ISSN:0975-9646

Survey on Text Classification (Spam) Using Machine Learning

Neetu Sharma
PTURC
SSIET, Derabassi, Punjab

GaganpreetKaur
A.P., CSE
SSIET,Derabassi

Ashish Verma
Asst. Professor ,Deptt. of CSE & IT
SSIET, Dera Bassi, Punjab

Abstract— E-mail spam is a very serious problem in today's life. It has many conséquences like it causes lower productivity, occupy space in mail boxes, extend viruses, Trojans, and materials containing 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. There are various classifiers used for e-mail spam detection like Naïve Bayes Classifier, KNN Classifiers etc. We will use SVM classifier for e-mail spam detection which has not been used till now for detecting e-mail spam.

Index Terms—Spam, Text Classification ,Spam Classifier Methods

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]

E-mail is an effective, fast and cheap communication way. Therefore spammers prefer to send spam through such kind of communication. Nowadays almost every second user has an E-mail, and consequently they are faced with spam problem. E-mail Spam is non-requested information sent to the E-mail boxes. Spam is a big problem both 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 average global spam rate for the year was 89.1%, with an increase of 1.4% compared with 2012. The proportion of spam sent from botnets was much higher for 2013, accounting for approximately 88.2% of all spam. Despite many attempts to disrupt botnet activities throughout 2013, by the end of the year the total number of active bots returned to roughly the same number as at the end of 2012, with approximately five 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 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. According to a report from Ferris Research, the global sum of losses from spam made about 130 billion dollars, and in the USA, 42 billion in 2012. [4] Besides expenses for acquisition, installation, and service of protective means, users are compelled to defray the additional expenses connected with an overload of the post traffic, failures of servers, and productivity loss. So we can do such conclusion: spam is not only an irritating factor, but also a direct threat to the business. Considering the stunning quantity of spam messages coming to E-mail boxes, it is possible to assume that spammers do not operate alone; it is global, organized, creating the virtual social networks. They attack mails of users, whole corporations, and even states.

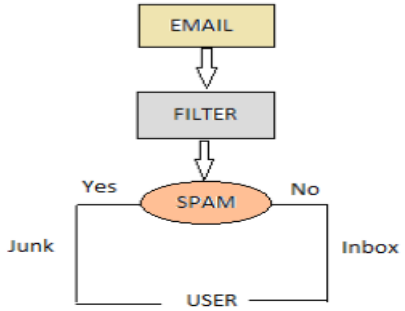


Figure 1 : Block diagram of spam filter

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, proposed parameter optimization and feature selection to reduce processing overheads with guaranteeing high detection rates. In previous papers, either parameter optimization or feature selection, but not both. Parameters optimization is to regulate parameters of spam detection models to figure out optimal parameters of the detection model. Feature selection is to choose only important features or feature set out of all the features. Feature selection enables to eliminate irrelevant features to avoid processing overheads. [7]

Razmara et.al in his work, present a novel solution toward spam filtering by using 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 representation of texts has been offered in Salton's works. In the elementary case, the vector model assumes comparison to each document of a frequency spectrum of words. The dimension of space is reduced by rejection of the most common words that increases thereby percent of the importance of the basic words in more advanced vector models. The possibility 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].

The comparison of advanced topics like multi-objective and ensemble-based evolutionary clustering; and the overlapping clustering are also mentioned in that paper. 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.

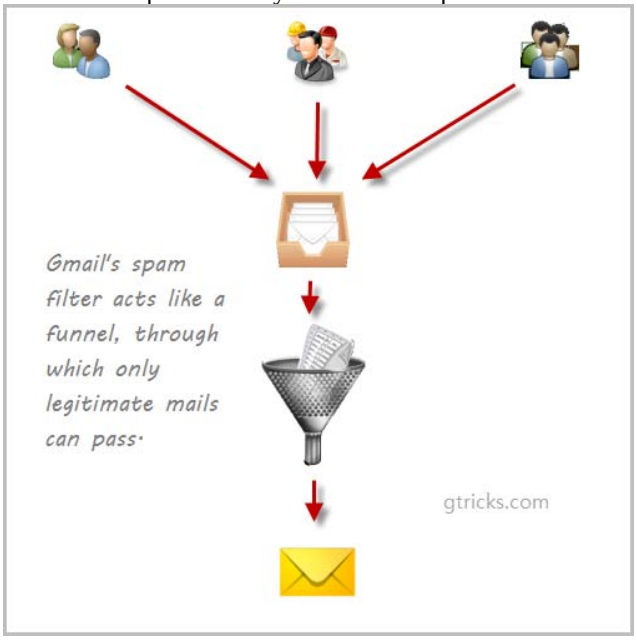


Fig No. 2 Text Classification Process [13]

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 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". Following the recommendation of [16], the information gain criterion will be used in this paper to select a subset of features. Finally, from IR it is known that scaling the dimensions of the feature vector with their inverse document frequency (IDF) [15] improves performance. Here the \tfc" variant is used. To abstract from different document lengths, each document feature vector is normalized to unit length.

IV. Spam Classifier Methods

There are several methods for spam detection. These methods include SVM, KNN, 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.

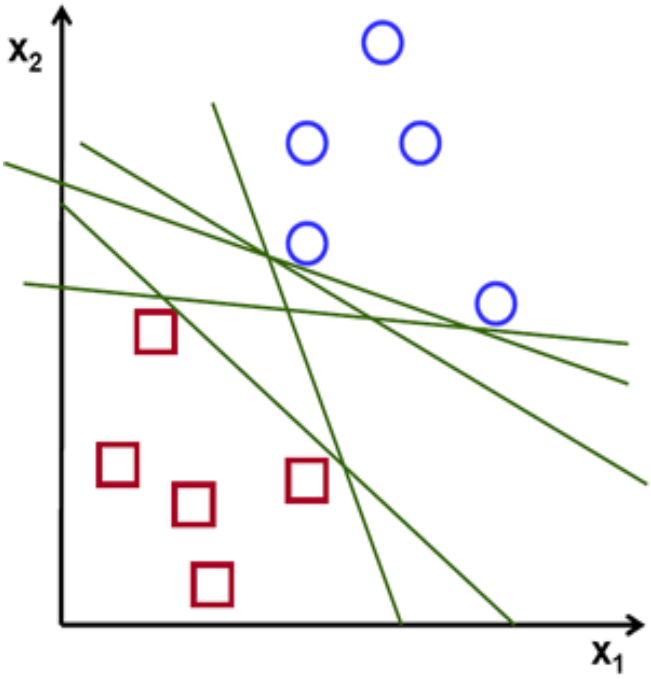


Fig no.3[17]

In the above picture you can see that there exists 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.

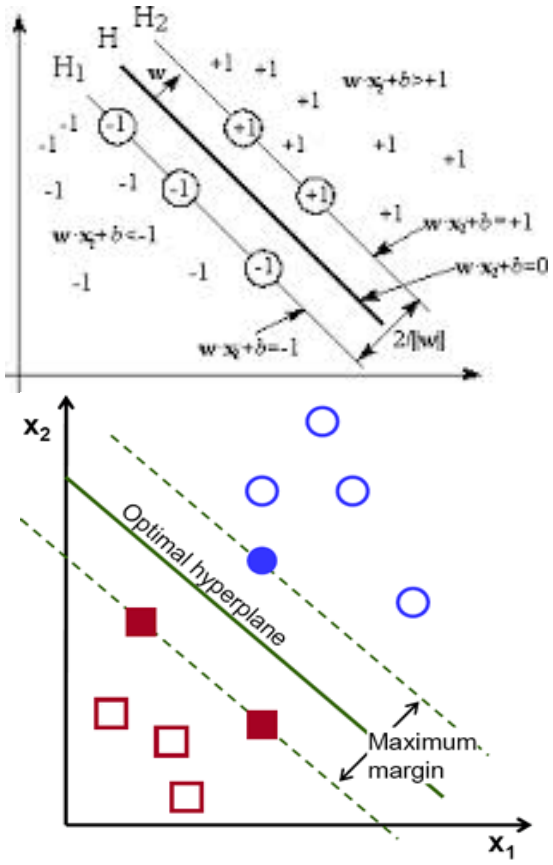


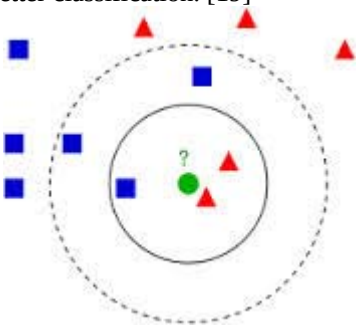
Fig No. 4[17]

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]

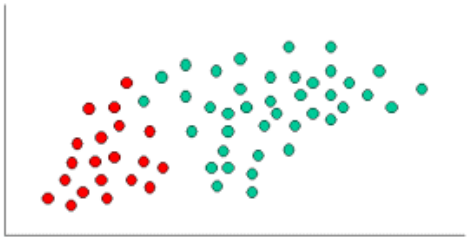
KNN: Instead of assigning to a test pattern the class label of its closest neighbor, the K Nearest Neighbor classifier finds k nearest neighbors on the basis of Euclidean distance. Square root of $((x_2-x_1)^2-(y_2-y_1)^2)$

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



Naive Bayes Classifier:

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.



$$P(y|f_1,\ldots,f_m) = \frac{P(f_1,\ldots,f_m|y)P(y)}{P(f_1,\ldots,f_m)} = \frac{P(y)\prod_{i=1}^{m}P(f_i|y)}{P(f_1,\ldots,f_m)}$$

$$\arg\max_y P(y|f_1,\ldots,f_m) = \arg\max_y \frac{P(y)\prod_{i=1}^{m}P(f_i|y)}{P(f_1,\ldots,f_m)} = \arg\max_y P(y)\prod_{i=1}^{m}P(f_i|y)$$

Fig No. 5 [14]

To demonstrate the concept of Naïve Bayes Classification, consider the example displayed in the illustration above. As indicated, the objects can be classified as either GREEN or RED. Our task is to classify new cases as they arrive, i.e., decide to which class label they belong, based on the currently existing objects.

Since there are twice as many GREEN objects as RED, it is reasonable to believe that a new case (which hasn't been observed yet) is twice as likely to have membership GREEN rather than RED. In the Bayesian analysis, this belief is known as the prior probability. Prior probabilities are based on previous experience, in this case the percentage of GREEN and RED objects, and often used to predict outcomes before they actually happen.

Thus, we can write:

$$Prior\ probability\ for\ GREEN \propto \frac{Number\ of\ GREEN\ objects}{Total\ number\ of\ objects}$$

$$Prior\ probability\ for\ RED \propto \frac{Number\ of\ RED\ objects}{Total\ number\ of\ objects}$$

Since there is a total of 60 objects, 40 of which are GREEN and 20 RED, our prior probabilities for class membership are:

$$Prior\ probability\ for\ GREEN \propto \frac{40}{60}$$

$$Prior\ probability\ for\ RED \propto \frac{20}{60}$$

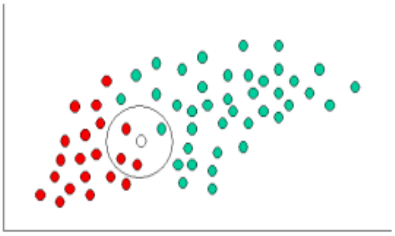


Fig No. 6[14]

Having formulated our prior probability, we are now ready to classify a new object (WHITE circle). Since the objects are well clustered, it is reasonable to assume that the more GREEN (or RED) objects in the vicinity of X, the more likely that the new cases belong to that particular color. To measure this likelihood, we draw a circle around X which encompasses a number (to be chosen a priori) of points irrespective of their class labels. Then we calculate the number of points in the circle belonging to each class label. From this we calculate the likelihood:

$$Likelihood\ of\ X\ given\ GREEN \propto \frac{Number\ of\ GREEN\ in\ the\ vicinity\ of\ X}{Total\ number\ of\ GREEN\ cases}$$

$$Likelihood\ of\ X\ given\ RED \propto \frac{Number\ of\ RED\ in\ the\ vicinity\ of\ X}{Total\ number\ of\ RED\ cases}$$

From the illustration above, it is clear that Likelihood of X given GREEN is smaller than Likelihood of X given RED, since the circle encompasses 1 GREEN object and 3 RED ones. Thus:

$$Probability\ of\ X\ given\ GREEN \propto \frac{1}{40}$$

$$Probability\ of\ X\ given\ RED \propto \frac{3}{20}$$

Although the prior probabilities indicate that X may belong to GREEN (given that there are twice as many GREEN compared to RED) the likelihood indicates otherwise; that the class membership of X is RED (given that there are more RED objects in the vicinity of X than GREEN). In the Bayesian analysis, the final classification is produced by combining both sources of information, i.e., the prior and the likelihood, to form a posterior probability using the so-called Bayes' rule (named after Rev. Thomas Bayes 1702-1761).

$$Posterior\ probability\ of\ X\ being\ GREEN \propto$$

$$Prior\ probability\ of\ GREEN \times Likelihood\ of\ X\ given\ GREEN$$

$$= \frac{4}{6} \times \frac{1}{40} = \frac{1}{60}$$

$$Posterior\ probability\ of\ X\ being\ RED \propto$$

$$Prior\ probability\ of\ RED \times Likelihood\ of\ X\ given\ RED$$

$$= \frac{2}{6} \times \frac{3}{20} = \frac{1}{20}$$

Finally, we classify X as RED since its class membership achieves the largest posterior probability.

V. CONCLUSION

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 defense 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.

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, "Modeling Suspicious Email Detection using Enhanced Feature Selection", April 11,2012 .

[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://www.statsoft.com/Textbook/Naive-Bayes-Classifier

[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 Edn.