Identifying Text Spam Based on Supervised Machine Learning Models: A Review

Pooja Sharma[1] Dr. Manish Vyas[2]

***Abstract*— Web users are continuously bombarded with spamming attacks. Sometimes, such spamming attacks may successfully re-direct the user to a malicious web link which is generally termed as re-direction spam. However, genuine re-directions are also common in case the web-servers are overloaded with requests which are more that that can be processed. It can be challenging to distinguish redirection spam from actual web-redirections. Lately, artificial intelligence has been used for redirection spam classification using the design of various models of artificial neural networks. The performance parameters are generally the accuracy and mean square error. This paper presents a comprehensive survey on redirection spamming attack detection using artificial intelligence based approaches so as to thwart spamming attacks for time critical applications. Various models have been discussed with their pros and cons.**

Keywords—Mobile Spam Classification, Machine Learning, Classification Accuracy.

I. INTRODUCTION

With increasing number of users using web services, the problem of spamming attacks has become very serious among both web and mobile applications. The most common form of spamming attack encountered is the redirection spam attack on cellular users. The main challenge is to detect whether a redirection is actually spam or not. With the vast amount of data and its complexity, manual classification of redirection spam in time critical applications is infeasible. This paper presents the basics of re-direction spam classification using artificial intelligence based techniques. It introduces the necessity of spam classification along with the various approaches used to classify them using artificial neural networks.

Generally it is difficult to classify based on auto-redirect or auto-refresh tag because when web servers are heavily loaded, they may introduce such measures to release load and avoid web server crashes. Hence it becomes mandatory to look for techniques which can classify with high accuracy in time critical aspects and situations. The following section presents the basics of artificial neural networks used for spam classification.

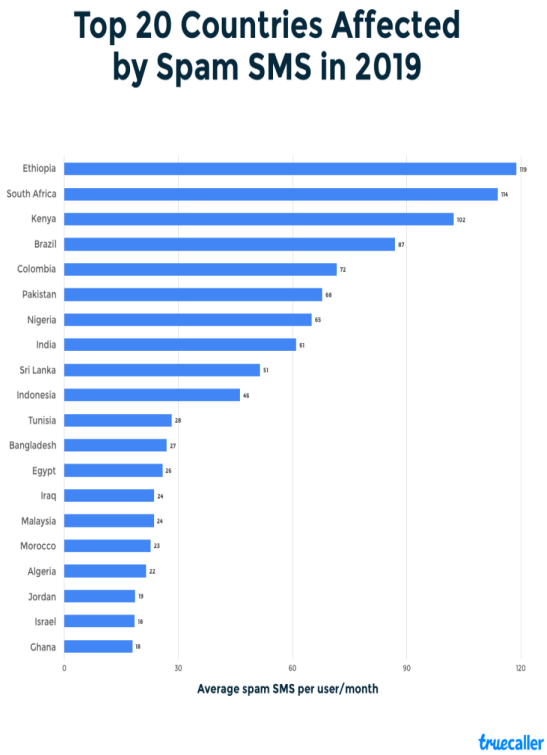

Fig.1 Countries Affected By SMS Mobile Spams (Source: Truecaller) https://dazeinfo.com/2020/08/24/indias-spam-sms-problem-are-these-smart-sms-blocking-apps-the-solution/

Some of the spamming attacks may be benign while others may be malignant trying to redirect mobile users to malicious websites where user security may be

compromised. Since the amount of data is staggering large and complex, off late machine learning based approaches are becoming common to filter out spams. One of the challenges which machine learning based approaches face for mobile spamming platforms is the limited computational and processing capabilities of hand held mobile devices. This makes is necessary to design and test algorithms which are compatible with various versions of mobile operating systems and also supported by limited memory and processing hardware as there exists a lot of diversity in the mobile hardware of different devices.

II. RELATED WORK

Various approaches have been devised for mobile spam classification.

AK Jain et al. proposed an approach for the detection of spam messages. We have identified an effective feature set for text messages which classify the messages into spam or ham with high accuracy. The feature selection procedure is implemented on normalized text messages to obtain a feature vector for each message. The feature vector obtained is tested on a set of machine learning algorithms to observe their efficiency.

KS Adewole et al. proposed a unified framework is proposed for both spam message and spam account detection tasks. Authors utilized four datasets in this study, two of which are from SMS spam message domain and the remaining two from Twitter microblog. To identify a minimal number of features for spam account detection on Twitter, this paper studied bio-inspired evolutionary search method. Using evolutionary search algorithm, a compact model for spam account detection is proposed, which is incorporated in the machine learning phase of the unified framework. The results of the various experiments conducted indicate that the proposed framework is promising for detecting both spam message and spam account with a minimal number of features.

Aliaksandr Barushka et al. proposed a technique based on integrated distribution-based balancing approach for spam classification. The concept of deep neural networks is used in this paper. The major advantage of this approach is the distribution mechanism makes the computation of different parameters for classification simpler. Deep learning makes the classification accuracy higher.

Surendra Sedhai et al. proposed a technique that used semi-supervised approach for spam redirection classification mechanism. The concept used the training rules to be governed by supervised learning with an adaptive weight changing mechanism. However, the approach had the liberty of letting the weight adaptation fall into the purview of the training algorithm used.

Chao Chen et al. proposed a technique for the classification of drifted twitter spam based on statistical feature based classification. The major issues addressed in this paper, were the use of statistical features for spam classification. Drifted spam is often the result of several attached web links leading to the drifting mechanism of the tweets in social media applications with malicious URLs that can cause the spamming attacks on the web mails.

Nida Mirza et al. proposed a technique for spam classification based on hybrid feature selection. The major advantage of this approach was the fact that the hybrid parameters can be an amalgamation of both textual features and non-textual features. The evaluation of the performance of the proposed system was done on the basis of mean square error, hit rate and the accuracy. The performance of hybrid feature selection was shown to be better than the average features computation algorithms.

Hammad Afzal et al. in proposed a mechanism for the classification of bi-lingual tweets using machine learning algorithms. The methodology of the system was the use of natural language processing and thereafter the use of deep neural networks with multiple hidden layers. The learning rates were dependent on the differential changes in the architecture of the neural network used

Hailu Xu et al. proposed a technique for efficient spam detection across social platforms. The main problem with classification problems as presented by the authors is the lack of correlation between variables. This often leads to low accuracy in prediction. Hence this approach lacks any expert view on the apparent relation between the feature values and the outputs. The authors tried to address the problem of expert view exclusion in their work to enhance the accuracy.

Nadir Omer eta al. proposed a technique based on the use of support vector machine for spamming attacks. The authors tried to exhibit the fact that the support vector machine (SVM) used the concept of the Hyper-plane for the classification of the multi-dimensional data exposed to spamming attacks.

Tarjani Vyas et al. used the techniques of supervised learning for the classification of spamming attacks. The supervised learning mechanism was shown to have a different level of accuracy as compared to unsupervised learning. The classification process was however characterized by the computation of probabilities for classification as spam or ham.

Nishtha Jatana et al. proposed a technique for efficient radix encoded approach for the differentiation of SPAM and HAM based on the Bayesian Classifier. The classifier was used to classify the data set used for testing when the probability of spam and non-spam was computed based on the concept of the conditional probability.

Kamalanathan Kandasamy et al. used natural language processing (NLP) in conjugation with machine learning. The database was the twitter database. The paper presents an extremely interesting approach based on behavioural economics along with time series prediction. The authors considered twitter data (tweets) to access the mood of the society at large as an additional feature along with spamming data. The performance evaluation parameter was the mean absolute percentage error.

Navneel Prasad et al. proposed a comparative technique for spam classification based on back propagation and resilient networks. The authors cited that the problem addressed was the low accuracy in prediction by feed forward networks. Such networks do not have an error feedback mechanism for training to occur with the error as one of the inputs affecting the weights. Similar attributes of resilient and reinforced learning were used.

Wojciech Indy et al. used the MapReduce technique for spam classification mechanism. The MapReduce technique often finds similarities in data sets based on the spurious nature of spamming attacks often resembling actual ham. The performance indices were the accuracy and sensitivity.

Ashwin Rajadesingan et al. proposed a technique based on Comment Analysis whose data is often extracted from Comment-Blog Post Relationships. The approach is often useful when the blog part of the webmail produces the necessary textual data which can be either spam or non-spam.

Alper Kursat Uysal et al. proposed a mechanism for SMS Spam filtering. The authors used a hybrid of Artificial Neural Networks and particle swarm optimization (PSO) to reach desired values of the objective function. The authors used the Radial Basis Function (RBF) which have advantages of easy design, good generalization, strong tolerance to input noise, and online learning ability. The particle swarm optimization used along with it helped the authors to attain high accuracy of prediction.

D. Karthika Renuka et al. proposed a supervised approach for spam redirection classification mechanism. The concept used the training rules to be governed by supervised learning with an adaptive weight changing mechanism. The fully supervised learning mechanism made the accuracy increase.

Safvan Vahora et al. used the naïve Bayesian Classifier for spam classification. In straight forward terms, a naive Bayesian classifier assumes that the value of a specific feature is unrelated to the presence or absence of the other feature, given the category variable. Thus training it with the data set automatically trains the Bayesian Classifier to classify the data.

Lourdes Araujo et al. presented a quantified link analysis for spam filtering. Conventional techniques suffered from classification problems as presented by the authors is the lack of correlation between variables. This often leads to low accuracy in prediction. Hence this approach lacks any expert view on the apparent relation between the feature values and the outputs. The quantified link establishes the accuracy measure.

Sang Min Lee et al. tried out spam classification based on feature selection and data optimization. The major problem addressed in the paper was the fact that over fitting in the datasets for training for neural networks. Over-fitting introduces noise effects in the training and increases prediction errors. The authors proved that a function (training data set) with finite discontinuities can be approximated with a simpler neural network. The performance metrics are training time and mean square error.

Chih-Hung Wu et al proposed a technique based on back propagation and feature selection. The problem that the paper addressed was: number of iterations to be as less as possible for the

neural network are often high in case of large datasets with low correlation and stability of the algorithm is often less Both were proved to be achieved to a certain extent in this approach.

Chi-Yao Tseng et al. proposed an incremental SVM technique for spam classification of spam. The approach used the feature selection to be mapped on the hyperplane and then being classified by the SVM. The performance metric was the accuracy, precision and recall.

III. CLASSIFICATION AND PERFORMACNE METRICS

The need for probabilistic classifiers arise from the fact that the classification problem often encounters data sets with overlapping vectors. The major challenges in spam classification are:

1) It is very difficult to detect malicious redirections because redirections are also made intentionally for non-harmful purposes like load balancing.

2) If successful redirection is not employed, then Web Server may crash in case requests received becomes much more than request handling capacity.

3) It is very difficult to actually detect a malicious spam and differentiate it from a load balancing redirection.

The feature selection mechanism is also important for the computation of the various parameters that include the mean square error and accuracy. However, the addition of features makes the accuracy increase at times but also increases the complexity of the training.

4) Moreover, general machine learning techniques for spam classification are prone to poisoning attacks.

Depending on the implementation, Bayesian spam filtering may be susceptible to Bayesian poisoning, a technique used by spammers in an attempt to degrade the effectiveness of spam filters that rely on Bayesian filtering. A spammer practicing Bayesian poisoning will send out emails with large amounts of legitimate text (gathered from legitimate news or literary sources).

Spammer tactics include insertion of random innocuous words that are not normally associated with spam, thereby decreasing the email's spam score, making it more likely to slip past a Bayesian spam filter. However, with (for example) Paul Graham's scheme only the most significant probabilities are used, so that padding the text out with non-spam-related words does not affect the detection probability significantly. Words that normally appear in large quantities in spam may also be transformed by spammers. For example, «Viagra» would be replaced with «Viaagra» or «V!agra» in the spam message. The recipient of the message can still read the changed words, but each of these words is met more rarely by the Bayesian filter, which hinders its learning process. As a general rule, this spamming technique does not work very well, because the derived words end up recognized by the filter just like the normal ones.

The overlapping vectors make its challenging to find a clear boundary for the classification problem and often there exists only a fuzzy or non-clear boundary to demarcate among the data classes. In such overlapping classes, the final categorization of a new data vector 'X' is done based on the maximum mutual probability given by:

$$P(X) = Max\{\frac{X1}{U}.\frac{X2}{U}.\ldots..\frac{Xn}{U}) \quad (1)$$

Here,
X1, X2….Xn are the multiple classes
U is the universal set containing all the classes
P(X) is the maximum probability of a data sample to belong to a particular category.

The final classification accuracy is computed as:

$$Ac = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Here.
TP represents true positive
TN represents true negative
FP represents false positive
FN represents false negative

CONCLUSION: It can be concluded from previous discussions that the spam classification is a non-trivial task based on the amount and the complexity of data mobile and web servers receive in real time situations. It can be inferred from the discussions made so far that AI and ML based approaches are appropriate to cater to the needs of the web services. However, the challenging aspect in spam classification remains the accuracy that needs to be met for real life applications which may be challenging.

References

[1] AK Jain, D Goel, S Agarwal, Y Singh, G.Bajaj, "Predicting Spam Messages Using Back Propagation Neural Network", Journal of Wireless Personal Communications, Springer 2021, vol. 110, pp. 403-422.
[2] KS Adewole, NB Anuar, A Kamsin, "SMSAD: a framework for spam message and spam account detection", Journal of Multimedia Tools and Applications, Springer 2021, vol. 78, pp. 78, 3925–3960.
[3] Aliaksandr Barushka, Petr Hajek, "Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks", Springer 2018
[4] Surendra Sedhai, Aixin Sun, "Semi-Supervised Spam Detection in Twitter Stream", IEEE 2018
[5] Chao Chen, Yu Wang, Jun Zhang, Yang Xiang, Wanlei Zhou, Geyong Min, "Statistical Features-Based Real-Time Detection of Drifted Twitter Spam", IEEE 2017
[6] Nida Mirza, Balkrishna Patil ,Tabinda Mirza ,Rajesh Auti, "Evaluating efficiency of classifier for email spam detector using hybrid feature selection approaches",IEEE 2017
[7] Hammad Afzal ,Kashif Mehmood, "Spam filtering of bi-lingual tweets using machine learning",IEEE 2016
[8] Hailu Xu ,Weiqing Sun ,Ahmad Javaid," Efficient spam detection across Online Social Networks", IEEE 2016
[9] Nadir Omer Fadl Elssied,Othman Ibrahim ,Ahmed Hamza Osman," Enhancement of spam detection mechanism based on hybrid kkkk-mean clustering and support vector machine",SPRINGER 2015
[10] Tarjani Vyas , Payal Prajapati , Somil Gadhwal," A survey and evaluation of supervised machine learning techniques for spam e-mail filtering",IEEE 2015
[11] Nishtha Jatana ,Kapil Sharma," Bayesian spam classification: Time efficient radix encoded fragmented database approach", IEEE 2014
[12] Kamalanathan Kandasamy ,Preethi Koroth," An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques", IEEE 2014
[13] Navneel Prasad ,Rajeshni Singh ,Sunil Pranit Lal," Comparison of Back Propagation and Resilient Propagation Algorithm for Spam Classification",IEEE 2013
[14] Wojciech IndykEmail author, Tomasz Kajdanowicz, Przemyslaw Kazienko,Slawomir Plamowski," Web Spam Detection Using MapReduce Approach to Collective Classification", SPRINGER 2013
[15] Ashwin Rajadesingan, Anand Mahendran," Comment Spam Classification in Blogs through Comment Analysis and Comment-Blog Post Relationships", SPRINGER 2012
[16] Alper Kursat Uysal ,Serkan Gunal ,Semih Ergin ,Efnan Sora Gunal, "A novel framework for SMS spam filtering", IEEE 2012
[17] D. Karthika Renuka , T. Hamsapriya ,M. Raja Chakkaravarthi, P. Lakshmi Surya, "Spam Classification Based on Supervised Learning Using Machine Learning Techniques", IEEE 2011
[18] Safvan Vahora ,Mosin Hasan ,Reshma Lakhani, "Novel approach: Naïve Bayes with Vector space model for spam classification", IEEE 2011
[19] Lourdes Araujo, Juan Martinez-Romo, "Web Spam Detection: New Classification Features Based on Qualified Link Analysis and Language Models", IEEE 2010
[20] Sang Min Lee, Dong Seong Kim , Ji Ho Kim ,Jong Sou Park, "Spam Detection Using Feature Selection and Parameters Optimization", IEEE 2010
[21] Chih-Hung Wu, Chiung-Hui Tsai, "Robust classification for spam filtering by back-propagation neural networks using behavior-based features", SPRINGER 2009
[22] Chi-Yao Tseng, Ming-Syan Chen, "Incremental SVM Model for Spam Detection on Dynamic Email Social Networks", IEEE 2009