International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 04 | April -2020 ISSN: 2582-3930

A NEW SEMANTIC-BASED FEATURE SELECTION WITH DEEP LEARNING METHOD FOR SPAM FILTERING

PAVITHRA R[1]

Asso.Prof.Mr. J.JAYAPANDIAN

Krishnasamy College of Engineering and Technology, Cuddalore.

Abstract: The Internet emerged as a powerful infrastructure for the worldwide communication and interaction of people. Some unethical uses of this technology (for instance spam or viruses) generated challenges in the development of mechanisms to guarantee an affordable and secure experience concerning its usage. This study deals with the massive delivery of unwanted content or advertising campaigns without the accordance of target users (alsoknown as spam). Currently, words(tokens) are selected by using feature selection schemes; they are then used to create feature vectors for training different Deep Learning (ML) approaches. This study introduces a new feature selection method able to take advantage of a semantic ontology to group words into topics and use them to build feature vectors. To this end, we have compared the performance of nine well-known Machine Learning approaches. Results have shown the suitability and additional benefits of topic-driven methods to develop and deploy high-performance spam filters.

1.Introduction

According to report from Kaspersky lab, in 2015, the volume of spam emails being sent reduced to a 12-year low. Spam email volume fell below 50% for the first time since 2003. In June 2015, the volume of spam emails went down to 49.7% and in July 2015 the figures was further reduced to 46.4% according to anti-virus software developer Symantec. This decline was attributed to reduction in the number of major botnets responsible for sending spam emails in billions. Malicious spam email volume was reported to be constant in 2015. The figure of spam mails detected by Kaspersky Lab in 2015 was between 3 million and 6 million. Conversely, as the year was about to end, spam email volume escalated. Further report from Kaspersky Lab indicated that spam email messageshaving pernicious attachmentssuch as malware, ransomware, malicious macros, and JavaScript started to increase in December 2015. That drift was sustained in 2016 and by March of that year spam email volume had quadrupled with respect to that witnessed in 2015. In March 2016, the volume of spam emails discovered by Kaspersky Lab is 22,890,956. By that time the volume of spam emails had skyrocketed to an average of 56.92% for the first quarter of 2016. Latest statistics shows that spam messages accountedfor 56.87% of e-mail traffic worldwide and the most familiar types of spam emails were healthcare and dating spam. Spam results into unproductive use of resources on Simple Mail Transfer Protocol (SMTP) servers since they have to process a substantial volume of unsolicited emails. The volume of spam emails containing malware and other malicious codes between the fourth quarter of 2016 and first quarter of 2018 is depicted .To effectively handle the threat posed by email spams, leading email providers such as Gmail, Yahoo mail and Outlook have employed the combination of different machine learning (ML) techniques such as Neural Networks in its spam filters.

Content Based Filtering Technique: Content based filtering is usually used to create automatic filtering rules and to classify emails using machine learning approaches, such as Naïve Bayesian classification, Support Vector Machine, K Nearest Neighbor, Neural Networks. This method normally analyses words, the occurrence, and distributions of words and phrases in the content of emails and used then use generated rules to filter the incoming email spams.

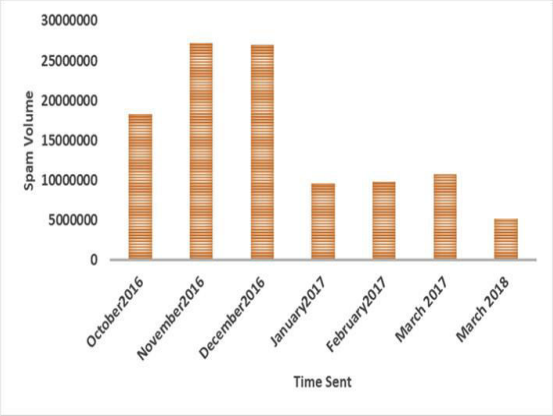
Previous Likeness Based Spam Filtering Technique: This approach uses memory-based, or instance-based, machine learning methods to classify incoming emails based to their resemblance to stored examples (e.g. training emails). The attributes of the email are used to create a multi-dimensional space vector, whichis used to plot new instances as points. The new instances are afterward allocated to the most popular class of its K-closest training instances.

Case Base Spam Filtering Method: Case base or sample base filtering is one of the popular spam

filtering methods. Firstly, all emails both non-spam and spam emails are extracted from each user's email using collection model. Subsequently, pre-processing steps are carried out to transform the email using client interface, feature extraction, and selection, grouping of email data, and evaluating the process. The data is then classified into two vector sets.

Heuristic or Rule Based Spam Filtering Technique: This approach uses already created rules or heuristics to assess a huge number of patterns which are usually regular expressions against a chosen message. Several similar patterns increase the score of a message. In contrast, it deducts from the score if any of the patterns did not correspond. Any message's score that surpasses a specific threshold is filtered as spam; else it is counted as valid. While some ranking rules do not change over time, other rules require constantupdating to be able to cope effectively.

Fig 1: The volume of spam emails 4th quarter 2016 to 1st quarter 2018.



2. Related work

There is a rapid increase in the interest being shown by the global research community on email spam filtering. In this section, we present similar reviews that have been presented in the literature in this domain. This method is followed so as toarticulate theissues that are yet to be addressed and to highlight the differences with our current review presented a brief survey to explore the gaps in whether information filtering and information retrieval technologybe operational in an efficient way. However, the survey did not present the details of the Machine learning algorithms, the simulation tools, the publically available datasets and the architecture of the email spam environment. It also fails short of presenting the parameters used by previous researches in evaluating other proposed techniques.

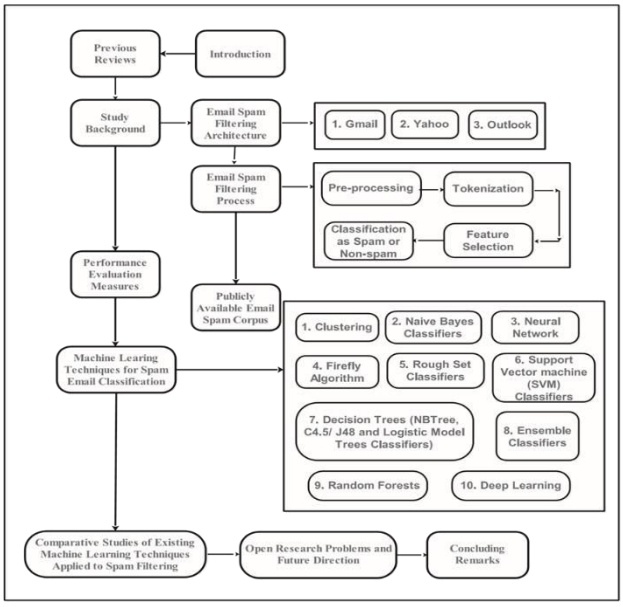


Fig 2: Pictorial Representation of the Structure of this paper.

3.Background: Here we discussed the architecture of email server and the stages in processing email. We explained the different stages involved in pre-processing and feature selection.

3.1. Email spam filtering architecture: Spam filtering is aimed at reducing to the barest minimum the volume of unsolicited emails. Email filtering is the processing of emails to rearrange it in accordance to some definite standards. Mail filters are generally used to manage incoming mails, filter spam emails, detect and eliminate mails that contain any malicious codes such as virus, Trojan or malware.

Spam filters are deployed by many Internet Service Providers (ISPs) at every layer of the network, in front of email server or at mail relay where there is the presence of firewall.

3.2. **How Gmail, Yahoo and Outlook emails spam filters work**:Different spam filtering formulas have been employed by Gmail, Outlook.com and Yahoo Mail to deliver only the valid emails to their users and filter out the illegitimate messages. Conversely, these filters also sometimes erroneously block authentic messages. It has been reported that about 20 percent of authorization based emails usually fail to get to the inbox of the expected recipient.The mechanisms are used to decide the risk level of each incoming email. Examples of such mechanisms include satisfactory spam limits, sender policy frameworks, whitelists and blacklists, and recipient verification tools.

3.3**Gmail filter spam**. Google's data centre's makes use of hundreds of rules to determine whether an email is valid or spam. Every one of these rules depicts specific features of a spam and certain statistical value is connected with it, depending on the likelihood that the feature is a spam. The weighted importance of each feature is then used to construct an equation.

3.4 **Yahoo mail filter spam**. Yahoo mail is the first free webmail providers in the world with over 320 million users. The email provider has its own spam algorithms that it uses to detect spam messages. The basic methods used by Yahoo to detect spam messages include: URL filtering, email content and spam complaints from users. Unlike Gmail, Yahoo filter emails messages by domains and not IP address.

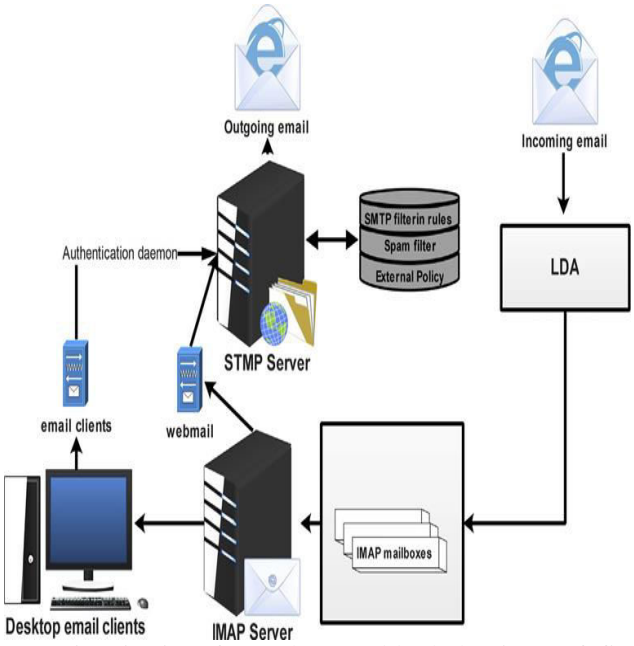
3.5 **Outlook email spam filter.** After Gmail and Yahoo mail, we discussed Outlook from Microsoft in this section and how it handles spam filtering. In 2013, Microsoft changed the name of Hotmail and Windows Live Mail to Outlook.com. Outlook.com was patterned after Microsoft's Metro design language and directly imitates the interface of Microsoft Outlook. Outlook.com is a collection of applications from Microsoft, one of which is Outlook webmail service.

4.**Email spam filtering process:**An email message is made up of two major components which are the header and the body. The header is the area that have broad information about the content of the email. It includes the subject, sender and receiver. The body is the heart of the email. It can include information that does not have a pre-defined data. Examples include web page, audio, video, analog data, images, files, and HTML mark up processing before the classifier can make use of it for filtering below depicts a mail server architecture and how spam filtering is done.

Fig.3: Email server spam filtering architecture.

4.1 **Firefly algorithm:**

The firefly algorithm (FA) is a population based metaheuristic algorithm proposed by. He got his



inspiration from the sparkly behaviour of fireflies. The algorithm preserves and increase several candidate solutions by means of population physiognomies to direct the search [69]. The design of the algorithm was founded on the study of the concept of communication among fireflies at the time they are getting ready to copulate, and immediately they are exposed to danger.Therefore, a sparkling light exuding from a firefly gets a response from fireflies around it within a visual range of the flash.

Algorithm: Email spam classification algorithm using Rough Set

1. Input Email Testing Dataset (Dis_ testing dataset), Rule (RUL), b

2. for x 2 Dis T E do

3. while RUL (x) ¼ 0 do

4. suspicious ¼ suspicious [{x};

5. end while

6. Let all r 2 RUL (x) cast a number in favor of the non-spam class

7. Predict membership degree based on the decision rules;

8. R ¼ r 2 RUL (x)jr predicts non-spam;

9. Estimate Rel (Dis_T E j x 2 non-spam);

10. Rel (Dis_T E j x 2 non-spam) ¼ Pr 2 R Predicts (non-spam)

11. Certaintyx ¼ 1/cer Rel (Dis_T E j x 2 non-spam);

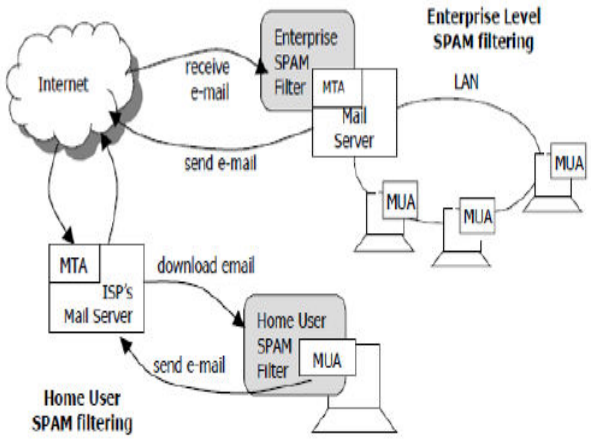
12. while Certaintyx1 – b do

13. suspicious ¼ suspicious [{x};

14.end

15.spamreturn¼Final Email Message Classispam [{x}; fication (Spam/Non-spam/Suspicious email)

16. end



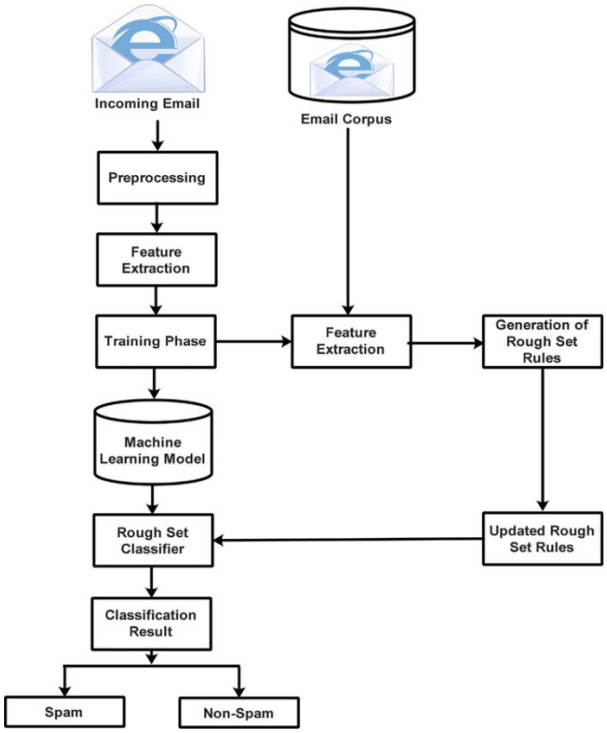


Fig 4 : Architecture of neural network (NN) Classifier.

5.What is Spam?

Spam is unsolicited and unwanted email from a stranger that is sent in bulk to large mailing lists, usually with some commercial objective.1 Some would argue that this definition should be restricted to situations where the receiver is not especially selected to receive the email – this would exclude emails looking for employment or positions as research students for instance. Spam is junk email; junk postal mail and junk faxes are also a problem. However,because of the special nature of the Internet, there are two reasons why junk email is a particular problem.

5.1Spam Filtering: Spam filtering in Internet email can operate at two levels, an individual user level or an enterprise level (see Figure 1). An individual user is typically a person working at home and sending and receiving email via an ISP. Such a user who wishes to identify and filter spam email installs a spam filtering system on her individual PC. This system will either interface directly with their existing mail user agent (MUA) (more generally known as the mail reader) or more typically will act as a MUA itself withfull functionality for

composing and receiving email and for managing mailboxes.

Fig 5: Alternatives for spam filtering in Internet e-mail.

5.2CorpusPre-processing: Not all information present in an e-mail is necessary or useful. Eliminating the less informative and noisy terms lowers the feature space dimensionality and enhances classification performance in most cases [Guzella and Caminhas, 2009], [et al, 2003] and [Shi et al, 2012]. Corpus pre-processing is a process that involves transforming the mail corpus into a uniform format that is more comprehensible to the machine learning algorithms [Zhang et al, 2004], [Katakis et al, 2007]. Due to the adversarial nature of spam, spam filters need to constantly adapt to changing spam tactics, particularly in feature extraction and feature selection aspects.

5.3Lexical Analysis (Tokenization): The string of text representing a message is tokenized in order to identify the candidate words to be adopted as relevant spam or ham terms. Headers, attachments, and HTML tags are stripped, leaving behind just the e-mail body and subject line text.

Stop-word Removal:Stop-word removal involves removing frequently used non-informative words, e.g. 'a', 'an', 'the', and 'is', etc. Obscure texts or symbols may also be removed in subsequent steps.

Stemming: Word-stemming is a term used to describe a process of converting words totheir morphological base forms, mainly eliminating plurals, tenses, gerund forms, prefixes and suffixes.

Representation: Involves the conversion of an email message into a specific or structured format as needed by the machine learning algorithm being employed. [Androutsopoulos et al, 2000a] studied the effect of corpus size, lemmatization, and stop-lists while in [Androutsopoulos et al, 2000c].

6.**Methods for Mitigating E-mail Spam:**Although there are 'social' methods like legal measures and personal measures (e.g. never respond to spam, never forward chain-letters) to fight spam, they have had a narrow effect on spam so far is seen by the number of spam messages received daily by users. Technical measures seem to be the most effective in countering spam. Prior to machine learning techniques, many different technical measures were employed for spam filtering, like - rule-based spam filtering, white lists, black lists, challenge-response (C/R) systems, spam filtering, honey pots, OCR filters, and many others, each with its own merits and drawbacks.

6.1Heuristic Filters: Initial spam filters followed the 'knowledge engineering' approach and were based on coded rules or heuristics Sanz [2008]. A content-based heuristic filter analyses the contents of a message M and classifies it to spam or ham based on the occurrence of 'spammy' words like 'viagra' or 'lottery' in it. They were designed based on the knowledge of regularities or patterns observed in messages Guzella and Caminhas [2009]. Cohen's Cohen [1996] was one of the earliest attempts to use learning machines that classify e-mail.

6.2 Blacklisting: A blacklist of E-mail addresses or IP addresses of the server from which spam is found to originate is created and maintained either at the user or server level. If a user receives an e-mail from any of these addresses, the message is automatically blocked at the SMTP connection phase. This method requires only a simple lookup in the blacklist every time.

6.3 **Whitelisting:** Whitelisting is the reverse of blacklisting. An e-mail whitelist is a list of pre-approved or trusted contacts, domains, or IP addresses that are able to communicate to a mail user. All e-mails from fresh e-mail addresses are blocked by this method. This restrictive method may introduce an extremely high false positive rate instead of reducing it. Such a method may be good for instant messaging environments but is not a good choice as it prohibits establishing new contacts through e-mail.

7. **Machine Learning Approach to E-mail Spam filtering:** The Algorithms: Spam filtering is a binary classification task, in which legitimate (good or ham) e-mails are treated as negative (-) instances, and spam as positive (+) instances [Song et al, 2009]. Machine Learning is a subfield of computer science that explores the design and development of computer systems that automatically improve their performance in a task based on experience. Automatic e-mail classification uses statistical approaches or machine learning techniques and aims at building a model or a classifier specifically for the task of filtering spam from a user's mail stream.

8**. Spam Filter Inputs and Outputs**:We have defined a spam filter to be an automated technique to

identify spam. A spam filter with perfect knowledge might base its decision on the content of the message, characteristics of the sender and the target, knowledge as to whether the target or others consider similar messages to be spam, or the sender to be a spammer, and so on. But perfect knowledge does not exist and it is therefore necessary to constrain the filter to use well defined information sources such as the content of the message itself, hand-crafted rules either embedded in the filter or acquired from an external source, or statistical information derived from feedback to the filter or from external repositories compiled by third parties.

8.1 **Typical Email Spam Filter Deployment:**The typical use of an email spam filter from the perspective of a single user. Incoming messages are processed by the filter one at a time and classified as ham (a widely used colloquial term for non-spam) or spam. Ham is directed to the user's inbox which is read regularly. Spam is directed to a quarantine file which is irregularly (or never) read but may be searched in an attempt to find ham messages which the filter has misclassified. If the user discovers filter errors either spam in the inbox or ham in the quarantine he or she may report these errors to the filter, particularly if doing so is easy and he or she feels that doing so will improve filter performance.

9. **Conclusion**: In this paper, we reviewed machine learning approaches and their application to the field of spam filtering. A review of the state of the art algorithms been applied for classification of messages as either spam or ham is provided. The attempts made by different researchers to solving the problem of spam through the use of machine learning classifiers was discussed. The evolution of spam messages over the years to evade filters was examined. The basic architecture of email spam filter and the processes involved in filtering spam emails were looked into. The paper surveyed some of the publicly available datasets and performance metrics that can be used to measure the effectiveness of any spam filter.

10.References:

[1] M. Awad, M. Foqaha, Email spam classification using hybrid approach of RBF neural network and particle swarm optimization, Int. J. Netw. Secur. Appl. 8 (4) (2016).

[2] Visited on May 15, 2017, Kaspersky Lab Spam Report, 2017, 2012, https://www. securelist.com/en/analysis/204792230/Spam_Report _April_2012.

[3]E.M. Bahgat, S. Rady, W. Gad, An e-mail filtering approach using classification techniques, in: The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Springer International Publishing, BeniSuef, Egypt, 2016, pp. 321–331.

[4]. "You might be an anti-spam kook if...," http://www.rhyolite.com/antispam/you-might-be.html.

[5] A. J. Alberg, J. W. Park, B. W. Hager, M. V. Brock, and M. Diener-West,

"The use of overall accuracy to evaluate the validity of screening or diagnostic

tests," Journal of General Internal Medicine, vol. 19, no. 1, 2004.

[6] I. Androutsopoulos, E. F. Magirou, and D. K. Vassilakis, "A game theoretic

model of spam e-mailing," in CEAS 2005 — The Second Conference on Email

and Anti-Spam, 2005.

[7] Androutsopoulos, I., Koutsias, J., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C., &

Stamatopoulos, P., (2000) Learning to filter spam e-mail: A comparison of a naive bayesian

and a memory-based approach. in Workshop on Machine Learning and Textual Information

Access, at 4th European Conference on Principles and Practice of Knowledge Discovery inDatabases (PKDD).