

A Multi-Classifer Based Prediction Model for Phishing Emails Detection Using Topic Modelling, Named Entity Recognition and Image Processing

C. Emilin Shyni¹, S. Sarju², S. Swamynathan³

¹Department of Information Technology, KCG College of Technology, Chennai, India

²Department of Computer Science, St. Joseph's College of Engineering and Technology, Kerala, India

³Department of Information Science and Technology, Anna University, Chennai, India

Email: shyniedwin@gmail.com, shyniedwin@yahoo.co.in, swamyns@annauniv.edu

Received 31 March 2016; accepted 21 April 2016; published 26 July 2016

Copyright © 2016 by authors and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Phishing is the act of attempting to steal a user's financial and personal information, such as credit card numbers and passwords by pretending to be a trustworthy participant, during online communication. Attackers may direct the users to a fake website that could seem legitimate, and then gather useful and confidential information using that site. In order to protect users from Social Engineering techniques such as phishing, various measures have been developed, including improvement of Technical Security. In this paper, we propose a new technique, namely, "A Prediction Model for the Detection of Phishing e-mails using Topic Modelling, Named Entity Recognition and Image Processing". The features extracted are Topic Modelling features, Named Entity features and Structural features. A multi-classifier prediction model is used to detect the phishing mails. Experimental results show that the multi-classification technique outperforms the single-classifier-based prediction techniques. The resultant accuracy of the detection of phishing e-mail is 99% with the highest False Positive Rate being 2.1%.

Keywords

Phishing, Conditional Random Field Classifier, Latent Dirichlet Allocation, Natural Language Processing, Machine Learning, Image Segmentation, Image Processing

1. Introduction

The internet has great influence in people's daily lives. The use of internet-based services, such as Online Banking and Online Purchasing has increased manifold in the past few years. The use of social networking sites and other similar services has also increased greatly in the last decade. Taking advantage of this dependence, social engineering schemes use spoofed emails to steal personal information (Identity Theft) from users. The email directs the user, via a hyper-link, into a fake web page owned by attackers that looks very similar to a legitimate site. Once the user enters any personal and financial information in the directed web page, it becomes available for attackers to access, and this is used to commit fraud and carry out illegal financial transactions. Technical subterfuge schemes trigger users to download malware onto their computers, by clicking on a link embedded in a spoofed email. Using these malware, attackers steal users' credentials from their own devices. Anti-Phishing Working Group [1] reported that there were at least, 74,127 unique phishing websites detected between January 1, 2013 and March 31, 2013.

As part of the research in this paper, Topic Modelling features, Named Entity features and Structural features were utilized to detect phishing emails. Images from the legitimate site and phished sites are extracted and an image processing technique is used to compare the similarity of that images. The Topic Modelling features were extracted using the GibbsLDA, while the Named Entity features were extracted using the CRF Classifier. A total of 61 features were extracted and used for training the classifiers. The multi-classifier prediction model is built by using Random Forest (RF), Support Vector Machines (SVM) and LogitBoost. The dataset includes a corpus of 5260 e-mails including phished e-mails and legitimate e-mails. Performance is evaluated using the different measures like Precision, TPR, FPR, F-Measure and Recall. The dataset contains different combinations of phished and legitimate mails.

The rest of the paper is organized as Related Works in Section 2, Proposed Method in Section 3, Experiments and Results in Section 4 and Discussion based on the Experiments conducted along with work planned for the future in Section 5.

2. Related Work

Phishing e-mails are a particular sort of spam mails that are used to get the personal and financial related data from the users, so its recognition and incapacitation obliges higher necessity than alternate sorts of the spam mail. Phishing mail has some remarkable characteristics contrasted with the legitimate mail. For instance, it is not intended for any specific user (an exception is the spear phishing mails), it is usually focused on a financial institution, and the content of the phished e-mail often includes terms associated with finance and any emergency.

Emails are not well structured documents, they are semi structured. Chandrasekaran [2] has shown the ease of use of the structural properties of the email to differentiate between a phished e-mail and a legitimate one. They have utilized 23 style marker features, two structural property characteristics and 18 functional words to classify e-mails. The exactness of the model is assessed using the Support Vector Machine (SVM) classifier. At the same time, using functional words did not help in effectively characterizing e-mails, on the grounds that the attackers may utilize the synonyms of the words.

Attackers use distinctive systems to defeat phishing discovery mechanisms, utilizing the frequency of words related to finance and emergency. Therefore alternate solutions must be used to detect phishing mails. Topic modelling is a machine learning and natural language processing technique that we can use to distinguish the topics in a given e-mail. For instance, the topic "finance" contains monetary terms such as "cash", "money" and "amount". As opposed to discovering the frequency of the monetary words, we discover the frequency of the topic from the given mail. Landauer [3] presented another Topic Modelling system called Latent Semantic Analysis (LSA), which aggregates the words into distinctive topics dependent upon Singular Value Decomposition (SVD) of the term/document matrix. Hofmann [4] proposed Probabilistic Latent Semantic Indexing (PLSI), an alternate topic modelling procedure with a strong statistical foundation. Latent Dirichlet Allocation (LDA) is the topic modelling technique presented by Blei [5] dependent upon the generative probabilistic model. LDA assembles topics dependent upon the context of the words that is it has the ability to differentiate between a "river bank" and a "financial bank".

The majority of the phishing mails are not specifically targeted on any individual and generally phished mail targets fiscal organizations. Named Entity Recognition (NER) names the given content into predefined labels,

for example, individual and organization names, this characteristic of NER might be used to recognize phished mails. Erik [6] proposed a language independent named-entity recognition called CoNLL-2003, capable of labelling the words identified with name of an individual, location and organization. Nadeau [7] carried out a survey of NER and classification, and recognized that CoNLL-2003 is well suited for labelling English and German words.

Spatial layout similarities of web pages are also used [8] to distinguish between a legitimate site and a phished site. An R-tree is constructed and special queries are used to compare the similarity of the pages.

The goal of this work is to use the combination of structural features, topic modelling features and Named Entity Recognition features for phishing email detection and thereby improving the accuracy of the detection mechanism.

3. Proposed Method

In this section, we introduce a new methodology that incorporates natural language processing, machine learning and image processing in the detection of phished emails as shown in **Figure 1**.

3.1. Feature Construction

Each phishing mail in the Multipart Internet Mail Extension (MIME) format is parsed in to an html file to extract structural features; An HTML parser is then used to convert the html file into plain text, which in turn is used to extract the named entity features and Topic Modelling features. A Topic Modelling feature is extracted using GibbsLDA [9] and the Named Entities are extracted using the CRF Classifier. A total of 61 features are used to detect a phishing email here. The accuracy of the detection model is evaluated using different machine learning classification algorithms.

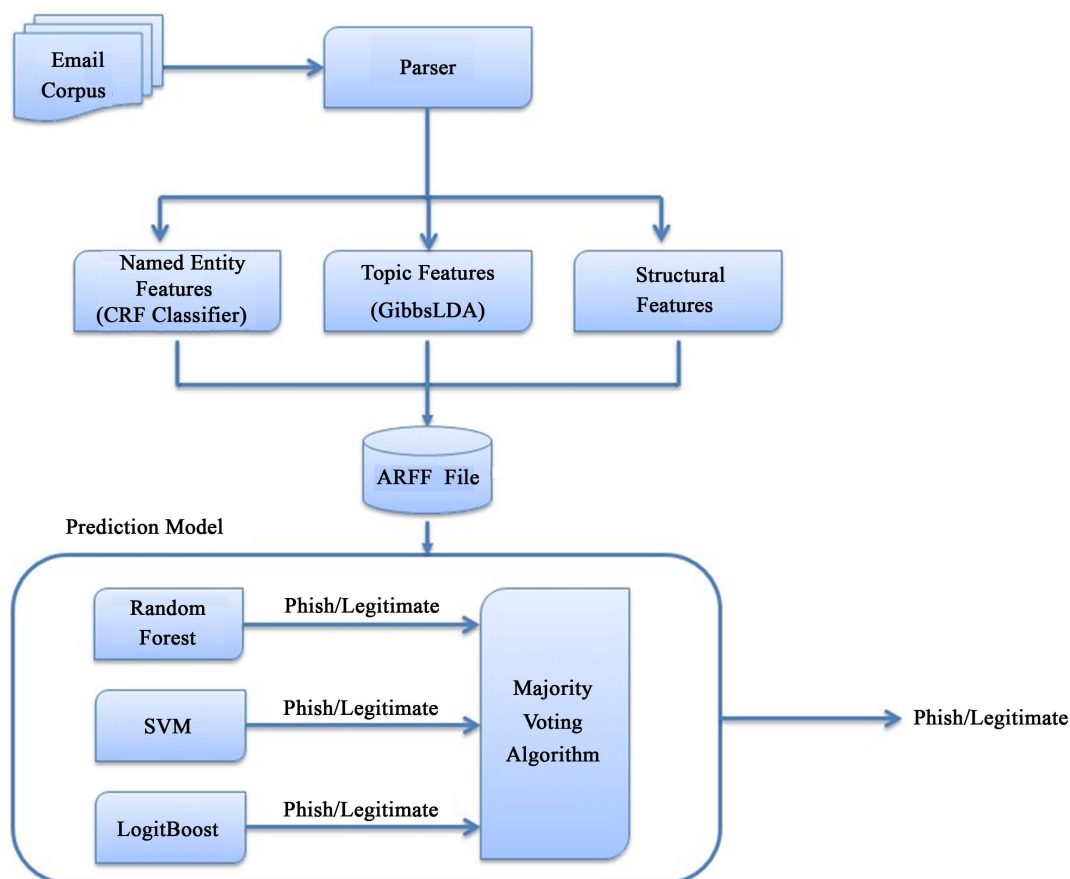


Figure 1. Methodology for phishing detection in corpus email.

Raw email data are typically present in the MIME format. In this paper, words and hyperlinks present in the body of the email are used to extract the features. Thus, the body text with the hyperlink is extracted using the parser. The two types of parsers used are the MIME parser and HTML parser.

MIME parser: -The Apache James Mime4 [10] is used in the development of the parser for extracting the content from e-mail message streams in plain Multipart Internet Mail Extension (MIME) format. It only deals with the structure of the message stream and has been designed to be extremely tolerant towards messages violating these standards. Structural features are extracted from the parsed document.

HTML Parser: -MIME messages containing HTML documents are included as multipart/HTML subpart in the email body. When the MIME parser detects a HTML subpart, it invokes the HTML parser to separate the text, style-sheets, hyperlinks and scripts. This output is given to the CRF Classifier for Topic Modelling.

3.2. Named Entity Recognition (NER)

The NER tags series of words in a text that should be the names of stuffs (nouns), such as individual and corporation names, or genetic material and protein names. The Conditional Random Field (CRF) is used to extract such named entities from the text of the email using the NER software written by Stanford's Natural Language Processing Group [11]. Ramanathan [12] gives a detailed usage of the Named Entity Recognition for phishing detection.

Conditional Random Fields

Conditional Random Fields (CRFs), used in machine learning for structured prediction, are a class of statistical modelling methods.

Given vector of input variables $X = \{X_1, X_2, \dots, X_n\}$ and a vector of output variables $Y = \{Y_1, Y_2, \dots, Y_n\}$ a model (discriminative) assesses the conditional probability $P(Y/X)$, and a generative model approximates the $P(Y, X)$. The CRF is a discriminative model, unidirectional that does not include a model of $P(X)$. Lafferty [13], defined the probability of a particular label sequence Y , given that the observation sequence X is a normalized product of latent functions, denoted as

$$\exp\left[\sum_j \lambda_j T_j(Y_{i-1}, Y_i, X, i) + \sum_k \mu_k S_k(Y_i, X, i)\right] \quad (1)$$

where $T_j(Y_{i-1}, Y_i, X, i)$ states the transition feature function and the tags at sites i and $i - 1$ in the tag sequence; $S_k(Y_i, X, i)$ is a state feature function of the tag at site i and the observation sequence; τ_j and μ_k are parameters to be estimated from the training data.

The probabilities of a tag sequence Y given an observation [14] sequence X to be written as

$$P(Y/X, \tau) = \frac{1}{Z(X)} \exp\left(\sum_j \tau_j F_j(Y, X)\right) \quad (2)$$

$Z(X)$ is called the Normalization Factor. The log-likelihood of CRF, is given by

$$L(\tau) = \sum_k \left[\log \frac{1}{Z(x^{(k)})} + \sum_j \lambda_j F_j(y^{(k)}, x^{(k)}) \right] \quad (3)$$

Conditional Random Fields Classifier (CRF Classifier)

Stanford's NER [11] software is published with a pre-trained model that has been trained on CoNLL, MUC6, MUC7, and ACE datasets. The CoNLL 2003 English training is the data-set used in this work. The CRF Named Entity Labeller component identifies and labels each word to one of three entities, namely, location, organization, and person. The output from the CRF Named Entity Labeller is used for extracting the Named Entity, which is the first set of features used for phishing detection. Figure 2 shows the Named Entity Recognition result, in which it labels the Sites, Corporate and individual.

Topic Modelling

The LDA [5] is a Natural Language Processing (NLP) method which is used to extract Topics from the collection of documents. They are modelled via a hidden Dirichlet random variable that specifies probability distribution on a latent, low-dimensional Topic space. Documents are represented as random mixtures over latent Topics and each Topic is represented by distribution over words.

To: [Fork] From:[Hallgeimsson] Subject: Marketing Manager Position. May be <ORGANIZATION>UC</ORGANIZATION> include any internship experience at <ORGANIZATION>Lewis & Clark College </ORGANIZATION>course work certifications, volunteering , co-curricular activities at <LOCATION>Berkey</LOCATION>, the <ORGANIZATION> University of California </ORGANIZATION>how and when the follow-up will be done

Figure 2. Named entity recognition from the e-mail.

Given α and β are the parameters, the joint distribution of a Topic mixture which is given by θ , a set of N words w , and a set of N Topics z is given by:

$$p(\theta, z, w | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta) \quad (4)$$

where $p(z_n | \theta)$ is θ_i for the unique i such that $z_n = i$. The marginal distribution of a document is obtained by Integrating on θ and summing on z , which is expressed as follows:

$$p(w | \alpha, \beta) = \int p(\theta, \alpha) \left[\prod_{n=1}^N \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \beta) \right] d\theta \quad (5)$$

The probability of the corpus, D is obtained by taking the product of marginal probabilities of single documents, and is expressed as follows:

$$p(D | \alpha, \beta) = \prod_{d=1}^M \int p(\theta_d | \alpha) \left[\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta) \right] d\theta_d \quad (6)$$

The parameters α and β are corpus level parameters. The document level variables θ_d are sampled once per document. The word level variables w_{dn} and z_{dn} are sampled once for each word in the document. Several algorithms have been developed to solve LDA that requires estimation of the posterior probability distribution of hidden Topic variables.

GibbsLDA

Topic Modelling is implemented by using JGibbLDA [9]. The parameter inference process requires less computational time than parameter estimation, JGibbLDA with the focus on inferring hidden/latent Topic Structures of unseen data upon the model estimated using GibbLDA++. This component consists of the following sub components.

The email extracted from the HTML parser is given to the Topic Modelling module after the pre-processing of the text document; a Term Document Frequency (TDF) matrix is created. This TDF matrix is used to train the LDA Model. LDA requires the number of Topics, K , to be initialized; in addition, LDA requires Dirichlet parameters, α , parameter of the Dirichlet prior to the per-document Topic distributions, and β , parameter of the Dirichlet prior on the per-topic word distributions, to be specified in advance.

The LDA Topic Probability Extractor extracts word/topic and topic/document distribution probabilities computed by the LDA model inference sub-component. Topic/Document distribution probabilities are used as the second set of features to build the classifier. By using these probability distributions instead of actual words, the classifier is expected to be quite robust in detecting phishing attacks. **Table 1** shows the topic distribution in a given email.

3.3. Structural Features

Emails have different Structural features, in which 10 of these structural features are used in this paper as the third set of features for detecting phished emails. **Table 2** shows the extracted structural features.

The targeted URL from the email is extracted and is checked for the legitimate site. All the images from the original site and the targeted sites are used for the similarity measures. All the images are resized into 300×300

Table 1. Topic distribution using LDA.

Topic 0	Account	Paypal	Yahoo	Business	companies
Topic 1	Subject	from	email	send	contact
Topic 2	Contact	People	Unsubscribe	Information	personal
Topic 3	Investment	money	amount	Bank	account
Topic 4	click	urgent	invalidate	important	verify

Table 2. Structural features extracted.

Feature Description	
1	Binary feature indicating whether the word “Dear” is present or not
2	Binary feature indicating whether a HTML tag is present or not
3	Binary feature indicating whether JavaScript has been used or not
4	Binary feature indicating whether the tag “ahref” is present or not
5	Binary feature indicating whether CGI has been used or not
6	Binary feature indicating the opening tag of table
7	Binary feature indicating whether OnClick event is present or not
8	Number of HTML opening comment tags
9	Binary feature indicating whether the text colour has been set to white
10	Binary feature indicating whether a URL contains “&” , “%” or “@”
11	Binary feature indicating whether a URL contains an IP address
12	Binary feature indicating the image similarity between an original site and a phished one, using image segmentation

pixels and are segmented into 25 RGB triplets as shown in the **Figure 3**. Each segment has a 30×30 pixel size and 25×3 feature vector is created. The similarity is calculated by using the Euclidean distance.

The distance from feature vector A to feature vector A will be zero. The maximum dissimilarity is calculated as the equation number 7.

$$D = \sqrt{(255-0) \times (255-0) + (255-0) \times (255-0) + (255-0) \times (255-0)} \quad (7)$$

where D is the dissimilarity value. This dissimilarity value is also used as the one of the features to detect the phished mails. **Figure 4** and **Figure 5** shows the targeted page in the mail and its original page. The dissimilarity measure between two pages is as shown in **Table 3**.

3.4. Prediction Model


Multiple learning systems try to exploit the local different behavior of the base learners to enhance the accuracy of the overall learning system. Multiple classifiers are a set of classifiers whose individual predictions are combined in some way to classify new examples. Combining classifiers solves three problems, and **Figure 6** shows the classifier combining steps.

The Prediction model is used to predict class label (Phished/Legitimate) of the given mail based on the training set which is constructed from the publicly available email corpus. The features that are extracted from the given mail are used to construct the ARFF file, which is the input to the prediction model and the output is the class label. Three classifiers are used to construct the prediction model; they are Random Forest (RF), Support Vector Machine (SVM) and LogitBoost. Each classifier predicts the category to which the mail belongs, and finally a decision is taken based on the majority voting algorithm. The F_k in the figure represents the feature sets; the public email corpus is used to collect the features and is used to train the classifiers. When a new mail arrives, the prediction model is capable of detecting a class label based on the training set.

The majority voting algorithm is shown in **Figure 7**, the parameters to the algorithm are Classifiers (C) and class Labels (L). The algorithm returns the majority class label.



Figure 3. Image.

HSBC  The world's local bank

Site map | Contact us | HSBC Group

HSBC United Kingdom


Personal Business

Financial Planning HSBC Premier HSBC Plus Current accounts Savings Investments Credit Cards Loans Mortgages Insurance International

Manage your money Grow your money Investment news Protect your lifestyle My plan Why HSBC?

Helping you achieve your vision of the future

An introduction to HSBC Financial Planning










CLICK TO PLAY 00:54

HSBC is dedicated to helping you secure the future you want

- Manage your money**
Start getting your finances in order.
- Grow your money**
Explore the whys and hows of saving and investing.
- Investment news**
Make the most of your portfolio.
- Protect your lifestyle**
Protect you and your family.
- Financial Health Check**
What are your financial priorities?
- My plan**
Store your personalised results
Video transcript

What stage of life are you at?

 **Single / Pre-Family**
 **Child Free Couple**
 **Young Family**
 **Older Family**
 **Children left home**
 **Approaching retirement**
 **Retirement**

Read the latest news and market commentary

FTSE 100	5196.98	1.06
NASDAQ 100	1730.76	0.56
S & P 500	1069.30	0.25

Sunday newspaper round-up: Barclays, Lloyds, Cadbury Gordon Brown yesterday threw his weight behind a "Tobin tax" on financial transactions as a ... [Read more](#)


Weekend tips round-up: Shanks, Tate & Lyle, Rentokil Waste disposal space is running out and the continued use of landfill is impractical - ... [Read more](#)

Read all news and commentary

This is a solution provided by Digital Look Corporate Solutions incorporating their prices, data and news on this site. All share prices and market indexes are delayed by at least 15 minutes. [Terms & Conditions](#).

Start your Financial Health Check

What are your financial priorities? Our Financial Health Check can help you work out where to start.



Other useful calculators and tools

- Budget Calculator
- What are my savings priorities?
- Investment portfolio modeller
- Retirement modeller

Personal Internet Banking

InvestDirect Sharedealing

Take the next step

Need help choosing the right products for your financial future?

08456 100 235

Or you can book an appointment to meet an adviser

Web chat

Got a question? Ask us online

My plan

Register for My plan and you can:

- Save the results of your **financial health check**
- Store all the topics that interest you

Already registered?

Figure 4. Legitimate page.

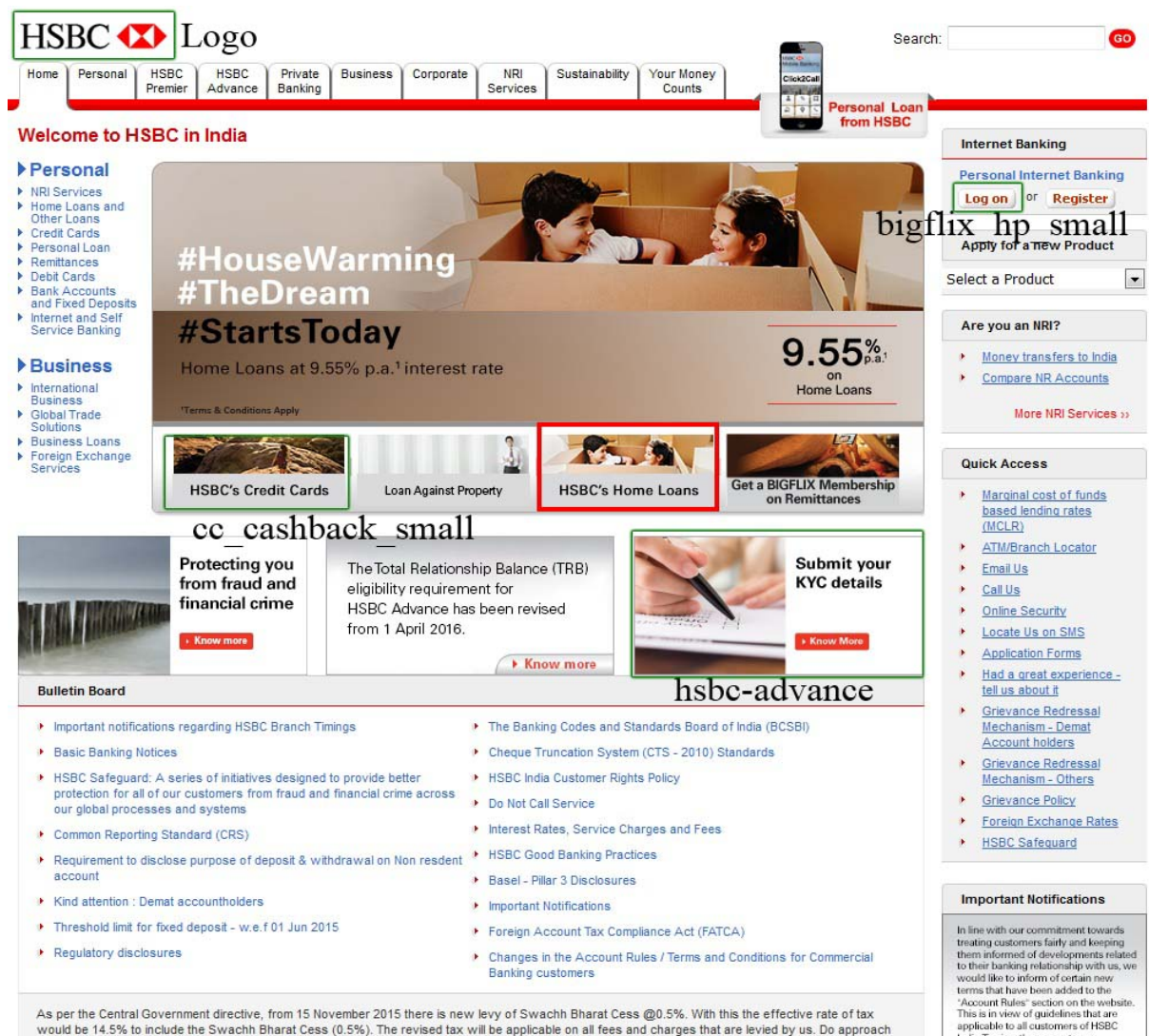


Figure 5. Phished page.

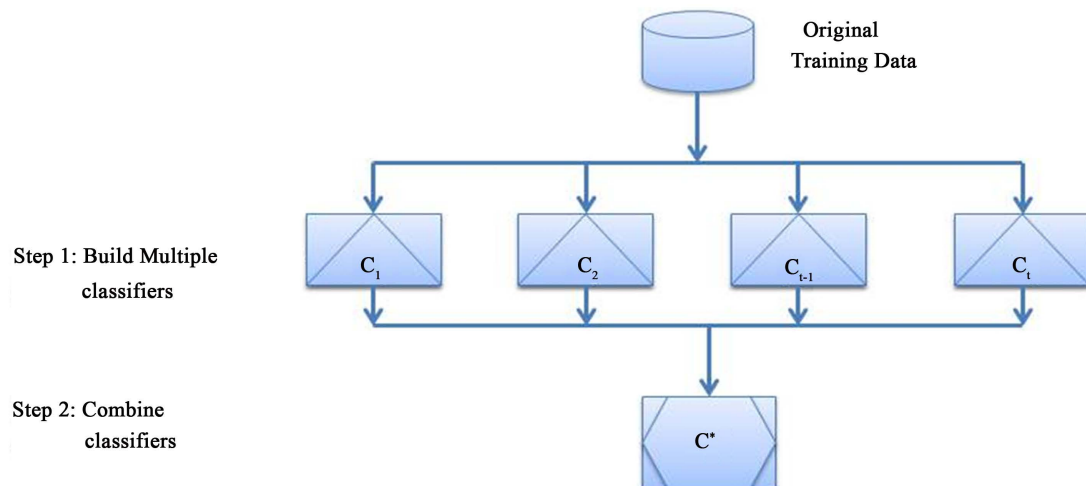


Figure 6. Classifier combining steps.


```

MajorityVote(C,L)
  Initialize
  countL1, countL2=0
  for each c ∈ C
    do if( Labelc = L1)
      Increment countL1
    else
      Increment countL2
  end for
  if(countL1>countL2)
    Return L1
  else
    Return L2
  end

```

Figure 7. Majority voting algorithm.

Table 3. Dissimilarity calculation.

Legitimate Page Image Name	Phished Page Image Name	Dissimilarity
Logo	logo	0
bigflicx_hp_small	bigflicx_hp_small	15.234
cc_cashback_small	cc_cashback_small	0
hsbc-advance	hsbc-advance	7.892
PB_Bigfix	PB_Bigfix	3.529
security_device	security_device	0
Total Dissimilarity	26.655	

4. Experiments

In this section, the performance of the proposed methodology is evaluated and the results are reported. The methodology is evaluated, using openly available standard datasets containing phishing and non-phishing data. The Evaluation of the phishing detection is carried out on email datasets using different classifiers named below.

4.1. Data Set Description

The data used for the evaluation of the proposed system has been obtained from the data sets available in the public domain [15]–[17]. The data set contains 5260 emails in all, and this includes both phished and legitimate mails. The composite mixture of the phished and legitimate mails is given as the data set (Table 4), and the features are extracted. These features are used as the input to the classifiers, for measuring their performance.

4.2. Training and Testing

The CRF Classifier (Stanford NER, 2013) is trained using the CoNLL 2003 English training data. Topic modelling is done by using the JGibbLDA with the following parameters, Dirichlet prior on the per-document Topic distributions (α), Dirichlet prior on the per-topic word distribution (β), Number of Topics (k) and Number of iterations (i), as shown in Table 5. Structural features were also extracted from the email. All 61 features were used to construct the ARFF file, which is the input file format of the WEKA. The k-fold cross validation was used to build the classifier, with a “ k ” value of 10. Thus 90% of the data is used to build the model, and the remaining 10% used as testing data.

Table 4. Data sets.

Data Set 1	Phished	50%
	Legitimate	50%
Data Set 2	Phished	40%
	Legitimate	60%
Data Set 3	Phished	30%
	Legitimate	70%
Data Set 4	Phished	20%
	Legitimate	80%
Data Set 5	Phished	10%
	Legitimate	90%

Table 5. LDA parameter values.

Parameter	Value
Per-Document Topic Distributions (α)	0.5
Per-Word Topic Distribution (β)	0.1
Number of Topics (k)	5
Number of Iterations (i)	100

4.3. Performance Analysis

The classification performance of the phishing detection is evaluated using the standard measures of performance described as follows. True Positive (TP) means the actual and predicted categories are positive and False Positive (FP) means the predicted value should have the negative classified instead of positive. Other performance metrics used in classifications are accuracy, precision, recall and F-measure. Receiver Operating Characteristic (ROC) represents the different trade-off between false positives and false negatives

$$\text{True Positive} = \frac{\text{PM to PM}}{\text{PM to PM} + \text{PM to LM}} \quad (8)$$

$$\text{False Positive} = \frac{\text{LM to PM}}{\text{LM to PM} + \text{LM to LM}} \quad (9)$$

where PM indicates phished mail and LM indicates legitimate mail.

4.4. Training and Testing

Results obtained from experimental setup are shown in **Table 6**, where the TPR and FPR for all the classifiers using the different data sets are shown. Each classifier algorithm gives dissimilar results for different datasets (**Table 4**). From the results it has been identified that the Multi-classifier gives good results when compared to a classifiers individual performance.

Figure 8 shows the accuracy of the Multi-Classifier and individual classifiers for phishing email detection. Evaluation of the figure gives a clear picture of the performance, and helps conclude that the Multi-Classification based methodology has a higher accuracy when compared to the others. In every data set, it gives an accuracy of above 96% and it reaches 99%. SVM, Random Forest and LogiBoost gives an accuracy of above 93%, but the Multi-classifier reaches above 96%.

Comparison of classifiers based on the Precision (P) and Recall (R) is shown in the **Table 7**. In all the data sets the Multi-classification gives a higher recall rate when compared to the individual classifiers. **Figure 6**

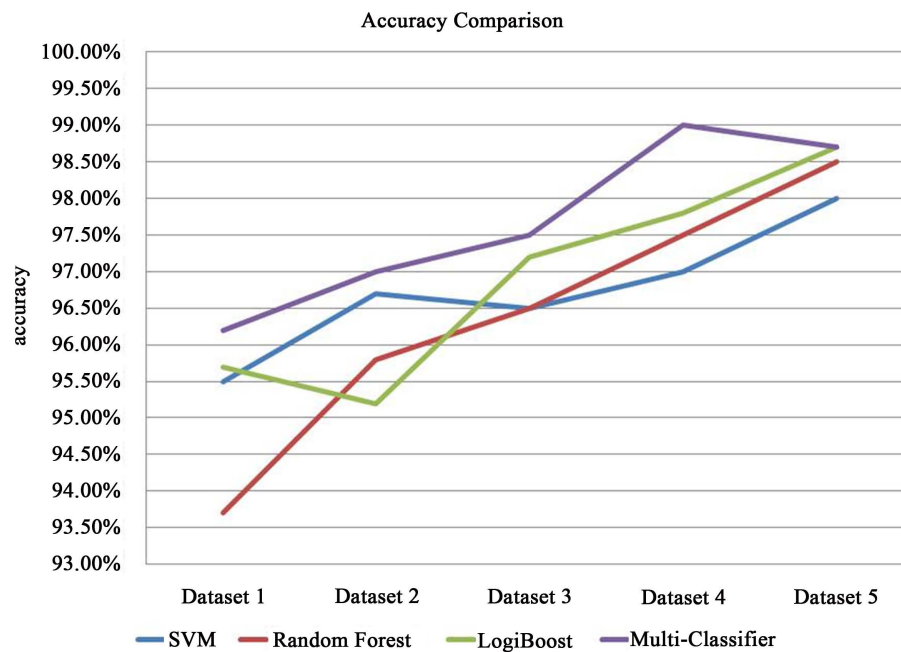


Figure 8. Comparison of accuracy of individual classifiers and multi-classifier.

Table 6. Comparison of TP (True Positive)-FP (False Positive) rate of individual classifiers and multi-classifier.

Classifier Used	Data set 1		Data set 2		Data set 3		Data set 4		Data set 5	
	TP (%)	FP (%)	TP (%)	FP (%)	TP (%)	FP (%)	TP (%)	FP (%)	TP (%)	FP (%)
SVM	95.5	4.5	96.8	4.0	96.5	5.3	97.0	7.3	98.0	13.6
Random Forest	93.8	6.3	95.8	4.3	96.5	2.5	97.5	5.3	98.5	4.6
LogitBoost	95.8	4.3	95.3	5.7	97.3	5.5	97.8	4.3	98.8	11.3
Multi-Classifier	96.3	3.8	97.0	3.7	97.5	3.9	99.0	2.1	98.8	9.0

Table 7. Comparisons of precision and recall of individual and multi-classifiers.

Classifier Used	Data set 1		Data set 2		Data set 3		Data set 4		Data set 5	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
SVM	95.5	4.5	96.8	4.0	96.5	5.3	97.0	7.3	98.0	13.6
Random Forest	93.8	6.3	95.8	4.3	96.5	2.5	97.5	5.3	98.5	4.6
LogitBoost	95.8	4.3	95.3	5.7	97.3	5.5	97.8	4.3	98.8	11.3
Multi-Classifier	96.3	3.8	97.0	3.7	97.5	3.9	99.0	2.1	98.8	9.0

shows the comparison of all classifiers based on the Precision. While considering the precision, it is found that the multi-classifier out performs the individual classifier.

Comparison of classifiers based on the Precision (P) and Recall (R) is shown in the [Table 7](#). In all data sets the multi-classification gives a higher recall rate when compared to the individual classifiers. [Figure 9](#) shows the comparison of all classifiers based on the Precision. While considering the precision, it is found that the multi-classifier out performs the individual classifier.

Finally, the performance of classifiers SVM, LogitBoost and Random Forest is compared, using the area under Receiver Operator Characteristics (ROC) curve. From the results it is clear that multi classifier prediction outperforms the individual classifier performances. [Figure 10](#) shows the ROC for individual classifiers and [Figure 11](#) shows the ROC for Multi classifiers.

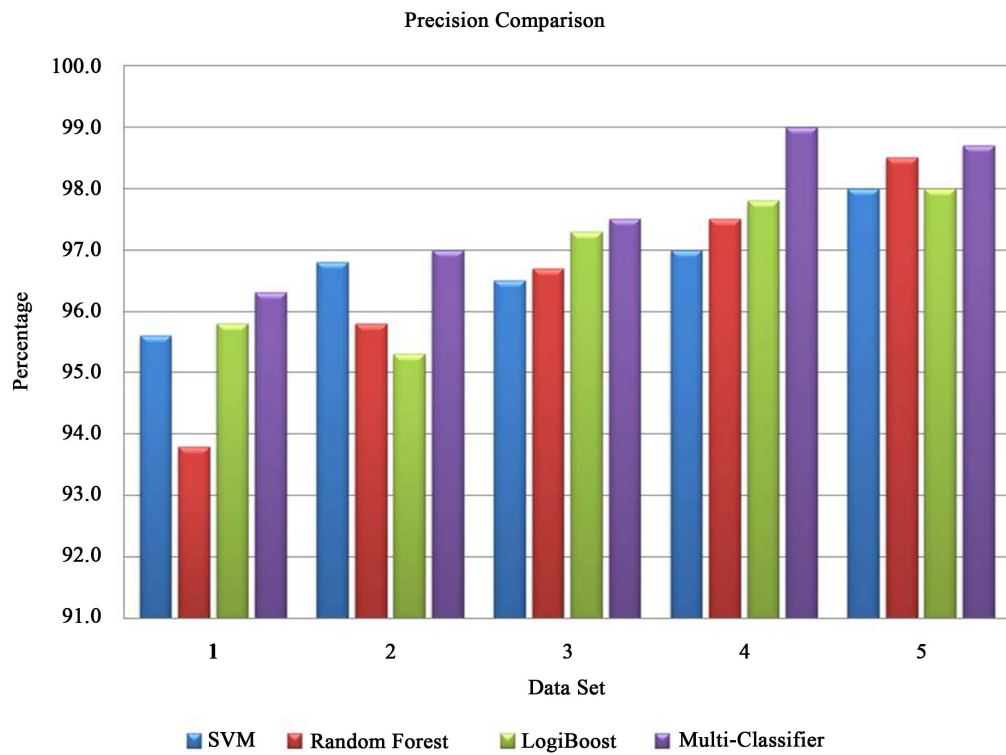


Figure 9. Comparison of F-measure of classifiers.

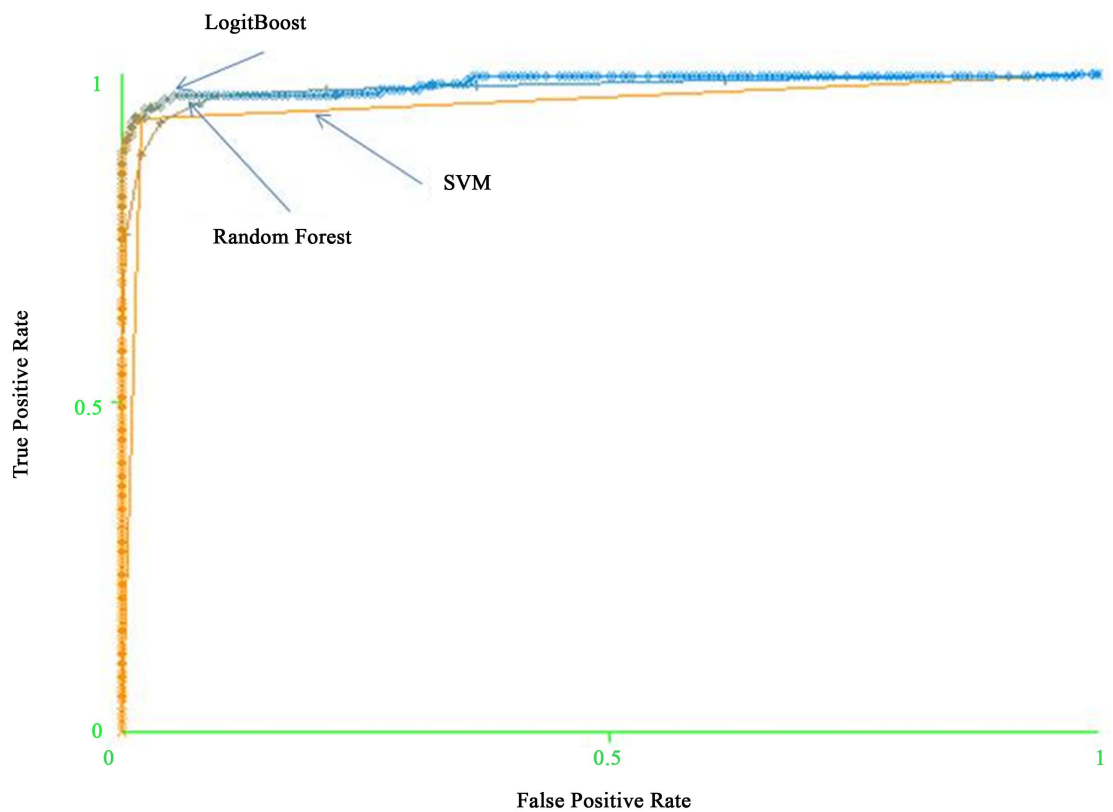


Figure 10. ROC for individual classifiers.

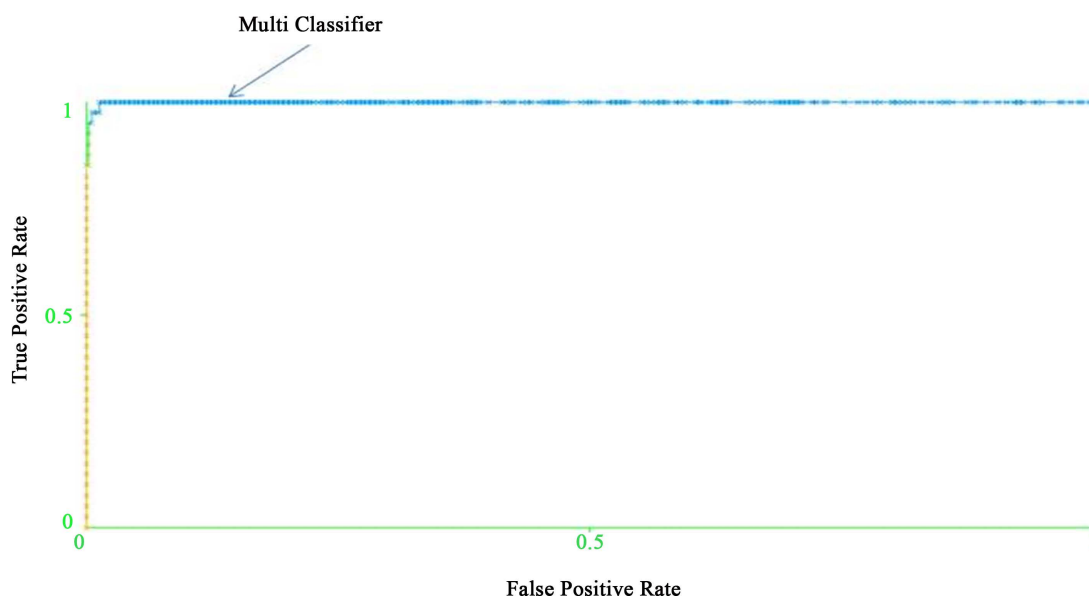


Figure 11. ROC for multi classifiers.

Considering all the experimental results, the Multi-Classifier withstands scrutiny with respect to the detection of phishing mails, and is capable of overcoming the flaws of using classifiers separately.

5. Discussion and Future Work

The present work has detailed phishing detection techniques, using the Gibbs LDA, CRF classifier and Image Processing. In addition, the multi-classifier prediction technique overcomes the drawbacks of individual classifiers.

Using the LDA and CRF improves the performance of detecting phished emails. The CRF's ability to automatically extract Named Entities from the body of the emails was greatly instrumental in determining the legitimacy of a given mail. As the CRF extracts the name based on the context in which the word appears, it is a very useful tool in combating the schemes adopted by phishers. The LDA is capable of discovering hidden Topics from the phishing messages, and is also efficient in handling synonyms. The dataset used contains various proportions of phished and legitimate mails, useful in the evaluation of the performance of the Classifiers, which help identify the most accurate ones available. The addition of structural features also improves the efficiency of phished mail detection. The image segmentation techniques improve the overall performance of the phishing email detection. The proposed methodology preserves an accuracy of 99% with an FP rate of 2.1% for detecting phishing mails. It achieves high accuracy. In future work we can have the segmentation in image processing technique and find the accuracy for detecting the phishing e-mails.

References

- [1] APWG (2013) Anti Phishing Working Group. <http://www.antiphishing.org>
- [2] Chandrasekaran, M., Narayanan, M. and Upadhyaya, S. (2006) Phishing Email Detection Based on Structural Properties. *Proceedings of 9th Annual NYS Cyber Security Conference*, Albany, 14 June 2006, 2-8.
- [3] Landauer, T.K. and Dumais, S.T. (1997) A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, **104**, 211-240. <http://dx.doi.org/10.1037/0033-295X.104.2.211>
- [4] Hofmann, T. (1999) Probabilistic Latent Semantic Indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, 15-19 August 1999, 50-57. <http://dx.doi.org/10.1145/312624.312649>
- [5] Blei, M., Andrew, Y. and Michael, I. (2003) Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, **3**, 993-1022.

- [6] Sang, E.F.T.K. and De Meulder, F. (2003) Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of the Seventh Conference on Natural Language Learning*, Edmonton, Canada, 31 May 2003, 142-147.
- [7] Zhang, W., Lu, H., Xu, B. and Yang, H. (2013) Web Phishing Detection Based on Page Spatial Layout Similarity. *Informatica*, **37**, 231-244.
- [8] Nadeau, D. and Sekine, S. (2007) A Survey of Named Entity Recognition and Classification. *Linguisticae Investigationes*, **30**, 3-26. <http://dx.doi.org/10.1075/li.30.1.03nad>
- [9] Gibbs LDA (2013) LDA Using Gibbs Sampling. <http://jgibbllda.sourceforge.net/>
- [10] Apache James Mime4J Parser (2013). <http://james.apache.org/mime4j>
- [11] The Stanford Natural Language Processing Group (2013). <http://nlp.stanford.edu>
- [12] Ramanathan, V. and Wechsler, H. (2013) Phishing Detection and Impersonated Entity Discovery Using Conditional Random Field and Latent Dirichlet Allocation. *Computers and Security*, **34**, 123-139. <http://dx.doi.org/10.1016/j.cose.2012.12.002>
- [13] Lafferty, McCallum, A. and Pereira, F. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of International Conference on Machine Learning*, San Francisco, 28 June-1 July 2001, 282-289.
- [14] Wallach, H.M. (2004) Conditional Random Fields: An Introduction. Technical Report MS-CIS-04-21.
- [15] Phish Tank (2013). <http://www.phishtank.com>
- [16] Phishingcorpus Homepage (2013). <http://monkey.org/~jose/wiki/doku.php?id=PhishingCorpus>
- [17] SpamAssassin (2013). <http://spamassassin.apache.org>



Scientific Research Publishing

Submit or recommend next manuscript to SCIRP and we will provide best service for you:

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc
A wide selection of journals (inclusive of 9 subjects, more than 200 journals)
Providing a 24-hour high-quality service
User-friendly online submission system
Fair and swift peer-review system
Efficient typesetting and proofreading procedure
Display of the result of downloads and visits, as well as the number of cited articles
Maximum dissemination of your research work

Submit your manuscript at: <http://papersubmission.scirp.org/>