Theory and methods of email classification

Sociology, Communication

Although nowadays, we are using instant message, mobile application and social network for communication but email is consider as one of the most important mode of communication in today’s world. Usage of emails has increased substantially for both personal as well as professional use. Email is convenient to send complex information including images, videos, documents, URLs etc. With this increasing email volume, it is essential to automate identification of spam emails.

The aim of this paper is to highlight the important techniques and methodologies employed in spam email identification. This paper provides a review of theory and methods of email classification, focusing on existing literature.

Email is the most important and effective way of communication for professional as well as personal communication. Usage of email is increasing rapidly around the world. In 2017, the total number of emails sent and received per day is 269 billion. This figure expected to continue to grow with average annual rate of 4. 4% for next four years, and will reach 319. 6 billion by the end of 2021. According to statistics, spam emails accounted for 48. 16% of total email traffic worldwide. Email user’s fruitful time is getting wasted to filter spam emails. Thus, email classification is very essentials to detect whether incoming email is legitimate or spam.

Plethora of uninvited and irrelevant emails sent every day, many of these mails are auto-generated by robots. A considerable amount of productive time of employees is get wasted to deal with these spam emails. According to study report by Kaspersky Lab Inc., average user wastes approximately 5

to 6 hrs per month to handle spam emails. These spam emails are sent for purpose of advertising, phishing, ransomware, malicious websites or apps, rootkits, spreading backdoors or malicious programs. Many users will be victimise by these spam emails. As users are using email for security to bank accounts, social networking accounts etc., so attacker can attack on these accounts linked with email account.

According to McAfee research, the transmission and receipt of unwanted email gobbles up 33bn kilowatt-hours of electricity a year, which is equivalent of the electricity used by 2. 1m US homes. Spam takes up memory space on mail servers, which is wastage of storage space and incur additional cost to company or user or service provider.

Thus, spam is waste of time, waste of storage, waste of energy, waste of communication bandwidth and thread to user’s privacy.

Due to these issues, spam email identification is challenge for researchers. Many researchers published the solutions to email classification using different machine learning techniques like Support Vector Machine, Naïve Bayes, Hidden Markov Model, Decision Tree etc. and non-machine learning techniques like Black/White List, Mail Header Checking Heuristic, Signatures etc. In this paper, we are observing various approaches to classify email as spam or ham. Our intention is to motivate other researchers to identify opportunities and gap in literature.

To classify incoming email as spam or not spam, different techniques are used some are based on rules and some on content of email. In case of Rules

based technique, a set of rules are used to classify email as spam or ham. For example, domain or IP from which mail is received is blacklisted or not. Gomes[4] characterisation found significant differences in spam and not-spam traffic patterns. They have examine email traffic attributes like email arrival process, size of email, number of recipient per email, popularity and temporal locality among the recipients. Blacklisting and whitelisting are another simple filtering techniques for email classification. Blacklisting also called block listing filters out emails received from specific senders. Whitelist also called allow lists do exactly opposite, automatically allowing emails from specific senders only. Blacklist or whitelist is prepared at server level or user level.

In content based email classification, which uses different technique for classification using different email fields such as To, From, Cc/Bcc, Message-ID in email header. Many of content based email classification techniques takes the email body into consideration with keyword extraction, keyword frequency and punctuations. Firte et al. extracted features like number of recipients, number of replies, number of attachments, subject length, message size, number of links and sources along with 50 top frequency words used in spam and applies KNN Algorithm and Resampling technique for spam detection. The proposed system by Firte et al. constantly updates dataset and top frequency words in spam messages.

Graph mining approach converts emails into graphs and then substructures of graph are used for matching. Manu et al proposed eMailSift System for email classification based on graph mining, and found that accuracy get

increased as increase in size of folder. However, according to later work by Sharma et al show that email classification accuracy decreased as number of folders increased.

Support Vector Machine (SVM) are learning models associated with learning algorithms which analyses data and recognize patterns, used to classify new sample. Drucker et al compare SVM with Ripper, Rocchio, and boosting decision trees algorithms. They experimented with two datasets, and reported that SVM performs best over binary features and require significantly less training time over other three algorithms.

Research Work

This section consists of analysis of main papers with their email classification approaches to solve spam identification problem. In Table 1 we have given parameter wise comparison of research papers. The papers, which we are going to consider for discussion, are as follow:

A. A Proposed Data Science Approach for Email Spam Classification using Machine Learning Techniques

Proposed model in this paper uses machine learning for email classification by using parameters from email header such as To field, From field, Cc/Bcc etc. and email body. Commonly used keywords and punctuations are extracted from text in email body. Proposed architecture consist of four stages i. e. preparation of data, data analysis, assessment and deployment. To create dataset, emails are retrieved from mail servers which are then scanned to retrieve parameters in preparation of data stage. This dataset is

then analyze in data analysis stage, and machine algorithm model is get trained to recognize patterns. Results of training are analyse to find outliers which divides dataset into two types of outputs i. e. ambiguous and proper outputs. Proper outputs are last time get assessed in assessment stage, where dataset and real time results get compared to improve accuracy. Lastly, system is deployed in deployment stage.

This paper prepares dataset independently for each user, as an email that is spam for one user may not be a spam for other in rare cases. As paper is based on machine learning model, accuracy of algorithm is based on number of samples in dataset in training phase. And as size of dataset is different for different user, accuracy will be different, hence accuracy value is not shown in table 1 for this paper. A proposed solution is self-learning system, customised per user. Accuracy of proposed system depends on number of emails available in user’s email account for training.

B. A Comparative Approach to Email Classification Using Naïve Bayes Classifier and Hidden Markov Model

Gomes et al. pre-process the dataset before applying classification filter to dataset. If emails contain HTML, then first of all text is extracted from such emails. NLP techniques like stemming, lemmatizing and stop words removal techniques are used to remove irrelevant data from emails in dataset. In this paper two classifiers are used i. e. Naïve Bayes classifier and Hidden Markov Model. In Naïve Bayes Classifier, a list of most frequent words from spam and important emails is prepared called word_features. Existence of these words is checked in email, if the word is exists then word: label is pair is maintain

where label is important or spam. This collection of word: label pair is called featureset. Then two classify email as important or spam, importantwords and spamwords are counted from this featureset. In this case, email can not be get classified if both importantwords and spamwords are equal. Out of the different combinations of NLP techniques, a combination of removing stopwords and lemmatizing gives most accurate that is 79. 19% results.

In case of Hidden Markov Model, an emission probability set is maintain which contain probability of occurrence in spam or important email for words in word_feature list. HMM also cannot determine email category if important words and spam words are equal. In HMM also, author execute different combination of NLP techniques. HMM with stemming give more accuracy i. e. 91. 28%.

HMM out performs well compared to Naïve Bayes Classifier in identifying spam and important emails.

C. Spam Filtering Email Classification (SFECM) using Gain and Graph Mining Algorithm

Proposed solution in this paper works in three stages: Email Preprocessing, Feature Extraction and Email Classification. In email preprocessing, POS tagger converts email text into email features. Feature extraction stage, filters spam emails and removes sign-off words, keywords, greetings etc. Then graphs are generated from emails. In last stage these template graphs are compared with new email graph to classify as spam or important.

Compared to existing solution, proposed solution in this paper gives absolutely 100% accuracy and no significant difference in processing time.

D. Statistical-based Bayesian Algorithm for Effective Email Classification

Bayes theorem is used to calculate the probability of an email to be spam. Initially probability of email to be spam and probability of email to be legal is considered as 0. 5 in Bayes Theorem. While proposed algorithm calculates actual priori probability of spam, as the proportion of legitimate messages and spam is extremely high or low. Divide each mail from spam corpus and legal corpus into token set and create respective hash-spam table and hash-legal table. Then map the token to occurrences in hash tables. Proposed system improved the selection rules of tokens. Like to avoid evade detection, they introduce phrase into token. Deleted rarely used tokens. Proposed system support Chinese spam detection, by extracting tokens from emails in Chinese language using segmentation methods. Hash probability table is maintained which is mapping of token ti to probability of spam contains token ti. To check incoming email is spam or not, calculate probability of email is spam. This calculated probability is compared with threshold value, and if it exceeds then email is declared as spam.

This paper also considered spam detection based on blacklisted URLs and Image spams. Spams with blacklisted URLs are detected by using behaviour-based blacklist filtering. To evade the text based spam filters, spammers use image spam, in which text of message is saved as JPEG, GIF or PNG image. OCR scanning, source identification and fingerprint algorithm are used to deal with image spams. The accuracy of proposed algorithm in this paper is 99. 2% which is better than existing solution.

E. Improved Email Classification Method Using Integrated Particle Swarm Optimization and Decision Tree

Kaur et al. found that use of unsupervised filtering is ignored by most of researchers and many are focusing on limited features of email for spam detection. UCI dataset is pre-processed for data normalization as quality of result depends on quality of data. Particle Swarm Optimization is used, which search for optimal solution on collection of random particles. After every iteration particles update their best and global value. After finding these two values, particles update its velocity and positions.

While experimenting author used 50% of database for training the system and 50% for testing purpose. PSO and J8 are applied on dataset to get results. This result then compared with k-means and support vector machine with and without unsupervised filter. According to analysis, proposed system gives more accurate result i. e. 98. 32%.

F. An Empirical Study on Email Classification Using Supervised Machine Learning in Real Environment

Instead of using existing dataset, Wenjuan et al. collected emails from real users that to from different environment i. e. Research Institute, Academic University and Commercial Company. Among different Supervised Machine Learning algorithm for classification Naïve Bayes, k-nearest neighbour (KNN), Neural Network and Support Vector Machine (SVM) are selected for experiment. 14 features are finalized as per author’s previous study.

While analysing the spam traffic, it is found that different environment have different spam traffic. Spam traffic depends on how network administrator applies strict rules and how many services are subscribed by email account.

Company environment has lowest spam traffic (28. 5%) while University environment has highest spam traffic (58. 3%). Experiment is conducted in two phases. In first phase, random 60% samples are used for training and remaining for testing. It is notice that SML classifiers perform differently with environment. Decision tree and support vector machine achieve better accuracy compared to other three classifiers. Also, company environment gives better accuracy and university environment gives worst accuracy for all classifiers.

In the second phase of experiment, all dataset samples are used to train classification model which is then tested in real environment for two weeks. It is observe that accuracy is increased than phase1 result, as size of training dataset is more. Observation are same as phase1 result that is, decision tree and support vector machine performs well than other three classifiers and company environment give better accuracy. Maximum accuracy achieved is 95. 7%.

Conclusion

By analyzing above papers we come with following conclusions

- The accuracy of classification algorithm depends on quality, size of dataset while training.
- Unsupervised learning algorithms are used by few researchers to classify emails.
- The environment factor significantly affects the accuracy of classification model.

- Classification model must be tested in real environment rather than portion of dataset, as real environment gives more diverse and complex spam samples.
- Involvement of users while testing classification model.