Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Survey on E-mail Spam Detection Using NLP

Er. Seema Rani
Research Scholar (CGC, Gharuan)
Computer Science & Engineering
PTU, India

Er. Sugandha Sharma
Assistant Professor (CGC, Gharuan)
Computer Science & Engineering
PTU, India

Abstract— E-mail spam causes a serious problem for the internet user. It has a lot of conséquences. It reduces productivity, takes extra space in mail boxes, extend viruses, Trojans, and materials that contains potentially harmful information for a certain category of users, destroy stability of mail servers, and as a result users spend a lot of time for sorting incoming mail and deleting undesirable correspondence. So it is necessary to detect the spam so that its consequences can be reduced. There are various classifiers used for e-mail spam detection like Naïve Bayes Classifier, KNN, and SVM Classifiers etc. Some of these methods are discussed in this paper.

Keywords— Spam, E-mail Spam, Text categorization, Spam Classifier Methods, Vector Space Model

I. INTRODUCTION

Spam is an unwanted communication intended to be delivered to an indiscriminate target, directly or indirectly, notwithstanding measures to prevent its delivery. Spam filter is an automated technique to identify spam for the purpose of preventing its delivery. [1] The motivation behind spam is to have information delivered to the recipient that contains a *payload* such as advertising for a (likely worthless, illegal, or non-existent) product, bait for a fraud scheme, promotion of a cause, or computer malware designed to hijack the recipient's computer. Because it is so cheap to send information, only a very small fraction of targeted recipients — perhaps one in ten thousand or fewer — need to receive and respond to the payload for spam to be profitable to its sender. [2] The main characteristics of spam are unwanted, indiscriminate, disingenuous, payload bearing. Unwanted spam means spam messages are not wanted by vast majority of people. Indiscriminate spam means Spam is transmitted outside of any reasonable relationship or prospective relationship between the sender and the receiver. In general, it is more cost effective for the spammer to send more spam than to be selective as to its target. [1] Disingenuous spam means because spam is unwanted and indiscriminate, it must disguise itself to optimize the chance that its payload will be delivered and acted upon. [1] The payload of a spam message may be obvious or hidden; in either case spam abatement may be enhanced by identifying the payload and the mechanism by which actions triggered by it profit the spammer. Obvious payloads include product names, political mantras, web links, telephone numbers, and the like. These may be in plain text, or they may be obfuscated so as to be readable by the human but appear benign to the computer. Or they may be obfuscated so as to appear benign to the human but trigger some malicious computer action. The payload might consist of an obscure word or phrase like "gouranga" or "platypus race" in the hope that the recipient will be curious and perform a web search and be delivered to the spammer's web page or, more likely, a paid advertisement for the spammer's webpage. Another form of indirect payload delivery is *backscatter:* The spam message is sent to a non-existent user on a real mail server, with the (forged) return address of a real user. [1]
Email spam, also known as junk email or unsolicited bulk email (UBE), is a subset of electronic spam involving nearly identical messages sent to numerous recipients by email. Clicking on links in spam email may send users to phishing web sites or sites that are hosting malware. Spam email may also include malware as scripts or other executable file attachments. E-mail is a good, quick and a low cost communication approach. So spammers always opt to send spam through e-mail. Today every second user has an E-mail, and they are faced with spam problem consequently. E-mail Spam is non-requested information sent to the E-mail boxes. Spam may be a massive drawback for users and for ISPs. The causes are growth of value of electronic communications on the one hand and improvement of spam sending technology on the other hand. By spam reports of Symantec in 2013, the typical world spam rate for the year was 89.1%, with a rise of 1.4% compared with 2012. The proportion of spam sent from botnets was a lot of higher for 2013, accounting for roughly 88.2% of all spam. Despite several makes an attempt to disrupt botnet activities throughout 2013, by the top of the year the overall variety of active bots came back to roughly an equivalent variety as at the top of 2012, with just about 5 million spam-sending botnets in use worldwide. [3] Spam messages cause lower productivity; occupy space in mail boxes; extend viruses, Trojans, and materials containing potentially harmful information for an explicit class of users, destroy stability of mail servers, and as a result users pay a plenty of time for sorting incoming mail and deleting undesirable correspondence. In step with a report from Ferris Analysis, the worldwide add of losses from spam created regarding 130 billion dollars, and within the USA, forty two billion in 2012. [4] Besides expenses for acquisition, installation, and repair of protective means, users are compelled to pay the extra expenses connected with an associated degree of the post traffic, failures of servers, and productivity loss. Thus we are able do such conclusion: spam is not solely an irritating factor, however a direct threat to the business. Considering the beautiful amount of spam messages

coming to E-mail boxes, it is possible to assume that spammers don't operate alone; it's world, organized, making the virtual social networks. They attack mails of users, whole firms, and even states.

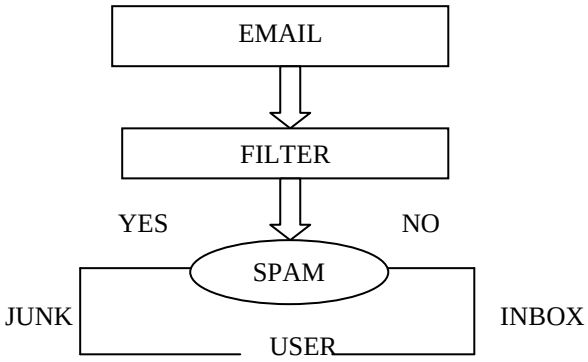

Fig No. 1 Block diagram of spam filter [21]

II. LITERATURE SURVEY

Dredze et al proposes a new and simple methodology to detect phishing emails utilizing Confidence-Weighted Linear Classifiers. They use the contents of the emails as features without applying any heuristic based phishing specific features and obtain highly accurate results compared to the best that have been published in the literature. Phishing is a criminal mechanism employing both social engineering and technical subterfuge to steal consumers' personal identity data and financial account credentials. Dredze et al. recently proposed confidence weighted linear classifiers (CWLC), a new class of online learning method designed for Natural Language Processing (NLP) problems based on the notion of parameter confidence. Online learning algorithms operate on a single instance at a time, allowing for updates that are fast, simple and make few assumptions about the data, and perform well in wide range of practical settings. Online algorithm processes its input piece-by-piece in a serial fashion, i.e., in the order that the input is fed to the algorithm, without having the entire input available from the start. [8]

Lee et.al in his paper, for spam detection, planned parameter optimization and has choice to cut back process overheads with guaranteeing high detection rates. In previous papers, either parameter optimization or feature selection are used, however not each. Parameter optimization could be a method that regulates parameters of spam detection models to work out optimum parameters of the detection model. Feature selection could be a method that chooses solely necessary options or feature commenced of all the options. Feature selection allows eliminating orthogonal options to avoid process overheads.

Razmara et.al in his work, present a novel solution toward spam filtering by employing a new set of features for classification models. These features are the sequential unique and closed patterns which are extracted from the content of messages. After applying a term selection method, we show that these features have good performance in classifying spam messages from legitimate messages. The achieved results on 6 different datasets show the effectiveness of our proposed method compared to close similar methods. Authors outperform the accuracy near +2% compared to related state of arts. In addition this method is resilient against injecting irrelevant and bothersome words.

This method is outlined as the following steps:

- Preprocessing and stemming datasets
- Selecting best discriminating terms based on a term selection method
- Looking for frequent sequential patterns in corpus
- Using patterns as features
- Feature selection and classification

The vector model for illustration of texts has been offered in Salton's works. Within the elementary case, the vector model assumes comparison to every document of a frequency spectrum of words. The dimension of space is reduced by rejection of the foremost common words that increases thereby % of the importance of the essential words in additional advanced vector models. The chance of ranging of documents according to similarity in vector space is the main advantage of vector model. Applied Computational Intelligence and Soft Computing Clustering is one of the most useful approaches in data mining for detection of natural groups in a data set. The up-to-date survey of evolutionary algorithms for clustering, such as partition algorithms, is described in detail in [12].

Nizamani et. al in his paper described the comparison of advanced topics like multi-objective and ensemble-based evolutionary clustering; and the overlapping clustering . Each of the algorithms that is surveyed is described with respect to fixed or variable number of clusters; cluster-oriented or non-oriented operators; context-sensitive or context-insensitive operators; guided or unguided operators; binary, integer, or real encodings; and graph-based representations. Clustering of spam messages means automatic grouping of thematically close spam messages. This problem becomes complicated necessity to carry out this process in real-time mode in case of information streams as E-mails. There are different methodologies that use different similarity algorithms for electronic documents in case of a considerable

quantity of signs. When classes are defined by clustering method, there is a need of their support as spam messages constantly changes, and the collection of spam messages replenishes. In this paper, the new algorithm for definition of criterion function of spam messages clustering problem has been offered. Genetic algorithm is used to solve the clustering problem [11].

Genetic algorithms are the subjects of many scientific works. For example, in a survey of genetic algorithms that are designed for clustering ensembles, the genotypes, fitness functions, and genetic operations is presented and it concludes that using genetic algorithms in clustering ensemble improves the clustering accuracy. In this work, the k-nearest neighbor method is applied for the classification of spam messages, and for the determination of subjects of spam messages, clusters will be applied to a multi-document summarization method offered in papers [12].

Huang et al., proposed a complex-network, which is based on SMS filtering algorithm that compares an SMS network with a phone- calling communication network. Although such comparison can provide some new features, that obtains well-aligned phone-calling networks and SMS networks that can be aligned perfectly is difficult in practice. In this paper, we present an effective SMS anti-spam algorithm that only considers the SMS communication network. We first analyze characteristics of the SMS network, and then examine the properties of different sets of meta-features including static features, temporal features and network features. We incorporate these features into an SVM classification algorithm and evaluate its performance on a real SMS dataset and a video social network benchmark dataset. We also compare the SVM algorithm to a KNN based algorithm to reveal the advantages of the former. Our experimental results demonstrate that SVM based on network features can get 7%-8% AUC (Area under the ROC Curve) improvement as compared to some other commonly used features. [2]

Spectral clustering method is applied to the set of spam messages collected by Project Honey Pot for defining and tracing of social networks of spammers. Social network of spammers is represented as a graph, nodes of which correspond to spammers, and social relations between spammers are represented by a corner between two junctions of graph as. In this paper, the document clustering method is applied for clustering and analyses of spam messages. In the text documents are E-mails in the text form. Instead of this fact that there are a number approaches for representation of text documents, the vector model is the most common of them. [5]

III. TEXT CATEGORIZATION

The goal of text categorization is the classification of documents into a fixed number of predefined categories. Each document can be in multiple, exactly one, or no category at all. Using machine learning, the objective is to learn classifiers from examples which perform the category assignments automatically. This is a supervised learning problem. Since categories may overlap, each category is treated as a separate binary classification problem.

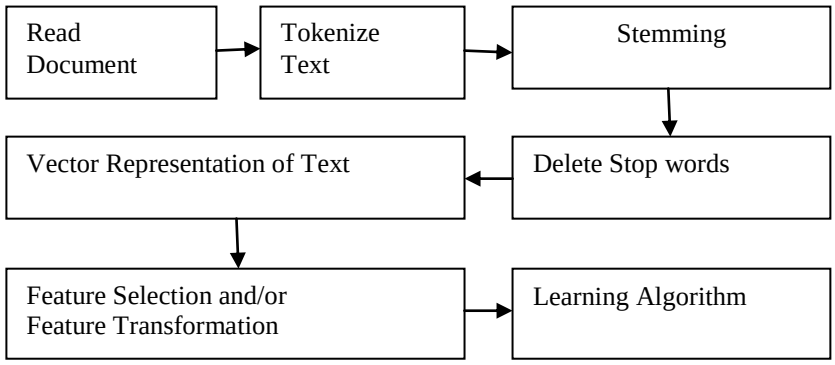

Fig No. 2 Text Classification Process [13]

The first step in text categorization is to transform documents, which typically are strings of characters, into a representation suitable for the learning algorithm and the classification task. Information Retrieval (IR) research suggests that word stems work well as representation units and that their ordering in a document is of minor importance for many tasks. This leads to an attribute-value representation of text. Each distinct word w_i corresponds to a feature with the number of times word wi occurs in the document as its value. To avoid unnecessarily large feature vectors, words are considered as features only if they occur in the training data at least 3 times and if they are not \stop-words" (like \and", \or", etc.). This representation scheme leads to very high-dimensional feature spaces containing 10000 dimensions and more. Many have noted the need for feature selection to make the use of conventional learning methods possible, to improve generalization accuracy, and to avoid \over fitting". Vector representation is used to represent the text in a document.

IV. SPAM CLASSIFIER METHODS

There are several methods for spam detection. These methods include SVM, KNN, and Naïve Bayes etc.

SVM Method: A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labeled training data (*supervised learning*), the algorithm outputs an optimal hyper plane which categorizes new examples. For a linearly separable set of 2D-points which belong to one of two classes, find a separating straight line.

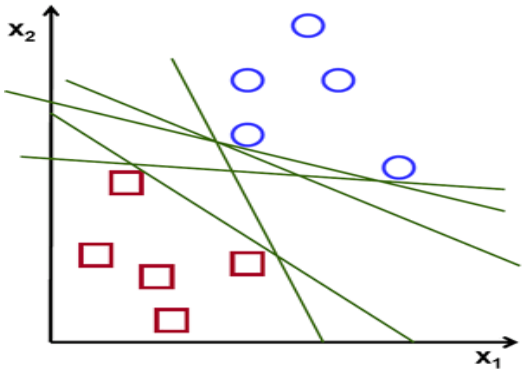

Fig no. 3 [17]

In the above picture you can see that there exist multiple lines that offer a solution to the problem. If any of them better than the others, we can intuitively define a criterion to estimate the worth of the lines:
A line is bad if it passes too close to the points because it will be noise sensitive and it will not generalize correctly. Therefore, our goal should be to find the line passing as far as possible from all points.
Then, the operation of the SVM algorithm is based on finding the hyper plane that gives the largest minimum distance to the training examples. Twice, this distance receives the important name of **margin** within SVM's theory. Therefore, the optimal separating hyper plane *maximizes* the margin of the training data.

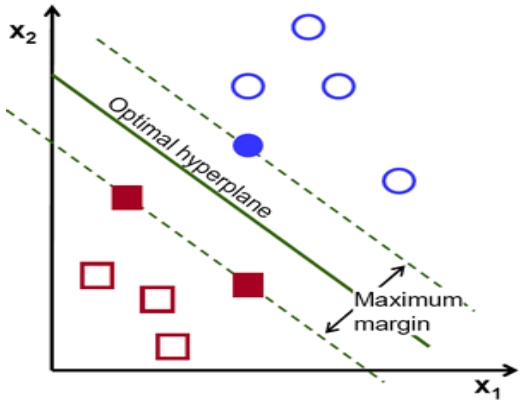

Fig No. 4 [17]

How is the optimal hyper plane computed for SVM?
Let's introduce the notation used to define formally a hyper plane:

$$f(x) = \beta_0 + \beta^T x,$$

Where β is known as the *weight vector* and β_0 as the *bias*

The optimal hyper plane can be represented in an infinite number of different ways by scaling of β and β_0. As a matter of convention, among all the possible representations of the hyper plane, the one chosen is

$$|\beta_0 + \beta^T x| = 1$$

Where $\mathbf{x}$ symbolizes the training examples closest to the hyper plane. In general, the training examples that are closest to the hyper plane are called **support vectors**. This representation is known as the **canonical hyper plane**.

Now, we use the result of geometry that gives the distance between a point $\mathbf{x}$ and a hyper plane (β, β_0):

$$distance = |\frac{\beta_0 + \beta^T x}{||\beta||}|$$

In particular, for the canonical hyper plane, the numerator is equal to one and the distance to the support vectors is

$$\text{distance}_{\text{support vectors}} = \frac{|\beta_0 + \beta^T x|}{||\beta||} = \frac{1}{||\beta||}.$$

Recall that the margin introduced in the previous section, here denoted as M, is twice the distance to the closest examples:

$$M = \frac{2}{||\beta||}$$

Finally, the problem of maximizing M is equivalent to the problem of minimizing a function $L(\beta)$ subject to some constraints. The constraints model the requirement for the hyper plane to classify correctly all the training examples x_i. Formally,

$$distance_{support\ vector} = \frac{|\beta_0+\beta^T x|}{||\beta||} = \frac{1}{||\beta||}$$

Where y_i represents each of the labels of the training examples

This is a problem of Lagrangian optimization that can be solved using Lagrange multipliers to obtain the weight vector β and the bias β_0 of the optimal hyper plane.

KNN Classifier:

K Means: The aim of K Means is to partition the objects in such a way that the intra cluster similarity is high but inter cluster similarity is comparatively low. A set of n objects are classified into k clusters by accepting the input parameter k. All the data must be available in advance for the classification. [18]

KNN: Instead of assigning to a test pattern the class label of its closest neighbor, the K Nearest Neighbor classifier finds k nearest neighbors on the basis of Euclidean distance.

Square root of $((x_2-x_1)^2-(y_2-y_1)^2)$

The value of k is very crucial because the right value of k will help in better classification. [19]

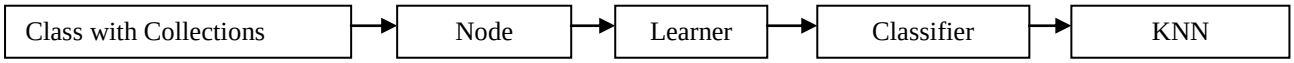

Fig No. 5 Steps in KNN Classifier [23]

This is a simple classifier that bases its decision on the distances between the training dataset samples and the test sample(s). Distances are computed using a customizable distance function. A certain number (**k**) of nearest neighbors is selected based on the smallest distances and the labels of these neighbouring samples are fed into a voting function to determine the labels of the test sample.

Training a kNN classifier is extremely quick; as no actual training is performed as the training dataset is simply stored in the classifier. All computations are done during classifier prediction. [23]

Naive Bayes Classifier:

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Given a class variable y and a dependent feature vector x_1 through x_n, Bayes' theorem states the following relationship:

$$P(y|x_1,\dots,x_n) = \frac{P(y)P(x_1,\dots x_n|y)}{P(x_1,\dots,x_n)}$$

Using the naive independence assumption that

for all i, this relationship is simplified to

$$P(y|x_1,\dots,x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1,\dots,x_n)}$$

Since $P(x_1,\dots,x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1,\dots,x_n) \propto P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y)\prod_{i=1}^{n} P(x_i \mid y),$$

and we can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i \mid y)$; the former is then the relative frequency of class y in the training set.

The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(x_i \mid y)$.

In spite of their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering. They require a small amount of training data to estimate the necessary parameters.

Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

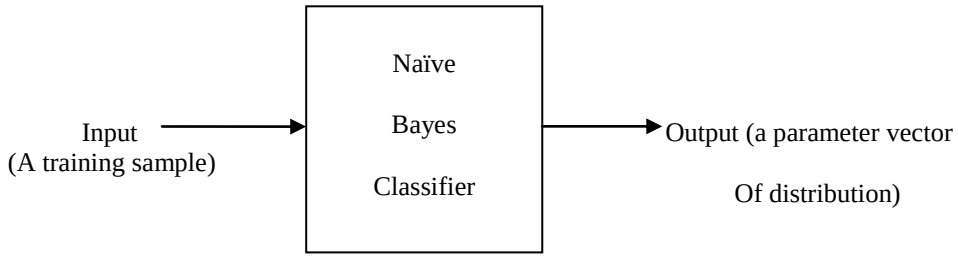

Fig No. 6

SGD Classifier:

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to discriminative learning of linear classifiers under convex loss functions such as (linear) Support Vector Machines and Logistic Regression. Even though SGD has been around in the machine learning community for a long time, it has received a considerable amount of attention just recently in the context of large-scale learning.

SGD has been successfully applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing. Given that the data is sparse, the classifiers in this module easily scale to problems with more than 10^5 training examples and more than 10^5 features. [22]

The advantages of Stochastic Gradient Descent are:

Efficiency

Ease of implementation (lots of opportunities for code tuning). [22]

The disadvantages of Stochastic Gradient Descent include:

SGD requires a number of hyper parameters such as the regularization parameter and the number of iterations.

SGD is sensitive to feature scaling. [22]

V. VECTOR SPACE MODEL

The **tf–idf** weight (term frequency–inverse document frequency) is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf–idf weighting scheme are often used by search engines to score and rank a document's relevance given a user query.

The *term frequency* in the given document is simply the number of times a given term appears in that document. This count is usually normalized to prevent a bias towards longer documents (which may have a higher term frequency regardless of the actual importance of that term in the document) to give a measure of the importance of the term *t i* within the particular document.

$$tf = \frac{n_i}{\sum_k n_k}$$

With n_i being the number of occurrences of the considered term, and the denominator is the number of occurrences of all terms.

The *inverse document frequency* is a measure of the general importance of the term (it is the logarithm of the number of all documents divided by the number of documents containing the term).

$$idf = log \frac{|D|}{|(d_i \supset t_i)|}$$

With

- |D| : total number of documents in the corpus
- $|(d_i \supset t_i)|$: Number of documents where the term *t i* appears.

Then

$$tfidf = tf.idf$$

A high weight in tf–idf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weight hence tends to filter out common terms.

For each document d_j and keyword k_i "tf-idf" is defined by the weight "w".

VI. CONCLUSIONS

Spam Detection is an automated technique to identify spam for the purpose of preventing its delivery. There are several different methods that spammers use to get your email address so that they can flood your inbox. Spam detection takes

these methods into account and uses that information to set up a line of defence against these annoying, unsolicited emails. There are many advantages of spam detection like it saves space in mail boxes, provides security against viruses, Trojans, and materials containing potentially harmful information for a certain category of users, saves time of users that they spend for sorting incoming mail and deleting undesirable correspondence. So spam detection is very useful. Various methods have been used for spam detection till now that has obtained a great success in spam detection.

ACKNOWLEDGMENT

This research paper is made possible through the help and support from my thesis guide.

First and foremost, I would like to thank Er. Sugandha Sharma for her most support and encouragement, she kindly read my paper and offered invaluable detailed advices on grammar, organization, and the theme of the paper.

Finally, I sincerely thank to my college HOD, teachers and friends. The product of this research paper would not be possible without all of them.

REFERENCES

[1] Gordon V. Cormack, *David R. Cheriton, "*Email Spam Filtering: A Systematic Review*",* Foundations and Trends ®in Information Retrieval Vol. 1, No. 4 (2006) 335–455©2008.

[2] M. Mangalindan, "For bulk E-mailer, pestering millions offers path to profit," *Wall Street Journal*, November 13, 2002.

[3] Fabrizio Sebastiani. "Machine learning in auto-mated text categorization- ACM Computing Surveys", 34(1):1-47, 2002.

[4] Qian Xu, Evan Wei Xiang and Qiang Yang, "SMS Spam Detection Using Non-Content Features" publication in IEEE Intelligent Systems, Nov.-Dec. 2012 (vol. 27 no. 6)pp. 44-51.

[5] K. S. Xu, M. Kliger, Y. Chen, P. J.Woolf, and A. O. Hero, "Revealing social networks of spammers through spectral clustering", in Proceedings of the IEEE International Conference on Communications, (ICC '09), Dresden, Germany, April 2013.

[6] Mohammad Razmara, Babak Asadi, Masoud Narouei, Mansour Ahmadi, "A Novel Approach toward Spam Detection Based on Iterative Patterns", 2012, IEEE.

[7] Sang Min Lee, Dong Seong Kim, Ji Ho Kim, Jong Sou Park, "Spam Detection Using Feature Selection and Parameters Optimization", pp. 883-888, 2010,IEEE.

[8] Ram B. Basnet, Andrew H. Sung, "Classifying Phishing Email Using Confidence-Weighted Linear Classifiers", pp. 108-112, 2010 IEEE.

[9] Juan Martinez-Romo, Lourdes Araujo, "Detecting malicious tweets in trending topics using a statistical analysis of language", Expert Systems with Applications 40 (2013) 2992–3000, Elsevier.

[10] Joshua Goodman, Gordon V.Cormack, and David Heckerman, "Spam and the Ongoing Battle for the Inbox", Communications of the ACM, February 2007/Vol.50,No.2.

[11] Sarwat Nizamani, Nasrullah Memon, Uffe Kock Wiil, Panagiotis Karampelas, "Modelling Suspicious Email Detection using Enhanced Feature Selection", April 2012.

[12] Qian Xu, Evan Wei Xiang and Qiang Yang, "SMS Spam Detection Using Non-Content Features" publication in IEEE Intelligent Systems, Nov.-Dec. 2012 (vol. 27 no. 6)pp. 44-51.

[13] M. IKONOMAKIS, S. KOTSIANTIS, V. TAMPAKAS, "Text Classification Using Machine Learning Techniques" WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, August 2005, pp. 966-974

[14] http://scikit-learn.org/stable/modules/naive_bayes.html

[15] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513{523, 1988.

[16] Y. Yang and J. Pedersen. "A comparative study on feature selection in text categorization." In International Conference on Machine Learning (ICML), 1997

[17] http://docs.opencv.org/2.4.5/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

[18] Sudipto Guha, Adam Meyerson, Nina Mishra, Rajeev Motwani and Liadan O'Callaghan, "Clustering Data Streams", IEEE Transactions on Knowledge & Data Engg., 2003

[19] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques ", Second Edition.

[20] http://www.iicm.tugraz.at/cguetl/courses/isr/opt/classification/Vector_space_Model.html

[21] Sahil Puri, Dishant Gosain, Mehak Ahuja, Ishita Kathuria, Nishtha Jatana, "Comparison And Analysis Of Spam Detection Algorithms", IJAIEM, Volume 2, Issue 4, April 2013

[22] http://scikit-learn.org/stable/modules/sgd.html

[23] http://www.pymvpa.org/generated/mvpa2.clfs.knn.kNN.html

