International Journal on Applications in Information and Communication Engineering
Volume 6 : Issue 2 : July 2020 , pp 39 –42 www.aetsjournal.com ISSN (Online) : 2394 - 6237

FRAUDULENT E-MAIL CLASSIFICATION USING DIFFERENT ALGORITHMS – A SURVEY

Balakiruba J, Bharathi S, Gajalakshmi S
Department of Information Technology
Rathinam Technical Campus, Coimbatore, Tamilnadu, India

Abstract— The dataset obtained from incoming emails is compared in this paper's proposed classification of fraudulent email. Combining the time- and memory-intensive k-means clustering technique with the low-accuracy SVM classification algorithm when working with big amounts of data. Using the author minor method, it can be tough to determine the features utilised for a certain collection of emails. Typically, the objective of fraudulent emails is to deceive the recipient by posing as helpless in an attempt to get compassion. In addition to emails that fall under this category, the dataset also includes emails that fall under the category of "normal" emails. Using the C5.0 decision tree classification method, we are constructing a dataset of real and fraudulent emails and classifying them. After constructing a decision tree, classification rules that utilise previously unidentified class labels can be built and applied to the classification of new cases. In order to build classification rules in the form of decision trees, the C4.5 method is utilised. The C4.5 method resolves problems from earlier papers and is more precise than previous algorithms.

Index Terms— Fraudulent mail, decision tree, Regular mail

I. INTRODUCTION

Email is a viable way of written communication in the modern era. Email is a key component of web data transfer. Spam is a lucrative business opportunity made more prevalent by the use of email. Spam is unwanted content that an online user receives via email or text message. For commercial purposes, unsolicited bulk mail is delivered to an undetermined set of recipients. The purpose of spam filters is to identify unwanted and unsolicited email and prevent it from entering users' inboxes. The decision-making criteria used by a spam filter are comparable to those used by other forms of filtering software. As a result of the increasing volume of spam emails, Internet Service Providers, Internet users, and the Internet's entire backbone network are all suffering significant problems.

Denial of service is an example, in which spammers flood an email server with traffic, thereby delaying the delivery of messages to their intended recipients. In addition to being a waste of storage space, processing traffic, and energy, spam emails may contain fraudulent schemes, bogus offers, and other schemes. Due to the ongoing change and evolution of spam communications, a single model cannot adequately address the issue. In addition, since these spams are typically tailor-made, accurate identification is made more difficult. In addition to being utilised in multiple attacks, these spam messages expand network memory and communication capacity. These attacks modify the user's identity or erase his data or information. Typically, fake emails deceive the receiver by depicting their plight in an attempt to elicit compassion. The collection contains emails that we would regard to be typical as well as emails that are deceptive.

The fake websites and emails used to detect phishing attacks are designed to resemble well-known banks, credit card companies, and e-commerce websites in an attempt to deceive users into divulging personal information. These emails, which we define as false, are included in the data set, while the remaining emails are usual. Utilizing various feature sets and classification techniques, evaluate experiments. We can identify fraudulent messages by removing the subject, body, and content from incoming email and adding a number of features to the content. Using the C5.0 decision tree classification method, we are constructing a dataset of real and fraudulent emails and classifying them. Once a decision tree has been established, it is straightforward to generate classification rules and categorise additional examples with undetermined class labels. A typical technique known as C4.5 represents classification rules with decision trees. Because it use decision tree techniques to eliminate unnecessary data from a dataset. When wrong datasets are utilised, the c4.5 algorithm's precision is improved.

II. A BOOK REVIEW

A literature analysis was conducted to explore the various strategies for detecting bogus emails. There are various available approaches. This chapter provides a summary of the distinctions between the employed algorithms.

Rekha SandeepNegi (2014) provides ways for detecting various types of spam because spam messages can be used to launch attacks and strain the memory capacity of communication networks.

These types of attacks are capable of stealing user information or exposing user identity. Methods for detecting spam are presented in this article [8]. In addition to squandering resources like as bandwidth, computing power, and storage, spam emails may contain fraudulent schemes, bogus offers, and other schemes.

Various techniques, such as Whitelist/Blacklist, Bayesian analysis, keyword checking, mail header analysis, etc., can be utilised to determine whether incoming messages are spam. No one can guarantee a flawless outcome utilising any of the above strategies. The false positive and false negative response rates of some procedures are quite high. For text and multimedia messages, there is much space for distinguishing spam from real emails.

Waheeb Abu-Ulbeh, Nadir Omer FadlElssied, 1othman Ibrahim (2014). In this study, he proposed a method for detecting email spam that relies on a predetermined number of clusters [5] and combines SVM and K-means clustering. Spam detection is used to identify spam emails and prevent their delivery to users' inboxes. For the purpose of evaluating the practicability of the proposed technique, an experiment was undertaken utilising a spam-based standard dataset. Support Vector Machine (SVM) is frequently employed for the detection of spam email. Dealing with a large amount of data needs considerable time, memory, and precision.

A comparison has been performed between SVM spam detection and a K-means clustering and SVM hybrid. The proposed mechanism includes advantages such as improved classification accuracy, fewer false positives, and time savings. The time-cost is 63.09 seconds, the false-positive rate is 0.04 percent, and the accuracy of classification is 98.01 percent. This hybrid technique provides improved classification precision, fewer false positives, and reduced time costs.

Farkhund In email forensics, Rachid Hadjidj, Benjamin C.M. Fung, and Mourad Debbab (2008) suggest mining write-prints as a new method for authorship attribution. In this study, the authors describe a novel data mining strategy for finding combinations of elements that commonly appeared in a suspect's emails, as well as a method for recording each suspect's unique write-print[2]. In this essay, the writers explain the problems with authorship identification and divide them into three subproblems: (1) to extract evidence to support the conclusion of authorship. (2) To recognise the handwriting of each suspect [2] (3) Determine the sender of the malicious email. The first stage is to identify a collection of qualities of a writer's style that are evident in the majority of their works.

The classifier is trained on the collected writing style characteristics to generate a model, which is then used to allocate the disputed email to the suspect with the best writing style.

The majority of previous contributions have focused on improving the accuracy of email authorship classification. Based on the concept of recurrent patterns and email authorship attributions, they provide an innovative way for generating new ideas for printed print. Benefits include justifiable proof, adaptive writing styles, feature optimization, and wide application.

Sugandha Sharma, Er. Seema Rani (2014). As a result of spam emails, Internet users face significant problems. It has a multitude of impacts [1]. As a result of the rise of viruses, Trojan horses, lost productivity, excess space used up in inboxes, and materials containing damaging information for a fraction of users, users must devote a significant amount of time to eliminating needless emails and organising incoming mail. A spam filter uses an automated technique to identify spam in order to prevent its delivery.

Spam increases the spread of viruses, consumes mailbox space, and decreases productivity.

Trojans and materials containing potentially hazardous information for damaging the reliability of mail servers, certain classes of users; as a result, users must spend a significant amount of time filtering and eliminating undesired email. They apply SVM technique. In addition to saving time and inbox space, there is virus protection for correspondence.

MaringantiHimaBindu, Jitendra Shrivastava (2013). This research asserts that combining GA with other email filtering techniques could result in a more reliable solution[4]. Here, a very high percentage; only emails with malicious attachments are considered, while spam emails containing links to malicious websites are ignored. Support vector machine and content-based spam filtering techniques are utilised. Formerly, malicious users relied on fake notifications from social networking sites, web hosting providers, delivery services such as courier, etc., as well as messages from non-government and government organisations. The dataset and GA parameters influence the efficiency of the procedure. The algorithm's efficiency exceeds 82 percent. Using GA in conjunction with other email filtering techniques results in a more precise SAPM filtering technique. It is feasible to reduce false positive and false negative outcomes.

Mrunal Mahajan, ShrutiRachh, Chaitali Shah, and Narendra. M. Shekokar (2015). In this study, they propose a phishing detection and prevention strategy[6] that combines URL-based and similarity-based detection. As part of the URL-based phishing detection procedure, both the visible and actual URLs are extracted. The Link Guard Algorithm is used to evaluate the two URLs before proceeding to the subsequent phase of the procedure based on the algorithm's outcome. The images in this publication [6] must be modified, demanding comparisons. There are other additional transformations available, including DFT, DCT, and cross-correlation.

By analysing the website's overall layout and the hyperlinks in the email page's source code, they proposed a way for detecting phishing to improve website security. This technique provides a superior and more dependable alternative to other costly site security programmes.

Gaganpreet Kaur and JaNeetu Sharma (2014). During this investigation, he presented a Text Classification. Using the Support Machine Learning method, a new algorithm for establishing the criterion function of the clustering problem for spam messages has been proposed[6]. A genetic technique is utilised to overcome the clustering issue. Spammers have used a variety of methods to get our email addresses in order to fill your mailbox. Detailed information on how to defend against these annoying, unsolicited e-mails, taking into account various spam detection systems.

Spam detection has a lot of benefits, including as reducing space in inboxes, safeguarding users from viruses, Trojans, and potentially harmful content, and saving users' time filtering through incoming mail and deleting unnecessary correspondence. In this paper[3], spam filter technology provides protection against e-mails containing viruses while conserving mailbox capacity and saving time.

Nasrullah Memon, Sarwat Nizamani (2013)

In this study, a cluster-based classification technique is used to construct an email authorship identification model [9]. It augments the usual set of features with the useful capability of mail authorship identification.

The proposed CCM-based e-mail authorship identification model beats both the state-of-the-art support vector machine (SVM)-based models and the method provided by Iqbal et al. (2010, 2013) [9]. They were presented with a CCM-based EAI model, which is a cluster-based classification model for mail-authorization identification. The objective of this paper[9] is to assess the compatibility of baseline-stylometric data with extra features for the EAI task.

The development of a new model for the EAI task and the selection of Information Gain features based on content characteristics. They were capable of authorship identification, but in the future, the model will need to be expanded to include authorship verification so that emails with unknown authors may be identified as such.

Loay E. George and Noor Ghazi M. Jameel, 2013)

He proposed a phishing detection model that employs extracted email characteristics to identify phishing emails. These characteristics are located in the email's header and body. The mechanism consists of three primary steps: pre-processing, feature analysis, and application of phishing detection using Feature Decisive Value criteria (FEFDV) and Feature Existence. In this manner, the visual appearances of the 19 adopted discriminating features were evaluated, and only the most potent and widely-applied features were included [7]. Utilizing binary values of 0 or 1 to implement the functionalities. With a value of 0 indicating the absence of this feature and a value of 1 indicating the presence of this feature in the tested email. In this method of detection, the features are balanced and analysed prior to employing the most beneficial features to classify the emails. They are using FEFDV Criteria with a system based on features. Using only 8 of the total 19 email properties, our new algorithm achieved a 97.79 percent accuracy rate. A single email was evaluated in 0.0004 milliseconds, a period of time that is extremely quick.

Nasrullah Memon, Sarwat Nizamani (2012). The research proposes a model for detecting suspicious emails with enhanced feature selection [10, 11]. You can send emails to particular individuals or groups. In a matter of seconds, a single email has the ability to reach millions of people. Most people now find it nearly impossible to envisage a world without email. Due to these circumstances, email has also gained popularity among terrorists as a form of communication. After the tragic events of September 11, 2001, numerous researchers focused on counterterrorism, attempting to predict terrorist plans based on suspicious communications. This motivated us to contribute to this field as well. This study proposes using feature selection tactics and classification approaches to identify terrorist mail. Email content analysis and email traffic analysis were the two primary foci of email analysis research.

III. CONCLUSION

In the present world, email is regarded as the most practical method of written communication. It must be an economical and efficient form of communication. The increase in email usage has expanded commercial options.

Spam is unsolicited information sent by email or text message to visitors of a website.

Emails can also be sent anonymously, without revealing the identify of the sender. The accuracy of the task for detecting fake e-mails is dependent on the selection of beneficial features, according to the research. In these trials, several classification methods, including C4.5, SVM, and CCM, were utilised. Regardless of the classification approach, frequency-based features obtain a high degree of accuracy for the detection of fraudulent emails. In this study, we employed the characteristics retrieved from the text and subject of incoming emails.

REFERENCES

[1]. "Survey on E-mail Spam Detection Using NLP," by Er. SeemaRani and Er. Sugandha Sharma

[2]. Volume 4, Issue 5 of the International Journal of Advanced Research in Computer Science and Software Engineering, May 2014

[3]. Rachid Hadjidj, Mourad Debbab, Benjamin C.M. Fung, and Farkhund Iqbal A novel method of authorship attribution in email forensics using write-print mining, ELSVIER, digital investigation 5 (2008)S42–S51

[4]. "Survey on Text Classification (Spam) Using Machine Learning," International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014, 5098-5102, by JaNeetu Sharma and Gaganpreet Kaur.

[5]. MaringantiHimaBindu, JitendraNath Shrivastava I.J. Intelligent Systems and Applications, 2014, 02, 54–60, "E-mail Spam Filtering Using Adaptive Genetic Algorithm"

[6]. An improved spam e-mail classification mechanism using k-means clustering was described by Nadir Omer FadlElsied, 1othman Ibrahim, and Waheeb Abu-Ulbeh in Journal of Theoretical and Applied Information Technology on February 28, 2014, Vol. 60 No. 3.

[7]. "An ideal approach for detection and prevention of phishing attacks," ELSVIER, Procedia Computer Science 49 (2015) 82–91, by Narendra. M. Shekokar, Chaitali Shah, Mrunal Mahajan, and Shruti Rachh

[8]. "Detection Phishing E-mails Using Features Decisive Values," International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013, by Noor Ghazi M. Jameel and Loay E. George

[9]. International Journal of Engineering Trends and Technology (IJETT), Volume 11 Number 6, May 2014, RekhaSandeep Negi, "A Review on Different Spam Detection Approaches."

[10]. Egyptian Informatics Journal (2014) 14, 239–249 SarwatNizamni, Nasrullah Memon, "CEAI: CCM-based e-mail authorship identification model"

[11]. Modeling Suspicious E-mail Detection Using Enhanced Feature Selection by SarwatNizamani and NasrullahMemon Vol. 2, No. 4, August 2012, International Journal of Modeling and Optimization

[12]. Savita Ajay, Suraj Prasad Keshari, and A.S. Zadgaonkar International Journal of Science and Modern Engineering (IJISME), ISSN: 2319-6386, Volume 1 Issue 6, May 2013, "A Model for Identifying Phishing E-Mail Based on Structural Properties"

[13]. Rajesh S. Prasad, Ph.D. and Mubin Shaukat Tamboli, "Authorship Analysis and Identification Techniques: A Review." Volume 77, Number 16, September 2013, International Journal of Computer Applications (0975 - 8887)

[14]. In their article "Comparative study of Authorship Identification Techniques for Cyber Forensics Analysis," published in the International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 4, No. 5, 2013

[15]. SmitaNirkhi and Dr. R. V. Dharaskar ScienceDirect, AASRI Procedia 4 (2013) 125–131 DhanalakshmiRanganayakulu, "Detecting Malicious URLs in E-Mail- An Implementation"

[16]. "E-mail classification for Spam Detection using Word Stemming," by Dr. T. Hamsapriya and Mrs. D. Kartika Renuka, 2010 International Journal of Computer Applications (0975 - 8887) Volume 1 - No.

[17]. "E-mail Filter for Spam Mail: A Review," International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 3, March 2013, by Amar V. Sable and Prof. Vijay S. Gulhane

[18]. E-mail Spam Detection Framework with Underlined Coefficient Equality Technique," International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064, by Harikrishna K. and Y. Siva Prasad (2012) Impact Factor: 3.358

[19]. Filter for Spamming E-mail by Using Ontology, Anand G. Sharma and Mr. Vedant Rastogi, International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 3, May 2014.

[20]. "Minimizing The Time Of Spam Mail Detection By Relocating Filtering System To The Sender Mail Server," International Journal of Network Security & Its Applications (IJNSA), Vol. 4, No. 2, March 2012.

[21]. by Alireza Nemaney Pour, RahelehKholghi, and Soheil Behnam Roudsari Phishing & Anti-Phishing Techniques, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 20. Jyoti Chhikara, Ritu Dahiya, Neha Garg, and Monika Rani

[22]. Research Trend Analysis Using Word Similarities and Clusters, International Journal of Multimedia and Ubiquitous Engineering, Vol. 8, No. 1, January, 2013, by KyoJoong Oh, Chae-Gyun Lim, Sung Suk Kim, and Ho-Jin Choi

[23]. "Spam and Marketing Communications," ScienceDirect, Procedia Economics and Finance 12 (2014) 265–272, by Kim Janssensa, Nico Nijsten, and Robrecht Van Goolen

[24]. Francis Fortin, Son Dinh, Taher Azeb, DjedjigaMouheb, and Mourad Debbabi "Detection, analysis, and investigation of spam campaigns" Digital Investigation 12 (2015) S12eS21, ELSVIER