ISSN:0975-9646

E-mail Spam Detection and Classification using SVM

Shivam Pandey, Ashish Taralekar, Ruchi Yadav, Shreyas Deshmukh and
Prof. Shubhangi Suryavanshi
Department of Computer Engineering
G.H.Raisoni Institute of Engineering & Technology,
Wagholi, Pune – 412207
shivampandey.pandey8@gmail.com

***Abstract*— here we present an inclusive review of recent and successful content-based e-mail spam filtering techniques. Our focus is majorly on machine learning-based spam filters and variants which inspired from them. We report on relevant ideas, techniques, major efforts, and the state-of-the-art in the field. The initial interpretation of the prior work shows the basics of e-mail spam filtering and feature engineering. In this we conclude by studying techniques, methods, evaluation benchmarks, and explore the promising offshoots of latest developments and suggest lines of future investigations.**

***Keywords*—— SVM Classifier, Spam Email Classification, Data Mining, Data Science, Machine Learning.**

I. INTRODUCTION

Recently unsolicited commercial / bulk e-mail, also known as spam, becomes a major problem on the Internet. Spam is a waste of time, storage space and communication bandwidth. The problem of spam and fraud e-mail has been increasing for years. In recent figures, 40% of all mail is spam that emails about 15.4 billion emails per day and costs Internet users about $ 355 million per year. Automatic e-mail filtering is the most effective way to deal with spam at the moment and there is a fierce competition between spammers and spam filtering methods. Now Spammers began using several tricky methods to overcome filtering methods such as using random sender addresses and / or adding random characters to the beginning or end of the message subject line.

Knowledge engineering and machine learning are two common approaches used in e-mail filtering. The knowledge engineering approach consists of specifying a set of rules according to which email is classified as spam or ham. A set of such rules must be created either by the user of the filter, or with some other authority (such as a software company that provides a special rule spam filtering tool). By applying this method, none show promising results because the rules should be necessary. Constant updates and maintenance are done, which is a waste of time and is not convenient for most users.

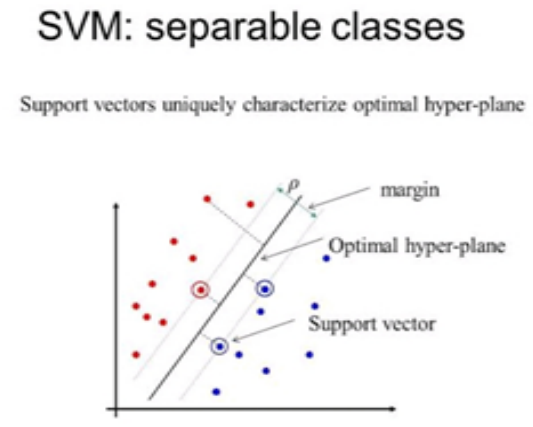
Machine learning approaches are more efficient than knowledge engineering approaches. It does not need to specify any rules. Instead, a set of training samples is used, these samples are a set of pre-classified e-mail messages. Machine learning approaches have been widely studied and a lot of algorithms can be used in e-mail filtering. These include Naive Bayes, Support Vector Machines,Neural Networks, K-nearest neighbours, rough sets and artificial immune systems.

II. PRELIMINARY AND PROBLEM STATEMENT

A Spamming is one of the major and common attacks that accumulate a large number of compromised machines by sending unwanted messages, viruses, and phishing through email. We have chosen this project because now there are many people who are trying to fool you just by sending you fake e-mails, as if you have won 1000 dollars, deposit this amount in your account as soon as you open this link. Once done, they will track you and try to hack your information. Sometimes relevant e-mail is considered spam email.

Unwanted email is harassing Internet consumers in ways such as:

- Important email messages were missed and / or delayed.
- Consumers seek ISP's frequent email delivery changes all the time.
- Internet performance and bandwidth loss.
- Millions of compromised computers.
- Loss of billions of dollars worldwide.
- Identification of theft.
- Increase in several viruses and Trojan horses.
- Spam can crash and affect the mail server and fill the hard drive



III. PROPOSED SYSTEM AND METHODOLOGY

A support vector machine (SVM) can be used when our information is completely in two classes. An SVM classifies information by detecting an ideal hyperplane that separates all information purposes of a class from an alternative class. The hyperplane for SVM means the largest difference between the two classes. The margin shows that the section parallel to the hyperplane has a maximum width with no internal information points. SVM has verified to be one of the most important economic kernel strategies. The success of SVM is mainly due to its high generalization capability. Not like many learning algorithms, SVM results in sensible demonstrations whereas previous data need not be included. In addition, the employment of positive fixed kernels within SVM can be taken as the associate degree embedding of the input field to the higher dimensional feature region where the classification is met, while the exploitation does not explicitly use this feature area has been done. Therefore, the case for selecting a design for neural network application is replaced by the case of selecting an acceptable kernel for a support vector machine. The support vector machine has shown power in binary classification. It's Wise Theoretical Foundation and Well Perfect Learning Algorithm Rules. This leads to stable information classification. The only disadvantage is this is it's time and memory once the size of the information is large.

In the following section we will discuss the proposed methodology for email spam detection technique. The Fig.1 shows the workflow

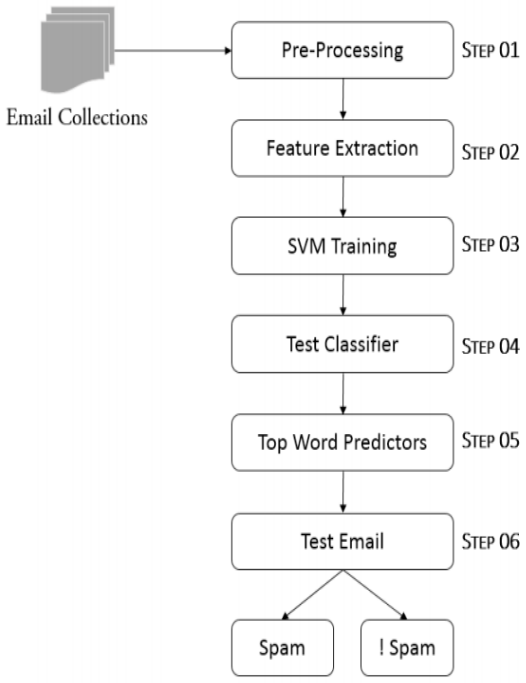


Fig. 1: Proposed Workflow Architecture

a. Pre-processing

The pre-processing step is used to remove noise from emails that are irrelevant and need not exist .Pre-processing phase includes:

1. Removing Numbers
2. Remove special symbol
3. URL deletion
4. HTML tags separating task
5. Performing Word Steaming

b. Feature Extraction

Feature extraction technique is used to extract important and relevant features from the email body. Feature replaces email 2D vector space features numbers. These features are mapped from the dictionary list.

c. SVM Training

Email spam is used for training purposes. Training datasets contain spam content and are trained using this classifiers. After training, the classifier is ready to classify spam emails.

d. Test Classifier

The classifier is tested with several training data to test it accurately. Until the proposed solution is obtained which gives 98% accuracy in classifying email.

e. Test Email

After the training phase is completed, the classifier is given a sample email as input to classify the email. Classifier Generates output in forms of 0 or 1, 1 means it is spam and 0 means it is not spam.

IV. FUTURE SCOPE

However, the experiment has made efforts towards solving the problem of spam e-mail. Proposed solutions using legislative, behavioural and technical measures are not a complete solution. The problem of spam e-mail and anti-spam solutions is game like cat and mouse, every day spammers will come up with new techniques Send spam e-mail. This work has given possible directions for classification. Spam e-mail Future efforts will be extended to:

1. Obtaining accurate classification, zero percent (0%) with abortion of ham E-mail as spam and spam as e-mail ham.

2. Many Efforts will be implemented to block phishing e-mail, which carries phishing Attacks and now days which is a matter of concern.

3. Also, work can be extended to keep it away from the Denial service attack (DoS) Now which has emerged in distributed fashion, is called distributed Denial Service Attack (DoS).

V. CONCLUSION

In this study, we reviewed the general application in the field of machine learning approach and spam filtering. A review of the state of the art algorithm has been implemented to classify the message as either spam or ham. Efforts made by various researchers to solve the problem of spam through the use of machine learning classifiers were discussed. The development of spam messages was investigated over the years to avoid filters. The basic structure of the email spam filter and the processes involved in filtering spam emails were noted. The paper surveyed some of the publicly available datasets and performance metrics that can be used to measure the effectiveness of any spam filter. The challenges of machine learning algorithms in efficiently handling the threat of spam were pointed out and a comparative study of machine learning techniques available in the literature. We also revealed some open research problems related to spam filters. In general, the amount and amount of literature we reviewed suggests that significant progress has been made and will still be made in this area. After discussing open problems in spam filtering, further research needs to be done to increase the effectiveness of spam filters. It will develop spam filters to continue an active research area for academics and industry practitioners researching machine learning techniques for effective spamming. Our hope is that research students will use this paper as a spring board to conduct qualitative research in spam filtering using machine learning, deep learning, and deep adversarial learning algorithms.

VI. REFERENCES

[1] Abduelbaset M. However, Tarik Rashed, Ali S. Elbekaie, and Husien A. Alhammi, “An Anti-Spam System Using Artificial Neural Networks And Genetic Algorithms” (A Neural Model In Anti Spam).

[2] Er. Seema Rani, Er. Sugandha Sharma, “Survey on E-mail Spam Detection Using NLP”, International Journal of Advanced Research in Computer Science and Software Engineering, India, Volume 4, Issue 5, May 2014.

[3] Masurah Mohamad, Khairulliza Ahmad Salleh, “Independent Feature Selection as Spam-Filtering Technique: An Evaluation of Neural Network”, Malaysia.

[4] El-Sayed M. El-Alfy, “Learning Methods For Spam Filtering”, College of Computer Sciences and Engineering King Fahd University of Petroleum and Minerals, Saudi Arabia.

[5] Upasna Attri & Harpreet Kaur, “Comparative Study of Gaussian and Nearest Mean Classifiers for Filtering Spam E-mails”, Global Journal of Computer Science and Technology Network, Web & Security, USA, Volume 12 Issue 11 Version June 2012.

[6] Alia Taha Sabri, Adel Hamdan Mohammads, Bassam Al-Shargabi, Maher Abu Hamdeh, “Developing New Continuous Learning Approach for Spam Detection using Artificial Neural Network (CLA_ANN)”, European Journal of Scientific Research, ISSN 1450-216X Vol.42 No.3 (2010), pp.511-521.

[7] Enrique Puertas Sanz, José María Gómez Hidalgo,José Carlos Cortizo Pérez, “Email Spam Filtering”, Universidad Europea de Madrid Villaviciosa de Odón, 28670 Madrid, SPAIN.

[8] Ravinder Kamboj, “A rule based approach for spam detection” ,Computer Science and Engineering Department, Thapar University, India, July 2010.

[9] Vandana Jaswal, Nidhi Sood, “Spam Detection System Using Hidden Markov Model”, International Journal of Advanced Research in Computer Science and Software Engineering, India, Volume 3, Issue 7, July 2013.

[10] Sahil Puri, Dishant Gosain, Mehak Ahuja, Ishita Kathuria, Nishtha Jatana, “Comparison And Analysis Of Spam Detection

[11] Algorithms”, International Journal of Application or Innovation in Engineering & Management (IJAIEM), India, Volume 2, Issue 4, April 2013..