IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An Analytical Study On An Email Spam Detection

Prof. Jyoti Choudhary

Assistant Professor, Nirmala Memorial Foundation College of Science & Commerce

Abstract:

Too many people today rely on emails and messages sent by strangers. Since anyone can leave emails and messages, spammers have a great opportunity to spam about our diverse interests. Spam fills your inbox with a bunch of ridiculous emails. Your internet speed will drop significantly. It steals useful information such as contact list details. Identifying these spammers and spam content can be a hot topic for research and tedious work. Spamming is the act of sending a large number of messages via email. Since the cost of spam is primarily borne by the recipient, it is effectively postage-based advertising. Spam is an economically viable type of commercial advertising as email can be a very cheap medium for senders. Using Bayes' theorem and a naive Bayesian classifier, this proposed model can declare whether a given message is spam or not, and often finds the sender's IP address as well.

Keywords - Term Frequency, Naive Bayes, spam, ham, Language Toolkit.

Introduction:

E-mail services have become widespread due to information communication via the Internet. These email features are also causing problems for users with electronic junk mail. These areknown as spam emails.

Spam emails are unsolicited and unwanted emails sent in bulk to a random list of recipients. Spam is usually sent for commercial purposes. It can be widely distributed by botnets, which are networks of infected computers. Spammers often access data under the guise of legitimate institutions. In doing so, they may create a fake email address that looks like your real email address. For example, PayPal spam emails can come from addresses ending with @paypai.com.

Phishing spam emails pretend to come from legitimate and trusted sources, such as banks, in an attempt to obtain your personal information. This information is used for fraudulent purposes.

The advantage of email is to provide people with standard, archived, one-to-many electronic communication. Standard means that it is based on SMTP, an international standard. Therefore, email sent from any email client application in any language in the world will be delivered to any other email client. Archiving is the natural storage of email-by-email servers and email clients for later use in a more organized and robust manner, usually than other forms of electronic communication such as SMS or instant messaging. means that One-to-many means that the email is designed to be delivered from her one sender to many recipients. This is also more natural than supporting text and messaging applications. Faster, cheaper and more convenient than traditionalpostal delivery. We also deliver electronic documents.

Malware is software designed to harm your device by releasing viruses to damage it or steal yourpersonal data.

One of the negative effects of e-mail is its susceptibility to abuse through methods such as spam. Spam can harm communication by cluttering valuable communication with a large amount of unnecessary communication. The chart on the left shows spam delivery attempts to employee accounts of a large US company. A second negative effect can affect privacy. Because e-mails are naturally archived, your private communications may be available to others who may have access to your emails.

The increase in unsolicited e-mail (known as spam) has increased the need to develop more reliable and robust anti-spam filters. Successfully detect and filter spam emails using new machine learning techniques. We present a systematic overview of some of the common machine learning-based email spam filtering approaches. Our review includes an overview of key concepts, efforts, efficiencies, and research trends in spam filtering. The preliminary discussion behind the study looks at the application of machine learning techniques to the email spam filtering processes of major Internet Service Providers (ISPs), such as email spam filters in Gmail, Yahoo, and Outlook.

Objectives :

1. To understand and identify spam mails.
2. To analyze the difficulties faced by the people.
3. To study the impacts of spam mail in the future.
4. To study and find the solution to spam mail.

Literature Review

(Razak, Mohamad, 2013) highlighted several features contained in the email header used to efficiently identify and classify spam messages. According to his study all the features are found in Yahoo Mail, Gmail, and Hotmail, so he proposes a generic spam message detection mechanism for all major email providers.

(Rathod, Pattewar, 2015) stated that a new strategy-based approach of using word repetitions was used. Key phrases in incoming emails, key phrases containing keywords need to be tagged, then the grammatical roles of whole words in the sentence are determined and finally put together into a vector. Similarity between incoming emails. The K-Mean algorithm is used to classify incoming emails. Vector determination is the method used to determine which categoryan email belongs to.

(Alurkar, Ranade, Joshi, Sonewa, Mahalle, Deshpande, 2017) mentioned that the proposed system uses machine learning techniques to try to recognize patterns of repetitive keywords classified as spam. The system also suggests a classification for the email based on various other parameters contained in the structure, such as Cc/Bcc, domain, headers, etc. Each parameter is considered a feature when applied to a machine learning algorithm. The machine learning model will be a pre-trained model with a feedback mechanism to distinguish between correct and ambiguous outputs. This method provides an alternative to his architecture that can implement a spam filter. This paper also examines email bodies containing commonly used keywords and punctuation marks.

(Rathod, Pattewar, 2015) examined the use of the string-matching algorithm for spam email detection. In particular, in this work, on two different datasets, six known string-matching algorithms, i.e. longest common subsequence (LCS), Levenshtein distance (LD), Jaro, Jaro - Winkler, Bi-gram, and TFIDF efficiency are examined and compared. Enron corpus and CSDMC2010 spam records. They observed that the bi-gram algorithm performed the best spamdetections on both datasets.

PROPOSED SYSTEM:

In this system, to solve the problem of spam, a spam classification system was created to distinguish between spam and non-spam. Spammers can send spam messages multiple times, making it difficult to manually identify each time. Therefore, we use several strategies in the proposed system to detect spam. The proposed solution not only identifies the spam word, but also the IP address of the system through which the spam message was sent. This will blacklist messages sent by the proposed system the next time it detects spam. Identified directly based on IP address. In the proposed model, the web application runs on a dot mesh and spam detection isdone on machine learning. A web application consists of the following modules:

1. First-time users must register. By using the website, a user or person must register. This registration helps us maintain separate accounts for each user. User registration is required before login. The user logs into the main page using their registered name and password. Allowed page is displayed as soon as the user successfully logs in. Otherwise, an error message will be displayed. Registration is required.

Login: Users login to the main page using their registered name and password. After the user successfully logs in, the authorized page is displayed. If not, an error message will be displayed.Registration is required.

Registration: Users or Individuals must register with the Site the first time they use it.
Registering this to

helps us maintain a separate account for each user. User registration is required before login.

2. Compose

Input: The sender composes a new email. The sender should add the recipient's address, subject,and message.

Output: Email will be sent based on the address provided by the recipient.

3. Inbox: This page stores all emails received by the user. All emails received are listed by date. Input: The Inbox page accepts all incoming emails sent to individuals.

Output: Recipients can open and read email sent to the address.

4. Sent: This folder stores all emails sent by usersEnter: Select and delete all unwanted emails.

Output: All deleted emails are placed in the Recycle Bin. The Recycle Bin stores all deletedemails.

6. Voice Message Input: An email was sent from the sender in the form of a text message.

7. Offline Notification

Input: Sender will send email

Output: Recipient will receive offline notification as SMS in text format.

8. Delete For all

Inputs: Here the sender deletes the email sent

Output: The email has been deleted or deleted for both sender and recipient.9 Read message

Input: The recipient reads the email.

Output: The sender receives notification that the message was read by the sender. When a message is received in the Inbox, this message is exported to Records. If the Naive Bayes classifier was not used, this message would be detected as spam or Models need to be trained before recognizing whether received messages are spam.

So there are two types of data in this repository: ham (non-spam) and spam data. Ham data also has light and heavy data, which means that there is non-spam data that is very similar to spam data. This can make system decisions difficult.

Examine and analyze data

Let's look at the contents of an email message to get a basic understanding of the dataHam

This is a normal email to someone else who wouldn't be hard to classify as a ham. It looks likean email reply.

Hard Ham (More Tricky Ham Email)

Hard Ham is really hard to distinguish from spam data as it contains keywords like limited time orders, special back to school offers, and highly suspicious!

Spam

One of the spam training data looks like one of the following spam promotional emails in yourjunk folder.

Visualization

Wordcloud is a handy visualization tool that provides rough estimates of the most commonwords in your data.

One important thing to note is that word clouds only show word frequency, not necessarily word importance. Therefore, some data must be erased., punctuation, Removing stop words etc. from the data before visualizing it.

Visualizing N-Gram Models

Another visualization technique is to use bar charts to show the frequency of the most frequently occurring words. N-gram refers to the number of words that are considered a unit when calculating word frequency.

Algorithm Implementation TfidfVectorizer + Naive Bayes Algorithm

The first approach I took was to use TfidfVectorizer as a feature extraction tool and Naive Bayes Algorithm for prediction. Naive Bayes is a simple, probabilistic, traditional machine learning algorithm.

The Naive Bayes library provided by the sklearn library saves you a lot of trouble implementing this algorithm yourself. It's easy to do with just a few lines of code.

from sklearn.naive_bayes import GaussianNBclf .fit (x_train_features.toarray() ,y_train)

Score output is accuracy of prediction# Accuracy: 0.995

clf.score(x_train_features.toarray(),y_train)# Accuracy: 0.932 clf.score(x_test_features. toarray(),y_test)

We achieve an accuracy of 93.2%. But accuracy is not solely the metrics to evaluate the performance of an algorithm.

Finding & Suggestions:

Unsolicited commercial email messages sent in bulk, often using purchased (or stolen) mailing lists that include your address. Spam is not only annoying and wastes time sifting through unwanted messages, it can also cause serious harm by infecting users' computers with malicious software that can damage the system or steal personal information. may cause harm. It can also consume network resources. Spam emails are dangerous. It may contain malicious links that can infect your computer with malware. Do not click spam links. Dangerous spam emails often sound urgent, so you should take action. Download spam filtering tools and anti-virus software Like ZEROSPAM Cloud-based spam, ransomware, and phishing blocker.
SpamSieve Best email spam filter for Mac with adaptive spam detection over time. This reduces the risk of malware infecting your computer. Therefore, opt for spam filter tools and antivirus software with such features to reduce the hassle of decrypting email content.

Conclusion:

E-mail is now the primary medium for communication, and messages can be delivered anywhere in the world via an Internet connection. More than 4,444 270 billion emails are exchanged every day, and about 57% of these 4,444 emails are just spam. Spam emails (also known as not self) are unsolicited commercial or malicious emails that contain or hack personal information, such as: In addition to advertisements, these may contain links to phishing or malware-hosting websites that are set up to steal sensitive information.

Therefore, this system is designed to detect and prevent junk and junk mail, helping reduce spam messages. This is a huge win for both the individual and the system. It is also possible to add functions to existing systems. When it comes to spam email filtering, fitness features are important and should be chosen very carefully. SPAM databases are also very important for classifying emails as SPAM and HAM emails. It is not possible to preset thresholds for the fitness function. It depends on the data and the nature of the problem. Data dictionary words are used to classify SPAM and HAM emails and should be chosen carefully.

References:

[1] Shukor Bin Abd Razak, Ahmad Fahrulrazie Bin Mohamad

, "Identifying Spam Email Based on Information from Email Headers," 13th International Conference on Intelligent SystemsDesign and Applications (ISDA), 2013.
[2] Mohammed Reza Parsei, Mohammed Salehi „E-Mail Spam Detection Based on Part of Speech Tagging" 2nd International Conference on Knowledge Based Engineering and Innovation (KBEI), 2015.
[3] Sunil B. Rathod, Tareek M. Pattewar "Content Based Spam Detection in Email using Bayesian Classifier", präsentiert auf derIEEE ICCSP 2015 Konferenz.
[4] Aakash Atul Alurkar, Sourabh Bharat Ranade, Shreeya Vijay Joshi, Siddhesh Sanjay Ranade, Piyush A. Sonewa, Parikshit N. Mahalle, Arvind V. Deshpande "Proposed Data Science
Email Using Machine Learning Spam Classification Approaches,Techniques, 2017.
[5] Kriti Agarwal, Tarun Kumar, Email Spam Detection Using the Integrated Approach of Naive Bayes and Particle Swarm Optimization, Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018.
[6] How to design a spam filtering system with Machine Learning Algorithm
https://towardsdatascience.com/email-spam-detection-1-2-b0e06a5c0472

[7] Mujtaba, Gram, et al. "Research Trends in Email Classification: Reviews and Open Issues". IEEE Access 5 (2017).