Sri Lanka Institute of Information Technology

Email Spam Classification Using Machine Learning

Individual Assignment

Secure Software Systems - IE3042

IT20618872- Thisitha K.L.D

Date of submission

26-05-2023

Abstract— **E-mail is one of the most popular and frequently used ways of communication due to its worldwide accessibility, relatively fast message transfer, and low sending cost. Flaws in e-mail protocols, as well as an increase in electronic business and financial activities, directly lead to an increase in e-mail-based threats. Email spam is a big issue on today's Internet, causing financial harm to businesses and annoyance to individual users. Spam emails are invading users' mailboxes without their knowledge. They use more network capacity and time monitoring and eliminating junk emails. Most Internet users openly despise spam, but enough of them respond to commercial offers that spam remains a viable source of cash for spammers. While most of the users desire to avoid and eliminate spam, they require clear and easy directions on how to behave. Despite all of the precautions made to combat spam, it has not been eradicated. When the countermeasures are very sensitive, even legal emails are blocked. Among the methods devised to combat spam, filtering is one of the most essential. Many spam filtering studies have focused on more advanced classifier-related challenges. Machine learning for spam classification has recently become a significant research topic. The suggested work investigates and identifies the application of several learning algorithms for categorizing spam messages from e-mail. A comparison of the algorithms has also been published.**

Keywords: Machine Learning, J48, MLP, Naive Bayesian, Spam Classification, FBL, Feature Subset Selection

1. INTRODUCTION

The use of the internet has increased dramatically over the last decade and continues to rise. As a result, it is accurate to state that the Internet is progressively becoming a fundamental part of daily life. Internet usage is anticipated to rise more, and e-mail has evolved into a strong tool for exchanging ideas and information. Negligible time delay during transmission, data security, and inexpensive prices are just a few of the numerous advantages that e-mail has over traditional physical means. However, there are a few difficulties that impede effective email use. One of them is spam email. Spam email, or Unsolicited Bulk Email (UBE), has become a common problem on the Internet in recent years. Because spam email is so inexpensive to transmit, unsolicited communications are sent to a large number of individuals indiscriminately. When a large number of spam messages are received, it is required to take a long time to determine whether they are spam or not, and their email messages may cause the mail server to crash.

There are many different sorts of spam email today, such as advertisements for the aim of making money or selling something, urban legends for the purpose of propagating hoaxes or rumors, and so on. Furthermore, HTML emails may contain a web bug, which is a visual in an email message meant to track who is reading the message. As a result, even when we apply existing screening algorithms, some spam emails are classified as non-spam.In general, the sender of a spam message is attempting to accomplish one of the following goals: to advertise some goods, services, or ideas, to defraud users of their private information, to deliver malicious software, or to temporarily crash a mail server. Content spam is separated not only across many topics, but also into several genres, which result from replicating various types of real mail, such as memoranda, letters, and order confirmations.

2. LITERATURE REVIEW

Spam mail, also known as unsolicited bulk e-mail or junk mail, is unsolicited email delivered to a large number of recipients who have not requested it.

The goal of spam filtering is to eliminate unsolicited e-mails.

automatically from a user's mail stream. These unsolicited mails have already caused numerous issues, including overflowing mailboxes, engulfing important personal mail, wasting network bandwidth, and consuming users' time and energy to sort through, not to mention all of the other issues associated with spam (crashed mail-servers, pornography advertisements sent to children, and soon). According to a series of polls done by CAUBE.AU 1, the total amount of spam received by 41 email addresses has increased by a factor of six in two years (from 1753 spams in 2000 to 10,847 spams in 2001). As a result, developing spam filters that can successfully delete the increasing numbers of undesirable mails before they reach a user's mailbox is difficult.

Four distinct classifiers were used to classify email data. The experiment was carried out with various data sizes and feature sizes. If it is finally spam, the final categorization result should be '1'; otherwise, it should be '0'. This research demonstrates that a simple classifier that generates a binary tree can be efficient for datasets that can be classified as binary trees.

3. DATASET DESCRIPTION

The dataset utilized in this study was compiled over the course of two months from various e-magilids. Around 57 spam email attributes were detected and used in the sample. A few of the attributes used were from address to address, type of spam received, and organization from whom the spam was received. The UCI Machine Learning Repository contains datasets for machine learning techniques. UCI's spam dataset comprises of data gathered from 4601 email messages. Each instance in the Spam dataset has 58 attributes. Most of the properties show the frequency of a specific word or character in the email associated with the instance.

- Char freq cap: 3 attributes describing the longest length,

total numbers of capital letters and average length.

- Spam class: the target attribute denoting whether the email

was considered spam or no spam.

4. Methodology

The supervised learning techniques were used to analyze real-time datasets and forecast performance. Different algorithms use different biases to generalize distinct knowledge representations. As a result, they frequently make mistakes in various sections of the instance space. The combination of multiple methods may result in the correction of individual uncorrelated errors. Classifier Selection and Classifier Fusion are the two basic paradigms for dealing with an ensemble of diverse classification methods.
The former chooses a single method for classifying new instances, whereas the latter combines the decisions of all algorithms. This section summarizes the most essential methods from each category. Classifier Selection is a straightforward approach that yields Selection or Select Best. This method assesses each classification algorithm on the training set and chooses the best one for use on the test set. The Classifier Fusion technique can take numerous specialized classifiers as input and learn how well they perform and how their outputs should be blended using training data.

ii. CLASSIFICATION ALGORITHMS

. Spam emails have been filtered using text classification techniques. It comprises keyword, phrase, and character-based research. For filtering spam emails, machine learning for spam classification has been proposed. WEKA is a set of machine learning algorithms written in Java. WEKA is used to conduct a comparison of different learning algorithms for categorizing spam messages from e-mail.

5. RESULT EVALUATION

The data set was divided into two parts: one used as training data to create the prediction model, and the other as test data to test the correctness of our model. Each record's feature values and categorization are stored in the Training data set. The 10-fold cross validation method is used for testing.

6.Conclusion

The use of neural networks to detect spam is a potential technique that is currently being researched. However, in order to achieve peak performance, we must conduct extensive data analysis. Furthermore, this data analysis must be broad in order to detect a broader range of spam. The fundamental premise underlying every spam filtering strategy, whether heuristic or keyword-based, is the same: spam communications often look different than good messages, and recognizing these differences is an effective way to identify and eliminate spam. The main distinction between both systems is the difficulty in discriminating between these two types of email. In achieving this goal, the neural networks technique is more precise, mathematical, and possibly far more accurate and dependable.

7. Future Work

On the basis of this project, the following future work can be completed: - Detecting spam using various networks such as back propagation networks and RBF networks. - For added robustness, we can implement on two or more hidden layers. - Another key content-based strategy for distinguishing spam is fuzzy logic. A fuzzy logic approach to the same problem can provide some new insights. - To attain higher categorization rates, a combinational strategy (including header filters, content-based filters, and user-specific information) might be applied.

8. References

[1] C. Pu and S. Webb, “Observed trends in spam construction techniques: A case study of spam evolution”, Proceeding of 3 rd Conference on E-Mail and Anti-Spam, 2006.

[2] M. Embrechts, B. Szymanski, K. Sternickel, T. Naenna, and R. Bragaspathi,“Use of Machine Learning for Classification of Magnetocardiograms”, Proceedings of IEEE Conference on System, Man and Cybernetics, Washington DC, pp. 1400-05, 2003.

[3] Upasana Pandey and S. Chakraverty “A Review of Text Classification Approaches for E-mail Management”, in IACSIT International Journal of Engineering and Technology, Vol. 3, No. 2, 2011.

[4] IronPort Systems, “Spammers Continue Innovation: IronPort Study Shows Image-based Spam, Hit & Run, and Increased Volumes Latest Threat to Your Inbox,” June 2006. [Online]. Available: http://www.ironport.com/company/ironport_pr_2006-06-28.html

[5] Alessandro Rozza, Gabriele Lombardi, and Elena Casiraghi. Novel ipca based classifiers and their application to spam filtering. In Intelligent Systems Design and Applications, 2009. ISDA’09. Ninth International Conference on, pages 797-802. IEEE, 2009.

[6] Aziz Qaroush, Ismail M Khater, and Mahdi Washaha. Identifying spam email based-on statistical header features and sender behavior. In Proceedings of the CUBE

International Information Technology Conference, pages
771-778. ACM, 2012.