ISSN 2320-9984
Volume 7 Issue 4 | December 2019 ISSN: 2320-9984 (Online)
International Journal of Modern Engineering & Management Research
Website: www.ijmemr.org

Email Spam Detection using Text Classification

Hrithik Sanyal
Research Scholar
Department of Electronics & Telecommunication,
Bharati Vidyapeeth College of Engineering,
Pune (M.S.) India
Email: hrithiksanyal14@gmail.com

Rajneesh Agrawal
\Mentor,
Comp-Tel Consultancy,
Jabalpur (M.P.) India
Email: rajneeshag@gmail.com

Abstract— *Spam has been a major problem in any online system and is available in different formats. Majorly spam creates a lot of problems in the email system. Since spam is unexpected and unwanted and is sent by the spammers to promote, hack or send malicious contents to the recipients, therefore, they create major problems such as wastage of network resources, wastage of time, damage of PC's and laptops due to viruses security breach, mail quota problems, irritations to the recipients, ethical issues etc. Identifying and filtering SPAM mails from all the emails of the users has become a real problem as SPAM mails cause several problems for the users. SPAM mails recognition is a problem of data mining as a user may receive hundreds of emails in a day and a few of them are SPAM mails. This work is focused on clustering of Enron Data set for SPAM and NON-SPAM cluster formation from small data set to large data sets. This work finds the possibilities of providing high performance to real-time email service providers.*

Keywords:— *Text Processing, Email, Spam, Clustering, Classification, Ant Clustering*

1. INTRODUCTION

In this digital age, which is the era of electronics & computers, one of the efficient & power modes of communication is the email. Undesired or unsolicited emails can be a headache for the users; however, it may contain security threats as well. Like, a spam mail may contain a link which would intend to capture user's login credentials (identity theft, phishing), or links that would forward to a website which would indeed install malicious software on the computer.

Malware installed could be used to control information of users such as to send spam, host malware, host phish etc. While prevention of spam transmission would be ideal, detection allows users & email providers to address the problem today.

In the Section I, the introduction is detailed. Section II discusses on Machine Learning and Technologies revolving around it, Section III elaborates more on Clustering. In Section IV, Data Pre-Processing is deliberated. In Section V, the Existing Systems are illustrated. The Steps of Algorithms are mentioned in Section VI, Proposed Algorithms in Section VII. Section VIII briefs about the Results & Discussions and Section IX explains the Conclusion & Future Scope.

2. MACHINE LEARNING TECHNOLOGIES

Machine learning is a subfield of artificial intelligence (AI). The machine learning understands the structure of data and puts that data into new structural models that are understandable and useful for people. Machine Learning uses two types of techniques. These techniques are:

- Supervised Learning
- Unsupervised Learning

A. SUPERVISED LEARNING

Supervised learning means a kind of learning which trains a model on known input and output data. This helps in predicting future outputs accurately. On taking and learning from known inputs and outputs, it builds and trains a model that would make predictions based on evidence in the presence of uncertainty. Supervised learning is mostly used for prediction when the output of the data is known. Supervised learning uses classification and regression techniques for building up a predictive model.

Classification — Classification technique is used to predict discrete responses, e.g., whether a tumour is benign or malignant, whether an email is genuine or spam. This technique categorizes the input data into different categories. It is most useful when we can tag, categorize or separate data in classes or groups.

For performing classification some common algorithms are:

- Support Vector Machine (SVM)
- Neural Networks
- Naïve Bayes
- k-Nearest Neighbor
- Boosted Decision Trees
- Discriminant Analysis
- Logistic Regression

Regression — Regression techniques are used to predict responses which are continuous. Algorithms using regression are:

- Linear Model
- Non-Linear Model
- Regularization
- Adaptive Neuro-Fuzzy Learning
- Boosted Decision Trees
- Step-wise Regression

B. UNSUPERVISED LEARNING

Unsupervised learning is a type of machine learning which searches hidden patterns or structures in data. It helps to make inferences from datasets which are consisting of responses which are not tagged or labelled. Unsupervised learning mostly uses Clustering technique.

Clustering — It is the most used unsupervised learning technique. It is used for finding hidden patterns or groupings in datasets and thus analyzing them. For performing classification some common algorithms are:

- K-Means & k-Medoids
- Hierarchical Clustering
- Gaussian mixture models
- hidden Markov models
- fuzzy c-means clustering
- subtractive clustering
- Self-Organizing Maps

Some applications of classification are as follows:

- Gene sequence analysis,
- Market Research
- Object Recognition

3. CLUSTERING

Clustering can be best described as a group of multifaceted data objects which are also pigeon-holed as similar objects or can also be defined as dividing data-sets into multifarious groups using clustering of data based on cluster analysis which is further based on the similarity of the data. One such similar set or group of data is the cluster of data.

Soon after the division and classification of data into multifarious groups, they are all assigned a label; a different and unique label for each group of data-set. This

is very useful when trying to adapt to the changes due to classifications.

Clustering methods are used to identify groups of similar objects in a multivariate data sets collected from fields such as marketing, bio-medical and geo-spatial. They are different types of clustering methods, including:

- Partitioning methods
- Hierarchical clustering
- Density-based clustering
- Model-based clustering

4. DATA PRE-PROCESSING

The basic step done in data pre-processing is stopping and stemming.

The process of removing words that may be short in length, very frequently occurring ones or special symbols, is called Stopping. For other reasons such as grammar-related, documents will use different types of words such as organize, organizes or organizing.

For reduction of derivation forms to a common base, Stemming is mostly used. Both Stemming and Stopping algorithms could be helpful for these purposes. While, stemming removes any kind of words which are based on similar patterns or changed because of tenses, stopping removes words which are frequently being used.

Stopping and stemming are done to reduce the vocabulary size which helps information retrieval and classification purposes. These steps of stopping and stemming are applied to both the training phase and the testing phase of the system.

5. EXISTING SYSTEMS

In this recent era, email remains one of the best and cheapest modes of communication. It is still easy when compared to other modes, and is still used for any kind of official work, be it business, be it corporate, people still prefer emails for sharing official information and records. But people are again misusing this mode of communication by sending over useless and unwanted emails, generally regarded as spam mails. These spam emails do affect the user in some or the other way, like excessive usage and wastage of memory, etc. Thus, through this paper, an integrated machine learning approach based on Naïve Bayes algorithms and Particle Swarm Optimization (PSO) is used for detection of spam. While, Naïve Bayes algorithm is used for learning and classification purpose, which classifies email content spam or not, the PSO is used for distribution, optimization of the Naïve Bayes approach. For this work, Ling Spam dataset is used and then evaluated [13].

In the last decade, we have seen an emergence of the internet in the lives of people. It has become a vital part of everybody's lives. With the advancements and innovations in this field, there has been an increasing emergence of irrelevance as well. Nowadays we see that mobile device usage has increased exponentially which itself has increased the no. of spam messages as well. For example, in one survey it was reported that 96% of Indian received uncalled for text messages every day. This concludes that SMS Spam is any unwanted text in the form of a weblink, or promotional or irrelevant message or texts that are sent to mobile phones regularly. To counter this, a rule-based approach has been employed. According to this, protocols of rules are manually used on texts or messages by the authorities themselves. This method becomes daunting and thus not favoured by many. To overcome this arduous and daunting traditional method, deep learning methods may be used. These are efficient and easier when compared to the traditional approach. Deep learning requires a protocol of dataset samples (training samples) which would learn the rules by reading the SMS texts and then construct a text-based classifier that would efficiently classify SMS messages as spam or not. Through this paper,

a systematic review of deep learning methods employment is done. This paper mainly focuses on convolutional neural network and recurrent neural network on a huge corpus of SMS texts and thus builds a spam classifier that would classify messages as ham or spam. [14].

When we talk about spams, the most chronic problems we face in today's world is spam emails. Not only is it costly but very dangerous in nature, for networks and computer. Although we have seen an emerging scene of social media or internet-based information exchange, email remains one of the best and most trusted forms of communication till date. Albeit many spam detections filters have been developed for helping prevent spam emails, which restricts the spam emails from entering the inbox of the user, nonetheless, there is a lack of innovation or implementation focusing on text alteration. In the current scenario, Naïve Bayes is one of the best methods for spam classification, because it is simple and efficient as well. Although very accurate, but it cannot rightly classify emails as spam when they possess leetspeak or diacritics. So, with this paper, a new algorithm is being proposed, which enhances the accuracy of Naïve Bayes Spam Filter enabling it to correctly detect but also classify mails as ham or spam. This python algorithm is a combination of semantic & keyword-based and machine learning algorithms for the increment of accuracy. When compared to Spam assassin it is found that the enhanced accuracy of this new Naïve Bayes Spam Filter is more than 200%. Furthermore, a relationship of spam score and length of an email is jotted down, which indicates the Bayesian Poisoning, which itself is a very controversial topic [15].

Social Media is becoming the popular hub for the transfer of information, allowing users online to publish personal reviews and opinions. However, the introduction of notorious popular Spam comments affects the users seriously. For detecting comments which may be spam, in social Media of China, we propose a semantic analysis for building a self-expanding dictionary which keeps on updating and automatically adding new cyber words. This semantic analysis feature is very helpful for the classification of text which also based on the statistical analysis of microblogging comments. Selection of four characteristic text-based features of Chinese spam comments is done which uses spam dictionary and text-based features for constructing a classifier for the detection of spam comments. To conclude, through this work, an accuracy of 93.6% is received which states that this work is better than the existing spam comments detection methods [16].

In this paper, we propose two things. Firstly, we propose a concept, where motion-mode with certain constraints as a Control-Law Module (CLM) is encapsulated. Any control law with certain constraints is also called an Instance of the CLM. Secondly; we propose a control framework called RFM which is done by the decomposition of a feasible conformation-path with just one type of CLMs in each partition. This helps in designing and incorporation of partitions easily and interchangeably, which indeed makes the framework more adaptive, modular and flexible [17].

In recent years the utilization of web has increased exponentially, allowing business exchanges which indeed have allowed the online business to permit their clients to review products on their experience. This means surveys are although basic but very important as reviews are extremely important not only for the customers but also for the company itself. This is to keep the business-focused. But with every advantage, we get disadvantages as well. Spam reviews are becoming very notoriously popular these days. Many times, spam reviews are being used for posting fake reviews. This is why spam detection is the need of the hours. Through this paper, we concentrate on Twitter, where we utilize most of it, as it is one of the most recognized micro-blogging sites. We work to gather corpus for feeling

exploration and assess mining practices.

Utilizing these mass data, we build a classifier called feeling classifier which would indeed decide whether a post is positive, negative or neutral [18].

6. STEPS OF ALGORITHM

The algorithm works in following steps:

- It will first give the option to the user to add the Handmade Rules in the database for filtering.
- A Training set of emails shall be provided for training phase which will be applying the Detecting SPAMs
- In the testing phase, I will provide some pre-known emails to test for spam emails and check the accuracy of the system.
- Then the user will select the real emails to cluster spam mail and non-spam emails.
- In all phases, email data cleaning shall be performed.

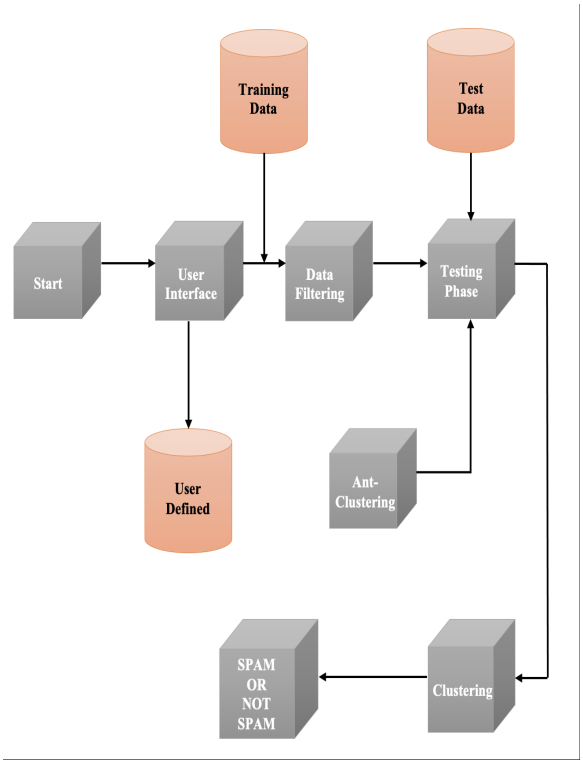

Figure 1: Flow Chart of the proposed algorithm

7. PROPOSED ALGORITHMS

This research is proposing to filter the spam mails based on the Handmade Rules and Automated Recognition of the SPAM using Ant Clustering.

The proposed algorithm works in two major phases, first training phase and second testing phase. In training phase emails from Enron Email Corpus is read email by email and a list of rules is applied to train the system. The rules are consisting of SPAM mail identification techniques such as handmade rules in which words of type SPAM and NON-SPAM both are loaded from the database created and updated by the users. The user also adds various email addresses which are considered as SPAM Emails, these are also added in the system during the training phase. Various automation rules are also added in the system during the training phase by working the training set of emails.

The testing phase uses the data collected during the training phase as the basis for SPAM filtering and in this phase testing data set is loaded which is a mix of the known count of SPAM and NON-SPAM emails. Data collected during the training phase is used to find the frequencies in the testing data and then weights are calculated as per the table below. These weights are used to cluster the emails by taking a certain percentage of the weight below which mails are considered as NON-SPAM mails and above which are SPAM mails.

The weight per cent has been decided by repeated execution of algorithms and finding the best threshold value to decide the SPAM and NON-SPAM. The weight per cent used has been applied to various datasets and found to produce good results

Accuracy Measurements

The evaluation measures which are used in the approach for testing process in our research work could be defined as follows:

True Positive (TP): This indicates whether a spam document is correctly classified as spam.

True Negative (TN): This indicates whether a non-spam document is correctly classified as non-spam.

False Positive (FP): This indicates whether a spam document is incorrectly classified as non-spam

False Negative (FN): This indicates whether a non-spam document is correctly classified as spam.

8. RESULTS & DISCUSSION

Table 1: Comparison of Metrics with the existing system and proposed system

Processing Type	Existing System	Proposed System
True Positives (TP)	2.0	4.0
False Positives (FP)	3.0	0.0
True Negatives (TN)	3.0	1.0
False Negatives (FN)	2.0	5.0
Precision (P) = TP / (TP+FP)	40.0	100.0
Recall (R) = TP / (TP+FN)	50.0	44.44
F-Measure = 2 * (P*R) / (P+R)	44.44	61.538

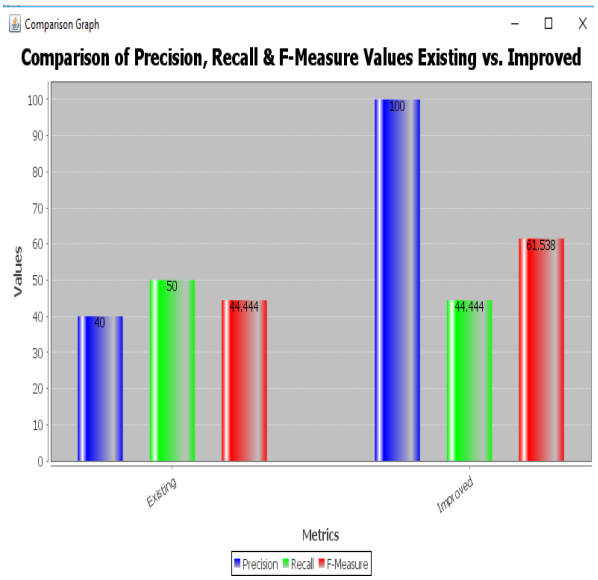

Figure 2: Comparison of Metrics with the existing system and proposed system

Inference: The above graph is a clear indication of the precision, recall and f-measure values of the existing and proposed system. From the graph, it is clear that the value of precision is higher for the proposed system in respect of the existing system. Similarly, f-Measure value is also improved considerably. Although the value of recall is a little lesser than the existing it is due to the impact of high precision values, which is desirably higher and hence the results are considered to be better than the existing work.

9. CONCLUSION & FUTURE SCOPE

In this work, an email clustering method is being proposed to efficiently detect the spam mails. The proposed technique will use the handmade rules and automated rules for training and clustering of the system. The proposed technique shall be implemented using Core JAVA, JDK 1.7 technology; Enron email corpus dataset was selected for the experiment. Different performance measures such as precision, recall, specificity & the accuracy, etc. will be observed.

Ant Clustering is shown to be best in both smaller data sets and large data sets. In ant clustering, several ants shall be created to search the dataset for clustering which provides high performance as the ants work in parallel for clustering of data. Data is scanned in parallel for the same data set therefore memory requirement is also less.

This work can be improved in future by the inclusion of more rules and will depend on the word selection for SPAM and NON-SPAM emails. Work can be further verified by testing on different data sets to check the accuracy, precision, recall and specificity of the proposed system. More weight calculation rules can be added for improving the accuracy of the system.

This work can be improved in future by the inclusion of more rules and will depend on the word selection for SPAM and NON-SPAM emails. Work can be further verified

by testing on different data sets to check the accuracy, precision, recall and specificity of the proposed system. More weight calculation rules can be added for improving the accuracy of the system.

REFERENCES:

[1] K. Agarwal and T. Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 685-690, doi: 10.1109/ICCONS.2018.8662957.

[2] S. Annareddy and S. Tammina, "A Comparative Study of Deep Learning Methods for Spam Detection," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2019, pp. 66-72, doi: 10.1109/I-SMAC47947.2019.9032627.

[3] W. Peng, L. Huang, J. Jia and E. Ingram, "Enhancing the Naive Bayes Spam Filter Through Intelligent Text Modification Detection," 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), New York, NY, 2018, pp. 849-854, doi: 10.1109/TrustCom/BigDataSE.2018.00122.

[4] Qiang Zhang, Chenwei Liu, Shangru Zhong and Kai Lei, "Spam comments detection with self-extensible dictionary and text-based features," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 1225-1230, doi: 10.1109/ISCC.2017.8024692.

[5] W. Li, "Notion of Control-Law Module and Modular Framework of Cooperative Transportation Using Multiple Nonholonomic Robotic Agents With Physical Rigid-Formation-Motion Constraints," in IEEE Transactions on Cybernetics, vol. 46, no. 5, pp. 1242-1248, May 2016, doi: 10.1109/TCYB.2015.2424257.

[6] C. Visani, N. Jadeja and M. Modi, "A study on different machine learning techniques for spam review detection," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 676-679, doi: 10.1109/ICECDS.2017.8389522.

[7] Viktor Mauch, Marcel Kunze and Marius Hillenbrand "High performance cloud computing", Future Generation Computer Systems, Elsevier, 2012.

[8] C. Baun, M. Kunze, T. Kurze, V. Mauch, High performance computing as a service, in: I. Foster, W. Gentzsch, L. Grandinetti, G.R. Joubert (Eds.), High Performance Computing: From Grids and Clouds to Exascale, IOS Press, 2011.

[9] Qi Zhang · Lu Cheng · Raouf Boutaba Cloud computing: state-of-the-art and research challenges Springer Published online: 20 April 2010

[10] Sanjay P. Ahuja, Sindhu Mani The State of High Performance Computing in the Cloud February 2012

[11] UC Berkeley Reliable Adaptive Distributed Systems Laboratory http://radlab.cs.berkeley.edu/ Above the Clouds: A Berkeley View of Cloud Computing February 10, 2009

[12] Amel Haji, Asma Ben Letaifa, Sami Tabbane "Implementation of a virtualization solution: SaaS on IaaS" 2011 IEEE

[13] Software and Information Industry Association, "Strategic Backgrounder: Software as a Service," 2001, http://www.siia.net/estore/pubs/SSB-01.pdf

[14] T. Dierks and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.2," 2008, IETF RFC5246, http://www.ietf.org/rfc/rfc5246.txt

[15] Netscape, "The SSL Protocol: Version 3.0," Netscape/Mozilla, 1996. http://www.mozilla.org/projects/security/pki/nss/ssl/draft302.txt

[16] Napper, J. and Bientinesi, P., Can Cloud Computing Reach the TOP500?,Proceedings of the Workshop on UnConventional High Performance Computing, in conjunction with The ACM International Conference on Computing Frontiers, 18-20 May 2009

[17] Labate, B. and Korambath, P., Use of Cloud Computing Resources in an HPC Environment - IDREHPC Research Projects, 2009 International Conference on Computing Frontiers, 18-20 May 2009

[18] Subramanian, V., Ma, H., Wang, L., Lee, E. and Chen, P., Azure Use Case Highlights Challenges for HPC Applications in the Cloud, HPC in the Cloud, (Web:http://www.hpcinthecloud.com) February 21, 2011.