also helps in eliminating possible noises introduced at training stage due to the inclusion of uninformative terms. We take a novel approach at the term selection stage which is different from Zhu et al. [1]. Preselected sets are initialized as empty sets at start. We separate the mail messages into spam and legitimate and extract the terms for each using DF. This results in two different preselected sets for spam and legitimate each. The high frequency terms from the two preselected sets are compared for their tendency to be found in the Spam or Legitimate mails which results in the creation of DS for Spam and Legitimate mails.

ALGORITHM 1

```
Initialize preselected set and DS as empty sets;
Generate terms set T_S and T_L from spam and legitimate tokenized emails;
for each term in the terms set T_S and T_L
do
   Calculate importance of the term according DF term selection method;
end for

Sort the terms in descending order of the importance;
Add the front m% terms to the preselected sets for Spam and legitimate each;
for each term in the preselected sets for Spam and legitimate do
   Calculate Tendency of the term according to number of Occurrences;
   if Tendency of Occurrence is high for Spam and Legitimate then
        Add the term to DSs and DS_L;
   else if Tendency of Occurrence is high for Spam
        then
        Add the term to DSs;
   else
        Add the term to DS_L;
   end if
end for
```

Algorithm 1: Split LC-based Novel Approach for Term Selection and DS Generation

According to our proposed Algorithm 1, we add the term with higher frequency of occurrence in spam than legitimate mails to the detector set (DSs), which represents the terms corresponding to spam genes. Similarly, the terms with higher frequency of occurrence in legitimate than spam are added to the detector set (DS_L) which represents the terms corresponding to legitimate genes. The two DS (DSs and DS_L) are used for generating the LC-based feature vectors explained in the section IV.C.

This method for DS construction is better than the probabilistic approach taken by Zhu et al. as our approach lays stress on splitting the spams and legitimate DS based on the occurrence of terms, and same term may remain in both DSs and DS_L. This eliminates the faulty training of the classifier due to discriminated DS. For the purpose of better understanding we present an example: both the spam and legitimate mails can have the following subject line: "*Urgent Notification needing your Attention*", now if we move all the 5 tokens "*Urgent", "Notification", "needing", "your", "Attention*" to either Spam or Legitimate DS, then based on the size of the data set and number of mails of each category, it would result in faulty training of the ML classifier. We can further claim that our novel Split LC-based feature extraction approach is more accurate based on the results obtained from experiments which is explained with the Result Analysis in Section VII.

C. LC-Based feature Vector Construction

We retain the approach of Zhu et al. [1] for the calculation of LC-based feature vector. Sliding window of w_n terms is moved over the messages with a step of w_n terms in each round. With each movement of the window, a spam gene concentration SC_i and a legitimate gene concentration LC_i based on DSs and DS_L and the terms of the message captured by the window is calculated as follows:

$$SC_i = N_s / N_t \quad (1)$$

$$LC_i = N_l / N_t \quad (2)$$

where N_t is the number of distinct terms in the window i, N_s is the number of the distinct terms in the window matched with DSs and N_l is the number of distinct terms in the window matched with detector set DS_L. With each round of window movement, a pair of spam and legitimate gene concentration {SCi, LCi} is generated. The combination of all sets of gene concentration forms the whole LC-based feature vector which is used as input for the ML classifier. Algorithm 2 shows the process of constructing feature vector [1].

D. Sliding Window strategy for defining Local Areas

Zhu et al. apply two distinct sliding window strategies for defining local areas in messages [1]. They use - fixed length sliding window (FL) and variable length sliding window (VL).

1) *Using Fixed-Length sliding Window:* With FL strategy, a predefined window size is applied to each round of LC-based feature vector calculations. For shorter messages, the latter part of feature vectors are padded with zeroes. For example, a feature vector, {(SC_1, LC_1), (SC_2, LC_2), (SC_3, LC_3)} is expanded as {(SC_1, LC_1), (SC_2, LC_2), (SC_3, LC_3), (0, 0), (0, 0)}. The other alternative which they use is to replicate the front features. For example, the same feature vector would be expanded as {(SC_1, LC_1), (SC_2, LC_2), (SC_3, LC_3), (SC_1, LC_1), (SC_2, LC_2)}. For longer messages, the dimensionality is reduced by truncating the terms at the end of the messages [1].
2) *Using Variable-Length sliding Window:* With VL strategy, the length of sliding window is set in proportion to the size of message. For example, for constructing 2N dimensional feature vector, the window size is set to M/N for a message of length M [1]. This helps to maintain the same dimensionality of the feature vector irrespective of the size of the message. VL accommodates both the small and large messages without padding or duplicating the feature vectors. This strategy produces unique feature vectors and extracts important information from the messages. FL may suffer from the problem of misleading or over-optimistic predictions because of the replication of feature vectors.

Based on the above discussed additional advantages of VL over FL and project constraints, we choose to apply VL window strategy in our project experiments.

ALGORITHM 2

Move a sliding window of w_n -term length over a given message with a step of w_n -term;
for each position i of the sliding window do
  Calculate the spam genes concentration SC_i of the window according to (1);
  Calculate the legitimate genes concentration LC_i of the window according to (2);
end for

Construct the feature vector likes:
{(SC_1, LC_1), (SC_2, LC_2),..., (SC_n, LC_n)}.

Algorithm 2: Construction of Split LC-based Feature Vector [1]

V. SIMULATION DETAILS

A. Experimental Corpora

Project experiments are conducted on benchmark corpora PU2. The corpora consists of 721 messages in total, out of which 142 are spam. The messages in this corpora are pre-processed with the removal of attachments, HTML tags, and email headers. All the legitimate and spam mails in the corpora are English messages with numerical encryption. The message files contain the subject line and the mail body. Duplicate mails are removed to prevent over optimistic and

faulty training of the spam classifiers. The corpora is available online at site: http://www.aueb.gr/users/ion/publications.html

B. *Simulation Environment*

Project experiments are simulated using MATLAB neural network toolbox and SVM libraries. The system used for implementing the simulations has Intel core i5 with 4 Gig RAM. The simulation is executed in three stages, pre-processing and generation of detector sets, generating LC-based feature vectors by running variable length sliding window over the messages, training and evaluating the performance of SVM and ANN classifiers. The details of the source code files used in the project can be found in the appendix section X.

VI. EXPERIMENT SETUP

A. *Choice of ML classifiers*

Several ML methods are available in the field of classifier design. Some of the prominent ML methods which can be utilized for spam filtering are: support vector machines (SVM), artificial neural networks (ANN), naïve bayes (NB), k-nearest neighbor (k-NN), and artificial immune systems (AIS). Kumar et al. [17] evaluate the performance of various ML classifiers in the field of spam filtering. Based on the predictive accuracy and consistency of available methods and to add wider experimental coverage, we select to utilize ANN and SVM classifiers for the experiments in this project.

1) ***ANN:*** ANN draws inspiration from the biological neural networks. It is constructed with input, hidden and output layer combination. Some of the key features of ANN which make it suitable for spam filtering in this project are: capability to deal with unseen patterns, capability to utilize generalized patterns from training sets, robustness to noise. Multi-Layer Perceptron (MLP) with feed-forward network with stochastic gradient descent architecture is utilized for this project. The activation function is sigmoid and the number of hidden layers in the network is 3. No more than 3 layers are required in MLP feed –forward networks as three-layer net is capable of generating arbitrary complex decision regions [18]. It may suffer from the issue of local minima or over optimization which is taken care by taking mean value of 20 readings for each experiment. We make use of MATLAB neural network toolbox for simulating the experiments.

2) ***SVM:*** SVM constructs set of hyperplanes for multi-dimensional, non-linearly separable data and separates them across maximum margin classifier. This method can be effectively deployed for spam filtering which can classify the data into spam or legitimate. SVM's robustness and ability to deal with large feature spaces, captivates the attention of researchers for spam filtering. Several kernel functions can be used for mapping the linearly inseparable data to higher dimensional feature space. We evaluate the performance of linear, polynomial, quadratic, radial basis function (RBF) and MLP kernels in our project. We make use of MATLAB SVM libraries for simulating the experiments.

B. *Evaluation Criteria*

Various evaluation criteria has been designed in the area of spam filtering [13]. We adopt the following four evaluation criteria for evaluating the performance of spam filters in this project:

1) ***Spam Recall:*** Measures the percentage of spam that can be filtered by a classifier model. Higher spam recall infuses trust in the classifier for effective spam filtering. Spam recall can be represented in terms of number of spams prediction count.

$$\text{Spam Recall} = \frac{N(s,s)}{N(s,s) + N(s,l)} \quad (3)$$

where:
N(s,s): no. of spams correctly classified as spam
N(s,l): no. of spams incorrectly classified as legitimate

2) ***Spam Precision:*** Measures the precision in spam prediction, which is the number of actual spam from the total classification figures. It also reflects the number of false positives for the legitimate mails. High spam precision reflects lesser number of legitimate misclassifications.

$$\text{Spam Precision} = \frac{N(s,s)}{N(s,s) + N(l,s)} \quad (4)$$

where:
N(s,s): no. of spam mails correctly classified as spam
N(l,s): no. of legitimate mails incorrectly classified as spam

3) ***Accuracy:*** Measures the overall prediction performance of the classifier. It measures the total percentage for the correct classification of spam and legitimate mails.

$$\text{Accuracy} = \frac{N(l,l) + N(s,s)}{N(s) + N(l)} \quad (5)$$

where:
N(l,l): no. of legitimate mails correctly classified as legitimate
N(s,s): no. of spam mails correctly classified as spam
N(s): no. of spam mails in the corpus
N(l): no. of legitimate mails in the corpus

4) ***F_β measure:*** Measures the combined result of Spam Recall and Spam Precision. Weight β is assigned to spam precision. F_β combines the different analysis measure used by Spam Recall and Spam Precision to present the overall performance with respect to spam filtering analysis.

$$F_\beta = (1+\beta^2) \times \frac{(\text{Spam Recall}) \times (\text{Spam Precision})}{\beta^2(\text{Spam Precision}) + (\text{Spam Recall})} \quad (6)$$

When β is set to 1, this evaluation criteria is known as F_1 measure which is used in this project.

VII. RESULT ANALYSIS

A. Performance Evaluation for SVM Classifier

We conduct spam filtering experiments with SVM classifier using the different kernel functions. In this section, we discuss the simulations results based on the evaluation criteria discussed earlier in Section VI.B. We consider the mean value of 20 readings in the result estimation. To re-iterate, all the experiments are conducted on PU2 corpora. The corpora details are present in Section V.A. SVM classifier with ten-fold cross-validation is used, and DF was used as the term selection method. The results of the experiments is shown in Table I. We further consider evaluating polynomial kernel function with degrees 1, 2 and 3. As from the Table I, we see that Linear and Polynomial (degree 1) and RBF performed well in terms of spam recall and it is interesting to note that Linear and Polynomial (degree 1) kernel function show identical performance in terms of number of correct and incorrect classifications for both spam and legitimate mails. Therefore, all the evaluation criteria values are identical. It is also to be noted that with increasing degree polynomial function, a downward trend in the performance is observed. Based on the results of the experiment, we can suggest the use of linear or polynomial (degree 1) kernel functions for spam filtering.

TABLE I

SVM PERFORMANCE EVALUATION FOR SPAM FILTERING USING DIFFERENT KERNEL FUNCTIONS

SVM Kernel Method	Spam Recall (%)	Spam Precision (%)	Accuracy (%)	F_1 (%)
Linear	97.61	87.23	96.71	92.12
Polynomial (degree 1)	97.61	87.23	96.71	92.12
Polynomial (degree 2)	92.85	86.66	95.77	89.65
Polynomial (degree 3)	88.09	86.05	94.84	87.06
Quadratic	92.85	84.75	95.31	88.63
RBF	97.61	82.00	95.31	89.13
MLP	83.33	61.40	86.62	70.70

Overall, the accuracy of SVM classification with linear kernel function is 96.71% and F_1 measure obtained is 92.12%. The split LC-based feature extraction approach slightly out performs than the simple LC-based approach [1] for the same mail corpora PU2.

B. Performance Evaluation for ANN Classifier

In this section, we discuss the results for ANN classifier model. We conduct ANN experiments using different number of hidden neurons (3, 4, 10) while designing spam filtering classifier models. The validation technique is maintained as 10-fold cross validation and stochastic gradient descent architecture with sigmoid function is used. The experiments are performed on the PU2 corpora and mean value of 20 runs is used for calculating the evaluation measures. The obtained result is presented in Table II.

TABLE II

ANN PERFORMANCE EVALUATION FOR SPAM FILTERING WITH DIFFERENT NUMBER OF HIDDEN NEURONS

No. of hidden neurons	Spam Recall (%)	Spam Precision (%)	Accuracy (%)	F_1 (%)
3	97.6	98.2	98.1	97.8
4	95.2	100.0	99.1	97.5
10	97.6	99.4	99.2	98.5

Based on the results shown in Table II, ANN performs consistently well for spam filtering. The average prediction accuracy is close to 98.5% which is substantially good. Also, we notice that the results do not seem to vary marginally for the different number of hidden neurons in ANN model. The accuracy figures are the mean values of 20 readings which eliminates the concern of over-optimized result prediction. Moreover, the variance values obtained for the 20 readings for the different number of hidden neurons (3, 4, 10) are 0.12, 0.1 and 0.12 respectively. This shows that the issue of local minima or maxima was not considerably encountered during the experiments.

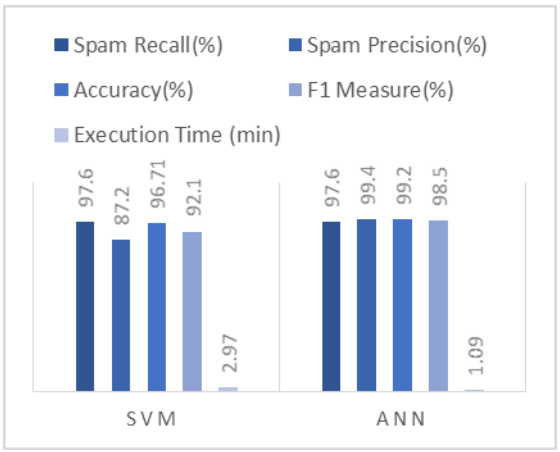

Fig. 4: Performance evaluation for SVM and ANN for spam filtering across PU2 data set

The bar graph presented in Fig. 4 summarizes the comparative performance analysis for ANN (10 hidden neurons) and SVM (linear kernel). It is clearly evident that ANN outperforms SVM for all the evaluation criteria outlined for spam filtering. We also include the classifier execution time for the same corpus data. ANN takes around 1.09 minutes for the whole operation (creation of Split LC-based feature vector time is not included) while SVM takes almost double time of 2.97 minutes for classifying data from 721 mails.
We have sufficient experimental proofs to preliminarily qualify ANN model as a good classifier choice for spam filtering using Split LC-based feature extraction method.

C. Experiments with LC Parameter Selection

Split LC-based feature extraction approach proposed in this project includes two parameter selection: (1) feature dimensionality and (ii) percentage of terms selected for constructing DS. We consider using VL sliding window which adjusts the size of window based on the feature dimensions as discussed in Section IV.D. By default, the percentage of terms selected for creating DS is 50%. We conduct some experiments to analyze the impact of varying these parameters and to further fine tune it to optimize the model. We conduct these experiments on the benchmark corpora PU2 with ten-fold cross validation and DF as the term selection method. The classifier is selected as ANN with 10 hidden neurons.

1) Selection of Proper Feature Dimensionality

Split LC-based model or the other existing LC-based model [1] brings out the biggest advantage of reduced feature-dimensionality without compromising on the performance as compared to the other existing feature extraction models like BoW. To decide on the optimal number of front sliding windows to be used for feature vector creation, we conduct this experiment with varying number of sliding windows. The feature dimensionality is two times the number of windows utilized as we generate two feature vectors (SC, LC) at each step as discussed in Section IV.C. The resulting graph is presented in Fig. 5 which shows that the model performs best with 3 or 5 utmost front windows. This represents that selection of 6 or 10-dimensional feature vectors is the optimal choice for better classification accuracy. Also, we notice computational overhead with reduced performance with increasing number of feature dimensionality. For remaining experiments, we consider using 6-diemensional feature vector for both ANN and SVM classifiers.

2) Selection of Percentage of Terms for constructing DS

Term selection determines the selection of more informative and distinctive tokens from the mail messages which is used further by the feature extraction model. Removal of less informative or tokens which do not have a positive weightage in the mail classification is necessary for optimizing the computational efficiency and as well as enhancing the predictive analysis of the classifier. These terms can add noise components which can result in the faulty training of the classifier. We conduct experiments to determine the optimal percentage of terms to be included for constructing Split LC-based feature vectors. Fig. 6 reflects the obtained results for various evaluation criteria for spam filtering with varying the percentage of terms. The best optimal result is obtained with 50-70% of the terms with VL window model. While experimenting with percentage of terms, we obtained few instances of results where 100% accuracy in spam predication was achieved with 60% term selection. Also, it is worth noting that with changes in percentage of term selection, the overall performance of the model is not impacted considerably. This is because of the Split LC-based feature extraction from these selected term which is used for the generating the feature vectors.

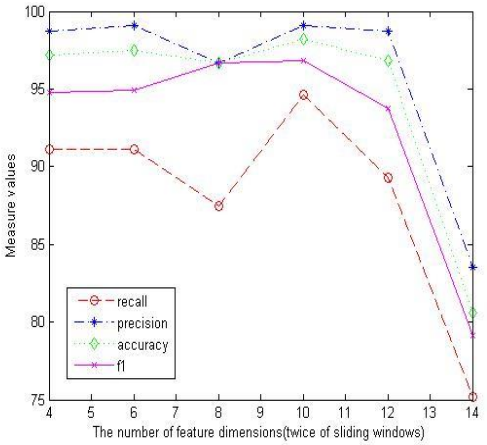

Fig. 5: Performance of Split LC-VL with different sliding windows which represent dimensionality.

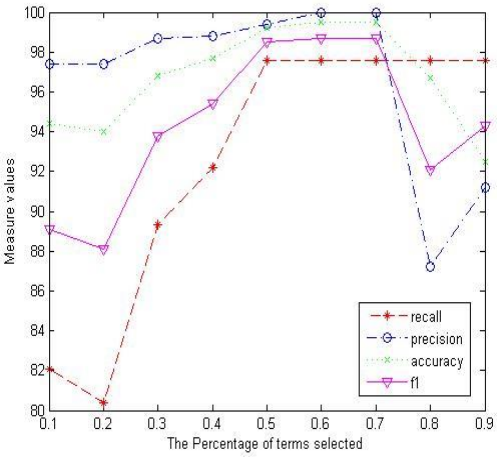

Fig. 6: Performance of Split LC-VL with different percentage of terms for constructing DS.

Further in the next section we present the comparative analysis of the performance of other feature extraction approaches with Split LC-based approach in detail.

D. Comparative analysis of Split LC Model and other approaches

To evaluate the performance of Split LC-based feature extraction approach, we conduct experiments to compare this model with other prevalent approaches like Naïve-Bayes-BoW, SVM-BoW, SVM-Global Concentration (GC), SVM-LC-FL and SVM-LC-VL [1]. We conduct the experiments on PU2 benchmark corpora and compare the results of other methods for the same data set presented by Zhu et al. [1]. SVM-GC is a special configuration of SVM-LC where the window size is set to infinite. Thus, SVM-GC reduces to a two-dimensional feature vector as the whole message represents a single feature vector. This special case is included to portray the capability of LC-based approach in extracting position related useful information from messages. The results are presented in Table III. Both the LC-based approaches perform better than SVM-GC which confirms the extraction of important information from messages for constructing feature vectors.

TABLE III

COMPARISON BETWEEN SPLIT LC-BASED MODEL AND OTHER PREVALANT MODELS

Approach	Spam Recall (%)	Spam Precision (%)	Accuracy (%)	F1 (%)	Feature Dimension
Naïve Bayes-BoW	90.00	80.77	63.66	85.14	600
SVM-BoW	79.29	88.71	93.66	83.74	600
SVM-GC	76.43	95.12	94.37	84.76	2
SVM-LC-FL	82.86	90.86	94.79	86.67	20
SVM-LC-VL	86.43	92.06	95.63	88.65	6
SVM-Split-LC-VL	97.61	87.23	96.71	92.12	6
ANN-Split-LC-VL	97.6	99.4	99.2	98.5	6

Results reflect the better performance of LC-based approach and Split LC-based approach over BoW in accuracy and F_1 measure. Split LC-based approach performs even better than the other LC-based approach [1]. These results are compared using SVM classifier. While comparing the performance of different classifiers in our experiments, ANN performs better than SVM for Split LC-based VL model with 99.2% accuracy. The proposed Split LC-based approach and the LC-based approach proposed by Zhu et al. [1] reduce the feature dimensions by huge margin. BoW has feature dimension count of 600 whereas the Split LC-based approach demonstrates better performance with just 6 dimensional feature vectors which higher precision and accuracy. In this aspect, SVM-LC-FL with reduced feature dimensionality of 20 and SVM-LC-VL with 6 or 10 has promising results as well. Overall, Split LC-based feature extraction approach with ANN classifier offers most convincing results. We do not consider implementing FL sliding window in our project as we can already see the better results and adaptability of VL over FL [1]. Moreover, VL sliding window can automatically accommodate both short and long messages without impacting performance.

VIII. CONCLUSION AND FUTURE WORK

Taking motivation from the AIS techniques for solving classification problems and the growing need for an efficient and intelligent spam filter, we selected to extend the work of Zhu *et al.* [1] and propose a new approach for extracting position correlated information from the text mail to create feature vectors for ML classifiers, which we call as Split LC-based model. To further support our insight, we test the Split LC-based feature vectors with a choice of ML classifiers: SVM and ANN. We conduct variety of experiments in depth before reaching to the preliminary conclusion that the proposed Split LC-based strategy seem to deliver promising results. Moreover, it possesses the advantage of efficiency while utilizing much lesser feature space. To summarize our findings:

i) Split LC-based feature extraction approach performs better than the current LC-based approach [1].
ii) Using VL sliding window, the spam predictive accuracy is higher and can accommodate mails of all sizes for appropriate feature extraction.
iii) Comparing the performance of various classifiers, ANN with 10 hidden neurons performs with high predictive accuracy of 99.2% for spam detection.
iv) LC strategies achieve better performance while involving lesser feature dimensions and less computational complexity compared to the legendary BoW model.

Overall, the proposed Split LC-based feature extraction approach with VL sliding window and ANN as the classifier is the right choice for Spam Filtering. Further, we can fine-tune the ANN classifier parameters like using 10 hidden neurons to extract even better results.

Considering the drifting trend in the spam designing and new methods adopted by spammers, we would like to extend the model to accommodate image spam filtering. The current model implements text based spam filtering and can be extended to include image spam, adaptive filtering model based on changing spam content and user preferences.

IX. REFERENCES

[1] Yuanchun Zhu, Ying Tan, "A Local-Concentration-Based Feature Extraction Approach for Spam Filtering" IEEE Transactions on Information Forensics and Security, Vol. 6, No. 2, June 2011.

[2] G. Ruan and Y. Tan, "Intelligent detection approaches for spam," in Proc. Third Int. Conf. Natural Computation (ICNC07), Haikou, China, 2007, pp. 1–7.

[3] I. Koprinska, J. Poon, J. Clark, and J. Chan, "Learning to classify e-mail," *Inform. Sci.*, vol. 177, pp. 2167–2187, 2007.

[4] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. Int. Conf. Machine Learning (ICML'97)*, 1997, pp. 412–420.

[5] T. Ouda and T. White, "Developing an immunity to spam," Lecture Notes Comput. Sci. (LNCS), pp. 231–242, 2003.

[6] T. S. Guzella, T. A. Mota-Santos, J. Q. Uchôa, and W. M. Caminhas, "Identification of spam messages using an approach inspired on the immune system," Biosystems, vol. 92, no. 3, pp. 215–225, Jun. 2008.

[7] Y. Tan, C. Deng, and G. Ruan, "Concentration based feature construction approach for spam detection," in Proc. IEEE Int. Joint Conf. Neural Networks (IJCNN2009), Atlanta, GA, Jun. 14–19, 2009, pp. 3088–3093.

[8] G. Ruan and Y. Tan, "A three-layer back-propagation neural network for spam detection using artificial immune concentration," Soft Comput., vol. 14, pp. 139–150, 2010.

[9] D. Dasgupta, "Advances in artificial immune systems," IEEE Comput. Intell. Mag., vol. 1, no. 4, pp. 40–49, Nov. 2006.

[10] Term Selection Wikipedia [Online]. Available: http://en.wikipedia.org/wiki/Tf%E2%80%93idf

[11] I. Koprinska, J. Poon, J. Clark, and J. Chan, "Learning to classify e-mail," Inform. Sci., vol. 177, pp. 2167–2187, 2007.

[12] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in Proc. Int. Conf. Machine Learning (ICML'97), 1997, pp. 412–420.

[13] T. S. Guzella and M. Caminhas, "A review of machine learning approaches to spam filtering," Expert Syst. Appl., vol. 36, pp. 10206–10222, 2009.

[14] E. Blanzieri and A. Bryl, A Survey of Learning-Based Techniques of e-mail Spam Filtering University of Trento, Information Engineering and Computer Science Department, Trento, Italy, Tech. Rep. DIT-06-065, Jan. 2008.

[15] W. S. Yerazunis, "Sparse binary polynomial hashing and the CRM114 discriminator," in *Proc. 2003 Spam Conf.*, Cambrige, MA, 2003.

[16] C. Siefkes, F. Assis, S. Chhabra, and W. S. Yerazunis, "Combining winnow and orthogonal sparse bigrams for incremental spam filtering," *Lecture Notes Comput. Sci.*, vol. 3202/2004, pp. 410–421, 2004.

[17] R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar, "Comparative Study on Email Spam Classifier
using Data Mining Techniques", Proceedings of the International MultiConference of Engineers and Computer Scientists, 2012 Vol I, IMESC 2012, Hong Kong.

[18] Richard P. Lippman, "Introduction to Computing Neural Nets", IEEE ASSP magazine, april 1987

[19] ANN Wikipedia: http://en.wikipedia.org/wiki/Artificial_neural_network

[20] SVM Wikipedia: http://en.wikipedia.org/wiki/Support_vector_machine

[21] C. Paulo, L. Clotilde, S. Pedro et al., "Symbiotic data mining for personalized spam filtering," in Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology, (IEEE/WIC/ACM), pp. 149–156, 2009.

[22] Yue Yang, Elfayoumy. S," Anti-Spam Filtering Using Neural Networks and Bayesian Classifiers", Proceedings of the 2007 IEEE International Symposium on Computational Intelligence in Robotics and Automation, Jacksonville, FL, USA, June 20-23,2007.

[23] Haiying Shen, Ze Li, "SOAP: A Social network Aided Personalized and effective spam filter to clean your e-mail box", Proceedings of 2011 INFCOM, pp. 1835 - 1843

[24] L.M. Ketari, M. Chandra, M.A. Khanum, "A Study of Image Spam Filtering Techniques", Proceedings of 2012 Fourth International Conference on Computational Intelligence and Communication Networks.

[25] A. Bratko, B. Filipic, G. Cormack, T. Lynam, and B. Zupan. "Spam Filtering Using Statistical Data Compression Models", the Journal of Machine Learning Research, pp., 2673–2698, 2006.

[26] Y. Gao, A. Choudhary and Gang Hua, A Comprehensive Approach to Image Spam Detection: From Server to Client Solution, Information Forensics and Security, IEEE Transactions on ,vol.5,no.4 (2010), pp.826-836.

[27] Symantec Corporation, Internet Security Threat Report Year: 2013, Vol: 18 http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_v18_2012_21291018.en-us.pdf

X. APPENDIX

List of MATLAB source code files developed for project simulation and available with report attachment:

i) project_9601_test1_pre_processing.m – This code file performs feature extraction, term selection and creation of spam/legitimate detector sets.
ii) project_9601_test2_sliding_window.m – This code file implements sliding window concept and generates feature vector and class vector which is used as input to the AIS classifiers.
iii) Project_9601_SVM.m – This code tests the SVM based classification for the spam mails.
iv) Project_9601_ANN.m – This code tests the ANN based classification for the spam mails.