A Social Network Spam Detection Model

Odukoya Oluwatoyin Helen

Abstract—The emergence of social networking sites in the early millennium has brought about another concern for a different spam type; social spam. Social spam emerged as a result of the posting and messaging capabilities of social networking site. The methodology used to develop this model is the Support Vector Machine which is used as a binary classifier to classify the postings from a social networking environment into both spam and non-spam. In the process of simulating the design of the model, the training dataset (representing postings from a social networking site) is used to train the SVM by supplying the training datasets in iterations into the classifying function (SMO). Based on the analysis of the result produced by testing the dataset supplied as the input into the SVM, it can be concluded that the SVM proved to be a good spam classifier with a high accuracy measure and True positive rate

Index Terms— Social spam, Support Vector Machine, accuracy, true positive.

1 INTRODUCTION

Social network sites according to [1] are webbased services that allow individuals to firstly construct a public or semi-public profile within a bounded system secondly articulate a list of other users with whom they share a connection and thirdly view and transverse their list of connections and those made by others within the system.

Social networks have independently and collectively gathered a huge number of monthly active users (MAUs) and this has made it easy to target huge chunk of audience on a single platform. The intense traffic pulled by various major online social networking sites (SNS) has made major social networks with credible traffic hit, a business hub for large companies, politicians, business owners, government agencies and ministries, and other organizations by implementing several advertisement and strategic marketing models and campaigns to increase their transactions and demanding business population. Concurrently, with the emergence of the high traffic driven by social networks, the attention of internet fraudsters and spammers has not been kept out of the story as the high traffic equally serves as an extremely useful tool in facilitating the core of their everyday objectives. Whilst the existence of spamming cannot be over-emphasized in today's technological era, owing to the need to pass information across to target audience via the most economic and rather efficient means (social networking), the need to ensure that users are secured while they spend time on social networks must not be left unattended to.

The emergence of social networking and its high traffic potentials (with Facebook having a Monthly Active Users count of 1.36 billion users in April, 2014) has attracted not only spammers and fraudsters, but also marketers with the need to meet up with sales targets and deadlines.

Spammers (mis)use the popularity and the high PageRank of social bookmarking systems for their purposes. All they need is an account; then they freely post entries which bookmark the target spam web site [2]

Detecting spam messages in online social networks is therefore required to provide security and guarantee users' safety against threats that emancipates out of spam messages. Such threats include phishing, pharming and other intruding / unauthorized information mining process.

This paper is organized as follows, the second section discusses the past and related works on social web spam, the third section the methodology, the forth section discusses the simulation approach; the fifth section discusses the evaluation. Finally, conclusion is drawn in the sixth section.

• *Odukoya Oluwatoyin Helen is currently Lecturing undergraduate and post graduate students in Computer Science and Engineering Department Obafemi Awolowo University, Ile-Ife Osun State Nigeria, PH-08139484145. E-mail: oodukoya@oauife.edu.ng*

2 LITERATURE REVIEW

2.1 Review Stage

The history of spam dates back to the beginning of electronic communication. As technologies evolved, spams have continued to associate with existing and newer communication technologies, from telephones, emails, instant messages and social networks. According to [3], there has been a growth percentage of 355% in the first half of 2013. This increase was observed in the social spams on an individual social media account. Spammers are turning to the fastest growing communication medium to circumvent traditional security infrastructures that were used to detect email spams.

Social media has led to new methods of delivering spams, such as spam-related apps, so-called 'like-jacking', social bots, and fake accounts. Spam-related apps offer to perform special task outside of social media networks original features. With like-jacking, instead of clicking on the malicious links, victims may be tricked into clicking on images that appear as likes or seemingly harmless buttons. Social bots and fake accounts are used to infiltrate the victim's social media network. Together, these new attack methods can significantly detract from a brand's social presence and their social marketing ROI [3].

In order to develop an effective detection system for social spam, it is of high importance to identify the motive and underlining behind social spams. The growth of social spam has

been exponential, owing to the traffic driven on social networks like Facebook, Twitter, Google Plus, Instagram, LinkedIn and the likes. Also, online social networks have become a primary alternative for communication locally and globally. The information dissemination characteristics have also added to the underlying notion that has attracted the interest of spammers.

Most previous work on social spam has focused on spam prevention on a single social network (e.g. Facebook [4], [5], My Space [6], Twitter [7]. A number of techniques have been implemented by [8] which include classification, collaborative filtering, behavoural analysis, and in some cases, friend-graph analysis. These techniques were implemented to cater for the robust nature of social spams which was generated by the wide range of motivations from spammers. The study by [9] focused on automatically detecting web spam using email spam or detecting Twitter spam using web pages. In addition to this, web spam classification methods were understudied by [9] and social profile spam detection (as demonstrated on My Space) methods used by [10]

The work of [11] survey the field of spam on social networking sites (SNSs), identifying several common approaches. Identification-based approaches identify spam to train classifiers based on labels submitted by users or trusted moderators. Rank-based approaches demote visibility of questionable content, while interface-based approaches apply policies to prevent unwanted behavior .This work groups classification-based approaches with detection, although classifiers can be used in conjunction with user information to prevent spam before it happens.

A large number of classifiers have been used in spam detection not choosing the right classifier and the most efficient combination of them is still a problem. Previous work by [12] proposes a Bayesian framework, which is a theoretically efficient and practically reasonable method of combination, when investigating the integration of text and image classifiers. Several novel classification approaches were proposed and implemented in cross-domain text classification. [13] presented semantics-based algorithm for cross domain text classification using Wikipedia based on co-clustering classification algorithm. [14] described a novel and efficient centroid-based algorithm Class-Feature Centroid Classifier (CFC) for cross-domain classification of weblogs; also they have discussed the trade-off between complexity and accuracy.

Many of the data mining algorithms used to detect spam and patterns of misuse on SNSs are designed with the assumption that the data and the classifier are independent. However, in the case of spam, fraud and other malicious content, users will often modify their behavior to evade detection, leading to degraded classifier performance and the need to re-train classifier frequently. Several researchers tackle this adversarial problem [15]

A modified Naïve Bayes classifier is proposed to detect and reclassify data taking into account the optimal modification strategy than an adversary could choose [16]. A framework for reverse engineering a classifier is provided in [17] to determine whether an adversary can efficiently learn enough about classifier to effectively defeat it.

Also, some URL spam filtering techniques have been promised by [18] to better address different web services such as social networks. They presented a real time URL spam filtering system named Monarch and demonstrated a modest deployment of this system on cloud infrastructure and its scalability.

In [19] a spam detection framework was implemented to detect spam on multiple social networks, developing a multiple framework that can be applied to multiple social networks with a resilient structure to evolution due to the spam arm-race. The implementation was poised for futuristic test and evaluation on live feeds from social networks. However, [19] failed to integrate the behavior of spammers into the developed framework; an aspect that was considered as a future work.

A Facebook application using data mining to detect spam was demonstrated in [20]; the blacklist, keyword blocking were applied first. Then the data mining model developed was used to detect spams further. The features used were the number of links, the number of words, and the length of posts etc. the precision and recall rate achieved was around 61-63%. These rates were said to be due to the small training sets which could be improved with more training data availability. The system integration was demonstrated where the user page could redirect to the web server for spam detection first.

3 METHODOLOGY

Most of the social networking sites have become vulnerable to users through the invasion of spam messages on these platforms. Architectural approaches attempt to solve this by creating a model capable of classifying spam contents in a social networking environment. This protection methodology is effective for many social networking sites architecture; however, some spams still find their way through as the system, as accurate as it is, is not a hundred per cent spam proof.

The software that will be used in the course of implementing this project is WEKA. WEKA is an open source under the GNU General Public License. System developed at the University of Waikato in New Zealand. "Weka" stands for Waikato Environment for Knowledge Analysis. The system is written using object oriented language Java. There are several different levels at which Weka can be used. Weka provides implementations of state-of-the-art data mining and machine learning algorithms. Weka contains modules for data preprocessing, classification, clustering and association rule extraction.

The WEKA will be used with Support Vector Machine. Some data sets, representing the Social networking postings (messages, comments etc), will be passed into the SVM as the input. This will be processed by WEKA to generate the output which will be the two classifications of the postings; that is, the one that is infected with spam and the other, otherwise

3.1System Architecture

The SNSs are web-based applications that run in the cloud. The SVM is used within WEKA for data mining. The model is first trained with a set of training data before the test data is run for classification using the SVM model within the WEKA simulation environment.

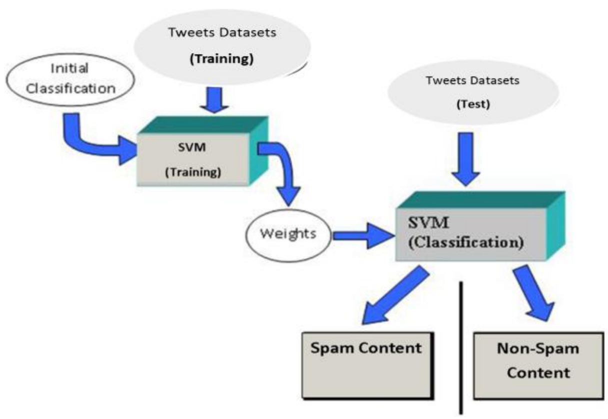

Fig. 1. The standard Support Vector Machine Architecture.

3.2 Support Vector Machines

The Support Vector Machine (SVM) is a widely used classifier. It is a state-of-the-art classification method introduced by [21]. It is mostly used in bioinformatics as a result of its high accuracy, flexibility in modeling diverse sources of data and ability to deal with high-dimensional data such as gene expression.
SVM are a new learning method relatively used for binary classification. The idea behind this is to find a hyper-plane that separates the dimensional data into two classes. SVMs belong to the general category of kernel method which is an algorithm that depends on the data through dot-products. When this happens, the dot product can be replaced by a kernel function that computes a dot product in some high dimensional feature space. There are two benefits of this: Firstly is the ability to generate non-linear decision boundaries through the use of methods designed for linear classifiers. Secondly is the use of kernel functions that allows user to apply a classifier to data which have no obvious fixed-dimensional vector space representation. Example of such data in bioinformatics are sequence, either protein or DNA, and protein structure.
In order to use the SVMs effectively, it is important to understand how they work. When training an SVM, a number of decisions are to be made by the practitioner including how to preprocess data, what kernel to use and also setting the parameters of the kernel and the SVM.

3.3 How the Support Vector Machine Works

Considering a real-valued instance space < $X=R^D$>, a linear support vector machine constructs the separating hyper-plane where (a, b) stands for the dot product of vectors a and b:

$$h(x) = \langle w, x \rangle + b \quad (1)$$

which maximizes the margin which is the minimal distance between the hyper-plane and the examples, when that kind of separating hyper-planes exist. A formulation, slightly more complex, involving the so-called slack variables x_i, is defined to deal with noise. [22].
The function to be optimized, the L_2 norm of the hyper-plane normal vector w, is quadratic: using Lagrange multipliers to account for the constraints gives rise to the so-called dual formulation. Assume the support vectors are those examples whose constraint is active (Lagrange multiplier α_i>0), then it becomes

$$h(x) = \sum y_i \alpha_i \langle x_i, x \rangle + b \text{ with } \alpha_i > 0; \sum \alpha_i y_i = 0 \quad (2)$$

The power of SVMs comes from the so-called kernel trick, naturally exporting the SVM approach to nonlinear hypothesis spaces. On mapping the instance space onto some feature space via mapping ϕ. If the scalar product on the feature space can be computed in X, then a linear classifier in the feature space (nonlinear with reference to X) is given as equation (2).
The only requirement is to use a positive definite kernel (ensuring that the underlying optimization problem is well posed). Again, this requirement can be relaxed in the evolutionary learning framework [23].
Figure (2) below shows the optimal hyper-plane that maximizes the minimal distance to the examples.

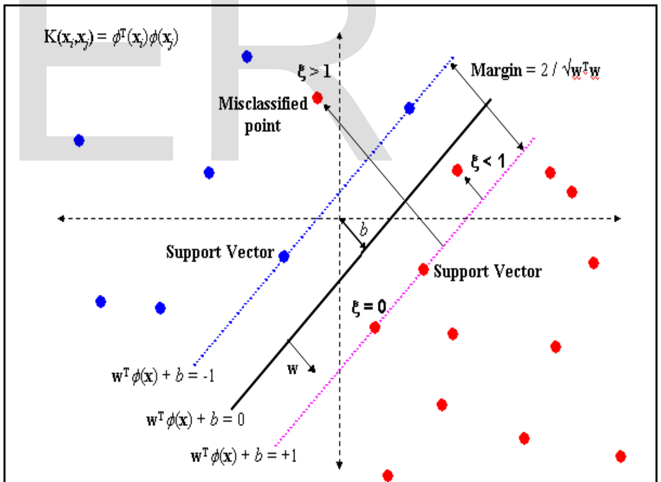

Fig. 2. The Standard SVM formulation

4 SIMULATION

The implementation was done using a classifying tool in WEKA, called the Support Vector Machine (SVM) which is majorly used for classifying 2-class problems. That is, problems that produces only two forms of output as we have in spam and non-spam contents in the classification phase.
The SVM is used to build the spam filter which will classify an input representing a social content, x, as spam (y=1) or as non-spam (y=0). These social postings are expected to be converted into a feature vector x £ R^n.

4.1 Simulation Procedures

In carrying out this project work, a set of data (datasets)

representing postings from a popular social networking environment (Twitter) was analyzed using WEKA (Waikato Environment for Knowledge Analysis). These datasets (representing train datasets and test datasets accordingly) were used to train and then test the dataset (representing tweets and related postings) using the Support Vector Machine as the classifier, in order to classify the post instances (postings representing tweets) as spam and/or non-spam contents

The Sequence Minimal Optimization Algorithm [12], which is an implemented class of Support Vector Machine in the WEKA environment, is called to classify the dataset for the training and the testing phases.

The training dataset is summarized in table (1) on the next page. The training dataset which comprises 400 instances was iterated ten times during the training period of the support vector machine (SVM). The trained model was then tested with the test data and the results obtained have been detailed in the later parts of this chapter.

The procedures for the simulation in Weka can be described with the following steps;

i. Prepare a training dataset (tweets)
ii. Open Weka software
iii. Open the training dataset in Weka
iv. Select SVM module in Weka (SMO)
v. Choose appropriate parameters SVM
vi. Selected test options
vii. Selected responses
viii. Results
ix. Prediction information (model estimators)
x. *And finally*, run Test data on trained SVM model

The results obtained from the simulation processes have been adequately discussed here with acute attention on the key model estimators

TABLE 1
SUMMARY OF THE PROPERTIES OF TRAINING DATASET

Number of Instances	400
Number of Attributes	43
Number of Kernel Evaluations	18625
Number of Classes	2

5 EVALUATION

5.1 Evaluation of Results

In discussing the results obtained from the simulation process in this chapter, key model estimators will be used and they are as follows;

Precision is best understood as the proportion of instances that are truly of a class divided by the total instances classified as that class.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall is described as the proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate).

$$\text{Recall} = \frac{TP}{TP + FN}$$

True Positive (TP) rate is the rate of true positives (instances correctly classified as a given class).

$$TruePositive(TP)rate = \frac{TP}{TP + FN}$$

False Positive (FP) rate is the rate of false positives (instances falsely classified as a given class).

$$FalsePositive(FP)rate = \frac{FP}{FP + TN}$$

While **Accuracy** measures the degree of correctness of the classifier; **Error rate** signifies the degree of deviation from correctness of the classified results/output.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The **Classification time** simply indicates how long it takes the system to classify an inputted dataset.

5.2 Results and Result Discussion

In training the SVM with the provided datasets, balanced datasets were fed into the Weka system in iterations (10). These iterations are seen in Table 4.2 which shows a summary of the results obtained based on some of the key model estimators.

The results from the training session revealed the model high accuracy ability with an accuracy measure of 98.5%. This accuracy measure is very reliable in binary classification. The accuracy is easily derived from the confusion matrix, where the TP and the TN are summed up over all other parameter in the matrix as seen below;

Also considering the high possibility of overfitting and the derived True Positive rate which measured consistently at 100% over nine (9) iterations, the reliability of the developed model to handle efficiently social spam classification despite the large volume of attribute (43) is very commendable. The confusion matrix reveals the TP and FP values as very reliable with a measure of 200 and 194 out of 200 apiece respectively. It was also observed that the classification time was remarkably negligible when compared with the corresponding classification time when using other classifiers besides implementations from SVM (like the SMO – as used in this case).

The F-measure estimated a remarkable value at 0.985 (the closer the F-measure value is to 1.0, the more reliable the system/model is).

The classification time was observed to be directly proportional to the size of the dataset / number of instances; implying that as the number of instances increased the classification time also increased.

The accuracy of the model measured at an index of 0.985 (equivalent to 98.5%) which is very remarkable for a classifying system.

The True Positive value showed that out of 200 spam contents, the system classified 200 correctly as spam and the False Positive value showed that out of 200 non-spam contents, the system classified 194 correctly as non-spam.

The precision rate which stood at 98.5% revealed the classifier's ability to determine the number of selected instances are relevant (spams).

The recall rate on the other hand, which reflected a rate of 98.5% showed the number of relevant instances that were selected by the classifier

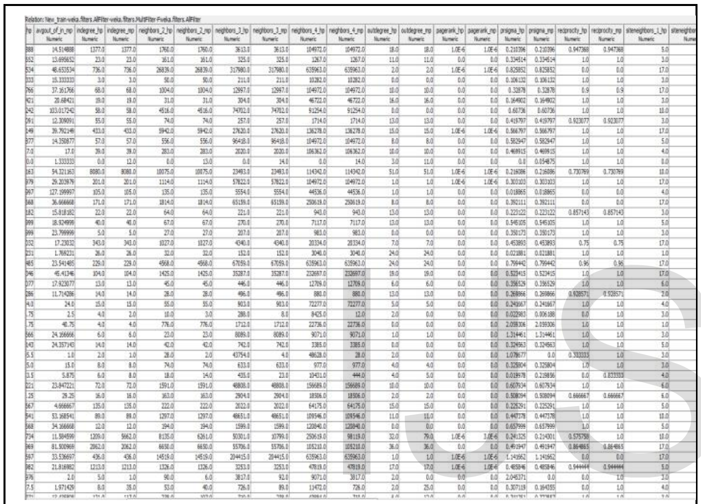

Fig. 3. Input training dataset (400).

ITERATED RESULTS OF TRAINING DATASET

No Of Instances	Accuracy %	Error%	TP rate %	FP rate %	Classification time (sec)
440	100	0	100	0	0.03
80	100	0	100	0	0.04
120	100	0	100	0	0.05
160	100	0	100	0	0.06
200	100	0	100	0	0.07
240	100	0	100	0	0.09
280	100	0	100	0	0.10
320	100	0	100	0	0.11
360	100	0	100	0	0.14
400	98.5	1.50	98.5	1.50	0.2

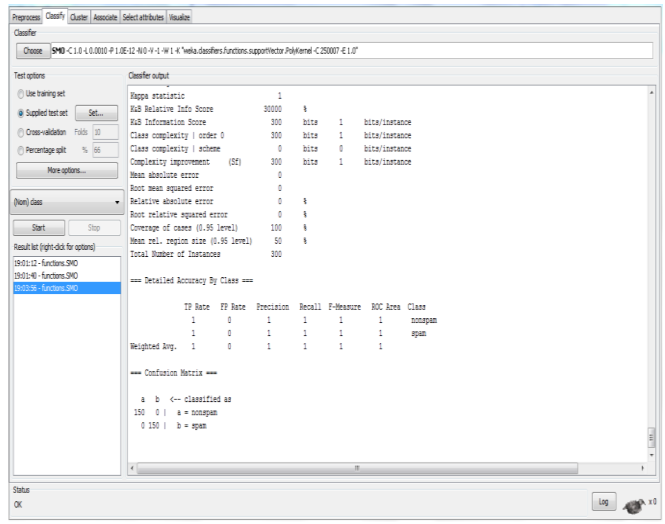

Fig. 4. Weka output interface of test dataset (300).

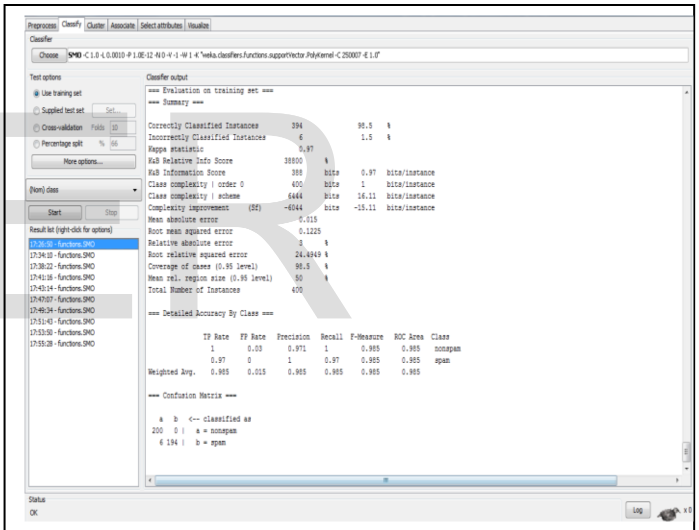

Fig. 5.Output Interface of the training dataset.

6 CONCLUSION

It can therefore be concluded that while a number of existing models exist for detecting spammy contents across social networking sites, this model displays a high reliability level based on the accuracy of the system. Further works in this line is therefore encouraged for further breakthroughs in detecting spams on social networking sites.

ACKNOWLEDGMENT

This research is funded by Center of Excellence in Software Engineering StepB/World Bank Grant. The grant is meant to set-up a world class center of excellence in Software engineering, software products development, high calibre man-power training and centre of academic excellence at the Obafemi Awolowo University, Ile-Ife.Nigeria.

REFERENCES

[1] Boyd, d. m. and Ellison, N. B. (2007), Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication, 13: 210–230. doi: 10.1111/j.1083-6101. 2007

[2] B. Krause, C. Schmitz, A. Hotho, and G. Stumme. The anti-social tagger: detecting spam in Social bookmarking systems. In Proc. 4th Intl. Workshop on Adversarial Information Retrieval on the Web (AIRWeb)

[3] Harold Nguyen: State of Social Media Spam, Research Report. Nextgate 2013

[4] Angie Beal (2009) Webopedia Listings Retrieved November 11, 2014 from http://www.webopedia.com/TERM/S/social_networking_site.html

[5] T. Stein, E. Chen, and K. Mangla. Facebook Immune System. In Proceedings of the 4th Workshop on Social Network Systems, SNS '11, pages 8:1{8:8, New York, NY, USA, 2011. ACM.

[6] F .Benevenuto, T. Rodrigues, V . Almeida, J. Almeida, and M. G on ¸calves. Detecting spammers and content promoters in online video social network s. In Proc. Of SIGIR, 2010.

[7] D. Wang, D. Irani, and C. Pu. A Social-Spam Detection Framework. In Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, CEAS '11, pages 46{54, New York, NY, USA, 2011. ACM.

[8] Dalvi, N. N.; Domingos, P.; Mausam; Sanghai, S.; and Verma, D. 2004. Adversarial classification. In KDD, 99–108.

[9] P. Heymann, G. Koutrika, and H. Garcia-Molina. Fighting spam on social web sites: A survey of approaches and future challenges. Internet Computing, IEEE, 11(6):36 {45, Nov.-Dec. 2007

[10] Giangluca Stringhini, Christopher Kruegel, Giovanni Vigna. Detecting Spammers on Social Networks (2010) ACSAC 2010 December 6-10 Austin, Texas, USA

[11] Antonio Lupher, Cliff Engle, Reynold Xin. Feature Selection and Classification of Spam on Social Networking Sites. 2012

[12] J. Platt: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning, 1998.

[13] M Soiraya, S Thanalerdmongkol and C Chantrapornchai. Article: Using a Data Mining Approach: Spam Detection on Facebook. International Journal of Computer Applications 58(13):27-32, November 2012.

[14] G. Brown, T. Howe, M. Ihbe, A. Praskash and K. Borders. Social Networks and content aware spa. In ACM Conference on Support Cooperative work, 2008

[15] S. Webb, J. Caverlee, and K . Lee. Uncovering social spammers: Social honey pots + machine learning. In Proc. of SIGIR, July 2010.

[16] Corinna Cortes and Vladimir Vapnik. Support Vector Networks. AT&T Laboratory Research, Kluwer Academic Publishers, Boston, USA 1995

[17] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg. Article: 10 Top Algorithm in Data Mining. In the IEEE International Conference on Data Mining (ICDM) in December 2006

[18] Mierswa, Ingo. Evolutionary Learning with Kernels: A Generic Solution for Large Margin Problems. In Proc. of the Genetic and Evolutionary Computation Conference (GECCO 2006), 2006.

[19] Techopedia.com, Retrieved October 2, 2014 from http://www.techopedia.com/definition/4956/social-networking-site-sns

[20] http://cuterank.net/what-is-link-spam/

[21] http://www.pcmag.com/encyclopedia/term/55316/social-networking-site

[22]