Automatic Genre Categorization of Emails into predefined categories using machine learning

Vinod Kumar Bhalla[1], Parag Nijhawan[2], Manish Kumar Singla[3]

[1,2,3] Thapar Institute of Engineering and Technology, Patiala

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: In today's dynamic world, there is a need for fast, efficient, and reliable means of communication. To meet these requirements email system was developed and it got popular with the invention of WWW. Now, the Email system has been used extensively for official, business, and personal communication. On average individual users receive 50-60 mails each day. It is becoming a burden to easily manage emails. So there is a need for effective and reliable means to organize the mails for easy and fast retrieval. An efficient approach is proposed in this paper to classify the mails based on the predefined genres. It has been observed in the proposed research that the classification of emails greatly improves efficiency and saves time and effort to manage them. The results obtained in this paper are very encouraging. Over 90 % of emails are categorized correctly. Email genres are predefined and corresponding keyword lists are generated. Frequency tf-idf of the keywords in the email decides the genre of mail. SVM is used as a multiclass classifier. In this paper need for negative training data has been removed as the proposed classifier works on the principle of one class against the rest.

Keywords: Email, Categorization, Multiclass-classifier, machine learning, classification, SVM

1. Introduction

Data classification has been done since ages for fast and easy retrieval of information. Earlier, amount of data and applications of data were limited. In non-digital era data managers were force to classify data by applying manual efforts. They succeeded to an extent to solve their local problems by putting cumbersome and laborious manual efforts. But, with the usage of digital technology, huge amount of data start getting generated. This corpus of data started posing challenge to the data managers. Engineers started putting efforts to automate the job of data classification. They applied various automatic and semi automatic techniques for classification [3]. The results obtained were not encouraging. Human intervention was still required to large extended to increase the faith in the classification procedure. The next big lead happened with the invention of the World Wide Web. In these days, daily TB of data is generated in the form of web pages/doc, blogs, articles, news items, educational content and emails. This enormous data puts forward further challenge to classify data in to various categories to gain maximum potential to use data in more meaningful ways. Data can be classified based on the predefined category such as – sentiment [2], subject, genre and functional category etc**.** Large number of applications can consume the classified data and get inherent benefits of categorization. Email application is one of the most prominent one to generate and consume data. Over 70-80% population is using this application for personal and official communication media for fast and effective ways. Daily, billions of emails are generated resulting in large data. There is the need to automatically classify [10] the contents of email to gain maximum benefits.

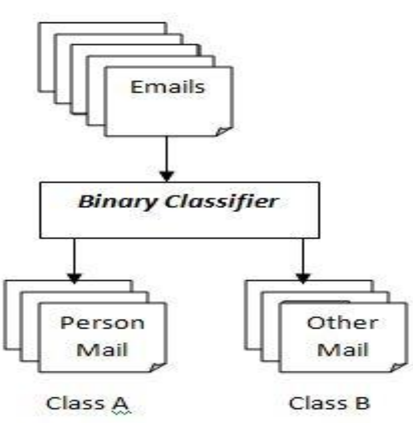


Fig1: Binary Classifier

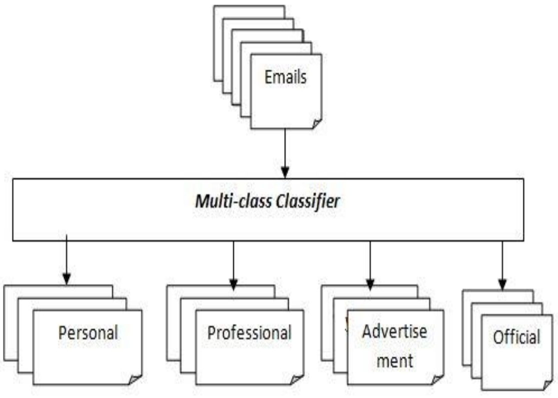


Fig2: Multiclass Classifier

Automatic classification of email data leads to efficiency in organization and performance of time large number of emails are gathered in the inbox and sent folder [8]. It becomes difficult to immediately organize, search and retrieve relevant mails. The purpose of the proposed research is to automate the process to classify the mails in to various categories for easy and fast access. It further helps in managing relevant and irrelevant mails. The author is convinced to develop a scheme which can classify mails to get the following benefits. Both binary and multiclass classification techniques as shown in fig1 and fig2 are used in the literature.

1.2. Organization

Section 2 discusses the literature survey corresponding to proposed work. Section 3 describes the motivation behind this work.. Section 8 explains various algorithms developed for this scheme. Section 10 discusses the experimentation results and discussion. Finally, Section 11 describe conclusion of the article with future scope.

2. LITERATURE SURVEY

Researchers Gomes *et. al.* (2017) [1] did a study on two approaches Naive Bayes and Hidden Markov Model (HMM). Combination of natural language processing techniques was tried on both techniques to compare the accuracy and find the best method. Saidani et.al. (2020) [2] used semantic analysis to enhance the spam detection performance. Their scheme was based on two semantic level analyses. In first step domain based email classification was done and in next step semantics features for each domain was applied to detect the spam emails. They conclude it is a better method for spam detection. Mohammad (2020) [3] used data mining and machine learning techniques and proposed an enhanced model ensuring lifelong spam classification model using Adjustable Dataset Partitioning (ELCADP). This method concluded enhanced performance in comparison to stream mining algorithms. Research work further emphasized that offline spam emails for creating lifelong classification systems.

Chen *et. al.* (2019) [4] also worked on spam detection using Long-Short-Term-Memory model. Active learning model was used to reduce the cost of labeling. Deep learning approach was applied to attain the better performance. Results concluded that this technique is better than classical CNN and RNN based models. Saini*et.al.* (2018) [5] used self-organizing map (SOM) in exploratory phase of data mining. Input data is projected as a lower dimensional map. This technique was based on based use of no labeled classification data. Cascaded SOM was capable of solving any multi-class classification problem that was not labeled. Li *et.al.* (2019) [6] proposed multi-view disagreement-based semi-supervised learning to reduce the threat of suspicious mails to Internet of Things (IoT). This method provides rich information for email categorization. In the opinion of researchers the multi-view data has the higher possibility to achieve higher accuracy in comparison to single view data. Bahgat*et. al.* (2018) [7] used syntactic feature selection. This is relatively new technique. The study concluded that this approach takes less time and a significant performance with higher accuracy.

Gupta et. al. (2017) [8] studied the issues of online websites and service providers using single mail-Id to address the issues and concerns of the customers. They applied artificial neural network (ANN) in their work correctly identify spam emails. This method showed text based classification of emails Kumaresan*et.al.* (2017) [9] used hybrid kernel based support vector machine learning in the study. The features are extracted from both text and images. TF-term-frequency is used for textual features and the image dependent wavelet moment is considered for classification of emails. Results claimed the accuracy up to 97.235%. Alkhereyf *et. al.* (2017) [10] work focused on extracting to lexical features in addition to data from social networks features. Enron and Avocado email data set are used for experimentation purpose. SVM and Extra-Trees classifiers are applied on these features to compare the results and it is concluded that SVM perform better in term of accuracy.

3. MOTIVATION

Email application users send and receive large number of mails on daily basis. Over the period

- Multiclass classifier is used. Hence reduced data set , training time, cost and efficiency
- More Genre can be easily incorporated
- No need of negative training data set, because it works one class against the other classes.
- To effectively manage the large number of emails
- To improve upon the existing methods of email classification.
- To save time in data organization
- To gain better user experience

4. PROPOSEDSCHEME

To overcome the efforts and problems of manual and semiautomatic classification methods, author proposes an efficient scheme for automatic classification of emails based on the predefined genres as shown in fig3. In the proposed scheme an algorithms is developed to extract the features from the text of email. Dataset used for this purpose is Enron email Dataset.

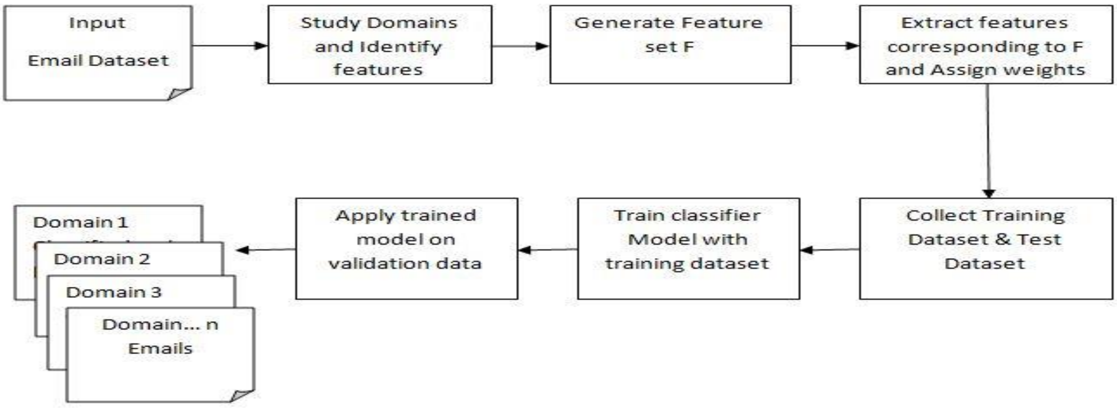

Input
Email Dataset
Study Domains and Identify features
Generate Feature set F
Extract features corresponding to F and Assign weights
Collect Training Dataset & Test Dataset
Train classifier Model with training dataset
Apply trained model on validation data
Domain 1
Domain 2
Domain 3
Domain... n Emails

Fig3: Proposed Methodology

5. FEATURES SOURCES

Mostly email data is in the form of bags of words present in the header "subject" and the body "content" of email. So the header and body of email are good source of features. Mostly text classification techniques are applied on such kind of data.

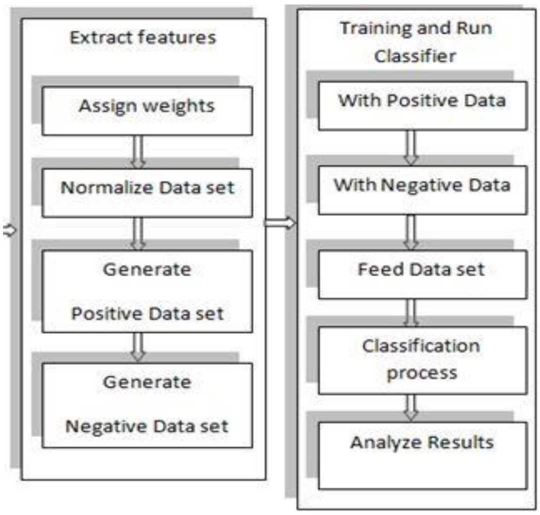


Fig4. Feature extraction and Training, Testing data generation

Feature from the header and body are extracted separately for the purpose of experimentation as shown in fig4 to analyze the contribution of just header in deciding the genre of mail Bags of words are extracted from the text and body by removing the stop words and punctuations. A list of words is created. Then these individual words are processed for obtaining the stemmed words. A final list of stemmed words is generated.

6. GENRES

Predefined genres are proposed for experimentation purpose. These genres are official, personal, promotional, confidential and others. Corresponding to each genre bags of words are generated in the form of genre keyword list. These lists are again stemmed to improve the results. Dimensions of the feature set are reduced by removing the irrelevant features by assigning these zero weight. This approach is iteratively applied on the model to select the most promising feature set to get high performance and accuracy.

7. CLASSIFIER MODEL

Support vector machines (SVM) are the unsupervised learning algorithm that learns some features from the dataset feed as training data on the basis of decision planes to generate decision boundaries. A decision plane separates between a set of items belonging to number of categories. SVM is capable to solve non-linear high dimensional and global optimum problems effectively. SVM is initially introduced by Vapnik et al. [11][14] as a semi supervised machine learning tools. It is extensively used for categorization and classification of data. The function f(x) : w_t x + b is defined; w is weight vector ; b is bias [11]. The value of b displaces f(x) away from the origin as shown in fig5.

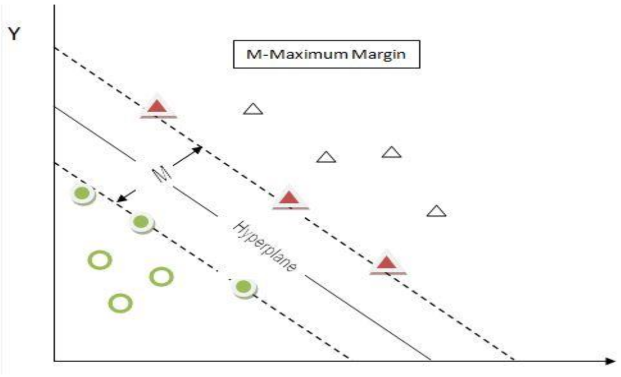


Fig5 Support vectors

New data is feed to use the previously learned features to decide the category of the data. Support vector machine (SVM) can also be used to train the multiclass classifier. In proposed work the unique model is trained as one class vs. other class. This will reduce the requirement of negative training data set.

8. ALGORITHM

Following four algorithms are developed for the proposed research work.

8.1. Algorithm1-Genre Selection

Genre keywords extraction

Input: Genre
Output: List of relevant Genre Keywords: $GKWL_i$

1. i=0;
2. Manually study the genre domain D_i,
3. Identify the most popular keywords corresponding to/used for genre based on term frequency-inverse document frequency (tf–Idf).
4. Store these words in genre keyword list (GKWL[i])
5. Apply word stemming (GKWL[i])
6. Select next genre; i++; Repeat until all predefined genre are traversed
7. Go to step 2
8. Stop.

8.2. Algorithm2

Feature extraction and feature sets generation

Input: Email text
Output: Set of features: EWL_i

1. Repeat
2. identify header, subject and body text

```
3.  Extract header, subject and Body text
4.  Remove stop words from each tag
5.  Generate list of words header:L1, Subject: L2, Body:L3
6.  Apply Word stemming  to (L1, L2 & L3)
7.  Generate Email Word List
EWL[i]={L1 U L2 U L3}.
I++;
8.  Until all dataset is traversed
9.  Stop.
```

8.3 Algorithm3

Feature vector generation

Input: Email Word List: EWL_i
Output: Feature vector weight generation :FW_i

```
1.  Let G ←{G1, G2, …Gn) ∀ Gi E Domain $D_i$
2.  Let $Fi \leftarrow \{f1, f2 ...fn\}$ ∀$Fi$ ∈Feature set where $fi$ represent feature
3.  Select $G_i$.
4.  Apply each genre word $GW_i$ ∀ GWL[i] list on EWL[i].
5.  Find the frequency of word in stemmed email dictionary word list EWL[i].
6.  Store the frequency of word in feature vector $f_i$ as weight $W_i$
7.  i++
8.  Repeat the step 2. – (for each genre list GWL[i] and generate the corresponding genre feature vector in form of word frequency count.)
9.  $Fi\ U\ \{f1, f2 ...fn\}$.
10. Stop
11. Separate training dataset and test dataset in SVM readable format
```

8.4. Algorithm 4

Multiclass classifier: Email genre categorization. It works one class against the other.

Input: Training set is represented as
$T_S = \{(X, Yi)|1 \leq i \leq n\}$
Output: To predict a label for each unknown dataset

```
1 Input space $X \leftarrow S^{Ts}$
2 Let $Y = \{1, 2, 3 ...Gn\}$ Possible labels
3 x$i \in X$  is a single instance and G$i \subseteq G$ is  the class set associated with $xi$
4 $Bi \leftarrow \{B1, B2 ...Bn\}$ $Bi$ ∈Binary classifier
5 Train $B_i$ ∀Gi
6 Generate a multi-class classifier based on maximum output
7 Go **for** j=1 ... Gn **do**
8 Use the sgn function argmax $gj$ where  $gj(x) = \sum_{i=1}^{n} yi\,\alpha ik(x, xi) + bj$
9 end for
10 Assign x to the class with largest confidence value.
11 Stop
```

Algorithm1 describes the method for predefined genre selection and generation of genre related keywords. Alogrithm2 is used for Feature extraction and feature sets generation from the email data set. Algorithm3 discuss the method of feature set generation and assigning weight to the features which are used for genre classification. Alogithm4 is developed to create multiclass classifier which operates on the principle of one genre against the other genre. Sigmoid function is applied with largest confidence value to decide the class of email. It use five-fold Cross validation technique to crosscheck the output.

9. DATASET ENRON

This dataset was prepared by the CALO Project. Enron contains data organized into folders and it is used as a resource for research purpose. This email dataset is in public domain [15]. The reason other datasets are not public is because of privacy concerns. Only few folder of this data set are used for the purpose of proposed work to collect approximately 5000 samples. Following genre labels as shown in table1 were used in the dataset

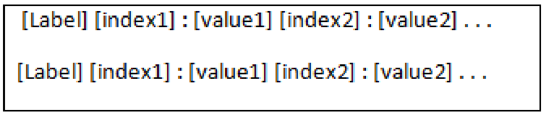
Table1: Genre Category label

Genre Label	Category
Official	1
Personal	2
Promotional	3
Confidential	4
Others	5

Dataset feature vector are scaled[14] in range [-1 , 1]. Scaled Training data set is feed to the SVM for training purposes.

10. EXPERIMENTATION RESULTS AND DISCUSSION

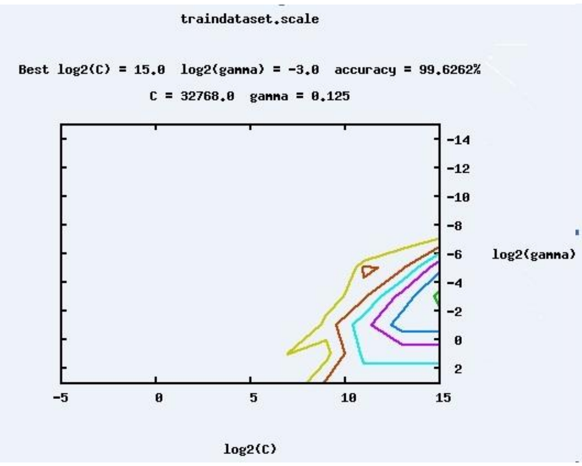
LIBSVM Support vector machine tool[12] is used to simulate the results and data is converted to SVM Data format. Training and test data sets are organized in the following format for the purpose of using LibSVM tool.



```
[Label] [index1] : [value1] [index2] : [value2] . . .

[Label] [index1] : [value1] [index2] : [value2] . . .
```

Researchers applied the liner, polynomial and Radial kernel function with different optimum parameter as shown in figures [6,7,8] respectively.



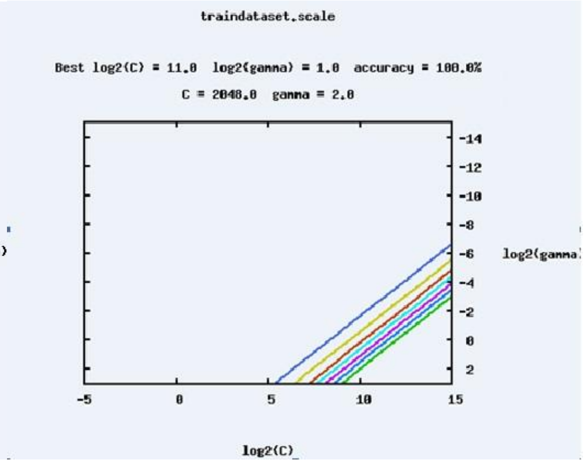


Fig6: Radial kernel performance

Fig7: Polynomial Kernel performance

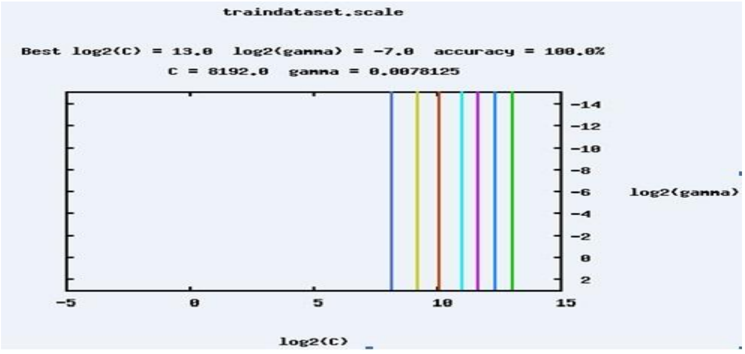


Fig8: Linear Kernel performance

Final the highest performing linear kernel is picked to further improve the performance by tuning the cost, gamma and error. Table2 and fig9 concludes the cross validation accuracy of linear kernel.

Table2: Cross Validation accuracy

Genre Label	Training Dataset CV Acc	Random Samples Data CV Acc
Official	97	91
Personal	87	82
Promotional	93	89
Confidential	92	79
Others	95	92

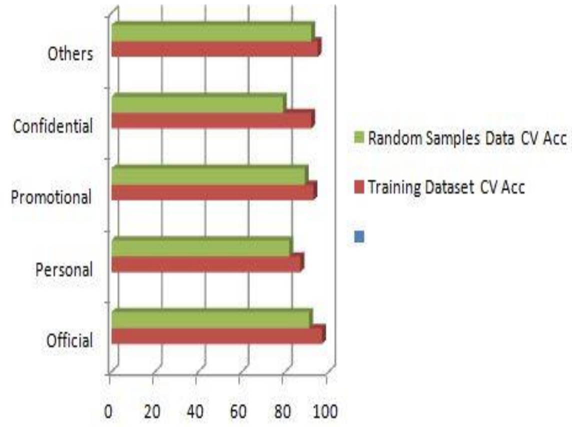


Fig9: CV Acc: Enron data and Random data

Multiclass classifier finally achieved satisfactory level cross validation accuracy with respect to predefined genre. In case of Enron data as shown in Table3 -the genre CV for the official email is 97%, personal is 87 %, Promotional is 93%, and confidential is 92%. Whereas random samples CV is 91% for official genre, 82% for personal, 89% promotional, 79% confidential and 92% others. The data is further analyzed and corresponding test data set is applied on the optimally trained model to find the evaluation metric precision (P), Recall (R), Accuracy (A) and F1-measure to rate the actual performance of model and to increase the confidence of the proposed work. The results described in fig10 show this work are able to achieve a higher level of accuracy.

Table3: Test Dataset1-Enron results

Domain Class	Test Dataset1-Enron			
	P	R	A	F1

Official	0.88	0.82	87	0.84
Personal	0.93	0.92	92	0.91
Promotional	0.89	0.78	88	0.84
Confidential	0.86	0.71	87	0.85
Others	0.91	0.89	91	0.89

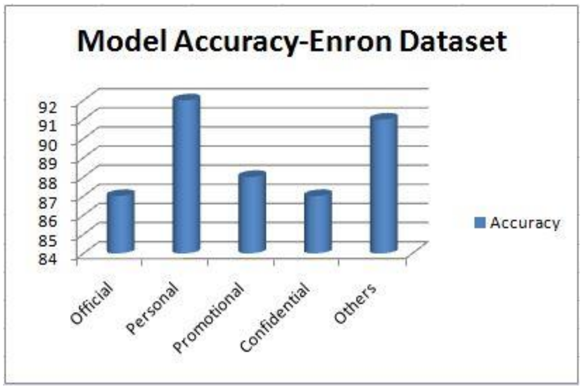


Fig10: Genre wise Accuracy-Dataset1

On the Enron- P, R , F1 and accuracy (Acc) is personal genre score is 0.93, 0.92, 0.91 and 92 % respectively. On random samples these values as shown in Table4 and fig11 are 0.90, 0.89, 0.89 and 89%.

Table4: Test Dataset2- Random Samples results

Domain Class	Test Dataset2-Random Samples (1200)			
	P	R	A	F1
Official	0.81	0.78	83	0.79
Personal	0.90	0.89	89	0.89
Promotional	0.91	0.88	89	0.89
Confidential	0.83	0.82	87	0.82
Others	0.89	0.88	90	0.88

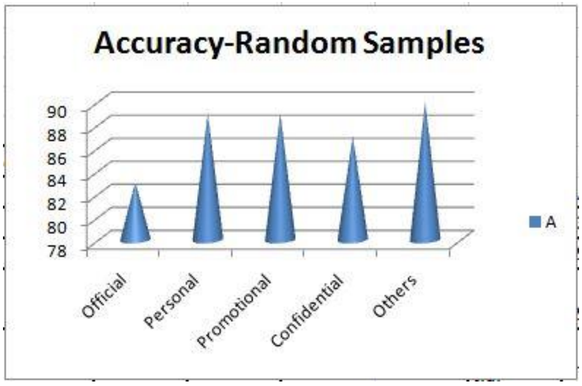


Fig11: Genre wise Accuracy-Dataset2

Moreover no negative data is required for classification because the multi classifier work on the unique principle of one class against the other classes. This multiclass classifier can be applied for spam [1] and fraud [7] email detection. This scheme can easily incorporate new genres.

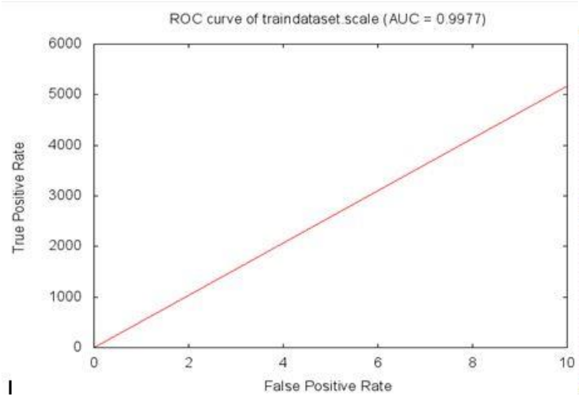


Fig12: AUC curve-training data

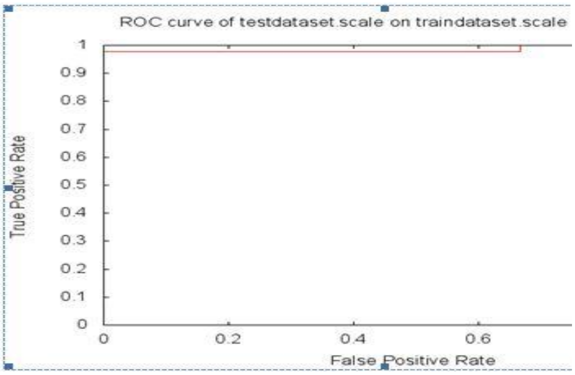


Fig13: ROC curve-test data

Hence, the research work may be used in categorization of large number of categories of emails. Overall results yield the superior performance as indicated in figures [12,13] through Receiver operating curve (ROC).

Similar research work is attempted by other researchers as discussed in section 2 and summarized in below table5.

Table5: Summary Of Review

Author& Citation	**Year**	**Technology used**	**Aspect of study**	**Outcome**
Mohammad *al.* [1]	2020	Human based spam classification approach	A technique that could highlight spam emails classifications using data mining and machine learning approaches.	Ensemble based Lifelong Classification using Adjustable Dataset Partitioning (ELCADP) out-performed all other contrasted stream mining algorithms
Saidani*et. al.* [2]	2020	Semantic-Based Classification Approach	To improve the accuracy of spam detection	Proposed method enables a better spam detection compared to existing methods based on Bag-of-Words (BoW) and semantic content
Gomes *et. al.* [3]	2017	Naive Bayes and Hidden Markov Model (HMM)	To identify the best technique among the two of them.	HMM turned out to be a better algorithm in comparison to the other available.
Chen *et. al.* [4]	2019	Long-Short-Term-Memory model	Correct classification of spam mails.	The results depicted that the proposed model performed better than standard CNNs and RNNs on email classification task
Saini *et.al.* [5]	2018	Cascaded self-organizing map (SOM) architecture	To overcome the issues related to no labeled data.	Proposed model performed well in email classification compared to standard classification.
Li *et. al.* [6]	2019	E-mail classification approach based on multi-view disagreement-based semi-supervised learning	To nullify the effect of spam mails on Internet of Things (IoT)	Results revealed that the use of multi-view data had the possibility to accurate email classification in contrast to single-view data, making this approach more effective.
Bahgat*et. al.* [7]	2018	Efficient email filtering approach	To identify the spam mails mainly use syntactic feature selection	The results depicted that the proposed approach worked had a highly significant performance with higher accuracy and less time

				compared to other related works.
Gupta *et.al.* [8]	2017	Artificial neural network (ANN) model	Issues related to online shopping websites or service provider that have single email-id where customers can send their query, concern etc.	As a result of the proposed system, typical Text Classification or Categorization can be identified.
Kumaresan*et.al.* [9]	2017	S-Cuckoo and hybrid kernel based support vector machine (HKSVM)	Threat of spam mails playing a vital role in recent days due to the uncontrollable growth happening in the electronic media.	Experimental results depicted that the proposed spam classification framework has outperformed other methods
Alkhereyf *et. al.* [10]	2017	Support vector machine	Extra-Trees classifiers developed and compared with SVM performance	A comparative between the performance of SVM and Extra-Trees classifiers have been presented within the study.

Some of the researcher were able to produce relevant results. But the proposed research work is producing the better results and it is more generic. It can be extended to easily incorporate new categories. Also this work does not require the negative training data.

10. CONCLUSION

Email system hugely contributed in transforming the world into global village by providing the fast and reliable source of the communication. All over the world people are dependent on various email services. Users of these services use email system for personal, professional, social, promotional communication. User’s email boxes are flooded with dozens of mails each day, which makes them uncomfortable in tracing out the important messages. In order to browse whole lot of mails user may miss or ignore some important communication and deadlines. So it becomes obvious to develop a scheme which can effectively categorize these mails into predefined genres to make enhanced user experience and improve the productivity. In this direction proposed scheme is developed to come up with a framework for automatic classification of emails into set genres. To achieve the target machine learning model is developed which is based on carefully selected variety of features to generate multi-attribute criteria. The experiment setup on sample and random data sets produced promising results in terms of overall accuracy of up 90% and efficiency. Proposed scheme used SVM as machine learning tool. Useful and most relevant feature set is the most important aspect for the success of the system and trained model.

11. FUTURE SCOPE

The scope of this work is going to be extended by considering more features and genres to improve the results. Further, work can be extended to accommodate regional language stuff. Language specific keywords and feature sets needs to be explored for this purpose. Idea is to eventually develop multilingual framework to fit the requirements of the modern societies by enhancing the user experience.

REFERENCES

1. R. M. A. Mohammad, A lifelong spam emails classification model, Applied Computing and Informatics.
2. N. Saidani, K. Adi, M. S. Allili, A semantic-based classification approach for an enhanced spam detection, Computers & Security (2020) 101716.
3. S. R. Gomes, S. G. Saroar, M. Mosfaiul, A. Telot, B. N. Khan, A. Chakrabarty, M. Mostakim, A comparative approach to email classification using naive bayes classifier and hidden markov model, in: 2017 4th International Conference on Advances in Electrical Engineering (ICAEE), IEEE, 2017, pp. 482–487.
4. Hiremath, Basavaraj, and S. C. Prasannakumar. "Automated Evaluation Of Breast Cancer Detection Using Svm Classifier." *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), 5 (1), 11* 20 (2015).

5. Z. Chen, R. Tao, X. Wu, Z. Wei, X. Luo, Active learning for spam email classification, in: Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence, 2019,pp. 457–461.

6. N. Saini, S. Saha, P. Bhattacharyya, Cascaded som: an improved technique for automatic email, classification in: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–8.

7. Danthala, S. W. E. T. H. A., et al. "Robotic Manipulator Control by using Machine Learning Algorithms: A Review." *Int J Mech Prod Eng Res Dev* 8.5 (2018): 305-310.

8. W. Li, W. Meng, Z. Tan, Y. Xiang, Design of multi-view based email classification for iot systems via semi-supervised learning, Journal of Network and Computer Applications 128 (2019) 56–63.

9. E. M. Bahgat, S. Rady,W. Gad, I. F. Moawad, Efficient email classification approach based on semantic methods, Ain Shams Engineering Journal 9 (4) (2018) 3259–3269.

10. D. K. Gupta, S. Goyal, Email classification into relevant category using neural networks, arXiv preprint arXiv:1802.03971.

11. BAGUL, PRIYANKA, and Leena Ragha. "OFFLINE SIGNATURE VERIFICATION USING HU'S MOMENT AND GABOR WAVELET TRANSFORM." *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN (P)*: 2249-6831.

12. T. Kumaresan, S. Saravanakumar, R. Balamurugan, Visual and textual features based email spam classification using s-cuckoo search and hybrid kernel support vector machine, Cluster Computing 22 (1) (2019) 33–46.

13. Jayaram, B., et al. "A Survey On Social Media Data Analytics And Cloud Computing Tools." *International Journal of Mechanical and Production Engineering Research and Development, 8 (3), 243* 254 (2018).

14. S. Alkhereyf, O. Rambow, Work hard, play hard: Email classification on the avocado and enron corpora, in: Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing, 2017, pp. 57–65.

15. Jayaram, B., et al. "A Survey On Social Media Data Analytics And Cloud Computing Tools." *International Journal of Mechanical and Production Engineering Research and Development, 8 (3), 243* 254 (2018).

16. Cortes C., & Vapnik V. (1995). Support vector networks. Machine Learning, 20, 273–297.

17. Chang C. C., & Lin C. J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2(27), 1–27. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

18. Boser B. E., Guyon I. M., & Vapnik V. N. (1992). A training algorithm for optimal margin classifiers. In D. Haussler (Ed.), 5th annual ACM workshop on colt (pp. 144–152), Pittsburgh, PA.

19. Jayaram, B., et al. "A Survey On Social Media Data Analytics And Cloud Computing Tools." *International Journal of Mechanical and Production Engineering Research and Development, 8 (3), 243* 254 (2018).

20. Chapelle O. (2007). Training a support vector machine in the primal. In L. Bottou, O. Chapelle, D. DeCoste, & J. Weston (Eds.), Large scale kernel machines (29–50), Cambridge: MIT Press.

21. [15] Enron-dataset-www.kaggle.com/wcukierski/enron-email-dataset