Suspicious Email Detection System

Karan Raj[1], Karunakar Mishra[2], Divyanshu Upadhyay[3], Swarnim Shikhar[4],
Mr. Sureshkumar M[5], Dr. Kavitha A S[6]

[1,2,3,4]Student, Dept. of Information Science and Engineering, Dayananda Sagar College of Engineering, Bangalore, Karnataka, India
[5]Assistant Professor, Dept. of Information Science and Engineering, Dayananda Sagar College of Engineering, Bangalore, Karnataka, India
[6]Associate Professor, Dept. of Information Science and Engineering, Dayananda Sagar College of Engineering, Bangalore, Karnataka, India

---***---

Abstract - *Electronic mail (email) is an easy and simple method of communicating with people using the Internet in electronic devices. Spam emails are the junk mails that are sent in bulk through the email system. The use of spam is becoming popular since the last decade and is a major problem faced by email users. Spam consumes 98% of billions of emails sent every day. Since there are many text-based spam filtering systems in the market, Spammers are using images and PDFs for sending spam. In this project, we are trying to filter spam emails containing images and PDFs and improve the performance of the existing spam filtering systems using Machine learning algorithms with acceptable Recall and Precision Value.*

Key Words: **Spam, Machine Learning, E-mail, Natural Language Processing, Image**

1. INTRODUCTION

Email is a way of transferring messages between people using electronic devices such as mobiles and PCs. As the internet provides an easy and fast platform for communication, almost every field uses email for data transfer. With 4.3 billion users, E-mail is nowadays the second most widely used form of communication, after mobile phones. In this era of technology, it is difficult to perform many tasks without the help of Email. There are many researchers concentrating in the field of suspicious email detection, to help prevent any malicious activities such as fraud, theft, and terrorism. According to recent reports, spam is being increasingly used to send viruses, spyware, phishing sites, etc. An average of 540 million spam email was sent each day. Spam email causes a decrease in the performance of the mail engine.

2. RELATED WORK

Bo Yu and Zong Ben Xhu categorized spam emails by trying 4 separate ML algorithms which are NB, Neural Network, SVM and RVM in 2008. Tiago and Akebo Yakamami in 2010 compared different ML algorithms using content-based spam-filtering. Their paper talked about 7 different types of NB Classifiers and checked those outcomes with Linear SVM on 6 non-identical big datasets. The accuracy of this model was 90%. Loredena, Camelia L and R Potolea in 2010 did a comparative analysis using the K-nearest Neighbour algorithm along with the resampling approach [1]. D Kartika Renuka, Dr T Hamsapriya, Mr Raja Chakravarti and P Lakshmi Surya in 2011 did relative research on e-mail spam categorization by applying non-identical ML algorithms. In this analysis, they compared three different ML classification algorithms which are NB classifier, J48 and MLP classifier. The results took too much time but had high accuracy. This modified NB had 91% accuracy [2]. Rushdi Shams and Robert Mercer used text and readability features for spam emails in 2013. This paper used 5 different classification algorithms [3]. Ms M. Rathi and Vikash Parik in 2013 did a comparison on different ML algorithms for email spam detection using Data Mining by doing research on different classifier model by selection and removal of the features [4]. Anirudh Hari Singhaney, Mr Amaan Dixit, Saurabh Gupta and Anuuja Aroraa in 2014 did relative research on images and text spam emails by deploying K-Nearest Neighbour, NB and Reverse Density-based spatial clustering of applications with noise (DBSCAN) algorithm for detection of spam email. Results showed that these 3 ML algorithms gave good results without doing the preprocessing of data. NB had the highest accuracy among the three algorithms [5]. Savitha Pundaalik and Santhosh Biraadar in 2014 did research on successful email categorization for spam and regular emails [6]. Ijjat Alsamadi and Ekdam Alhaami in 2015 did research on clusters of emails and email content categorization for spam emails. Their model applied the SVM for the categorization of emails which are obtained from the K Means clustering method. They used 3 kinds of categorization without taking off stopwords, taking off stop words and applying N-gram based categorization. N-gram based classification was the best among the three [7]. Ali Shafikh Aski and Naved Khalil Zadeh Sorati in 2016 build their model using machine learning algorithms. Their model utilized 3 ML algorithms which are Multi-Layer Neural Network, NB Classifier and J48 for categorization of spam emails from regular mails by applying twenty-three regulations. In their model, MLP had the highest accuracy with low execution time compared to NB [8].

3. DATA COLLECTION

The dataset which will be used in our model would be collected over a period of time from different sources. We can use various websites that provide datasets. Also, we can collect different spam emails from some random email IDs which can be either business-related or informal emails.

4. MODEL DESCRIPTION

The model proposed is a content-based spam filtering system that filters the emails containing images of all formats and PDFs. This is a binary classification model that predicts whether an email given by the user is a HAM or SPAM. The process includes data collection, data preprocessing with the steps of tokenization, stemming and removal of stopwords. *CountVectorizer* is used to get bag-of-words counts in vector form. After the counting, TF-IDF can be used for normalization and the term weighting. A pipeline is used to chain multiple estimators into one and hence, automate the machine learning process. This is extremely useful as there is often a fixed sequence of steps in processing the data. The final stage of the process is training and testing the model using the best machine learning algorithm for the given dataset.

5. FUTURE WORK

In the future, our algorithm can also be applied to the image and pdf embedded emails. Also, this algorithm can be strengthened to enhance the suspicious email detection job, using the upcoming better techniques. Our model is largely compatible. In the present scenario, the dataset is constructed using spam email collected from various sources, but later it can be applied to huge datasets too.

6. CONCLUSION

E-mail is an important way of communication. However, an increase in spam email causes traffic congestion due to which productivity has decreased and it has become a serious problem for industries. In Machine Learning, Support Vector Machine can play an important role in spam filtering and detection but SVM implementation has major problems with performance. Decision Tree classifier is the most popular classifier but it has one problem that it uses large memory space and it is not appropriate for the spam filtering method.

REFERENCES

[1] B. Yu and Z. Xu, “A comparative study for content-based dynamic spam classification using four machine learning algorithms”, in Knowledge-Based System-Elsevier, vol. 21, pp. 355–362, 2008

[2] D. K. Renuka, T. Hamsapriya, M. R. Chakkaravarthi and P. L. Surya, “Spam Classification Based on Supervised Learning Using Machine Learning Techniques”, in proc. IEEE- International Conference on Process Automation, Control and Computing, 2011, pp. 1–7.

[3] R. Shams and R. E. Mercer, “Classifying spam emails using text and readability features”, in proc. IEEE International Conference on Data Mining (ICDM), 2013, pp. 657–666.

[4] M. Rathi and V. Pareek, “Spam Email Detection through Data Mining-A Comparative Performance Analysis”, in International Journal of Modern Education and Computer Science, vol. 12, pp. 31-39, 2013.

[5] A. Harisinghaney, A. Dixit, S. Gupta, and Anuja Arora, “Text and image-based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN Algorithm”, in proc. IEEE-International Conference on Reliability, Optimization and Information Technology (ICROIT), 2014, pp.153-155.

[6] S. P. Teli and S. K. Biradar, “Effective Email Classification for Spam and non- spam”, in International Journal of Advanced Research in Computer and Software Engineering, Vol. 4, 2014.

[7] Alsmadi and I. Alhami, “Clustering and classification of email contents”, in Journal of King Saud University - Computer and Information Science -Elsevier, vol. 27, no. 1, pp. 46–57, 2015.

[8] A. S. Aski and N. K. Sourati, “Proposed efficient algorithm to filter spam using machine learning techniques”, in Pacific Science Review- A Natural Science Engineering- Elsevier, Vol. 18, No. 2, pp. 145–149, 2016.