Institute of Technology Carlow
Kilkenny Road,
Carlow
Bsc (Hons) Cybercrime and IT Security

The Spam Catcher
Research manual

Author: Aya Aloraimi
Student Number: C00212086
Supervisor: James Egan

Table of Contents

Abstract

E-mail is perhaps the most way to transmit messages from one person to another digitally. For each user, email spam is the very recent issue that faces individuals and companies. According to a study, individuals receive on average more than 50% of spam mails. In fact, the amount of spam messages obtained is increasing exponentially. Therefore, dealing with spam is very important [1]. There are different techniques were suggested by researchers to protect our spam mailbox from spammers. This paper presents different topics related to email spam such as types of email spam, the historical background of email spam. Also it illustrates different types of machine learning and spam filtering techniques. Furthermore, describing the way of implementing Naïve Bayes classifier in machine learning. This technique is used to reduce spam email by discriminate on the basis of incoming mails are spam or legitimate.

Introduction

In early 1990s the use of email spam has been increased and it's become an issue that faces email users. Email spam, otherwise called junk email, is unwanted bulk messages that are sent through email in huge amount. A spammer uses spambots which are programs that looks for email addresses over the Internet from different websites and chatrooms. Once spammers obtain list of emails they send an unsolicited email to lots of email addresses in order to expect interaction from them. In the past, the main aim was to flood mailing list with inappropriate messages while nowadays people tend to obtain money by tempting users to buy products and prohibitive nature, harvesting sensible information (from emails and passwords to bank accounts) via bank frauds and requests for help. According to recent reports [IBM, 2014], [Cyberoam,2014], [Symantec, 2014], spam is being increasingly used to distribute viruses, malware, links to phishing sites, etc. An average of 54 billion spam e-mails was sent worldwide each day [Cyberoam, 2014][2].

There are different types of spam that are used to conduct email fraud such as, job opportunities, online degrees, sell products and services in order to make money from receivers. Based on the Ferris Research (2009), spam can be classified into the following: 1. Health; such as fake pharmaceuticals; 2. Promotional products; such as fake fashion items; 3. Adult content; such as pornography and prostitution; 4. Financial and refinancing; such as stock kiting, tax solutions, loan packages[3]. One of the most popular form of email spam is phishing emails which are created to direct recipients to a fake version of bank website, login page where a user is prompted to enter the required details. Additionally, Malware is considered as a type of spam email which includes malicious files, scripts and links that contains malware and viruses to be installed in the user's PC. However, There are lots of spam emails that don't contain any type of viruses or malware and don't appear to have any rational whatsoever, this type is known as Nonsensical email.
By the end of 2006 the nature of spam had completely changed. While spam was mostly based on text, it began to look more graphic. Spammers take the advantage of images to avoid text-based filters, simply by hiding the textual spam message into images therefore, anti-spam program could only see pixels whereas end user receives a text-message.

Topic area

This topic provides the history of email spam and describes different email spam filtering techniques that are used to mitigate email spam. In addition, it provides implementation of some methods in Machine learning using different algorithms.

Overview of the history of email spam

Email Spam history is one that is closely tied to the Internet's own history and evolution. Before 2004, the historical background of email spam begins. The figure below describes the timeline sequence of email spam.

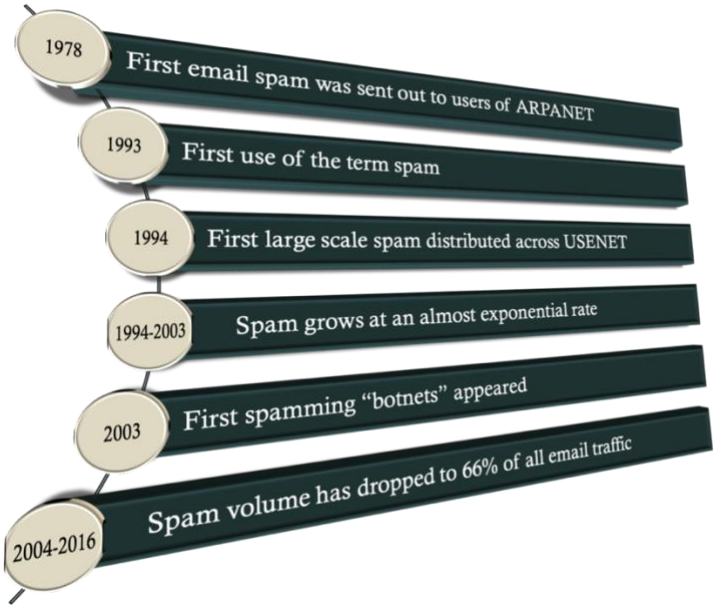

Figure 1- Timeline for the history of email spam from 1978-2016

The first occurrence of email spam was sent out to users of ARPANET in 1978, when a man named Gary Thuerk distributed an advertisement for a new digital computer and it went out to almost 400 users who had email accounts on ARPANET. In fact, the word "spam" would not come into use until 1993. It has been added to a message from the USENET" is a large collection of computers that share data with each other"[4] by Richard Depew to news.admin.policy, which was the product of a flaw in a software program which caused 200 messages to go out to the news group. In 1994, the first large-scale spam spread across USENET. A student posted a message with the title "Global Alert for All: Jesus is Coming Soon" to every newsgroup. From 1994 to 2003, spam rises at an almost unprecedented pace until it eventually accounts for the vast majority(80 to 85 percent) of the worldwide email messages. "Spam spread quickly among the UseNet groups who were easy targets for spammers simply because the e-mail addresses of members were widely available" (Templeton, 2003) [5]. After the email spam spread significantly, the first spamming "botnets" occurred in 2003. Recently, the tide has swung against spammers as according to Symantec Corporation's "2014 Internet Security Threat Report, volume 19, " spam volume has dropped to 66 percent of all email traffic.

The impact of email spam

Spam email generally occurs as a result of providing your email address on an unauthorized or unskilled websites. In fact the presence of the spam email has an effect on several things for instance, fills your inbox with number of useless emails, reduce the Internet speed to a

large extent, details can be stolen such as contact list and also change the result of your search on any search engine.

Email spam filtering techniques (non-machine learning)

Over the last few years, spam filtering techniques have advanced rapidly as spam emails have increased. For example, Blacklist is a common spam-filtering method which is used by organization's system administrator to create a pre-set list of senders in order to block messages come from them. The Blacklist basically is a list of email addresses or Internet Protocol (IP) addresses that have been already used to send spam. When an incoming message arrives, the spam filter checks to see if its IP or email address is on the blacklist; if so, the message considered spam and rejected. Likewise the Real-Time Blackhole List method which works identically as blacklist but it is maintained by a third party system which take the time to build comprehensive blacklists on the behalf of subscribers. There is another technique that lets users to specify email addresses from trusted senders which is known as Whitelist. However, using this technique would automatically block anyone who was not approved. Most anti-spam programs use an adaptive whitelist variant of this scheme. In this case, the email address of unknown sender is checked against a database ; if they do not have spamming history, their message is sent to the receiver's inbox and added to the whitelist. Greylist is a new technique for spam-filtering which take advantage of the fact that many spammers are only trying to send a lot of junk mail once. The receiving mail server rejects any mails come from unknown users and sends a message of failure to the originating server, if the mail server tries to send the message a second time- a move will be taken by the most legitimate servers – the Greylist believes the message is not spam and allows it to continue to the inbox of the receiver.
Rather than implementing all emails from a particular email or IP address across the board, content-based filtering method analyses common words or phrases contained in each message to decide whether an email is spam or legitimate, this type can produce false positive. For instance, if the filter is set to block all messages containing the word "discount", emails from legitimate senders offering a reduced price to your non-profit hardware or software may not reach their destination.

Heuristic filters take into account multiple terms contained in an email rather than blocking emails that contains a single phrase. It search incoming email content and allocate points to words or phrases. The most common words found in spam messages receive higher point, whereas terms often found in normal emails receive lower scores. Then the filter adds all points and the total score is calculated, after that the filter determines if the email receives a certain score or lower (determined by the administrator of the anti-spam application). This method works quickly to avoid latency and is quite efficient when enabled and configured. Nevertheless, it can produce false positive if a legitimate contact sends an email with a certain combination of words. Some spammers may learn which terms to avoid using the heuristic filter into thinking that they are successful senders.

Bayesian filter is the most advanced form of content-based filtering, it uses statistical probability laws to decide that messages are legitimate and which are spam. The end user must first "practice" it by flagging the message manually as either junk or legitimate. The filter takes words and phrases that are used in legitimate emails over time and adds them to a list; it does the same for spam expressions. The Bayesian filter checks the contents of the email to decide that incoming message are marked as spam, and then compares the text to its two-word lists to measure the probability that the message will be spam.

Usual structure of spam filter

In general, each message has (the subject, sender and receiver) and body (the actual message content of the message). These information is used by a classifier but before that certain steps are required and they are :

1- Tokenization: extracts words in the context of the email;
2- Lemmatization: decreases words to their root form for example(“taking” to “take”);
3- Removing stop-word (such as “the”) that a search engine has been programmed to ignore. Also removing those words that often arise in many messages (e.g., “to”, “a”, “for”);
4- Representation which translates the set of words in the message to the specific format needed by the algorithm used for machine learning.

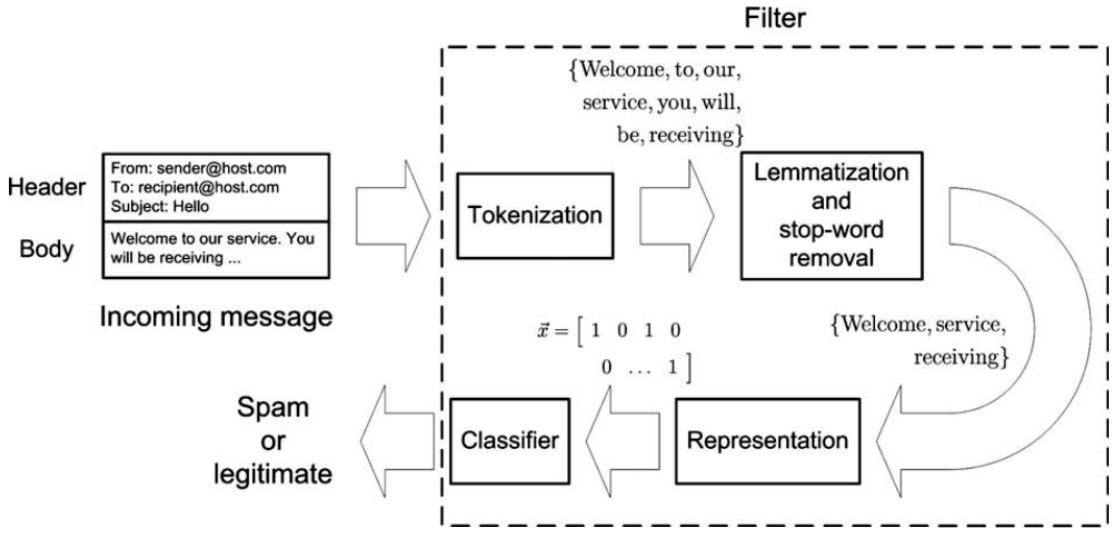

Figure 2- The main steps required in a spam filter[6]

Types of machine learning

Machine learning is an artificial intelligence (AI) technology which based on the idea that with minimal human input, machines can learn from information, recognize patterns and make decisions. “In 1959, Arthur Samuel described ML as the field of study that gives computers the ability to learn without being explicitly programmed”(Samuel 1959)[7].It focuses on creating computer programs that can navigate and use data to learn on their own. The performance of machine learning techniques in text categorization has led researchers in anti-spam filtering to explore learning algorithms. Therefore, machine learning algorithms are categorized as:

- Supervised machine learning algorithms: This algorithm consists of a target/ result variable (or dependent variables) to be calculated from a given set of predictors (independent variables). By using these set of variables we generate a function that maps inputs to desired outputs. There are lots of supervised learning examples, such as: Regression, Tree Decision, Random Forest, KNN, Logistic Regression, etc.

- Unsupervised machine learning algorithms: Unlike supervised machine learning, this machine learning aims to discover previously unknown patterns in data, but most of the time these patterns are weak approximations of what supervised machine learning can accomplish. Therefore, as you don't know what the outcomes should be, there is no way to determine how correct they are. Clustering is a good application to be used in this type of machine learning.

- Reinforcement machine learning algorithms: Various software uses this machine learning to find the best direction it should follow, take appropriate actions and discover errors or rewards in a particular situation.

Machine learning algorithms

There are different algorithms to be applied in machine learning and each algorithm can be used depending on data problem. Here is a list of some widely used algorithms:

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. Naïve Bayes
5. KNN
6. K-Means
7. Random Forest
8. SVM
9. Dimensionality Reduction Algorithms
10. Gradient Boost & Adaboost

Linear Regression

It is used to calculate real values based on continuous variable(s) such as house price and total sales. In this algorithm, a relation is established between two variables which are Independent and dependent variables by fitting a best line which is known as regression line and can be represented by a linear equation Y= a*X + b. In this equation:

- Y - Dependent Variable
- a - Slope
- X - Independent variable
- b - Intercept

There are two types of Linear Regression: Simple Linear Regression which is characterized by one independent variable and Multiple Linear Regression which is characterized by multiple variables.

Logistic Regression

Based on the set of Independent variable(s), this algorithm is used to estimate discrete values(Binary values like 0/1, yes/no, true/false). Basically, by fitting data to a logit function, it calculates the likelihood of an event occurring. Since the likelihood is expected, its output values range from 0 to 1.

Decision Tree

In supervised machine learning this algorithm is used for classification problems and it operates for both categorical and continuous dependent variables. In this algorithm, the

population is divided into two or more homogeneous sets to make as distinct group as possible based on the most important attributes/ independent variables.

Naïve Bayes

It is a technique of classification based on the theorem of Bayes with an assumption of independence among predictors. A Naïve Bayes classifier assumes that any other feature is unrelated to the presence of a particular feature in a class. This method is easy to construct and is especially useful for huge data sets.

KNN (K- Nearest Neighbours)

This algorithm is used in both classification and regression issues. Nevertheless, it is mostly used in industry classification problems. It stores all available cases and by a majority vote of its K neighbours classifies new cases. The case specified to the class is most common among its closest neighbours, measured by a distance function which can either Euclidean, Manhattan, Minkowski or Hamming distance.

K-Means

It is used in unsupervised machine learning to solve the clustering problem, by following a simple and easy way of classifying a given set of data through a number of clusters (assuming K clusters) and each cluster has its own centroid.

Random Forest

Random Forest is a trademark term for a group of decision trees, these decision trees is known as "Forest". Each tree classifies a new object based on attributes so, we say the tree "votes" for that class. The forest chooses classification with the majority of votes.

SVM(Support Vector Machine)

SVM is a classification method which works by plotting each data item as a point in n-dimensional space (n refers to the number of features you have) with the value of each feature being the value of a specific coordinate.

Dimensionality Reduction Algorithms

It is the method of reducing the number of random variables to be regarded by obtaining a set of major variables . it can be broken down into the feature selection and feature extraction.

Gradient Boost & Adaboost

This algorithm is used when dealing with a lot of data to make a forecast with a high power of prediction. Boosting is an ensemble learning algorithm that incorporates multi-base estimator prediction to improve robustness over a single estimator. It works well in data science competitions like Kaggle, AV hackathon, CrowAnalytix.

Similar Applications

Regarding to email spam, there are lots of researches, studies, surveys and projects that have been carried out by researchers and students from different universities. Most of them describe types of email spam in general and how to mitigate against them using different methods. They also examine different algorithms to classify incoming emails as spam or legitimate. Moreover, there are relevant applications that have been written such as SMS spam filtering, Detecting spam at the Network level, Mobile phone spam scams, etc.

"Email Classification using Naïve Bayesian Classifier" is the name of journal article from International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). This article illustrates the Naïve Bayesian Classifier method which is used to classify spam and non-spam mails. In this experiment, researchers used the Lingspam dataset that contains total 960 mails in which 700 train dataset and 260 test dataset. They used the word-count algorithm for extracting features from the dataset. "In this algorithm we pre-

process the dataset and remove the stop-words and non-words in dataset. And then it counts the total number of unique word out of the total word and finds the frequency of that word in a particular document." [8]. As a result out of 700 train dataset the 350 are spam mails and 350 are non-spam mails. Additionally, there are 130 spam mails and 130 non-spam mails in the 260 test dataset. They noticed that the Naïve Bayesian Classifier classifies word in a correct way and also it produces a better result when the number of dataset is increased compared to the Support Vector Machine.

Another journal article, " Symbiotic filtering for spam email detection", explains another spam filtering technique which is Symbiotic Filtering(SF). This technique aggregates separate local filters from multiple users to improve the overall spam detection efficiency. Additionally, it allows users in social networks with same or related interests have the opportunity to form mutually beneficial partnerships in order to enhance the techniques of spam detection. In this article, a realistic mixture of real spam and ham messages from Enron data were collected to be used in the Symbiotic filtering method and also to measure the effectiveness of SF and Content-Based Filtering. As a result SF has been shown to be more robust to word attacks such as dictionary and focused assaults.

Backend Technologies

In this project, the supervised machine learning will be used to classify email messages as spam or non-spam using Enron datasets. The Naïve Bayes classifier will be applied because it has faster performance, making use of algorithm in different classification fields very popular and it has been incorporated successfully in other machine learning approaches. Since the theorem of Bayes is based on "probabilistic classifiers", the probability equation would be:

$$\Pr(S|w) = \frac{\Pr(w|S)\cdot\Pr(S)}{\Pr(w|S)\cdot\Pr(S)+\Pr(w|\overline{S})\cdot\Pr(\overline{S})}$$

Figure 3- probability equation[9]

Where Pr(***S***) is the aforementioned likelihood to be set to the predicted spam ratio, $\Pr(w|S)$ $\Pr(w|\overline{S})$ are simply quantify the frequency of each word in spam and non-spam emails in the training data. Pr(***S***|w) is called the posterior likelihood, which can be determined using the previous probability of spam and the probability of a given word appearing in spam and non-spam emails. The classifier is trained using some data to determine the probabilities of these words, which can also be changed if a client considers a new email to be spam or vice versa. For implementation, Python will be applied as programming language and there are specific Python libraries which will be used in the implantation for this classifier such as pandas, numpy, io, os, CountVectorizer and MultinomialNB from sklearn.

Conclusion

Spam is becoming one of the most distracting problem in In the world. Current spam filter code can not accommodate large amounts of spam slipping past anti-spam defences. When spam issues intensify, they require effective and efficient resources to manage them. Machine learning techniques have provided a better way for researches to counter spam. In addition, the pre-processing steps added to the training and testing of vector functionality can improve the performance of the spam filter to decrease the amounts of spam. As mentioned in this paper various algorithms can be applied to mitigate against spam mails. However, Naïve

Bayes is the best algorithm that can be used in supervised machine learning as it produce better result and the error rate is very.

Bibliography

[1] Li, Y.,Fang, B. X., Guo, L., Tian, Z. H., Zhang, Y. Z., & Wu, Z. G. UBSF : A Novel Online URL-Based Filter,IEEE Symposium on (pp. 332-339).ISCC 2008

[2] Alexy Bhowmick, Shyamanta M. Hazarika.(January, 2018) *Machine Learning for E-mail Spam Filtering: Review,Techniques and Trends*. Tezpur University, Indian Institute of Technology Guwahati

[3] Subramaniam, Thamarai, Jalab, Hamid A, Taqa, Alaa Y.(2010) *Overview of textual anti-spam filtering techniques*. Computer System and Technology, Faulty of Computer Science and Information Technology, University Malaya, Malaysia.

[4] Mark Moraes. (1999). *A Primer on How to Work With the Usenet Community*. [online] Available at: http://faqs.cs.uu.nl/na-dir/usenet/primer/part1.html [Accessed 22 Oct. 2019].

[5] de Freitas, Sara Levene, Mark. (2004) *Spam on the Internet: can it be eradicated or is it here to stay?*

[6] Guzella, Thiago S, Caminhas, Walmir M.(2009) *A review of machine learning approaches to Spam filtering*. Department of Electrical Engineering, Federal University of Minas Gerais

[7] Hoffer, Eric.(2005) *Machine Learning*.(Seoul)

[8] Sao, Priyanka, Prashanthi, Pro Kare.(2015) *E-mail Spam Classification Using Naïve Bayesian Classifier*, College of Engineerin & Technology Bhilai (C.G.)

[9] Anon, (2012). [online] Available at: https://www.quora.com/Which-algorithms-are-best-to-use-for-spam-filtering-How-should-they-be-implemented [Accessed 30 Oct. 2019].

[10] Mehrotra, Kunal, Watave, Shailendra.(2009) *Spam Detection : A Bayesian Approach to Filtering Spam*

[11] Sable, Amar V, Gulhane, Prof Vijay S, Ntroduction, I.(2013) *Email Filter for Spam Mail : A Review*

[12] Symantec.(2017) *Email Threats 2017 An ISTR Special Report Analyst: Ben Nahorney Internet Security Threat Report*

[13] Androutsopoulos, Ion, Paliouras, Georgios, Michelakis, Eirinaios.(2004*) Learning to Filter Unsolicited Commercial E-Mail*

[14] Abdi, Asad.(2016) *Three types of Machine Learning Algorithms List of Common Machine Learning Algorithms*

[15] Lopes, Clotilde, Cortez, Paulo, Sousa, Pedro, Rocha, Miguel, Rio, Miguel.(2011) *Symbiotic filtering for spam email detection*

[16] n.a.(2014) *The History of Spam*

[17] Dada, Emmanuel Gbenga, Bassi, Joseph Stephen, Chiroma, Haruna,Abdulhamid, Shafi'I, MuhammadAdetunmbi, Adebayo Olusola, Ajibuwa, Opeyemi Emmanuel.(2019) *Machine learning for email spam filtering: review, approaches and open research problems*

[18] No, Issn, Revar, Pooja, Shah, Arpita, Patel, Jitali, Khanpara, Pimal.(2017) *Available Online at www.ijarcs.info International Journal of Advanced Research in Computer Science A Review on Different types of Spam Filtering Techniques*

[19] Mimecast.com. *Prevent Snowshoe Spam | Mimecast*. [online] Available at: https://www.mimecast.com/content/snowshoe-spam/ [Accessed 30 Oct 2019].

[20] SearchSecurity.. *What is email spam? - Definition from WhatIs.com*. [online] Available at: https://searchsecurity.techtarget.com/definition/spam [Accessed 20 Oct. 2019].

[21] Avira.com. *What is Email spam*. [online] Available at: https://www.avira.com/en/support-what-is-email-spam [Accessed 20 Oct. 2019].
[22] Group-mail.com.*What is Spam?*. [online] Available at: https://group-mail.com/email-marketing/what-is-spam/ [Accessed 20 Oct. 2019].
[23] WhatIs.com. *What is command-and-control server (C&C server)? - Definition from WhatIs.com*. [online] Available at: https://whatis.techtarget.com/definition/command-and-control-server-CC-server [Accessed 20 Oct. 2019].
[24] anti Spam Filter Software - Virtual Appliance for Enterprise and ISP. *Purpose Of Blank Spam Emails Remains Unclear*. [online] Available at: https://www.mailcleaner.net/blog/spam-world-news/purpose-of-blank-spam-emails-remains-unclear/ [Accessed 21 Oct. 2019].
[25] echsoupcanada.ca. *Ten Spam-Filtering Methods Explained | TechSoup Canada*. [online] Available at: https://www.techsoupcanada.ca/en/learning_center/10_sfm_explained [Accessed 21 Oct. 2019].
[26] Anon.[online] Available at: https://www.researchgate.net/publication/320703241_E-Mail_Spam_Filtering_A_Review_of_Techniques_and_Trends/link/5ae9992945851588dd8211da/download [Accessed 24 Oct. 2019].
[27] Aski, A. and Sourati, N. *Proposed efficient algorithm to filter spam using machine learning techniques*. [Accessed 24 Oct. 2019].
[28] Mimecast.com. *Prevent Snowshoe Spam | Mimecast*. [online] Available at: https://www.mimecast.com/content/snowshoe-spam/ [Accessed 25 Oct. 2019].
[29]OrganicWeb. *Why text on images sends your email to spam*. [online] Available at: https://organicweb.com.au/marketing/image-spam/ [Accessed 28 Oct. 2019].
[30] DataRobot *What Is Unsupervised Machine Learning? | DataRobot*. [online] Available at: https://www.datarobot.com/wiki/unsupervised-machine-learning/ [Accessed 1 Nov. 2019].