ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Visit: www.ijirset.com*

**Vol. 8, Issue 3, March 2019**

# Detection of Minimum Number of Features for Identification of Group Emails

Priti Kulkarni[1], Dr. H. S. Acharya[2]

Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed) University, Pune-16, Maharashtra, India[1]

Allana Institute of Management Sciences, Camp, Pune-1, Maharashtra, India[2]

**ABSTRACT**: Emails have become an integral part of business communication. Group emails are very common in corporate sectors as well as in general day to day communications between groups of peoples. Group emails contains list of email addresses collected under one name. Group emails are standard process and part of email policies within organization. Group emails are used to inform end users about message content in general regarding any basic or specific information, invitation, survey conducted across department or any kind of announcement. The intention here is to inform and educate target audience about the message content. Though group email is easy way to send the messages to multiple recipient, it can become dangerous if it is send to wrong group or outsider may use group email id to sent malicious email to target the multiple users. Such emails should get identified and filter at server level.
The objective of this paper is to:-
a) Find minimum number of features in the email header for group email identification.
b) The effect of this minimum number of identified features on the accuracy of classification of emails.
The information Gain (IG), Chi-squared, relief, correlation based and wrapper feature selection techniques are used to find number of features. The resultant numbers of features are used as input features to Naïve Bayes, Decision tree and K-nearest neighbour classification techniques. The effect of these features on accuracy of classification techniques is recorded.

**KEYWORDS**: group email, feature selection, email header, classification

## I. INTRODUCTION

***Group email is facility to send email to group of people using single email id. It facilitates all group members to communicate and collaborate to each other. An organisation can create and maintain multiple such groups for different purposes.*** Group mails are a specific category which facilitate in quick mass communication. But group mail as a mode of communication has few challenges associated with it. Sending wrong content to the group, wrong choice of target audience, negligence on sender part are few challenges, which may put organisation in big trouble and might results in legal complications. In corporate sector, email policy is implemented in such a way to educate employees about proper email usage within an organisation but sometimes it is ignored by the employees.
There are two major types with group mailing which organisations should be aware of:

1. Group mails sent or initiated by employees within organization but which do not **confirm to** email policy document of organization.

2. Mails received from outside network on valid group email, possibly unsolicited, sent with malicious intent must strictly be restricted and filtered at server level.

Thus the first issue causes ***managerial confusions,*** unnecessary overheads. Second type is the ***most dangerous.*** Unchecked external mails can effectively imply negative influence most of the time bypassing the authority of the

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

organisation. Identification of group email in systematic way and controlling it according to group email policy set by organisation is major challenge.

### A. Feature selection

Feature selection is process of reducing less significant features from email header and selecting minimal subset of features [3]. There are two ways to select features.

#### a) Filter method

Feature selection is carried out using filter method which is independent of learning algorithm. It depends upon general characteristics of training data to select features [8].

#### b) Wrapper Method

This method requires predetermined learning classifier. The subsets of features are selected by evaluating performance of classifier for particular training dataset. But they are computationally expensive than filter method [13].

## II. RELATED WORK

There are many studies of existing feature selection methods available in the literature. Mendez et al [5] have used information Gain, Document Frequency, Chi-squared Jin et al.[2], Mutual Information for spam classification and concluded that Mutual information performance is worst among all methods. Relief is proposed by Kira p [4] and suitable for classification. Hall Mark [1] proposed Correlation based feature selection techniques for feature selection in machine learning. It considered as most reliable method. Zhang et.al [10] have used chi-squared for feature selection in email classification and proved to be effective method. Lai [14] has empirically studied by comparing three machines learning algorithm NB, KNN, SVM, TFIDF+SVM and showed that SVM (94%) performs superior among all. Kiritchenko & Matwin [15] used correlation based feature selection and conclude that SVM outperform than Naive Bayes classifier. Awad et. al [17] have compared different algorithm such as SVM, KNN, NN, AIS (Artificial immune system) RS (rough sets) for performance evaluation on the SpamAssassin spam corpus. It is showed that a Naive Bayes and rough sets method has a very satisfying performance among the other methods. Youn [16] used tfidf feature selection technique and proved that decision tree, Naive Bayes performed well. Neural Network and SVM were not appropriate to make binary decision.

### B. Email Structure

An email contains headers and body. Email header field format is defined in RFC 822, RFC 2822. The message header contains control information such as sender address (From), recipients addresses (To, Cc, and Bcc), and content type. One may classify an email by identifying headers of an email. According to RFC 822/2822 email header contains useful information (Metadata). There are multiple email header features available for email. It is necessary to find minimum number of email header features that plays important role in group email identification.
The main challenge is to identify and automate group email classification. To automate the process, it is necessary to identify header features.

## III. DATA COLLECTION

We have used three datasets with different size. Total 6942 emails are extracted from personal inbox of Gmail by developing custom code using python programming language. The figure1 shows the output of email extraction program. Data is extracted and store into CSV file. Two more benchmark spam dataset $D_{cs}$, $D_{SA}$ [11][12] were used. Details are as follows,

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

**1.Description of datasets used in experiment**

| Dataset | Description | Volume | Group emails |
|---|---|---|---|
| $D_{in}$ | A personal inbox of Gmail domain | 6942 | 2449 |
| $D_{CS}$ | Benchmark dataset of CSDMC2010 | 4318 | 2245 |
| $D_{SA}$ | Benchmark dataset SpamAssian | 750 | 109 |

Following 37 features as present in our extracted personal dataset (Din) as follows,

UF37(s) = { Authentication-Results, bcc, cc, Content-Disposition, Content- Type, Date, DKIM-Signature, From, In-Reply-To, List-Archive, List-Help, List-ID, List-Owner, List-Post, List-Software, List-Subscribe, List-Unsubscribe, Mailing-List, Message-ID, Precedence, Received, Received-SPF, References, Reply-To, Resent-bcc, Resent-cc, Resent-Date, Resent-From, Resent-Message-ID, Resent-Reply-To, Resent-To, Return-Path, Subject, Thread-Index, Thread-Topic, To, X-Mailer}

If feature is present in email header, it is represented with bit “1” otherwise value to this feature is set to as “0”. In order to obtain a training corpus for supervised learning algorithms, emails were classified manually with domain knowledge.

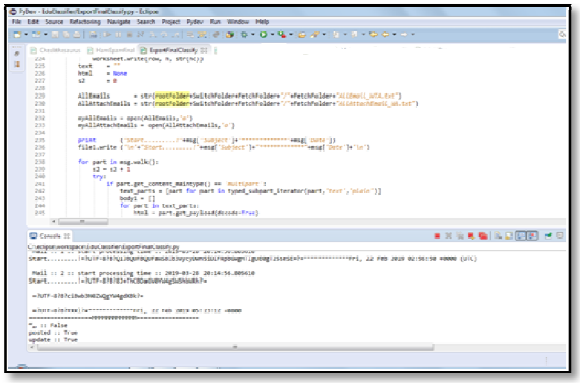

Figure 1 output screen for Email extraction

## IV. EXPERIMENT

The following steps were followed in order to identify email header features for group email identification**.**

1. Input: Em={f1,f2,f3,f4…...f37} where, fi=feature from email header set.
2. Apply Feature selection technique namely Information Gain (IG), Chi-squared, relief, correlation based, wrapper feature selection.

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

3. Output: List of selected features, Fs
4. Repeat step 2 to step 3 for different feature selection techniques
5. Apply Naïve Bayes, Decision tree and K-nearest neighbour classifiers on the each feature set generated in step 2
6. Record the accuracy
7. End

The feature selection techniques namely Information Gain (IG), Chi-squared, relief, correlation based as filter methods and wrapper feature selection were applied on datasets. The resulting set of features used as input to classifier Naive Bayes[7], Decision tree [6] and K- nearest neighbour [9] to find effects on accuracy on three datasets. The accuracy of resultant set of features was tested using three classification algorithms. So our experimental deign consists of,

3 datasets*5 feature selection techniques*3 classifiers= 45 runs

A data mining tool Weka is used to carry out all experiments. All the feature selection methods were adopted using the feature Selection facility provided in WEKA. The result of number of features generated by feature selection techniques on three datasets are reported as below in table 1,

Where, NB=Naïve Bayes, DT=Decision tree, KNN= K- nearest neighbour

**Table 1 output of number of features Vs accuracy of classifier for three datasets**

| **Datasets =>** | $D_{in}$ | | | | | $D_{SA}$ | | | $D_{CS}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **FSM** | **No of features generated** | **NB** | **DT** | **KNN** | **No of features generated** | **NB** | **DT** | **KNN** | **No of features generated** | **NB** | **DT** | **KNN** |
| Chi square | 27 | 99.74 | 100 | 99.65 | 28 | 97.19 | 100 | 99.95 | 13 | 99.33 | 99.73 | 100 |
| Correlation based | **3** | 99.88 | 99.88 | **99.91** | **8** | 99.79 | 99.95 | **99.97** | **3** | 99.6 | 100 | **100** |
| IG | 20 | 99.31 | 99.83 | 99.87 | 32 | 97.19 | 100 | 99.93 | 20 | 99.33 | 99.33 | 100 |
| Relief | 27 | 99.74 | 100 | 99.65 | 28 | 97.2 | 100 | 99.95 | 13 | 99.33 | 99.73 | 100 |
| Wrapper+ DT | 1 | 99.88 | 99.88 | 99.88 | 3 | 100 | 100 | 100 | 4 | 100 | 100 | 99.87 |

## V. RESULT AND DISCUSSION

As shown in Table1, For Din dataset, Naïve Bayes and KNN show the better performance with 3 numbers of features as compared to 27 features. Decision tree performs best with 27 features as compared to 3 features. Accuracy of decision tree classifier slightly reduces from 100% to 99.88% with 3 features whereas NB and KNN show improvement in the accuracy. So we have considered 3 features-List-Help, List-ID and Mailing-List as optimum feature set. Correlation based feature selection and wrapper provide minimum number of features. But wrapper method aims to find best features based on classifier as input. We have used decision tree as input classifier. Wrapper method had generated minimum feature from 1 to maximum 4 features.

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

The figure 2(a) and 2(b) shows the output for Decision tree and Naïve Bayes classifier

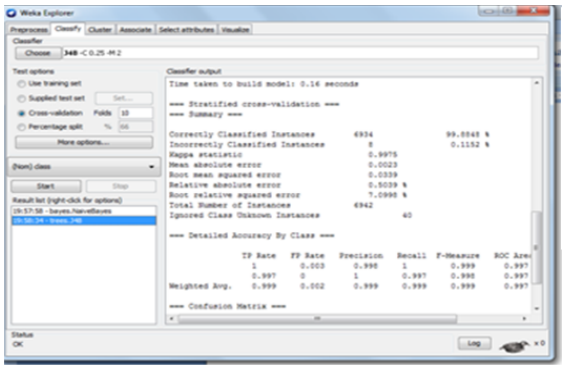

Figure 2 (a) output for Decision tree classifier

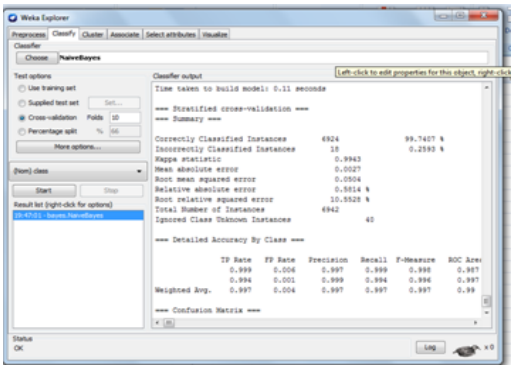

Figure 2(b) output for Naïve Bayes classifier

On benchmark dataset Dcs, KNN performs better with 20 features, 13 features and 3 features but when four features are considered than accuracy slightly reduces from 100% to 99.87%. On $D_{SA}$ dataset with minimum 8 features had shown better performance. Wrapper feature selection with decision tree classifier has generated 3 features. All three classifiers performed well showing up to 100% accuracy with 3 features.
Following Table 2 shows the comparison of minimum number of features generated by all feature selection techniques,

**Table 2. Comparison of minimum number of features generated by feature selection techniques**

| **Dataset Name and Size** | **Din ( Size =6942)** | | **$D_{SA}$, size=4318** | | **Dcs, size=750** | |
|---|---|---|---|---|---|---|
| **Feature selection technique** | Correlation based | Wrapper+ DT | Correlation based | Wrapper+ DT | Correlation based | Wrapper+ DT |
| **No of Features** | **3** | **1** | **8** | **3** | **3** | **4** |
| List-Help | Y | | | | | |
| List-ID | Y | Y | Y | Y | Y | Y |
| Mailing-List | Y | | | | | |
| Precedence | | | Y | | | Y |
| To | | | Y | | | |
| Received | | | Y | | Y | |
| Return Path | | | Y | | | |
| bcc | | | Y | | | |
| thread topic | | | Y | | | |
| List Subscribe | | | Y | Y | Y | Y |
| List Unsubscribe | | | | Y | | |
| Content Type | | | | | | Y |

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Visit: www.ijirset.com*

**Vol. 8, Issue 3, March 2019**

These generated features and its presence in our training datasets was studied. This study concludes that though 'precedence' feature has lower weightage, it is present in all datasets with value "bulk" or "List". Pairing of this header field with "List-id" helps to identify group emails.

## VI. CONCLUSION

Group emails have grown in proportion over the last decade. The experimental result confirms that header features "List-id" and "List-subscribe" both are adequate to identify group emails. Feature "precedence" also provides additional clue to detect group email. Wrapper feature selection technique can be good choice but this technique finds features based on input classifier. So number of features may vary. In case of filter techniques, the correlation based feature selection has generated minimum number of features. The KNN has performed better than NB and Decision tree. In case of maximum number of features chi square or relief feature selection with decision tree classifier is the best choice. The number of features generated also depends upon nature of datasets.

## REFERENCES

[1] Hall M. A. (1998). Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand.
[2] Jin, X., Xu, A., Bie, R., & Guo, P. (2006). Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In International Workshop on Data Mining for Biomedical Applications (pp. 106-115). Springer, Berlin, Heidelberg.
[3] Kumar, V., & Minz, S. (2014). Feature selection. SmartCR, 4(3), 211-229.
[4] Kira, Kenji and Rendell, Larry (1992), The Feature Selection Problem: Traditional Methods and a New Algorithm. AAAI-92 Proceedings.
[5] Mendez JR, Diaz F, Iglesias EL, Corchado JM (2006) A Comparative Performance Study of Feature Selection Methods for the Anti-spam Filtering Domain. In: Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining, Springer Berlin Heidelberg, pp 106–120
[6] Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International journal of computer science and applications, 6(2), 256-261
[7] Ron Kohavi.1996. Scaling up the accuracy of Naïve Bayes classifiers: a decision-tree hybrid, In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 202–207
[8] Sanchez-Marono, N., Alonso-Betanzos, A., & Tombilla-Sanromán, M. (2007). Filter methods for feature selection–a comparative study. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 178-187). Springer, Berlin, Heidelberg.
[9] Wang, A., An, N., Chen, G., Li, L., & Alterovitz, G. (2015). Accelerating wrapper-based feature selection with K- nearest-neighbour. Knowledge-Based Systems, 83, 81-91.
[10] Zhang, L., Zhu, J., & Yao, T. (2004). An evaluation of statistical spam filtering techniques. ACM Transactions on Asian Language Information Processing (TALIP), 3(4), 243-269.
[11] https://sites.google.com/site/ssudslab/datasets/csdmc2010spammailcorpus
[12] http://spamassassin.apache.org/old/publiccorpus
[13] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273-324.
[14] Lai, C. (2007). An empirical study of three machine learning methods for spam
[15] Kiritchenko, S., & Matwin, S. (2011, November). Email classification with co-training. In Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research (pp. 301-312). IBM Corp.
[16] Youn, S. a. (2006) , A Comparative Study for Email Classification . Journal Of Software , 2 (3), 1-13
[17] W.A, Awad and S.M, ELseuofi (2011) Machine Learning Methods for Spam E-Mail Classification. International Journal of Computer Science & Information Technology, Vol 3