An excellent classifier is evaluated not only by classification accuracy but also by its rapid false alarm detection with less number of features. This research presents the effects of using features selected by the four feature subset search methods, i.e. Genetic, Greedy Stepwise, Best First and Rank Search, on popular Machine Learning Classifiers like Bayesian, Naïve Bayes, Support Vector Machine, Genetic Algorithm, J48 and Random Forest. Tests were performed on three different publicly available spam email corpuses: "Enron", "SpamAssassin" and "LingSpam".

6.1 Best combination of Machine Learning and Features Subset Search

The description and experimental design of this study is given below:

6.1.1 Aim of this study

This study is done to evaluate a best pair of Machine Learning Classifier and Feature Subset Search Technique.

6.1.2 Corpora for this study

Three different corpuses taken from three different sources are preferred for this study discussed in Chapter 3. Our main tests are performed on "Enron email" corpus and further "SpamAssassin" and "LingSpam" corpuses are used for validation of the results coming from the first corpus.

6.1.3 Feature Subset Search Techniques

Genetic, Greedy Subset, Best first, and Rank search methods are used for this study to give a subset of less number of most informative features.

[8] Trivedi, Shrawan Kumar, and Shubhamoy Dey. "Interaction between feature subset selection techniques and machine learning classifiers for detecting unsolicited emails." *ACM SIGAPP Applied Computing Review* 14.1 (2014): 53-61.

6.1.4 Classifier for this study

Bayesian, Naïve Bayes (*NB*), Support Vector Machine (*SVM*), Genetic Algorithms (*GA*), Simple Decision Tree (*J48*), and Random Forest (*RF*) are preferred in this study.

6.1.5 System Design

After pre-processing, Greedy Stepwise, Best First and Rank Search methods are used for selecting the most informative features. All corpuses are split for performing the training and testing where 66% instances are taken for training on the selected features and remaining 34% kept for testing.

6.1.6 Evaluation Metrics

Classification accuracy, F-Value and False Positive Rate are the evaluation metrics of this study.

6.1.7 Results and Analysis

This study presents a comparative analysis of various Machine Learning Classifiers that are tested on different numbers of most informative features. Percentage Accuracy, F-Value and False Positive Rate are the measures used for analysis. For clear understanding, this analysis is divided into three segments. The first segment deals with the analysis of Machine Learning Classifiers, the second segment analyses the feature subset search methods, and the last segment uses the False Positive rates for evaluating sensitivity of the classifiers.

Table 31. Accuracy and F-value of Enron corpus

In Percentage		**Genetic**	**Bayesian**	**NB**	**SVM**	**J48**	**RF**
Genetic	Acc	80.4	85.6	84.8	87.1	86	86.6
	F-Value	80.4	85.6	84.8	87.1	86.1	86.6
Greedy	Acc	87.6	93.0	94.0	**94.2**	92.1	93.8
	F-Value	87.5	93.1	93.9	**94.3**	92.2	93.9
Best First	Acc	80.7	92.1	91.2	**94.1**	92.6	94.0
	F-Value	80.7	92.2	91.2	**94.2**	92.6	94.1
Rank	Acc	80.8	92.0	91.4	93.8	91.4	93.7
	F-Value	80.8	92.1	91.5	93.8	91.4	93.8

Table 32. Accuracy and F-value for SpamAssassin corpus

In Percentage		Genetic	Bayesian	NB	SVM	J48	RF
Genetic	Acc	95.2	91.9	91.2	96.2	95.7	96.5
	F-Value	95.2	91.9	91.2	96.2	95.8	96.5
Greedy	Acc	96.4	97.1	96.6	**97.8**	97.9	**98.4**
	F-Value	96.4	97.1	96.7	**97.8**	97.9	**98.4**
Best First	Acc	95	92.8	93.1	**97.9**	96.3	**98.2**
	F-Value	95.1	92.8	93.2	**97.9**	96.4	**98.2**
Rank	Acc	95.4	92.2	94.4	97.5	96.4	97.6
	F-Value	95.5	92.2	94.5	97.5	96.4	97.6

Table 33. Accuracy and F-value of classifiers tested on LingSpam corpus

In Percentage		Genetic	Bayesian	NB	SVM	J48	RF
Genetic	Acc	89.5	89.5	89.7	92	89.4	90.1
	F-Value	89.5	89.5	89.8	92.1	89.5	90.1
Greedy	Acc	93.2	97.7	97.8	**97.8**	95.7	96.5
	F-Value	93.3	97.8	97.8	**97.8**	95.7	96.6
Best First	Acc	93.1	97.8	97.1	**97.5**	96.5	96.6
	F-Value	93.1	97.8	97.2	**97.5**	96.6	96.6
Rank	Acc	93.2	96.2	96	96.9	94.2	95.1
	F-Value	93.2	96.3	96.1	97	94.2	95.1

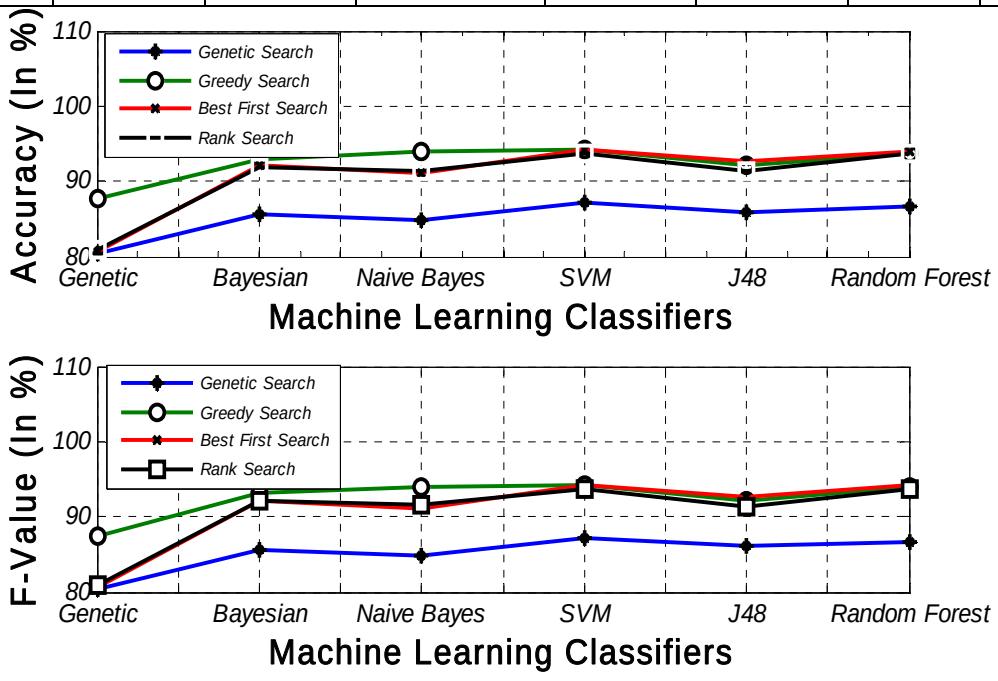

Fig 32. Accuracy and F-value for Enron Corpus

6.1.7.1 Analysis of Machine Learning Classifiers

The results of the classifiers tested on the Enron corpus is shown in Table 31 and Figure 32, which demonstrate that Support Vector Machine is the most accurate classifier amongst the other. In this case, the classification accuracy comes in between 87.1% and 94.2%. However, Random Forest classifier (with classification accuracy 86.6% to 94.1%) is found to be the second best classifier whose results were proximate to SVM. The Genetic Classifier is found to be the worst in terms of accuracy varying between 80.4% and 87.6 %. The Bayesian and Naïve Bayes are the third and fourth best classifier respectively with accuracy in between 85.6% and 93.1% for Bayesian and 84.8% and 93.9% for Naïve Bayes.

Testing of the same classifiers on the SpamAssassin corpus validate the results obtained from the Enron corpus. Results of the experiments on the SpamAssassin corpus are shown in Table 32 and Figure 33. In this case, again SVM and Random forest are proven to be excellent with classification accuracy in 96.2% to 97.8% for SVM and 96.5% to 98.4% for Random Forest. Genetic algorithm (with classification accuracy 95% to 96.4%) is again predicted to be weak amongst all.

The same test performed on LingSpam corpus also validates the above results. The results of LingSpam (Table 33 and Figure 34) show that SVM (with classification accuracy 92% to 97.8%) is the best classifier amongst all whereas the results of Random Forest (with classification accuracy 90.1% to 96.6%) are proximate to the best one. Genetic classifier (with classification accuracy) is continually showing poor performance.

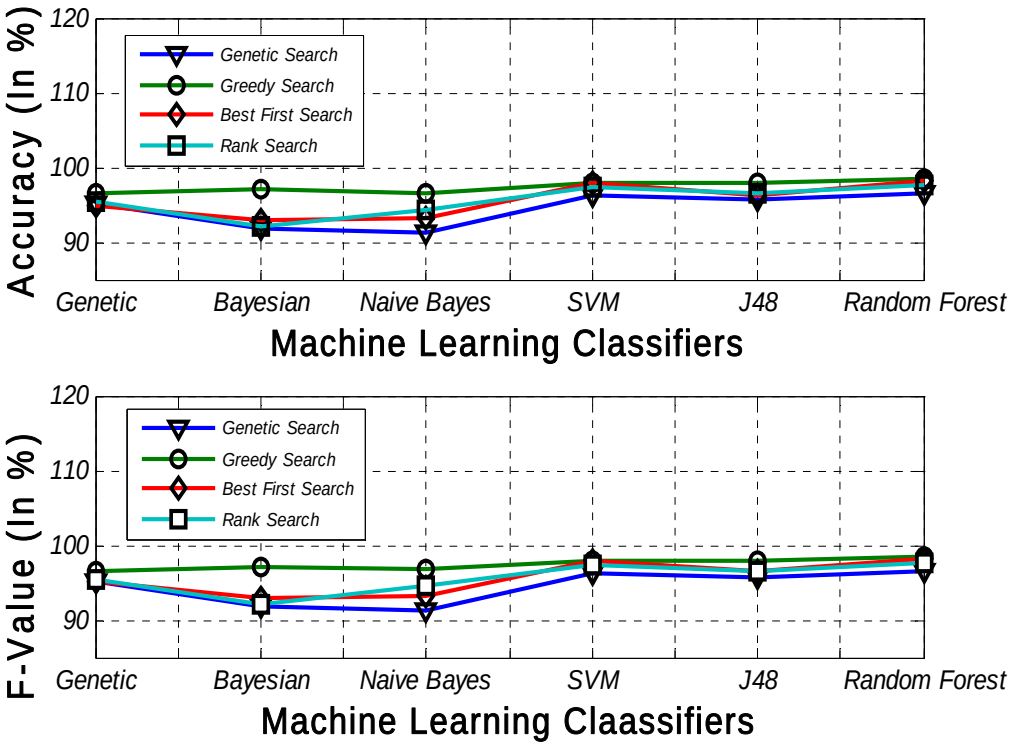

Fig 33. Accuracy and F-value for SpamAssassin Corpus

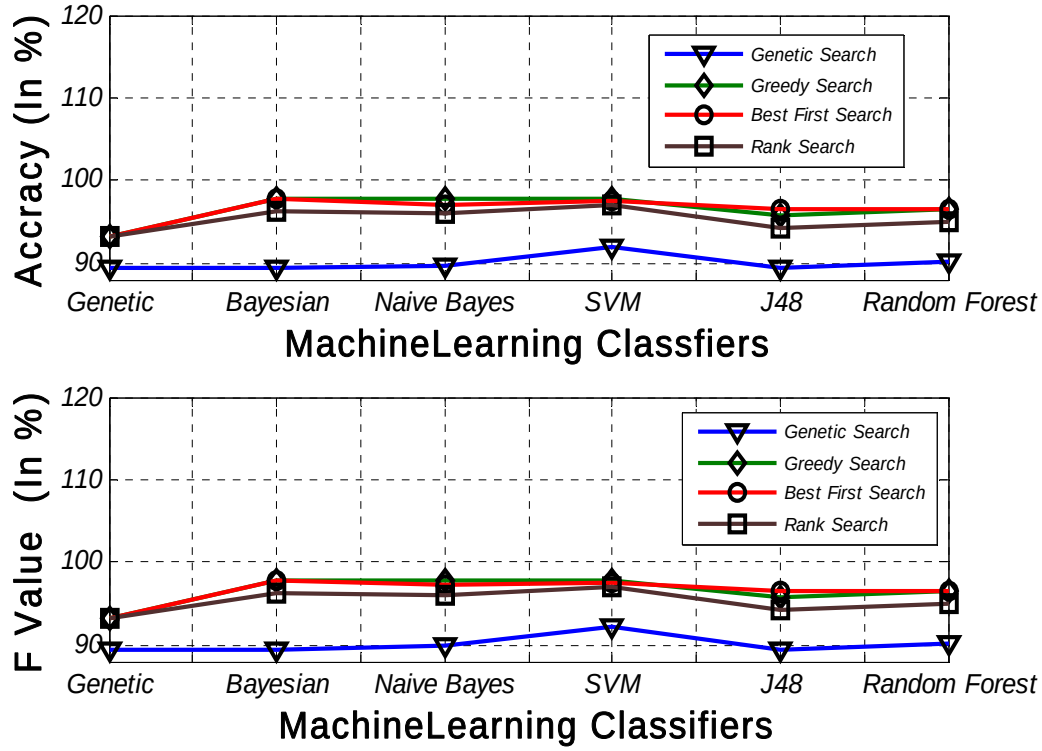

Fig 34. Accuracy and F-value for LingSpam Corpus

6.1.7.2 Analysis of Feature Selection Methods

As discussed in the preceding sections, four feature subset search methods are employed to obtain most informative feature subsets. By using Genetic, Greedy Step-Wise, Best First and Rank subset feature search method, the most informative features subset are selected. Initially, 48 best features out of 1500 initial created features for Enron corpus, 35 best features out of 1414 features for SpamAssassin corpus, and 50 best features out of 1658 features are selected for testing the concerned classifiers. The results present in Tables 31, 32, 33 and Figures 32, 33, 34, demonstrate that Greedy Step-Wise search method is the best for all the three corpuses with classification accuracy between 87.6% and 95.2% for Enron corpus, 96.4% and 97.8% for SpamAssassin corpus and 93.2% to 97.8% for LinSpam corpus. However, Best first search method is the second best amongst all with accuracy between 80.7% and 94.1% for Enron corpus, 92.8% and 98.2% for SpamAssassin corpus and 93.1% to 97.8% for LinSpam corpus.

However, features selected by Genetic search show weak results i.e. 80.4% to 87.1% for Enron corpus, 91.2% to 96.2% for SpamAssassin corpus and 89.5% to 90.1 for LingSpam corpus.

6.1.7.3 Analysis with False Positive Rate

Although some machine learning classifiers are observed to be excellent in overall classification accuracy but the possibility of misclassification of the positive instances could be high. Legitimate emails are considered to be important and if they get misclassified as Spam, it may lead to serious consequences. This problem can be well tackled by considering the False Positive Rate (FP Rate) that captures the rate of misclassified legitimate emails.

From Table 34 and Figure 35, it is clear that SVM and Bayesian Classifier perform better in terms of the FP rate. For these classifiers the FP Rate is low in all three corpuses (1.8% to

7.3% for Bayesian classifier and 2.6% to 7.3% for SVM on Enron corpus, 0.1% to 3.6% for Bayesian and 1% to 2.1% for SVM on SpamAssassin corpus as well as 0% to 18.1% for Bayesian and 1% to 7.3% for SVM on LingSpam corpus). The above results are for Genetic, Greedy Stepwise, Best First and Ranker search which indicate that the use of Greedy Stepwise search method for feature selection leads to lower FP rate.

Table 34. False Positive Rate for all corpuses

		Genetic	Bayesian	NB	SVM	J48	RF
Genetic Search	Enron	22.6	7.3	10.6	7.3	8.8	8.6
	SpamAssassin	3.5	0.1	0.4	2.1	3.6	3.1
	LingSpam	13.1	18.1	16.9	0.6	1.3	6.3
Greedy search	Enron	22.5	1.8	4.4	2.6	2.7	2.7
	SpamAssassin	2.4	3.6	4.5	1.0	2.2	1.7
	LingSpam	10.0	0.0	0.0	1.9	3.8	3.8
Best First Search	Enron	36.9	2.5	3.7	3.5	3.7	3.6
	SpamAssassin	4.5	0.7	1.1	1.9	3.2	1.5
	LingSpam	11.9	0.0	0.6	2.5	3.1	1.9
Rank Search	Enron	25.7	1.1	2.2	2.8	4.6	3.3
	SpamAssassin	4.1	0.1	0.6	2.4	3.0	2.2
	LingSpam	12.5	0.0	0.6	1.9	3.1	5.0

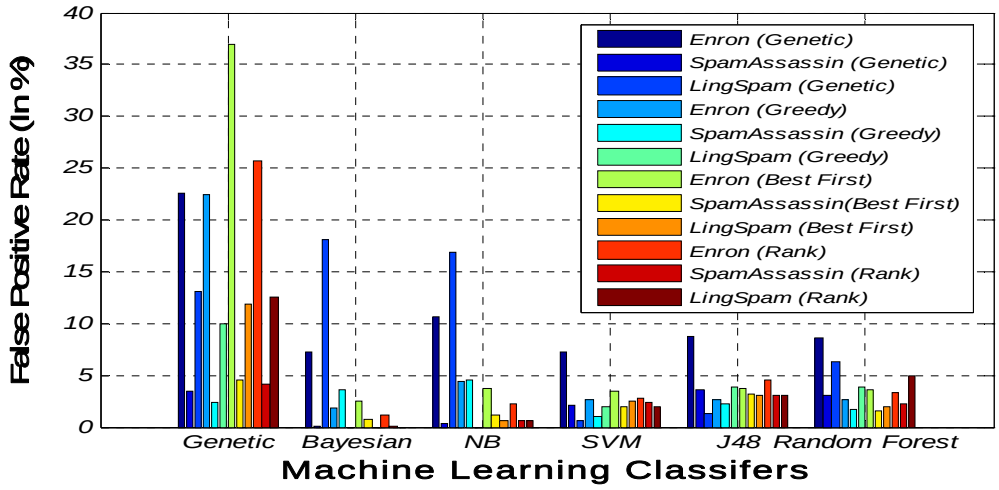

Fig 35. False Positive Rate for all corpuses

By considering classification accuracy and false positive rate together, SVM classifier with Greedy subset search method is identified to be an excellent combination.

6.2 Discussion

Achieving good classification accuracy of classifiers by using the minimum number of features has always been one of the primary research objectives in text classification. This study has done a comparative analysis of four feature subset search methods: Genetic, Greedy Stepwise, Best First, Rank search, and their interactions with some Machine Learning Classifiers, to find best pair of Feature Subset Selector and Machine Learning Classifier. The purpose of this study is successfully achieved. The results lead to the following conclusions: First, among the Machine Learning Classifiers examined, SVM has shown best classification accuracy and also the lowest False Positive Rate; Random Forest was second best; Second, Greedy Stepwise Search was found to be the best feature subset selector; Third, Greedy Stepwise Subset Selector with SVM classifier has been found to be an excellent pair for classification.

This chapter has presented a comparative analysis of various machine learning classifiers and observed the effect of different feature search methods on them. This study has archived its objective and obtained best pair of machine learning and feature subset search i.e. SVM with Greedy stepwise. Another problem with the machine learning research has been found weak learning during the training which happens due to sampling as well as learning errors. Next chapter demonstrates two different novel techniques for minimizing these errors. The first method presented a combining classifier with committee selection mechanism whereas second technique employs a novel ensemble-based approach.