VOL. 10, NO. 12, JULY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

A NOVELTY APPROACH ON TAMIL SPAM TEXT EXTRACTION BY USING TEXTON TEMPLATE BASED SUPPORT VECTOR MACHINE AND LP BOOSTING CLASSIFIER

A. Pandian[1] and Mohamed Abdul Karim Sadiq[2]
[1]Department of MCA, SRM University, Kattankulathur, Tamil Nadu, India
[2]Department of IT, Ministry of Higher Education, College of Applied Sciences, Sohar, Oman
E-Mail: pandian.a@ktr.srmuniv.ac.in

ABSTRACT

In this proposed method, the Tamil language texts are analyzed through the Morris-Pratt Algorithm as input image that filtered with Gabor filter for edge analysis. Then, it converted into unique strings from the text blocks. The text strings consist of text stroke to analyze the pattern. By using wavelet transform, the features of pattern are extracted and it undergoes for mapping with the texton patterns. It reduces the multiple dimensional signature patterns into reduced level. It then compared with the various image transformation methods such as DST (Discrete Shearlet Transform, DWT (Discrete Wavelet Transform and DCT (Discrete Cosine Transform that trained feature based on hybrid of SVM with Linear predictive boosting (LP boosting) algorithm. The effectiveness of the result is cross validated through confusion matrix and the result shows the proposed classifiers is more accurately predicts the tested Tamil text strings with reduced misclassification levels.

Keywords: Tamil text, search engine spam, support vector machine (SVM), boosting classifier, gabor filter, texton pattern.

1. INTRODUCTION

The amount of information available to the user in global internet for every year becoming, more and more, that includes growing number of web pages and number of email users. With the increasing amount of information on the Internet search engines have become accessible easily, the task of search engine for each search request to issue a ranked set of pages most relevant to the users. By spam content methods include adding artificial keywords on the page (in the headlines, Meta tags, text links, URL titles and text pages). Reference is spam formation reference structures that could affect the algorithms of the search engines for the purpose of achieve higher positions in the results Search for user queries. The phenomenon site in the top ten issuing popular search engines such requests provides significant influx of visitors that identified based on IP (Internet Protocol) address to the site and as a consequence a large number of buyers. In connection with that there is competition between the creators of websites for getting top positions. It leads to the fact that some site creators or business firm are trying to influence the outcome the algorithms used in search engines to unfairly increase the score by sending the personal scam message to unlimited number of users. This phenomenon is known as search spam [1]. The spamming regards as the main threats to the modern search engines that degrades the quality of search results and increases the load. It estimated up to 20 % of all Internet content is search spam [2]. Search engines use a variety of information to rank pages such as Filter page and the site on which it is located; links between pages and sites, etc. At present, there are several types of search engine spam, aimed at discrediting the various spam detection algorithms used in search engines [3]. In the generation of texts target of spammers is to hit the extradition request with a small amount of relevant pages. To maximize the number of clicks users such requests spammers have to create thousands of pages, each of which must be shown on one or more low-frequency queries. This spam is particularly dangerous for the search engines since such pages are likely to fall into the issue. Since the creation of a large number of pages with text manually is not possible, Spammers use automated algorithms for generating mass texts. At the same time they need to obstruct the detection of such texts from the search engine. There are two basic approaches to the mass generation of texts [3]: Copy existing natural texts and synthesis of texts on the basis of natural sample document. Currently, there are a number of effective methods for the detection of duplicates can detect copied texts wide Internet [4]. In connection with this broads widely used algorithms for automatic generation of texts. This work is devoted to detection algorithms massively generated unnatural texts. In the next section, the concepts to the review of existing methods for the detection of search engine spam and the corresponding spam in email are reviewed. In section 3, the description of the theoretical model of texts generated from samples undergoes to the various stage process. Section 4 describes a new proposed algorithm for detecting Tamil scam unnatural texts by using combined machine learning. In Section 5 shows the results of study on the applicability of the proposed algorithms to model data. Section 6 is devoted to testing the proposed algorithm on a real collection.

2. RELATED WORK

One of the common ways automatically creating a large number of texts in a single pattern is to generate text based on the Markov method. The text generation based on Markov chains first selected trained with set of patterns, then it generate a large number of meaningless, but locally connected texts. It describes the laws of generation connected meaningful texts with many known and unknown patterns. The Markov chains allow to simulate a local connectedness of the text and the common theme characteristics that developed by Haveliwala and Kamvar in [5]. The string pattern matching algorithms are primarily used for text extraction. It includes the Rabin-Karp string search algorithm [6]; it applied in threat signatures based on intrusion detection and prevention system in networking. The Rabin-Karp algorithm applied with hashing of various text pattern strings. It is generally useful to detect the plagiarism. Since, it applied with hash function, the substring as collected as number in ASCII (American Standard Code for Information Interchange) format. It helped to achieve higher optimized computation capability compared with other existing methods such as Finite-state automaton based search [7], Knuth-Morris-Pratt algorithm [8]. The state machine based string search possessed with various trigger event. In the existing approach by Scarpazza *et al*., in [7] that employed a dictionary consists of text information. The finite state string matching algorithm maps the optimal text that relevant to the dictionary data. It achieves maximum performance without implementing through the probabilistic string matching algorithms like bloom filters. It optimization reduced with maximum number of states. Similarly, the method based on Knuth-Morris-Pratt algorithm developed by Fu *et al* in [8] explains Unicode security attack problems. It induces the spam attacks, phishing attacks and web identity attacks. It analyze with the common pattern it replace the words based on semantic of similar words through phonetic substitution. It preprocessing and matching time are $\Theta(m)$ and $\Theta(n)$. It is very suitable for phishing prevent in web based applications. The input images are analyzed based on connectivity or geometrics correlations to form text regions. It achieved through the size of the text, aspect ratio that used with the rule-based heuristics that applied by Fletcher *et al* [9]. It create the text blocks that are aggregated through the Hough transformation through the Linear Filter that map the edge regions of an image through dilation and it applied by LeBourgeois in [10]. It generates the coherent pattern of text regions and it follows certain texture pattern in a connected component. The texture patterns are analyzed with texton patterns [11] that connect the similar regions. It then undergoes for higher extraction of the features based on histogram analysis by Liu *et al* in [12]. Similarly, the text segmentation based on text strokes is extracted in multiple phases into limited block by Wu *et al* [13]. The next phase involves machine learning to classify the patterns. The patterns are classified through the classifiers such as neural networks [14], Bayesian classifiers [15] and Support Vector Machines [16]. The neural networks analyze and recognize the text that based optical Character Recognition (OCR) technology. In this method, the text is automatically detects and extracted as a collection of strokes in rectangular boxes surrounding the text unique strings. It then converted into binarized text as strokes and fonts are analyzed. It compared with neural network rich set of training samples and segmented into horizontal and vertical cuts. It achieves better accuracy. Similarly, the Bayesian logistic regression methods classify the maximum level of text categorizations that estimated with the maximum likelihood method and combined with feature selected from the dictionary data. It is suitable for maximum data classification and also involves misclassification. Finally in the our existing work, the Support vector machine achieves maximum classification that analyzed through the Radial Basis Function (RBF) for Tamil language text extraction by Pandian *et al* in [17] and [18]. It analyzed through the Fisher's linear discriminant method that combined with Radial basis function network for conversion of 322 dimensional patterns into 2 dimensional patterns to achieve efficient that reduce misclassification for large scale system. In this proposed method, the Tamil language texts are analyzed as input image. Then, it converted into unique strings from the text blocks. The text strings consist of text stroke to analyze the pattern. By using wavelet transform, the features of pattern are extracted and it undergoes for mapping with the texton patterns. It reduces the multiple dimensional signature patterns into reduced level. It then compared with the trained feature based on hybrid of SVM with Linear predictive boosting (LPboosting) algorithm. The effectiveness of the result is cross validated through confusion matrix and the result is compared with various methods.

3. METHODOLOGY

3.1 Morris-Pratt algorithm for text analysis

It scans the pattern and the window of the text from left to right.

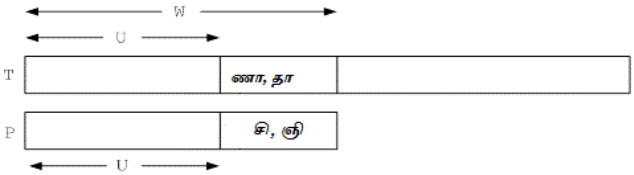

Figure-1. MP algorithm distributed over multiple Tamil text.

In Figure-1, the algorithm scans from the left and it get a complete match up to the prefix U of the window

W. That is, the prefix U of W is equal to a prefix of P and the next character "ணா, தா" to U of W is not equal to the next character "சி, ஞி" to U of P as shown in the above diagram. Therefore, it must shift P. Then, it iterates U as a partial window. If there is a suffix of U of W which is equal to a prefix of U of P, let V the longest one, as shown in Figure-2. In this case, it shift P to the right to such an extent that the prefix V of U of P is exactly aligned with the suffix V of U W, as shown in Figure-3.

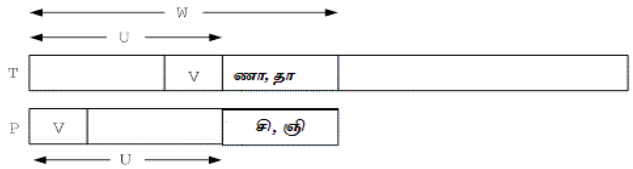

Figure-2. The suffix to prefix relationship in the partial window U for Tamil text.

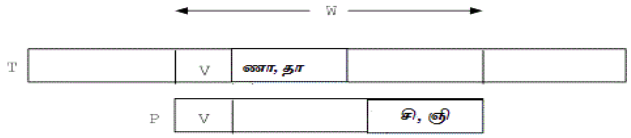

Figure-3. One case of moving P in the MP algorithm.

In the above figure, If there is no suffix of U of W which is equal to a prefix of U of P, we shift P to the right as shown and in the following Figure-4, it finally move down P.

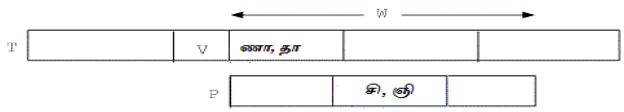

Figure-4. Then finally, it moving P in the MP algorithm.

3.2 Gabor filter

Gabor filters to extract textures of input frame Tamil text image of different sizes and orientations (i.e. Gabor-based texture feature). A Gabor filter is defined by a two-dimensional Gabor function, $g(x, y)$:

$$g(x,y)=\left(\frac{1}{2\pi\sigma_x\sigma_y}\right)\exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)+2\pi jWx\right] \quad (1)$$

where σ_x and σ_y denote the scaling parameters of the filter in the horizontal (x) and vertical (y) directions, and W denotes central frequency of the filter. The Fourier transform of the Gabor function $g(x, y)$ is defined as:

$$G(u,v)=\exp\left[-\frac{1}{2}\left(\frac{(u-W)^2}{\sigma_u^2}+\frac{v^2}{\sigma_v^2}\right)\right] \quad (2)$$

where $\sigma_u = 1/2\pi\sigma_x$ and $\sigma_v = 1/2\pi\sigma_y$.

The Gabor filters can be obtained by dilations and rotations of G(x, y) following a class of functions defined in:

$$g_{mn}(x,y)=a^{-m}G(x',y'),\ \ a>1,\ \ m,n=\text{integer}$$

$$x'=a^{-m}(x\cos\theta+y\sin\theta),\ \ and\ y'=a^{-m}(-x\sin\theta+y\cos\theta) \quad (3)$$

where θ is the orientation of the wavelet and is defined by $\theta = n\pi/K$, and K denotes the total number of orientations. We generated six different orientations of Gabor filters (K = 6). The size of the filter is defined by a and m in Equation (3). These Gabor filters transformed the region $I(x, y)$ into $X_{mn}(x, y)$:

$$X_{mn}(x,y)=\int I(x_1,y_1)g_{mn}{}^*(x-x_1,y-y_1)dx_1dy_1 \quad (4)$$

where * denotes the complex conjugate. Assuming that the local regions are spatially homogeneous, we can use the mean, u_{mn}, and standard deviation of these regions, σ_{mn}, as textural features.

$$\mu_{mn}=\iint|W_{mn}(xy)|\,dxdy$$
$$\sigma_{mn}=\sqrt{\iint|W_{mn}(xy)|-\mu_{mn})^2\,dxdy} \quad (5)$$

The general form of 2D Gabor wavelet with an identical modulation frequency of ω at both x and y directions and shift of m_x and m_y at x and y directions respectively can be as the product of Gabor wavelet in x and y directions in [20].

3.3 Texton co-occurrence matrix

The overall combination of the proposed system diagram is given in Figure-4. In a gray level image, the texton co-occurrence matrix (TCM) differentiates the features of pixel based on the interrelation to the textons. Let g be the unit vector corresponding to the G of the gray

level in the image, then the following vectors co-ordinate with the function f(x, y) [11, 12]:

$$u = \frac{\partial G}{\partial x} g \tag{6}$$

$$v = \frac{\partial G}{\partial y} g \tag{7}$$

The dot products to the above vectors are given below:

$$g_{xx} = u^{\mathbf{T}} u = |\frac{\partial G}{\partial x}|^2 \tag{8}$$

$$g_{yy} = v^{\mathbf{T}} v = |\frac{\partial G}{\partial y}|^2 \tag{9}$$

$$g_{xy} = u^{\mathbf{T}} v = \frac{\partial G}{\partial x} \bullet \frac{\partial G}{\partial y} \tag{10}$$

The $\theta(x, y)$ is the direction that changes with the vectors:

$$\theta(x, y) = \frac{1}{2} \tan^{-1}[\frac{2 g_{xy}}{(g_{xx} - g_{yy})}] \tag{11}$$

To identify the value ranges $\mathrm{C}(x, y)$ from lower value to higher value of 0 to 255, the $\mathrm{G}(x, y)$ is given below:

$$\mathrm{G}(x, y) = \{\frac{1}{2}[(g_{xx} + g_{yy}) + (g_{xx} - g_{yy}) \cos 2\theta + 2 g_{xy} \sin 2\theta]\}^{\frac{1}{2}} \tag{12}$$

The texton templates [11, 12] consists of five types of unique frames to identify the textons that illustrated in Figure-5.

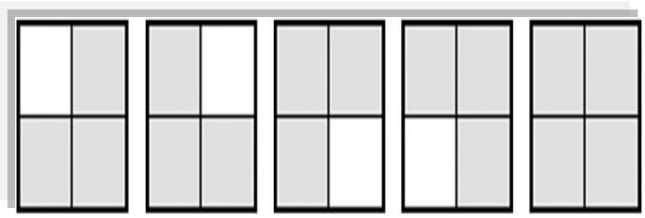

Figure-5. Texton templates with five unique types.

To identify the texton in the original image, the texton templates are morphed the input image on various texton locations that generate the five unique combinations of texton component images. Finally, the component images are combined together into texton identified image by enumerating the boundary for all morphed regions that shown in Figure-4. The texton image T with the adjacent pixels as well as and its corresponding weight of the pixels. Similarly, the orientation angle of the image indicated as and Then, the group of texton image are undergoes to wavelet transform that decompose the image. The texton structure map extraction process is shown in Figure-5. It has a four step process as described below

a) Divide the original image f (x, y) into 3 x 3 blocks

b) Move the 3x3 block horizontally and vertically from left to right and top to bottom throughout the original image f(x,y) with a step length of three pixels from the origin (0,0). Then, generate the texton structure map T_1(x,y), where $0 \le x \le M-1$,$0 \le y \le N-1$.

c) Repeatedly doing the step ii) from the origin (0,1),(1,0) and (1,1) and generate the texton structure

d) maps T_2(x,y), where $0 \le x \le M-1$,$1 \le y \le N-1$, T_3(x,y), where $1 \le x \le M-1$,$0 \le y \le N-1$ and T_4(x,y), where $1 \le x \le M-1$,$1 \le y \le N-1$ respectively.

e) Generate the final texton structure map T(x,y) using the equation ()

$$T(x, y) = Max\ \{T_1(x, y), T_2(x, y), T_3(x, y), T_4(x, y)\} \tag{13}$$

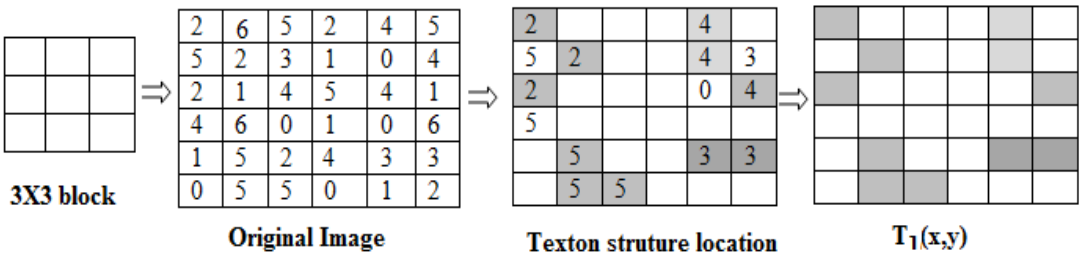

Figure-6. Texton structure map $T_1(x,y)$ extraction Process.

The formation of final texton structure T(x, y) using fusion of texton structure map $T_1(x,y)$,$T_2(x,y)$,$T_3(x,y)$ and $T_4(x,y)$ is shown in Figure-6.

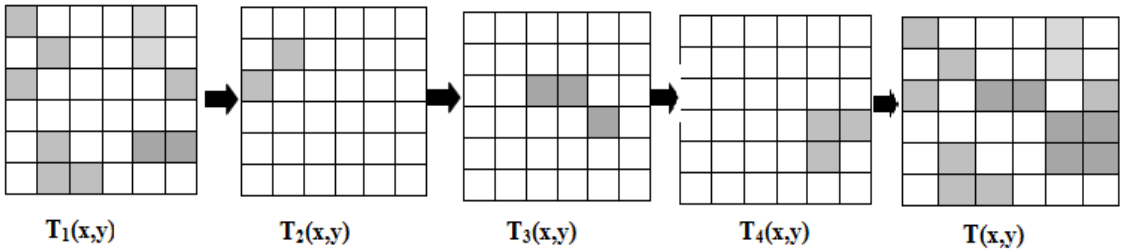

Figure-7. Formation of final Texton structure T(x,y) using Fusion of texton structure map $T_1(x,y)$,$T_2(x,y)$,$T_3(x,y)$ and $T_4(x,y)$.

The texton structure mask is applied to the original image. The pixels that do not match the mask are set to empty. The final texton structure image formation process using modified MTH is shown in Figure-7.

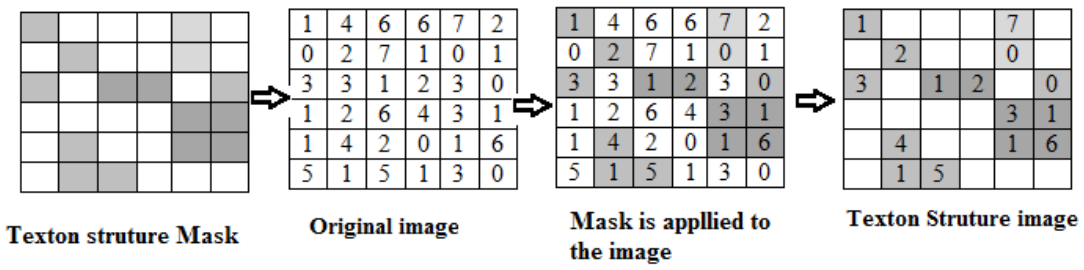

Figure-8. Texton structure Image extraction process.

a) Blocks count value of texton structure image

In this system the values of a texton structure image T(x,y) are denoted as $T(x, y) = w, w \in \{0,1,2,.....N-1\}$. In each 3X3 block of T(x,y) , $P_0 = (x_0, y_0)$ denotes the center position on it and let $T(P_0) = w_{0.}$, $P_i = (x_i, y_i)$ denotes the eight neighbouring pixels to P_0 and let $T(P_i) = w_i$, $i = 1,2,3,....8$. Let N denotes the co-occurring number of two values w_0 and w_i , and $\overline{N}$ denotes the occurring number of values w_0 . Moving the 3X3 block from top to bottom and left to right throughout the texton structure image. The texton structure image is defined as per equation (14)

$$H(w_0) = \begin{cases} \dfrac{N\{T(P_0) = w_0 \wedge T(P_1) = w_i \,||\, P_i - P_0 \models 1\}}{8\overline{N}\ (T(P_0) = w_0)} \\ where \quad w_0 = w_i, i \in \{1,2,.....8\} \end{cases} \quad (14)$$

Then, the block count value is calculated for each intensity value (1-255) on this image. The resultant vector $H(V_2)$ is obtained from the texton structure image

b) Concatenation

The computed vectors such as $H(V_1)$ and $H(V_2)$ are then concatenated to obtain the MMTH feature $H(V)$ for brain tumor classification. The concatenation process is similar.

3.4 Discrete wavelet transform

The DWT begin by defining the wavelet series expansion of function $f(x) \in L^2(\mathbf{R})$ relative to wavelet $\psi(x)$ and scaling function $\phi(x)$. It written as

$$f(x) = \sum_k c_{j_0}(k)\phi_{j_0,k}(x) + \sum_{j=j_0}^{\infty}\sum_k d_j(k)\psi_{j,k}(x) \quad (15)$$

where j_0 is an arbitrary starting scale and the $c_{j_0}(k)$'s are normally called the approximation or scaling coefficients, the $d_j(k)$'s are called the detail or wavelet coefficients. The expansion coefficients are calculated as

$$c_{j_0}(k) = \left\langle f(x), \tilde{\phi}_{j_0,k}(x) \right\rangle = \int f(x)\tilde{\phi}_{j_0,k}(x)dx \quad (16)$$

$$d_j(k) = \left\langle f(x), \tilde{\psi}_{j,k}(x) \right\rangle = \int f(x)\tilde{\psi}_{j,k}(x)dx \quad (17)$$

If the function being expanded is a sequence of numbers, like samples of a continuous function $f(x)$. The resulting coefficients are called the discrete wavelet transform (DWT) of $f(x)$.

$$W_\phi(j_0,k) = \frac{1}{\sqrt{M}}\sum_{x=0}^{M-1} f(x)\tilde{\phi}_{j_0,k}(x) \quad (18)$$

$$W_\psi(j,k) = \frac{1}{\sqrt{M}}\sum_{x=0}^{M-1} f(x)\tilde{\psi}_{j,k}(x) \quad (19)$$

for $j \geq j_0$ and

$$f(x) = \frac{1}{\sqrt{M}}\sum_k W_\phi(j_0,k)\phi_{j_0,k}(x) + \frac{1}{\sqrt{M}}\sum_{j=j_0}^{\infty}\sum_k W_\psi(j,k)\psi_{j,k}(x) \quad (20)$$

where $f(x)$, $\phi_{j_0,k}(x)$, and $\psi_{j,k}(x)$ are functions of discrete variable x = 0, 1, 2, ... , M − 1. The fast wavelet transform (FWT) is a computationally efficient implementation of the discrete wavelet transform (DWT) that exploits the relationship between the coefficients of the DWT at adjacent scales in [21].

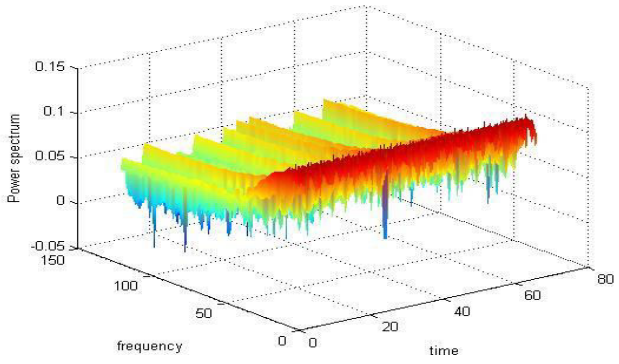

Figure-9. Wavelet transformation visualization factor.

4. PROPOSED WORK

4.1. Training phase

The output from the improved multi-texton is given as input to the training phase. The input function gives the set of values which are non- separable. All the possible separations of the point set can be achieved by a hyperplane. For that, a set of data drawn from an unknown distribution, $((x_1, y_1),...,(x_1, y_l), x_i) \in R^n$, $y_i \in \{-1,1\}$ and also a set of decision functions, or hypothesis space $f_\lambda : \lambda \in \wedge$ are given, where Λ (an index set) is a set of abstract parameters, not necessarily vectors. $f_\lambda : R^n \to \{-1,+1\}$ is also called a hypothesis. The set of functions f_λ could be a set of Radial Basis Functions or a multi-layer neural network. All the possible separations of the point set can be achieved by a hyperplane. In the Lagrange optimization formulation the optimal separating hyperplane normal vector can be found. A kernel is any function $K : R^n \times R^n \to R$. This corresponds to a dot product for some feature mapping and is given in equation (21)

$$K(X_1, X_2) = \phi(X_1).\phi(X_2) \text{ For some } \phi \quad (21)$$

The kernel function can directly compute the dot product in the higher dimensional space. Introduce kernel-based Lagrange multipliers is defined as per equation (22)

$$L_p = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i . x_j) \quad (22)$$

Minimize L_p with respect to w, b and maximize with respect to α_i , $\alpha_i \geq 0 \forall_i$

In a convex quadratic programming problem, the plane is a nonlinear combination of the training vectors and is defined as per equation (23)

$$w = \sum_{i=1}^{n} \alpha_i y_i K(x_i) \quad (23)$$

Thus, the hyperplane is separated into two clusters. The sample representation of this process is shown in Figure-10.

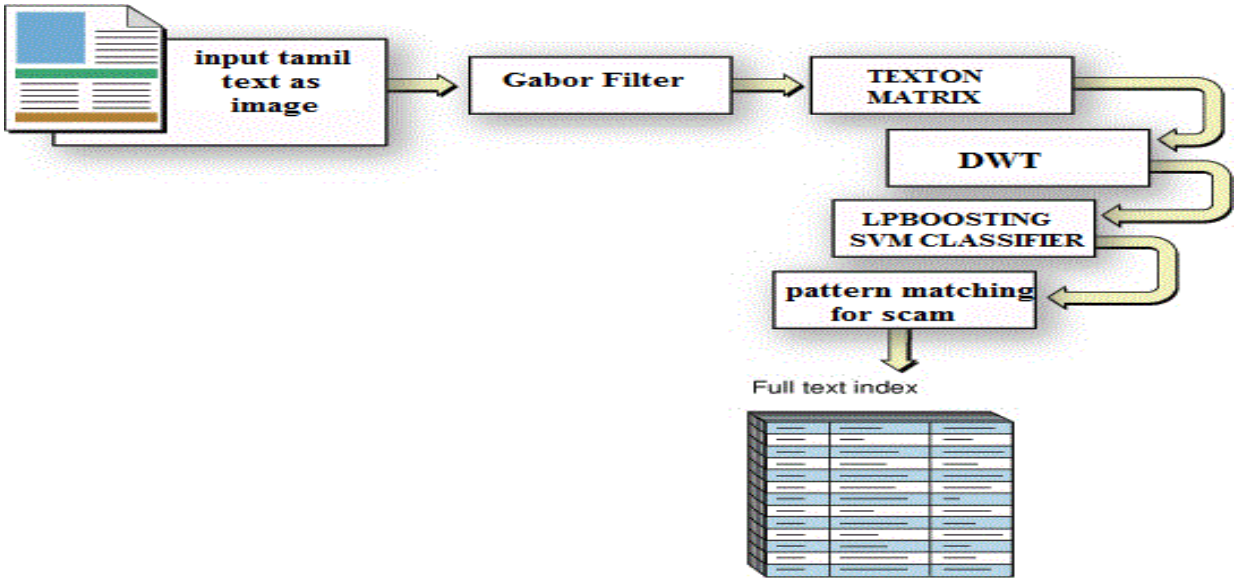

Figure-10. Proposed work flow method.

4.2 Testing phase

The various texture feature extracted using improved multi texton is given as an input vector to classifier in the testing phase and the classifier is identify the types of tumor is present in the input image.

The class of an input data x is then determined is defined as per equation (24).

$$class(x) = sign(\varphi(x).w - b) = sign\left(\sum y_i \lambda_i \varphi(x_i).\varphi(x) - b \right) \quad (24)$$

The integration of two kernels through addition or multiplication leads another kernel function (Chen *et al* 1991) in [25] and used them in the proposed work namely, RBF and quadratic function.

Radial basis function: The support vector will be the centre of the RBF and σ will determine the area of influence. This support vector has the data space. RBF kernel is defined as per equation (25).

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (25)$$

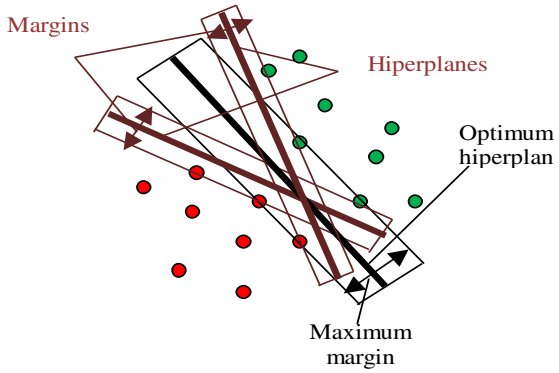

Figure-11. Sample representation of optimal hyperplane.

Quadratic kernel function: Polynomial kernels are of the form and is defined as per equation (26)

$$K(\vec{x},\vec{z}) = (1+\vec{x}^T\vec{z})^d \quad (26)$$

Where $d=1$, a linear kernel and $d=2$, a quadratic kernel are commonly used.

Let $k_1(RBF)\ and\ k_2(Quadratic)$ be kernels over $\Xi\times\Xi, \Xi\subseteq R^p$, and k_3 be a kernel over $R^p\times R^p$. Let function $\varphi:\Xi\to R^p$. The four kernel based formulations are represented is given in set of equations (27)

- $k(x,y)=k_1(x,y)+k_2(x,y)$ is a kernel

  $k(x,y)=\alpha k_1(x,y)$ is a kernel, when $\alpha>0$
- $k(x,y)=k_1(x,y)k_2(x,y)$ is a kernel
- $k(x,y)=k_3(\varphi(x)\varphi(y))$ is a kernel (27)

Substitute the four kernels as per equation (27) in Lagrange multiplier equation (5.5) and get the proposed hybrid kernel. It is exposed in equation (28).

$$L_p=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j(k_1(x_i.x_j)+k_2(x_ix_j))$$

$$L_p=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j a k_1(x_ix_j)$$

$$L_p=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j k_1(x_i.x_j)k_2(x_ix_j)$$

$$L_p=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j k_3\varphi((x_i.x_j)\varphi(x_ix_j)) \quad (28)$$

Substitute the four kernels as per equation (28) in Quadratic function equation (27) and get the proposed hybrid kernel. It is exposed in equation (29).

$$w=\sum_{i=1}^{n}\alpha_i y_i(K_1(x_i)+k_2(y_i))$$

$$w=\sum_{i=1}^{n}\alpha_i y_i\alpha K_1(x_iy_i)$$

$$w=\sum_{i=1}^{n}\alpha_i y_i\alpha K_1(x_i)k_2(y_i)$$

$$w=\sum_{i=1}^{n}\alpha_i y_i K_3\varphi(x_i)\varphi(y_i) \quad (29)$$

The boosting classifier optimizes the classification based on edges. The LP boosting strong classifier focuses on the weak classifier for the extracted features based on the shearlet transform relative entropy. It bounds as the edges of the strong classifier in which are lesser for minimum edges based on the convergence rate. The distributions of the edge margin are linear for training the set of images based on the similar features. The entropy regularized parameters for the feature vector, to update the distribution clearly. Based on the mini-max theory that eliminates the error in classifying though error matrix that shown in Figure-12.

	h_1	…	h_n	$\bar{d}$
x_1	$u_{1,1}$	…	$u_{1,n}$	d_1
…	…	…	…	…
x_m	$u_{m,1}$	…	$u_{1,n}$	d_n
$\bar{w}$	w_1	…	w_n	

Figure-12. Error matrix in the training sets.

In the Error matrix, the training sets X={ x1, x2, x3…xm } and is the distribution of various training sets from d1, d2, d3…dn with the distribution of the hypothesis from w1, w2, w3…wn that based on the features that manipulated with the hypothesis for each sample sets h1, h2, h3…hn. The minimax theory suggests the edge constraints based on its relative entropy through the feature extracted region. It helps to solve the weak classifier optimization effectively.

$$f^t(x)=\sum_{q=1}^{t}w_q h^q(x).$$
$$\min_{d_t,\gamma}\ \gamma+\eta.\Delta(\bar{d}_t,\ \bar{d}_0)$$
$$s.t\ \sum_{i=1}^{m}u_{i,j}d_i\le\gamma,\ \text{for } 1\le j\le t;$$
$$\sum_{i=1}^{m}H'_{t-1}(x_i)\,y_i\,d_i\le\gamma;$$
$$0\le d_j\le\frac{v}{m},\sum_j d_j=1; \quad (30)$$

The main advantages of using LP boost classifiers are it performs train sequentially for the weak classifiers based on the preceding rules. It reduces the complication based on its hypotheses. The discriminate functions based on the LP boosting classification reduce the redundancy and misclassification.

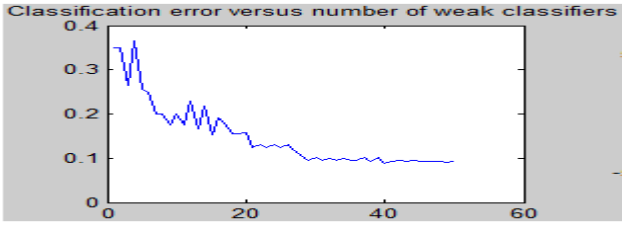

Figure-13. The misclassification to the weak classifier in LPboosting classifier.

Based on the error matrix, the misclassification reduced with constant iteration that shown in the Figure-12. The classification rates abruptly increases as the training sets feature increases. It predominantly shows the efficient classification in the training images. Similarly, the collection of test images which identified by the subset homogenous pattern for classification. It improves the calculation and classification performance which shown in the below Figure-13.

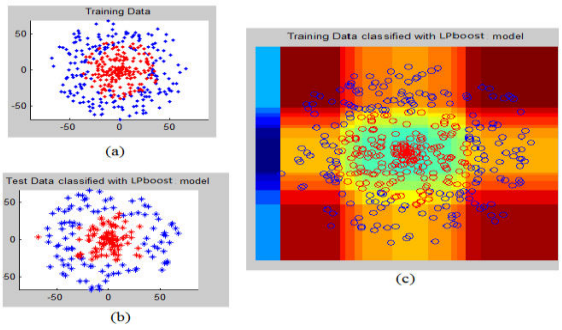

Figure-14. SVM_LP boosting classification on the texture images: (a) Training data, (b) Test data classified with LP boost model, (c) Training data classified with LP boost model.

5. EXPERIMENTAL RESULT AND DISCUSSION

In this heuristics on Tamil text strings of various parameters such as spacing, punctuation mark or other induction marks are not prioritized. In this proposed method, the tamil text are manipulated into images and the algorithm has the potential to analyzed with 100 various images fields (simplified into 10 classes)that includes magazines, OCR extracted images, newspaper and printed color advertisement. It consists of signature based Tamil font style, non- structured layout and it background texture pattern are analyzed in the stage-wise process. In the test Tamil images, there are 28420 characters and 5000 words are analyzed with trained images. Around 25262 characters and 3700 words are successfully analyzed through this proposed method. It is very reliable and robust and stable.

Table-1. Comparison of the various Evaluation metrics with the proposed system.

Evaluation metrics		SVM_LP boosting+DCT	SVM_LP boosting+DST	SVM_LP boosting+DWT
Input text strings texture images for various classes	TP	3700	3500	3800
	TN	800	800	900
	FP	200	200	100
	FN	300	500	200
	Sensitivity	0.925	0.875	0.95
	Specificity	0.73	0.62	0.9
	Accuracy	0.9	0.86	0.94
	Total error (%)	10	14	6

In this analysis, the text strings texture images of various set are taken and it divided into class 1 to class 10. The text strings image classification accuracy of the proposed system is evaluated using the evaluation metrics, such as sensitivity, specificity and accuracy that based Zhu *et al*. (2010) is defined. Based on the confusion matrix, the error in the LP boosting classifier is clearly shown for various Text strings texture image classes. It is noted that the performance of the algorithm efficiently improved when the machine classifier analyze the Text strings

texture images in the "class 5". The similarity of "class 5" to compare with other classes image are significantly reduced during the process of retrieving the "class 5" images.

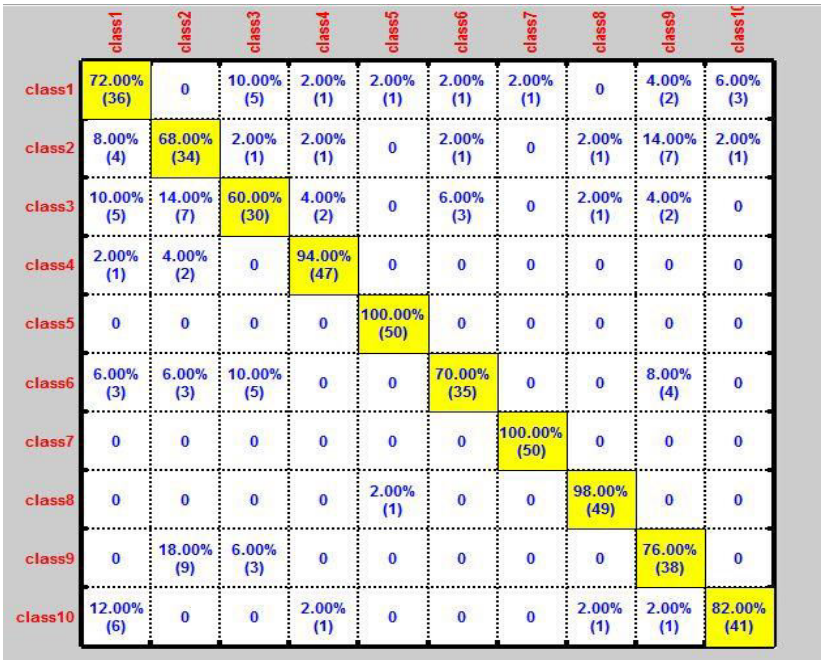

Figure-15. Confusion Matrix analysis for the proposed Tamil text extraction in dictionary.

The performance evaluations of the proposed texture classification system are identified. From each original image, 128x128 pixel sized images are extracted with an overlap of 32 pixels between vertical and horizontal direction. The performance are measured through the SAS (Sensitivity, Specificity and Accuracy) parameters

$$Sensitivity = TP/(TP + FN)$$
$$Specificity = TN/(TN + FP)$$
$$Accuracy = (TN + TP)/(TN + TP + FN + FP) \quad (35)$$

Where TP stands for True Positive, TN stands for True Negative, FN stands for False Negative and FP stands for False Positive. As suggested by above equations, sensitivity is the proportion of true positives that are correctly identified by a diagnostic test. It shows how good the test is at detecting a texture features based on the classifier. Specificity is the proportion of the true negatives correctly identified by a trained test. It suggests how good the test is at identifying normal (negative) condition. Accuracy is the proportion of true results, either true positive or true negative, in a population. It measures the degree of veracity of a diagnostic test on a condition that analyzed in [22]. In the below Table-3, the sensitivity of the proposed SVM_LP boosting+DWT approach is better compared to other methods SVM_ LP boosting+DST and SVM_LP boosting+DCT. The specificity for the proposed design LP boosting+DWT leads by 0.17% and 0.28% of the existing SVM_LP boosting+DST and SVM_LP boosting+DCT method respectively. Similarly, the accuracy of SVM_LP boosting+DWT is extremely higher than all other approaches. Based on the experimental results, the proposed system classification error rate is less than the other classifier; it is shown in Figure-17. It is seen that the proposed method error ratio is only 7.5% for text strings image datasets whereas the SVM_LP boosting+DCT and SVM_LP boosting+DST methods have error rate of 12.5% and 17.5% respectively. Compared to existing methods, the proposed SVM_LP boosting+DWT algorithm is much sophisticated for the classification of text strings texture images.

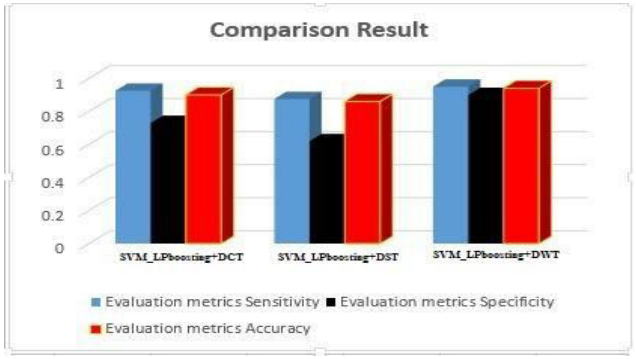

Figure-16. Comparison result analyses of SVM_LP boosting with DCT, DST and DWT.

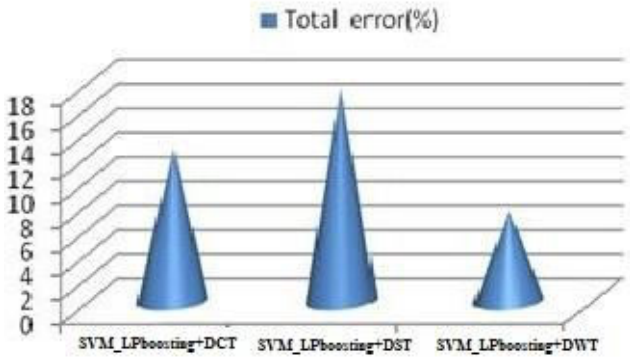

Figure-17. Error analysis of SVM_LP boosting with DCT, DST and DWT.

Table-2. Error analysis of various transform method based on SVM_LP boosting classifier.

Methods	Mean error
SVM_LP boosting+DWT	7.5
SVM_LP boosting+DCT	12.5
SVM_LP boosting+DST	17.5

6. CONCLUSIONS

This paper has introduced a new efficient hybrid of Support Vector machines with Linear Programming Boosting classifiers for classifying large set of text strings data. It reduces large number of character pattern into small dimensional form. The proposed algorithm having flexibility that ease the design of complexity for many applications. The Tamil language texts are analyzed through the Morris-Pratt Algorithm as input image. It features are extracted through the Gabor filter that filters edge regions and it process the image into text strings. It features are extracted through wavelet transformation that mapped with texton patterns. It reduces the multiple dimensional signature patterns into reduced level. It then compared with the various image transformation methods such as DST (Discrete Shearlet Transform, DWT (Discrete Wavelet Transform and DCT (Discrete Cosine Transform that trained feature based on hybrid of SVM with Linear predictive boosting (LP boosting) algorithm. our results appear to be at the state of the art in digital text pattern recognition for spam based application that yields 0.94% accuracy levels. In future, the proposed algorithm will be utilized for complex level bilingual text patterns.

REFERENCES

[1] Hayati P. and Potdar V. 2009, June. Toward spam 2.0: an evaluation of web 2.0 anti-spam methods. In Industrial Informatics, 2009. INDIN 2009. 7th IEEE International Conference on. pp. 875-880. IEEE.

[2] Ntoulas A., Najork M., Manasse M. and Fetterly D. 2006, May. Detecting spam web pages through content analysis. In: Proceedings of the 15th international conference on World Wide Web. pp. 83-92. ACM.

[3] Gyongyi Z., Berkhin P., Garcia-Molina H. and Pedersen J. 2006, September. Link spam detection based on mass estimation. In Proceedings of the 32nd international conference on Very large data bases. pp. 439-450. VLDB Endowment.

[4] Carpinter J. and Hunt R. 2006. Tightening the net: A review of current and next generation spam filtering tools. Computers and security. 25(8): 566-578.

[5] Haveliwala T. and Kamvar S. 2003. The second eigenvalue of the Google matrix. Stanford University Technical Report.\

[6] Kijewski P. 2006, June. Automated extraction of threat signatures from network flows. In 18th Annual FIRST conference.

[7] Scarpazza D. P., Villa O. and Petrini F. 2007, March. Peak-performance DFA-based string matching on the Cell processor. In Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International. pp. 1-8. IEEE.

[8] Fu A. Y., Deng X., Wenyin L. and Little G. 2006, July. The methodology and an application to fight against unicode attacks. In Proceedings of the second symposium on Usable privacy and security. pp. 91-101. ACM.

[9] L.A. Fletcher and R. Kasturi. 1988. A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images. IEEE Trans. on Pattern Analysis and Machine Intelligence. 10(6): 910-918.

[10] Lebourgeois F. and Emptoz H. 1999, September. Document analysis in gray level and typography extraction using character pattern redundancies. In Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on. pp. 177-180. IEEE.

[11] Liu G. H. and Yang J. Y. 2008. Image retrieval based on the texton co-occurrence matrix. Pattern Recognition. 41(12): 3521-3527.

[12] Liu G. H., Zhang L., Hou Y. K., Li Z. Y. and Yang J. Y. 2010. Image retrieval based on multi-texton histogram. Pattern Recognition. 43(7): 2380-2389.

[13] Wu V., Manmatha R. and Riseman E. M. 1999. Textfinder: An automatic system to detect and recognize text in images. IEEE Transactions on pattern analysis and machine intelligence. 21(11): 1224-1229.

[14] Lippmann R. P. 1989. Pattern classification using neural networks. Communications Magazine, IEEE. 27(11): 47-50.

[15] Genkin A., Lewis D. D. and Madigan D. 2007. Large-scale Bayesian logistic regression for text categorization. Technometrics. 49(3): 291-304.

[16] Joachims T. 1999, June. Transductive inference for text classification using support vector machines. In ICML. 99: 200-209.

[17] Pandian A. and Karim M. A. 2012. Detection of Fraudulent Emails by Authorship Extraction. International Journal of Computer Applications. 41.

[18] Pandian A. and M.A.K. Sadiq. 2014. Authorship categorization in email investigations using fisherâ™s linear discriminant method with radial basis function. J. Comput. Sci. 10: 1003-1014.

[19] Priya R. L. and Manimannan G. 2014. A Study of Ambiguous Authorship in Tamil Articles using Multivariate Statistical Analysis. International Journal of Computer Applications. p. 86.

[20] Ma L., Wang Y and Tan T. 2002. Iris recognition based on multichannel Gabor filtering. In Proc. Fifth Asian Conf. Computer Vision. 1: 279-283.

[21] Shensa M. 1992. The discrete wavelet transform: wedding the trous and Mallat algorithms. Signal Processing, IEEE Transactions on. 40(10): 2464-2482.

[22] Zhu W, Zeng N and Wang N. 2010. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations. NESUG proceedings: health care and life sciences, Baltimore, Maryland.

[23] Fang Y. K., Fu Y., Sun C. J. and Zhou J. L. 2010. LPBoost with Strong Classifiers. International Journal of Computational Intelligence Systems. 3(sup01), 88-100.

[24] Fang Y., Fu Y., Sun C. and Zhou J. 2011. Improved Boosting Algorithm Using Combined Weak Classifiers. Journal of Computational Information Systems. 7(5): 1455-1462.

[25] Chen S., Cowan C. F. N. and Grant P. M. 1991. Orthogonal least squares learning algorithm for radial basis function networks. Neural Networks, IEEE Transactions on. 2(2): 302-309.

[26] Ahmed N., Natarajan T. and Rao K. R. 1974. Discrete cosine transforms. Computers, IEEE Transactions on. 100(1): 90-93.

[27] Lim W. Q. 2010. The discrete shearlet transform: A new directional transform and compactly supported shearlet frames. Image Processing, IEEE Transactions on. 19(5): 1166-1180.

[28] Tong S. and Koller D. 2002. Support vector machine active learning with applications to text classification. The Journal of Machine Learning Research. 2: 45-66.

[29] Leopold E. and Kindermann J. 2002. Text categorization with support vector machines. How to represent texts in input space? Machine Learning. 46(1-3): 423-444.

[30] Drucker H., Wu S. and Vapnik V. N. 1999. Support vector machines for spam categorization. Neural Networks, IEEE Transactions on. 10(5): 1048-1054.