JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

NATURAL LANGUAGE PROCESSING FOR DETECTING SPAM E-MAIL

Rahul[1], Dr. Banita[2]
Research Scholar[1], Associate Professor[2]
Computer Science and Engineering
Baba MastNath University, Rohtak

Abstract

Email has been the most important medium of communication nowadays, through internet connectivity any message can be delivered to all aver the world. More than 270 billion emails are exchanged daily, about 57% of these are just spam emails. Spam emails, also known as non-self, are undesired commercial or malicious emails, which affects or hacks personal information like bank ,related to money or anything that causes destruction to single individual or a corporation or a group of people. Besides advertising, these may contain links to phishing or malware hosting websites set up to steal confidential information. Spam is a serious issue that is not just annoying to the end-users but also financially damaging and a security risk. Spam is any irrelevant and unwanted message or email sent by the attacker to a significant number of recipients by using emails or any other medium of information sharing. As a result, there is a huge need for email system security. Spam emails could contain Trojans, rats, and viruses. Attackers primarily employ this strategy to entice people to use internet services. They might send spam emails with multi-file attachments and URLs jam-packed with harmful and spammy websites, which could result in identity theft, financial fraud, and data breaches. Many email service providers let their customers create keyword-based rules for email filtering.

Keywords:- Natural Language Processing, Spam Filtering, machine learning

I. INTRODUCTION

In recent years, internet has become an integral part of life. With increased use of internet, numbers of email users are increasing day by day. This increasing use of email has created problems caused by unsolicited bulk email messages commonly referred to as Spam. Email has now become one of the best ways for advertisements due to which spam emails are generated.Spam e-mail identification is based on the identities of the e-mail senders. Whitelist and blacklist known as List based method which is used in this identification process. E-mail users have the provision to set up the whitelist and blacklist. The users can add e-mail addresses of their senders in the whitelist and blacklist. The users present in the whitelist are the senders who are allowed to send e-mails to the users' inbox whereas the e-mails sent by the users present in the blacklist

are moved into the spam box. The adaptive blacklist is maintained in the server side in the case of distributed adaptive blacklist method. The filter gets invoked when an e-mail is received by the e-mail transfer agent. Currently existing spam messages identification methods are discussed. The general classifiers used in the spam message identification are explained. Not only the selection of classifiers is essential but also the selection of features incorporating into the spam message filtering algorithms is also important.

The header part of the e-mail is checked and identifies the spam e- mails is a type of identity-based spam classification method. The e-mails those have no sender id in the 'From' section, and too many receivers id in the 'To' section is checked by this method.

Spam emails are the emails that the receiver does not wish to receive. Many identical messages are sent to several recipients of email. Spam usually arises as a result of giving out our email address on an unauthorized or unscrupulous website. There are many of the effects of Spam. Fills our Inbox with number of ridiculous emails. Degrades our Internet speed to a great extent. Steals useful information like our details on you Contact list. Alters your search results on any computer program .Spam is a huge waste of everybody's time and can quickly become very frustrating if you receive large amounts of it .Identifying these spammers and the spam content is a laborious task . even though extensive number of studies have been done, yet so far the methods set forth still scarcely distinguish spam surveys, and none of them demonstrate the benefits of each removed element compose.

Machine learning is one of the most important and valuable applications of artificial intelligence (AI), which gives computer systems the ability of automatically learning and enhancing their functionality without explicit programming. The primary purpose of machine learning algorithms is to build automated tools to access and use the data for training. Machine learning consists of three major kinds, used for numerous tasks. For the last decade, researchers have been trying to make email communication better than today. Spam filtering of emails is one of the most critical ways of protecting email networks. Many research articles have been published using various machine learning approaches to identify and process spam emails, but there are still some research gaps. Junk mail is one of the central, attractive research fields for filling the gaps.

Apart from the numerous advantages that emails can provide, unsolicited emails may sometimes emerge in a user's mailbox. On the Internet, spam emails have long been a concern. Their high volume on the Internet has a negative impact on the email server's memory. Furthermore, spam emails may result in financial loss as a result of malicious users' deception. Also, unsolicited mail recipients will be inconvenienced and waste time since they must read the whole content to decide whether it is unsolicited mail. They obstruct the timely transmission of legitimate messages. To detect Spam e-mail is a challenge for researchers trying to design an efficient filtering mechanism as a result of the aforementioned issues. Knowledge engineering and machine learning are two popular ways for filtering unsolicited emails. Knowledge engineering requires a set of rules. It is not a powerful mechanism as the rule set needs to be continually updated. Machine Learning is proved to be efficient for this purpose as rule set is not required. Machine learning uses set of training and test data. Training data consists of mails which are already classified as spam or ham. Natural Language Processing tasks play an important role in detecting whether an email is unsolicited or not. NLP converts unstructured

text of e-mail into structured text and helps in text analysis. NLP is used to increase the model's accuracy

II. PROPOSED METHOD

In this system, to solve the problem of spam, the spam classification system is created to identify spam and non-spam. Since spammers may send spam messages many times, it is difficult to id entify it every time manually. So we will be using some of the strategies in our proposed system to detect the spam. The proposed solution not only identifies the spam word but also identifies the IP address of the system through which the spam message is sent so that next time when the spam message is sent from the same system our proposed system directly identifies it as blacklisted based on the IP address.

In the proposed model ,the web application is done using dot net and spam detection is done using machine learning. The web application consists of following modules:

- **User Management**
- **Login**
- **Registration**
- **Compose**
- **Inbox**
- **Sent**
- **Trash**
- **Voice Message**
- **Offline notification**
- **Delete For everyone**
- **Read Message**

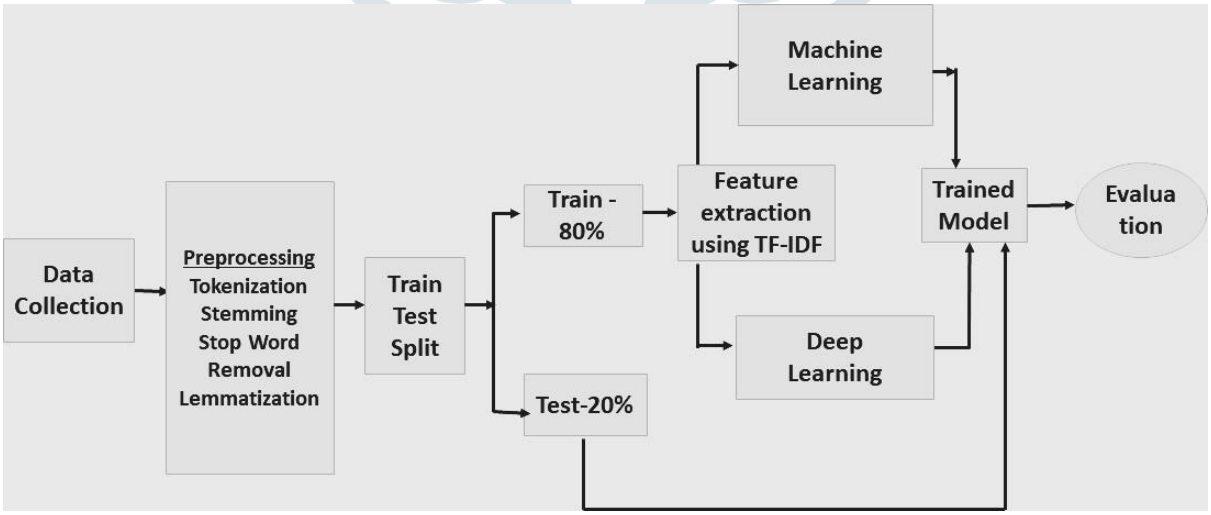

Fig.1: Data Flow Diagram

This Fig.1 describes a strategy for utilising Natural Language Processing (NLP) to anticipate Spam and Ham email based on the Spam Email Dataset. We begin by converting, pre-processing, and partitioning datasets to meet the needs

of the algorithms under consideration. The multiple models are then trained and evaluated, with performance measures used to evaluate and compare them.

III. SPAM MESSAGES

a) Standard filtering

The email spam definition is ambiguous since everybody has their views on it. At present, email spam is getting the attention of everyone. Email spam ordinarily includes particular spontaneous messages sent in mass by individuals you do not know. In the era of technology, the dodger/spammer shows a story where the unfortunate casualty needs forthright financial help so that the fraudster can gain a lot bigger total of cash, which they would then share.

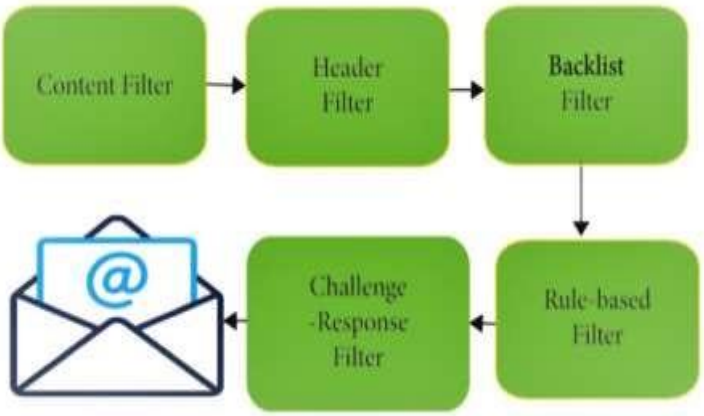

Fig.2: Filtering spam

The fraudster will either earn a profit or avoid communication when the unfortunate victim completes the instalments. Currently, various companies develop different techniques and algorithms for efficient spam detection and filtering. We address some filtering strategies in this section to understand the filtering process. Standard spam filtering is a filtering system that implements a set of rules and works with that set of protocols as a classifier. Figure 1 illustrates a standard method for filtering spam. In the first step, content filters are implemented and use artificial intelligence techniques to figure out the spam. The email header filter, which extracts the header information from the email, is implemented in the second step. After that, backlist filters are applied to the emails to clinch the emails coming from the backlist file to avoid spam emails. After this stage, rule-based filters are implemented, recognizing the sender using the subject line and user-defined parameters. Eventually, allowance and task filters are used by implementing a method that allows the account holder to send the mail,

IV. Use of Machine Learning for Spam Email Detection

Machine learning is one of the most important and valuable applications of artificial intelligence (AI), which gives computer systems the ability of automatically learning and enhancing their functionality without explicit programming. The primary purpose of machine learning algorithms is to build automated tools to access and use the data for training. Machine learning consists of three major kinds, used for numerous tasks. For the last decade, researchers have been trying to make email communication better than today. Spam filtering of emails is one of the most critical ways of protecting email networks. Many research articles have been published using various machine learning approaches to identify and process spam emails, but there are still some research gaps. Junk mail is one of the central, attractive research fields for filling the gaps. That is why, this paper is presented to make a summarized version of different existing machine learning models and approaches that are being used for email spam detection. This paper also evaluates the most common machine learning approaches like KNN, SVM, random forest, and Naive Bayes.

V. CONCLUSION

This study concludes that most of the proposed email and IoT spam detection methods are based on supervised machine learning techniques. A labelled dataset for the supervised model training is a crucial and time-consuming task. Supervised learning algorithms SVM and Naive Bayes outperform other models in spam detection. The study provides comprehensive insights of these algorithms and some future research directions for email spam detection and filtering.

REFRENCES:

[1] AbdulNabi and Q. Yaseen, “Spam email detection using deep learning techniques,” in Procedia Computer Science, 2021, vol. 184, pp. 853–858. doi: 10.1016/j.procs.2021.03.107.

[2] S. Madisetty and M. S. Desarkar, “A Neural Network-Based Ensemble Approach for Spam Detection in Twitter,” IEEE Trans. Comput. Soc. Syst., vol. 5, no. 4, pp. 973–984, Dec. 2018, doi: 10.1109/TCSS.2018.2878852.

[3] A. Salunkhe, “Attention-based Bidirectional LSTM for Deceptive Opinion Spam Classification,” Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.14789

[4] A. Hosseinalipour and R. Ghanbarzadeh, “A novel approach for spam detection using horse herd optimization algorithm,” Neural Comput. Appl., Mar. 2022, doi: 10.1007/s00521-022- 07148-x.

[5] P. K. Roy, J. P. Singh, and S. Banerjee, “Deep learning to filter SMS Spam,” Futur. Gener. Comput. Syst., vol. 102, pp. 524–533, Jan. 2020, doi: 10.1016/j.future.2019.09.001.

[6] K Agarwal abd T Kumar, "Email Spam Detection using Integrated approach of Naïve Bayes and Particle Spam Optimation, "IEEE Electron Devices Society, Institute of Electrical and Electronics Engineers, and Vaigai College of Engineering, Proceeding of the 2018 International Conference on Intelligent Computing and Control Systems (ICICCS) : June 14-15, 2018.

[7] E. M. Bahgat, S. Rady, W. Gad, and I. F. Moawad, "Efficient email classification approach based on semantic methods," Ain Shams Eng. J., vol. 9, no. 4, pp. 3259–3269, Dec. 2018, doi: 10.1016/j.asej.2018.06.001.

[8] W. Pan et al., "Semantic Graph Neural Network: A Conversion from Spam Email Classification to Graph Classification," Sci. Program., vol. 2022, 2022, doi: 10.1155/2022/6737080.

[9] J. Z. Pan, S. Pavlova, C. Li, N. Li, Y. Li, and J. Liu, "Content based fake news detection using knowledge graphs," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, vol. 11136 LNCS, pp. 669– 683. doi: 10.1007/978-3-030-00671-6_39.

[10] [41] Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., & Akbar, M. (2020). Fake news detection using deep learning models: A novel approach. Transactions on Emerging Telecommunications Technologies, 31(2), e3767.

[11] B. K. Dedeturk and B. Akay, "Spam filtering using a logistic regression model trained by an artificial bee colony algorithm," Appl. Soft Comput. J., vol. 91, Jun. 2020, doi: 10.1016/j.asoc.2020.106229.

[12] W. Feng, "2016 IEEE 35th International Performance Computing and Communications Conference, IPCCC 2016," 2016 IEEE 35th Int. Perform. Comput. Commun. Conf. IPCCC 2016, 2017.

[13] Amanoul, S. V., Abdulazeez, A. M., Zeebare, D. Q., & Ahmed, F. Y. (2021, June). Intrusion Detection Systems Based on Machine Learning Algorithms. In 2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS) (pp. 282-287). IEEE.

[14] T. Verma and D. C. Rana, "Data Mining Techniques for the Knowledge Discovery," Int. J. Eng. Technol., vol. 9, no. 3S, pp. 351–354, Jul. 2017, doi: 10.21817/ijet/2017/v9i3/170903s054.

[15] D. Tang, B. Qin, and T. Liu, "Document Modeling with Gated Recurrent Neural Network forSentiment Classification," Association for Computational Linguistics, 2015. [Online]. Available: http://ir.hit.edu.cn/

[16] A. Ishaq et al., "Extensive hotel reviews classification using long short term memory," J. Ambient Intell. Humaniz. Comput., vol. 12, no. 10, pp. 9375–9385, Oct. 2021, doi: 10.1007/s12652- 020-02654-z.