COMPUTATIONAL INTELLIGENCE AND MACHINE LEARNING
e-ISSN: 2582-7464

Computational Intelligence and Machine Learning
Vol-1, Issue-1, August 2020

Electronic Mail Classification System Based on Machine Learning Approach

Subhrajyoti Ranjan Sahu[1]*, J.Sunil Gavaskar[2]

[1] B. Tech.,ECE, Nalanda Institute of Technology, Bhubaneswar, India.
[2]Assistant Professor, Lord Jegannath College of Engineering & Technology, Ramanathichenputhur, Tamil Nadu, India.
*Corresponding Author Email: [1] subhrajyotiranjan@hotmail.com

Abstract

In current times, users depend comprehensively on electronic communication ways such as electronic mails as it is considered a foremost source of communication. A vast amount of time is investedin electronic mail for communication in the information technology field, due to which electronic mail management has become a prominent feature among the mailing applications. Electronic mail classification comes under this type of management which helps the expert to eliminate the time invested during un-necessary mail reading. Also, the content of electronic mail is further used in the analysis for future prediction and reading behaviors in which a good mail classification system would reduce a lot of time and resources. Conventionally many other systems or methods are present and widely popular in the market but there is no such system that achieves high accuracy. This paper proposes a novel electronic mail classification system that is based ensemble technique which combines the result of many classifiers to achieve good accuracy.

Keywords

Classifiers, Content Analysis, Electronic communication, Electronic mail, Feature extraction.

INTRODUCTION

Communications is part and parcel of everyday operations run smoothly, run smoothly. Good communication is not only within the company but rhe customer also extends. Telecommunications Consumer Care, E-mail and talk communications are primarily based today. Every e-mail a service ticket is considered the client. It refers to small or medium businesses might be enough for the whole support team to have a common email inbox works together on service tickets for consumers.

The method is not scalable though, the support team is also rising as the business expands. Take a look at a situation teams, each running a broad support team miscellaneous errands. To optimize performance, and minimize time the support ticket is spent in the system and incoming tickets must be sorted and assign right support team to them.

This task is time consuming and Intensive labor but automation is not a trivial task because of the complexity of the Natural languages which the software needs to understand. Any system that does the processing of natural languages, that is to say a language used by humans to Communicate, does Natural Language Processing (NLP).

Automating email labelling and sorting requires a model that can differentiate between different types of errands and support requests. Such models must be able to do this even if the email contains spelling mistakes, previous conversations, irrelevant information, different formatting or simply rubbish. The LSTM model is an extended version of the Recurrent Neural Network (RNN) network which is a sequential model often used in text classification. Word embedding models aim to model the words of a language in a vector space and placing words with similar semantic meaning close to each other.

Electronic Mail (E-mail) is commonly a procedure for transferring and receiving electronic messages by using electronic devices such as smartphones or laptops.Emails were accessible in the 1960's, at that period only administrators could only transfer applications on the same device, so early email networks needed to concurrently submit both the sender and the receiver electronically, equivalent to immediate texting. In 1971, Ray Tomlinson created a first program that could transfer mail between the uses on various servers around the ARPANET using the @ sign for connecting the username to a destination site and that was recognized as email in the mid-1970s.

Mail operates on computer networks, primarily uses the Internet. The email networks today are focused on a store-and-forward approach i.e. Allow, forward, send, and store messages from email servers. Neither of the consumers or their machines must be concurrently online, nor will they sign in normally with a mail server or a webmail system for writing, receiving or uploading information.Original ASCII text only, Internet email has been expanded to express text through certain character sets and interactive material attachments via Multipurpose Internet Mail Extensions (MIME). The history of current Internet email systems dates back to earlier ARPANET and specifications for email encoding released in 1973 (RFC 561). This is the consequence of the fact that foreign email addresses using UTF-8 are universal, although not generally accepted. In the early 1970's, a text address is identical to a foundational email received today.

Email was a popular and efficient communication mechanism increases the number of users on the Internet. Email control is also necessary and relevant the rising issue

for citizens and businesses since it's vulnerable to violence. The Blind Layout Spam is a spam-known unwanted email a mishandling case.

A vast volume of data from different channels such as existing / potential clients, suppliers and internal contact inside the organization, product/service requests, other private and government organizations, etc. is delivered to businesses including IT firms, company institutes (such as investment banks), production sectors, and process sectors.Such unstructured communications are handily categorized by most organizations, with the assistance of trained customer support personnel, depending on the skills needed to answer and respond to the information contents of the document. The vast complexity of incoming e-mails nevertheless renders this strategy difficult to handle, time-consuming and misunderstood.

Email content Analysis is the term given to the process of exploration of the content present in the electronic mails; it is a useful process and one of the foremost applications of this type of content analysis lies under digital forensic investigation for distinguishing the abnormal activities like crimes, frauds. Also one of the application of this type of analysis is organizations get to know about the behavior of the email users. For instance, if there is company named "Nutrition SPL" and it sends mails for advertising there products, now with the help of the email content analysis, the company would get to know about how customers react to the emails such as, they read it or the customers leave the mails unread.

For conducting a good content analysis, the electronic mails should be categorized properly. Though some electronic mail applications provide features like filters. In which the user applies a filter to get some specific types of mail. But if the vast amount of content is considered then it is not easy to filter out the mails and content according to some particular requirements.

The most common application of the email classification is spam classification, it has been observed in many surveys and feedbacks, that spam in the mail is considered as one of the furthermost complex problems in the email services.Spam e-mails are any unintended e-mails not meant for a single recipient and submitted for marketing reasons, fraud, hoaxes etc. In 2009-2010 about 97% are reported to have been spammed[1].Therefore, several academic publications or reviewing emails centered on this topic (e.g. spam classification). But there is an ongoing conflict between spammers and spam detection devices, in which each party attempts to develop different methods of beating the other's techniques.

Some local papers [2], [3]conducting a spam evaluation have shownthat the problem was realistic. Researchers also carried out studies to determine existing spam delivery status in KSA. Writers have sought to sum up crucial explanations for spam communications and e-mails including pornographic content, advertisement, phishing, faith, etc. Of course, overuse and bandwidth and resources for no good purposes are a major disadvantage of spam spread.

An e-mail spam classifier is not only required to identify spam as junk mail correctly but also to recognize non-spam or regular e-mails in this regard. It is when the criteria for determining its definition or its estimation are all known.Then the accuracy of the email forecast is measured by four forecast metrics. True Positive (TP) claims the spam detector method assumes spam is spam, it always was spam,and it is spam. True Negative (TN) means the device or email program predicts the email is usual, not spam and it was right. The method inappropriately predicts that spam (alleging false alarms) is a positive e-mail[4]–[6].

"Eventually False Negative (FN) often leads to a further mistake in which spam email is supposed to be ordinary. The identification method would also include the following values: TP 100%, TN 100%, FP 0% and FN 0%. It is difficult and unrealistic to attain this optimal condition. TP and FP balance each other by 100% (i.e. 100% of them total). Some email classification systems face the difficulty of limiting TP across multiple spam detection functions, but also many false alarms. On the other side, very lean rules that earn very high TN yet FN.Another challenge in emails' spam detection is speed. Insecurity, speed or performance is always in a trade-off withsecurity where too many roles may slow down the system." In addition to the classification that is based on spam messages, some papers have conducted researches in the content of the email, discussed other aspects for instance: automatic folder classification, contacts and email ID classification and alike.

The emails could be structured into chains and contain several emails sent back and forth between the client and the support staff. The models could use this in theory to determine the context and how the subject changes conversations. These chains are not included in the emails our model use, they are isolated and labeled individually. The model will be restricted to, and no classification or training will be done on any sort of document other than emails.

The field of machine learning is a substratum for the broad artificial intelligence field, aimed at ensuring that machines learn as human. Training implies that certain mathematical processes are known, identified and described.

Furthermore, this research paper is divided into various segments to address the issue and solution along with related work related to the email classification. Second segment of this paper describes and cites some related work like publications, patents and articles related to the field of email classification. Third segment of this paper discloses about the proposed work and fourth segment talks about the testing and relative results of the proposed work, lastly, the fifth segment of the researcher paper concludes all the work and addresses some advantages of the proposed work.

RELATED WORK

This segment of the research paper deals with some research work related to email classifications.

Table 1 : Some research work related to E-mail Classification

1. 2015- [4]	Discloses about frequency and term frequency combined feature selection method (DTFS) to improve the performance of email classification.
2. 2015- [5]	Discloses about classification task is called feature selection, which is used to reduce the dimensionality of word frequency without affecting the performance of the classification task.
3. 2016- [6]	Discloses about fuzzy logic techniques for email clustering. Extract concept and feature, same feature keyword goes into one cluster if a new keyword is found and not matched with any existing cluster than a new cluster is defined for that. Based on these clustering techniques authors wants to update that calendar for real time information and hassle free for reading unnecessary emails.
4. 2017-[7]	Discloses about two different approaches for classifying emails based on their categories. Naive Bayes and Hidden Markov Model (HMM), two different machine learning algorithms, both have been used for detecting whether an email is important or spam.
5. 2018-[8]	Discloses about different representation schemas for emails and a large set of features are also adopted for the purpose of experiment. Proposed Cascaded SOM based classification model performs well in email-classification compared to standard classification approaches and classical SOM based model.
6. 2019-[9]	Discloses about the use of semi-supervised learning can help leverage both labeled and unlabeled data. In the evaluation, we investigate the performance of our proposed approach with two datasets and in a real network environment.

As it is discussed that email classification is divided according to the situation requirements, so this segment cites paper according to the field such as spam-non-spam email classification, email data analysis research goals, ontology classification of email contents. Table 1 presents the work done related to email classification [7][8][9][10][11][12]. Figure 1, below, illustrates the distribution of research paper according to the email classification. The areas are categorized into five domains: spam, phishing, spam and phishing, multi-folder categorization, and others. Figure 2 depicts the frequency of email classification techniques according to the domains.

Categorization of single-label text (classification) is defined as the task of assigning a category to a document given a predefined set of categories. The goal is to approximate the representation of the text, so that it coincides with the text 's actual category. If a document will consist of multiple categories, then we need to adapt our algorithm to multiple category performance, which is called multilabel classification. The task then is to assign a appropriate number of labels that correspond to the document 's actual labels.

A fundamental categorization objective is to categorize documents in the same set that have the same context and documents that do not have the same context in separate sets. This can be related with various approaches involving algorithms for the machine learning. Machine learning algorithms learn to generalize categories from documents previously seen that are later used to predict the category of documents previously unseen.

The non-sequential models are used as a baseline for comparison with the LSTM network. The parameters of these models are not optimised and do not use preprocessing techniques such as lemming, stemming or stop word removal. The BoW or Average Word Vector (AvgWV) text representation models used for the non-sequential models are also not optimised. The BoW hyperparameters filter the numbers of words within a relative range, i.e sub- and supersampling. This is done to make the experiments possible with all non-sequential models

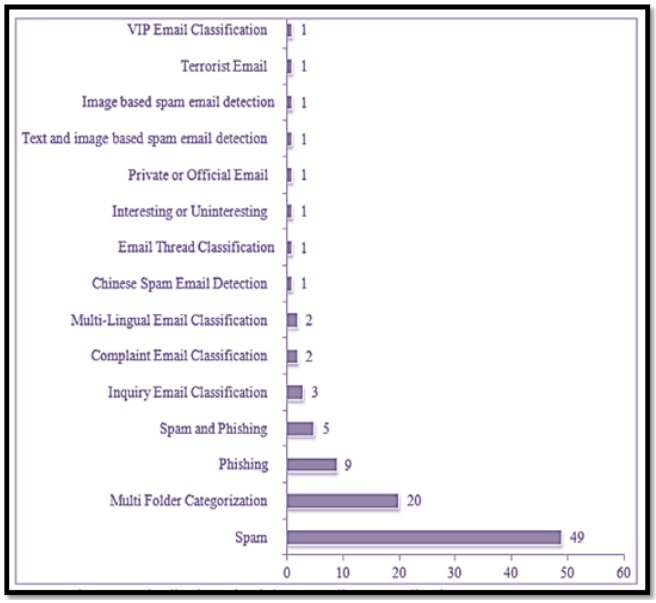

Figure 1: Distribution of Research paper According To Email Classification

An Indian financial services firm gets about 1.2 million emails a month. The firms externalized e-mail categorization to a help staff of 50 leaders. Every part of help had to read and categorize 2.27 emails every minute. Customers encountered many challenges, among them 24-hour SLA, the team collapse and less than 80 percent classification performance[3].

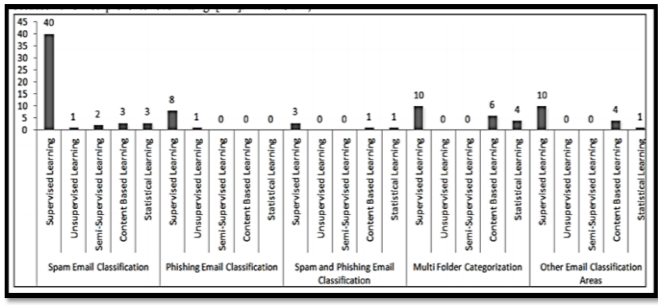

Figure 2: Frequency of email classification technique in different domains

Several types of email classification systems and methods are developed by using various techniques of machine learning. But still some challenges are still there, i.e. the challenges are dynamic, which is caused to the heterogeneous behavior of the content present in the email and some challenges are caused due to CPU limitations during implementation, due to which the required accuracy i.e. goal is not accomplished. Due to the aforementioned drawbacks, there is a need to develop an effective email classification system.

Several of the topics needed to effectively classify emails, i.e. models that interpret natural language and classifiers that use word relationships in a time series, are investigated. The bulk of the recent literature on the English language has been conducted while conducting similar studies and no study has been performed on the Swedish language. In terms of email classification and how to better use the NLP and machine learning models within that context, little work has also been done.

PROPOSED WORK

This segment of the research paper talks about the proposed system for email classification based on machine learning technique. Figure 3, shows the proposed system, various modules used in the system are explained below.The Python Application and Gmail API Bridge between the email client and the classification system connect to an email client and transmits an email to the rating system (in HTML form).

The same GUI

is used after grouping to route messages. Extractor software removes and blends email material, e.g. Email body and name. The purpose of the preprocessing module is to clean and prepare data that can be transmitted through the machine learning algorithm for the development of apps. Feature vectors are built from textual content in the function creation and representation module such that a learning algorithm can be understood (textual material must be expressed in number).The classifier is just a machine-learning algorithm to be equipped to do well in an unreleased dataset (unseen email). The professional classifier is used to determine in real time which e-mail ID a given e-mail will be routed to.

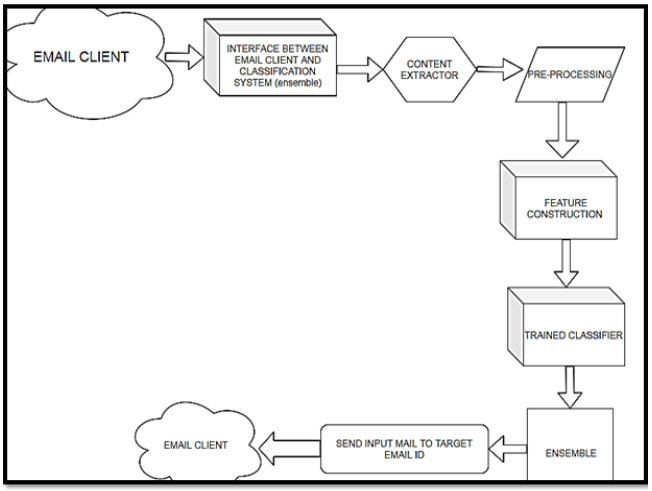

Figure 3 : Structure of machine learning-based email classification system

Additionally, the proposed system has an ensemble module for increasing accuracy. A series of learning algorithms are used by an actor to blend their results to simulate. There was a misunderstanding. This method works very well when the simple algorithms are unbalanced: they are typically unique, depending on the sub-set of the results.Ensemble learning as a whole is the way several models are systematically created and merged, such as classifiers or specialists, to solve a particular computational issue. The ensemble approach is used primarily for classification enhancement. Figure 4 shows the working of the ensemble module.

Ensemble approach/modeling is considered as a powerful technique to ameliorate the performance of the system. According to many subject experts relation machine learning, the ensemble is an art of integrating a diverse set of classifiers to ameliorate the stability and predictive power of the system. In simple terms, the prediction results of each classifier are combined and an average result is considered as the final result.

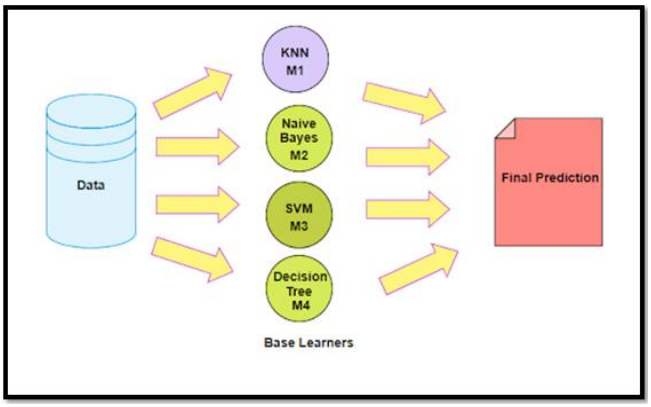

Figure 4 : Working of Ensemble Module

RESULTS

The simulations are performed to determine the correct e-mail classifier, function design and structural techniques. The impact of numerous preprocessing methods on the efficiency of the classifier will be studied (i.e. stopping words elimination, stamping, translating text to a single case of letters, different tokenization schemes).

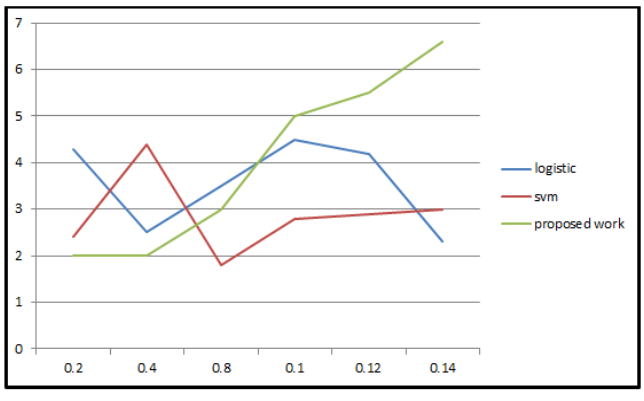

Figure 5 : Various proportion of training data in accordance to the techniques

The effect of connection terminology and the decrease in the classifier's output are also evaluated. Also, best attributes (i.e. hyper-parameter tuning) are used for each classifier along with grid search over a range of attributes. All the experiment work is conducted on two kinds of datasets i.e.

one of them is 20 newsgroups and the other is of demo email dataset that has more than 350 emails, which demonstrates the effectiveness of the proposed system in real-time. Figure 5 shows the graph representing the test error rate for various proportions of training data. In this case, the measure of test error rate is taken as log loss. Dataset used is 20 newsgroup datasets. Figure 6, illustrates a Bar chart showing a comparison between different classifiers based on score/accuracy.

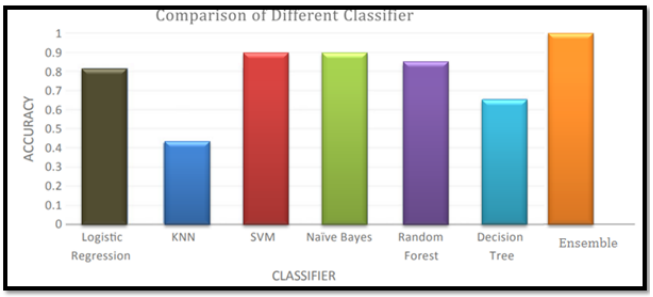

Figure 6 : Bar Chart showing a comparison between different classifiers based on score/accuracy

CONCLUSION

For big corporations, corporations, sectors, and businesses to evaluate vast e-mail details, an integrated real-time e-mail classification system would be quite useful. Once trained on a named dataset, the proposed email classification system can be used to accomplish real-time classification and even be compatible with any email user. A stand-alone framework that can be tailored to the needs of any enterprise is established as the approach suggested. It can be observed through the results that the addition done in the form of the ensemble in the proposed system defiantly boost up the accuracy of the system. The research can be further extended to analyze the efficacy of deep learning techniques to solve the classification problem because of the massive nature.

Extending the email classification to identify emotions can help the support team address angry or dissatisfied clients. That will improve customer service as support staff can cope with customer emotions. This would improve customer loyalty and raising the number of customers changing provider.

REFERENCES

[1] M. A. Al-Kadhi, “Assessment of the status of spam in the Kingdom of Saudi Arabia,” J. King Saud Univ. - Comput. Inf. Sci., 2011, doi: 10.1016/j.jksuci.2011.05.001.

[2] I. Alsmadi and I. Alhami, “Clustering and classification of email contents,” J. King Saud Univ. - Comput. Inf. Sci., 2015, doi: 10.1016/j.jksuci.2014.03.014.

[3] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, “Email Classification Research Trends: Review and Open Issues,” IEEE Access, 2017, doi: 10.1109/ACCESS.2017.2702187.

[4] K. Karthik and R. Ponnusamy, “Adaptive machine learning approach for emotional email classification,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011, vol. 6763 LNCS, no. PART 3, pp. 552–558, doi: 10.1007/978-3-642-21616-9_62.

[5] N. Sutta, Z. Liu, and X. Zhang, “A Study of Machine Learning Algorithms on Email Spam Classification,” 2020, vol. 69, pp. 170–159, doi: 10.29007/qshd.

[6] S. Kranthi Reddy, P. B. Tarun, S. Rushika, P. D. Redd, and E. Anjala, “Spam email classification using machine learning algorithms,” Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 7, pp. 1748–1752, 2019.

[7] Y. Wang, Y. Liu, L. Feng, and X. Zhu, “Novel feature selection method based on harmony search for email classification,” Knowledge-Based Syst., 2015, doi: 10.1016/j.knosys.2014.10.013.

[8] M. Mohamad and A. Selamat, “An evaluation on the efficiency of hybrid feature selection in spam email classification,” in I4CT 2015 - 2015 2nd International Conference on Computer, Communications, and Control Technology, Art Proceeding, 2015, doi: 10.1109/I4CT.2015.7219571.

[9] T. Suma and S. Y. S. Kumara, “Email classification using adaptive ontologies Learning,” in 2016 IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2016 - Proceedings, 2017, doi: 10.1109/RTEICT.2016.7808210.

[10] S. R. Gomes et al., “A comparative approach to email classification using Naive Bayes classifier and hidden Markov model,” in 4th International Conference on Advances in Electrical Engineering, ICAEE 2017, 2017, doi: 10.1109/ICAEE.2017.8255404.

[11] N. Saini, S. Saha, and P. Bhattacharyya, “Cascaded SOM: An Ameliorated Technique for Automatic Email Classification,” in Proceedings of the International Joint Conference on Neural Networks, 2018, doi: 10.1109/IJCNN.2018.8489584.

[12] W. Li, W. Meng, Z. Tan, and Y. Xiang, “Design of multi-view based email classification for IoT systems via semi-supervised learning,” J. Netw. Comput. Appl., 2019, doi: 10.1016/j.jnca.2018.12.002.