CHAPTER 7

RESULTS AND DISCUSSION

The important point of any research work is the performance evaluation of the algorithms. This part of the research methodology proves that the algorithms proposed meet the research objectives and help to improve the process of application. In this research, enhanced algorithms that perform feature engineering and classification were proposed to obtain high spam review detection accuracy. The experiments were designed with the aim of studying the effectiveness of the algorithms and optimization methods proposed in each phase and were conducted in three stages.

- Stage 1 : Evaluate the feature engineering algorithms
- Stage 2 : Evaluate the enhanced classifier and enhanced ensemble classifier
- Stage 3 : Evaluate the hybrid classification systems

This chapter starts with the discussion of the dataset, followed by the performance metrics used in each stage of experiment. Finally, the results of experiments are tabulated and discussed to identify the best method suited for each step.

7.1. DATASETS

The review dataset is one of the iconic products provided by Amazon. It has instances collected over a period of more than 2 decades; jillions of consumers have shelled out more than couple of hundred million reviews put into words of their opinions and describing their understanding and experience. This makes the Amazon dataset rich with information and is an ideal source for researchers and academicians. This dataset was constructed to represent a sample of customer reviews and has a collection of reviews written along with its metadata collected from 1995-2015. It also includes details regarding products, ratings along with the plaintext review. The dataset used was downloaded from http://liu.cs.uic.edu/download/data and has 42.8 million reviews.

The second dataset used during experiments is the Yelp dataset obtained from http://www.yelp.com. The yelp.com is a crowd-sourced website the reviews local businesses and is similar to social networking sites which allow users to interact with each other. The site has heavy traffic and has more than 132 million visitors per month, who have written around half a billion reviews. The Yelp dataset has 16282 reviews and was downloaded from http://liu.cs.uic.edu/download/yelp_filter. Both the datasets were obtained from Professor Bing Liu (https://www.cs.uic.edu/~liub).

The datasets obtained were preprocessed to suit the need of this research. The review set was created by manually annotating text with positive and negative opinion words. Some small modifications in certain fields were done to perform easy extraction and analysis. For example, date was stored as "5 21, 2014" format, which was converted to 21-May-2014". The final preprocessed dataset thus had details stored in a fashion that is easily accessible to the algorithms. The final dataset had 11 attributes, namely, Reviewer ID, Reviewer Name, Product ID, Review Text, Date, Time, User Rating, Overall User Rating, Rank in Sale, Average Rating and Helpful Feedback, Number of Feedbacks by the reviewer, Number of Feedbacks for a Product.

Further, all instances whose total number of review is less than five were removed as they do not have any noteworthy activities. Similarly, if the number of unique reviewers is less than six, all the reviews related to that product were removed. All instances with missing values in any one of the fields were considered incomplete and were removed. Thus, after removal of such reviews, the size of the amazon dataset was 37108864 million reviews and 12107 reviews with yelp reviews. In order to reduce the complexity of Amazon dataset, reviews in the period of 2000-2015 only were considered. This reduction mechanism was used, because testing with the whole dataset was cumbersome and extremely slow. This considerably reduced the size of the dataset into a manageable set. Now the size reduced to 4123844 reviews of 1985 products written by 3079542 reviewers.

7.2. PERFORMANCE METRICS

The data in the confusion matrix is commonly used for evaluating the performance of classification systems (Sokolova and Lapalme, 2009; Demsar, 2006). Aconfusion matrix for two possible outcomes p (positive) and n (negative) is given in Figure 7.1.

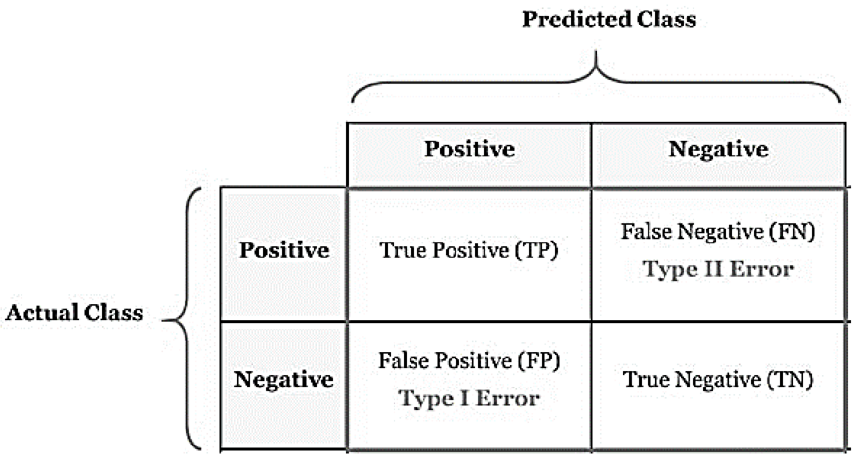

Figure 7.1 : Confusion Matrix

For evaluating findings, it is possible to use several measures related to classification. The most commonly used metrics are used to evaluate the algorithms presented in the research work. The various algorithms proposed are evaluated using five performance metrics, namely, precision, recall, F-measure, accuracy and speed. The first four of the performance metrics are estimated from the four measures obtained from the confusion matrix. They are, False Negative (FN), False Positive (FP), True Positive (TP) and True Negative (TN). In the definitions, 'positive' refers to 'correctly predicted' results, while 'negative' refers to 'wrongly predicted' results.

- **False Negative (FN) -** False negative is the ratio of number of signals where the algorithm fails to correctly identify the passing and failing students to the total number of students.

- **False Positive (FP) -** False positive is the ratio of number of students where the algorithm identifies the performance of a student wrongly to the total number of signals.

- **True Positive (TP) -** True positive is defined as the ratio of number of students performance correctly identified to the total number of students.

- **True Negative (TN) -** It is defined as the ratio of number of times the algorithm fails to correctly identify the performance category (fail/pass).

From the FN, FP, TN and TP, the four performance measuring values (precision, recall, F-measure, accuracy) are calculated (http://www.stat.psu.edu/online/courses/stat509/17_diagnos/17_diagnos_print.htm).

- **Precision**

Precision is the ratio of the quantity of exact predictions repossess to the total number of predictions. It is more often put across as a percentage and is calculated using Equation (7.1).

$$\text{Precision} = \frac{TP}{(TP + FP)} * 100 \quad (7.1)$$

- **Recall**

Recall (expressed in percentage) is the ratio of the number of correct predictions to the total number of correct predictions (Equation 7.2).

$$\text{Recall} = \frac{TP}{(TP + FN)} * 100 \quad (7.2)$$

- **F-Measure**

F-Measure is the harmonic mean between precision and recall and is measured using Equation (7.3).

$$F = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7.3)$$

- **Accuracy (%)**

Accuracy, expressed in percentage, is defined as the number of correct predictions to the total of predictions and is estimated using Equation (7.4).

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (7.4)$$

- **Speed**

The last performance metric is the speed parameter, which is used to calculate the execution time in terms of CPU time and is taken as the time period from the start of the algorithm execution until its termination. In other words, it is the time taken by the algorithm to finish its allotted task.

An ideal spam review detection system should have high detection accuracy. That is the ham/spam classification system should have high recall, precision, which in turn increases the F-Measure and accuracy. The main goal of the experiments is to find the combination of classification system that can achieve the above. To understand the amount of performance increase of the proposed algorithm when examine in contrast to the existing algorithms, another parameter called efficiency gain was used during analysis. Efficiency gain is defined as a measure that can numerically estimate the performance of two systems (generally, proposed and existing) and has been used several proposals (Al-Zoubi and Obeid, 2007). The efficiency gain, measured in percentage, is estimated using Equation (7.5).

$$\text{Efficiency Gain (\%)} = \frac{|\text{P - E}|}{\text{P}} x100 \qquad (7.5)$$

where P is the performance value obtained while using the proposed algorithm and E is the performance value obtained while using existing/conventional algorithms.

7.3. EVALUATION OF PHASE I ALGORITHMS

As mentioned earlier, stage 1 algorithms are the algorithms proposed in Phase I of the research methodology, which focused on feature engineering. It consists of two steps, the first is feature extraction and the second is feature selection. Three types of features, namely, review, reviewer and product, were extracted. An enhanced feature fusion and selection algorithm was proposed to construct the optimal feature vector. Thus, the predominant intension of stage 1 experiment is to understand the effect of using feature selection algorithm during ham/spam classification. The results obtained are compared with the existing and conventional counterparts in order to ascertain the efficiency of the algorithms in producing feature vectors which can effectively identify the type of review.

The feature selection algorithms are tested using three classifiers, namely, KNN, SVM and NB. The coding scheme used during discussion is given in Table 7.1.

TABLE 7.1

CODING SCHEME USED – FEATURE SELECTION

Code	Algorithm
NFS	No Feature Selection
FS_MRMR	Feature Selection Using MRMR Algorithm
FS_MIMG	Feature Selection Using MRMR_IG and MRMR_MI Algorithm
FS_ACO	Feature Selection Using ACO Algorithm
FS_ACO+GA	Feature Selection Using ACO + GA Algorithm
FS_MGA	Feature Selection Using FS_MIMG and ACO + GA Algorithm

The parameter settings used during the initialization of ACO and GA are given in Table 7.2.

TABLE 7.2

PARAMETERSETTINGS OF ACO AND GA

ACO		GA	
Parameter	**Value**	**Parameter**	**Value**
No. of Ants	10	Population Size	100
Max Iterations	500	Maximum Generation	500
Maximum Repeat Number	40	Mutation Probability	0.1
Trial Evaporation Rate	0.5	Crossover Probability	0.9
Pheromone Rate	2	Maximum Repeat Number	50
Desirability Rate	2	Selection Method	Roulette Wheel
Initial Pheromone Rate	2	Mutation Type	Swap Mutation
		Crossover Type	Single Point Crossover

Figures 7.2 to 7.5 show the precision, recall and F-Measure of the ham/spam detection system designed with KNN, SVM and NB classifiers when tested with Amazon

and Yelp datasets. The performance of the system was evaluated while varying the feature selection algorithm used. The results were compared with systems that used no feature selection algorithm and existing algorithms.

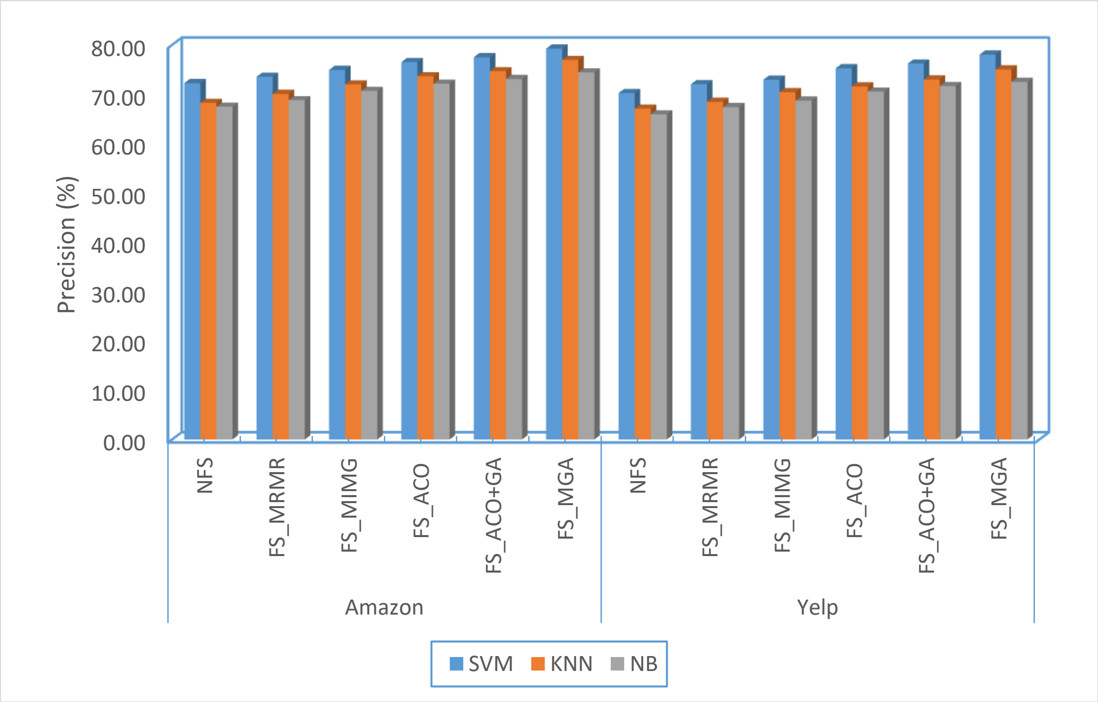

Figure 7.2 : Precision (%) of Feature Selection Algorithms

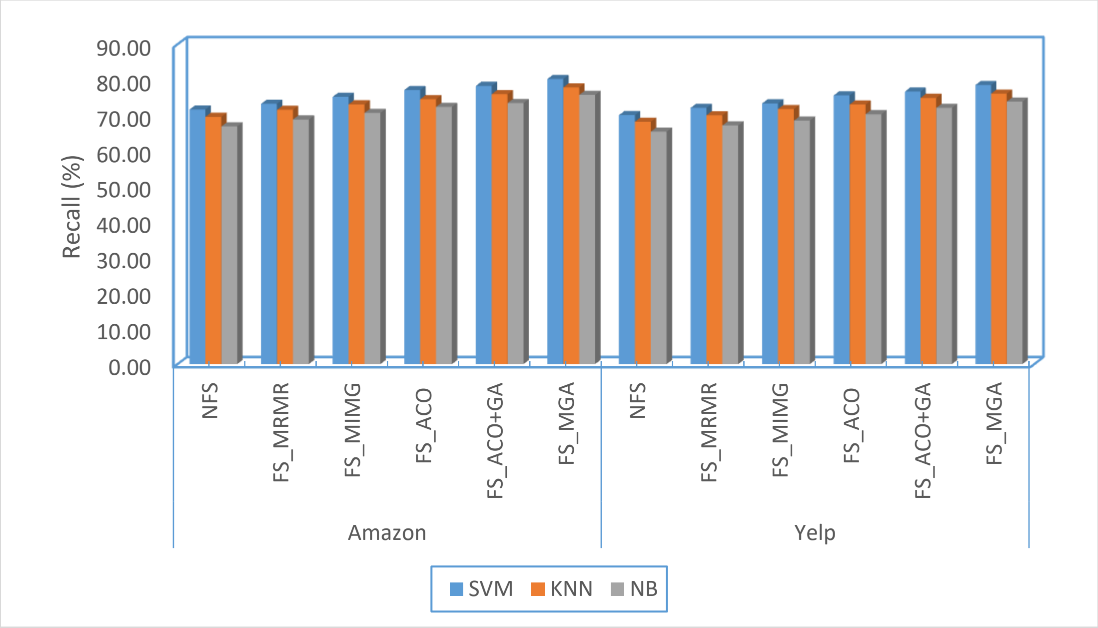

Figure 7.3 : Recall (%) of Feature Selection Algorithms

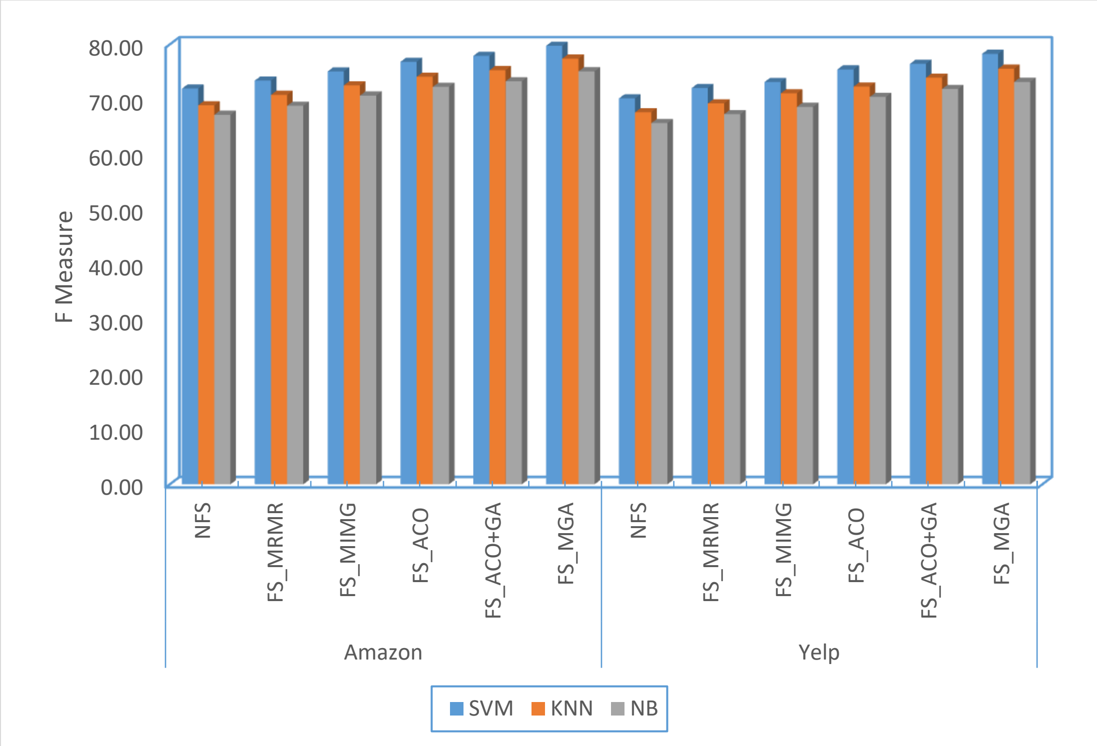

Figure 7.4 : F Measure (%) of Feature Selection Algorithms

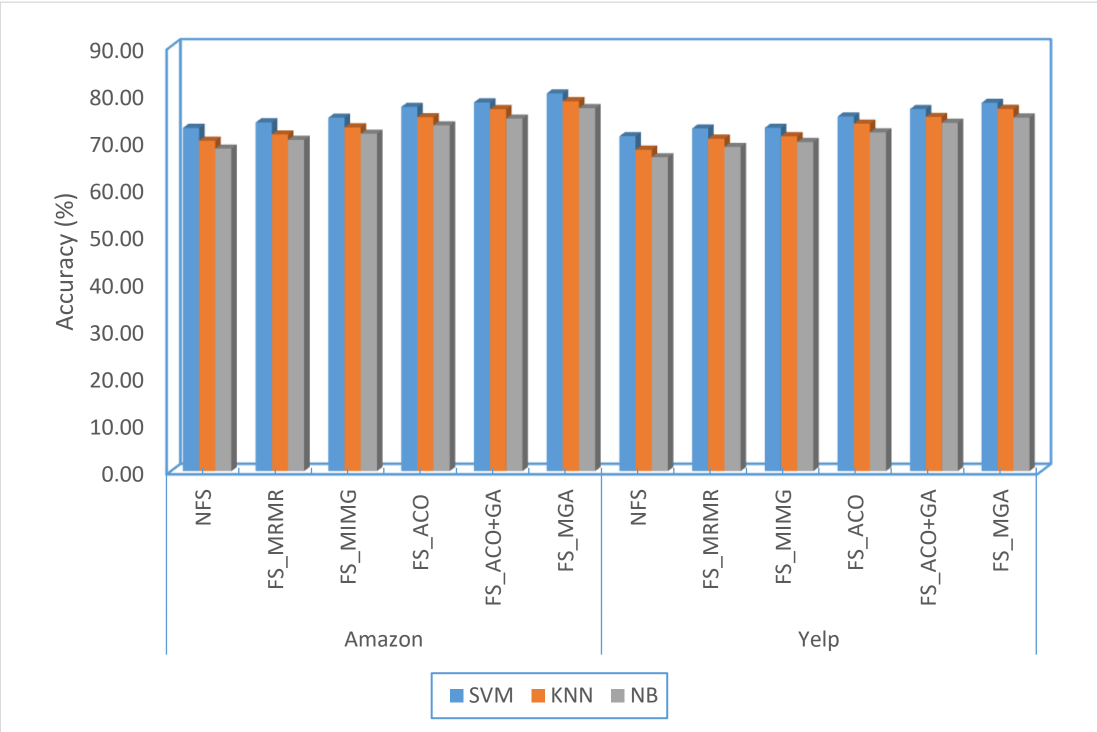

Figure 7.5 : Accuracy (%) of Feature Selection Algorithms

From the experimental results, it could be noted that the usage of any feature selection algorithm could improve the performance of OSRD system. However, maximum efficiency was obtained while using FS_MGA algorithm.

The OSRD system produced a precision of 72.27%, 68.27% and 67.50% (Amazon) and 70.26%, 67.11% and 65.94% (Yelp) when tested with SVM, KNN and NB classifiers respectively, when no feature selection algorithm was used. This increased to 79.25%, 76.94% and 74.43% (Amazon dataset) and 78.04%, 75.04% and 72.51% (Yelp dataset) respectively, while using the proposed FS_MGA algorithm.

Similarly, the OSRD system showed a high recall value of 71.68% (SVM), 69.60% (KNN) and 66.91% (NB) (Amazon) and 70.09% (SVM), 68.23% (KNN) and 65.43% (NB) (Yelp) was obtained when the raw feature set was used. The usage of the proposed MGA feature selection algorithm increased the recall value to 80.25% (SVM), 77.88% (KNN) and 75.76% (NB) (Amazon dataset) and 78.53% (SVM), 76.12% (KNN) and 73.87% (NB) (Yelp dataset) respectively.

F-measure performance metric is the harmonic mean of precision and recall. The precision and recall is considered in one number in order to determine single score that stabilizes both. The results related to this metric show that the proposed MGA feature selection algorithm is both precise (that is, how many review instances are classified correctly and robust (ability to produce stable results, where the testing results are close to the training results). While using Amazon dataset, the MGA algorithm produced an F-Measure of 79.75% (SVM), 77.41% (KNN) and 75.09% (NB). On the other hand, the same systems produced 71.97% (SVM), 68.93% (KNN) and 67.20% (NB) respectively. When tested with Yelp dataset, the F-Measure of MGA algorithm was 78.29%, 75.58% and 7318% when tested with SVM, KNN and NB classifiers respectively. On the other hand, the same classifiers showed a reduced performance of 70.18%, 67.66% and 65.69% respectively, when no feature selection algorithm was used.

A similar trend was envisaged with accuracy also. When no feature selection algorithm was used, the SVM, KNN and NB classifiers, while using Amazon dataset, produced an accuracy of 72.75% (SVM), 70.03 (KNN) and 68.35% (NB) and with Yelp dataset, the accuracy produced was 70.98% (SVM), 68.12 (KNN) and 66.52% (NB). A maximum accuracy of 80.08%, 78.40% and 76.92% with Amazon dataset and 78.07%, 76.76% and 74.94% with Yelp dataset, was envisaged while using the proposed MGA algorithm and when tested with SVM, KNN and NB classifiers respectively.

The above discussion proved the importance of feature selection algorithm on OSRD systems. The next part of analysis was concerned with the study of the effect of the enhancement methods proposed. As seen in Chapter 4, the first step of the proposed feature selection algorithm was candidate feature selection, which combined two versions of MRMR, namely, MRMR-IG and MRMR-MI.

From the results, it could be seen that the F-Measure improvement of MIMG algorithm over MRMR was 2.22% (SVM), 2.39% (KNN) and 2.62% (NB) when tested with Amazon dataset. The same algorithm produced an improvement of 1.49% (SVM), 2.59% (KNN) and 1.94% (NB) when tested with Yelp dataset. Similarly, while analysis is performed using accuracy metric, the efficiency gain obtained by MIMG over MRMR was 1.35% (SVM), 2.04% (KNN) and 1.80% (NB) when tested with Amazon review dataset. With Yelp dataset, the accuracy efficiency gain obtained was 0.22%, 0.73% and 1.43% respectively with SVM, KNN and NB classifiers. These increased in efficiency visualised prove that the combined MIMG feature selection to produce effective candidate feature set is effective in increasing the accuracy of classification.

The second part of analysis focuses on studying the effect of enhanced ACO algorithm. F-Measure result analysis revealed that the proposed ACO+GA algorithm showed an average efficiency gain of 1.39% (SVM), 1.60% (KNN) and 1.40% (NB) when compared with the conventional ACO based feature selection algorithm, while using Amazon dataset. The same with Yelp dataset, improved F-Measure by 1.33% (SVM), 2.19% (KNN) and 2.02% (NB). This shows that enhancement of ACO is successful. Accuracy analysis, on the other hand, showed that the MIMG algorithm improved the efficiency by 1.35% (SVM), 2.21% (KNN) and 1.88% (NB) while using Amazon dataset and 2.03 % (SVM), 1.91% (KNN) and 2.74% (NB) while using Yelp dataset, when compared with the existing MRMR feature selection algorithm.

The final analysis of Stage 1 is to study the effect of the combined MIMG and ACO+GA feature selection algorithm. From the results, it could be understood that this algorithm produced maximum effectiveness while detecting ham and spam reviews. The efficiency gain of the proposed FS_MGA algorithm over MIMG and ACO+GA is shown in Table 7.3.

TABLE 7.3

EFFICIENCY GAIN (%) OF FS_MGA ALGORITHM

	Amazon			**Yelp**		
	SVM	**KNN**	**NB**	**SVM**	**KNN**	**NB**
F-Measure Efficiency Gain (%)						
MGA Vs. MIMG	5.85	6.27	5.86	6.54	5.91	6.22
MGA Vs. ACO+GA	2.31	2.70	2.39	2.31	1.51	1.76
Accuracy Efficiency Gain (%)						
MGA Vs. MIMG	6.40	7.08	7.06	6.77	7.54	6.93
MGA Vs. ACO+GA	2.38	2.15	2.89	1.72	2.20	1.51

The speed performance metric was used to evaluate the feature selection algorithms in two manners. The first is to find how much time the feature selection algorithms take to produce the optimal feature subset. The second set of experiments is used to find the effect of feature selection on ham/spam identification. Both are measured in terms of seconds. The speed of the feature selection algorithms to construct the optimal feature vector is presented in Figure 7.6. The effect of using feature selection algorithms on the speed of classification systems, while classifying a single test vector is given in Figure 7.7.

Analysis of time taken to execution the feature selection algorithms, showed that the filter-based algorithms are the fastest. The ACO-GA algorithm while using the candidate set produced by MGA algorithm took on average, 13.01 seconds (Amazon) and 8.05 seconds (Yelp). This shows that the proposed algorithm is slow when compared with filter-based algorithm and ACO+GA algorithm in terms of time complexity. This is an obvious result, as the proposed MGA algorithm has to execute both candidate feature selection and optimal feature vector construction, which will need more computations.

However, the usage of this algorithm is still fast when compared to classification systems that used no feature section. As accuracy is more important during spam identification, it was decided to ignore this speed trend variation and use MGA feature selection algorithm during the design of hybrid classification systems.

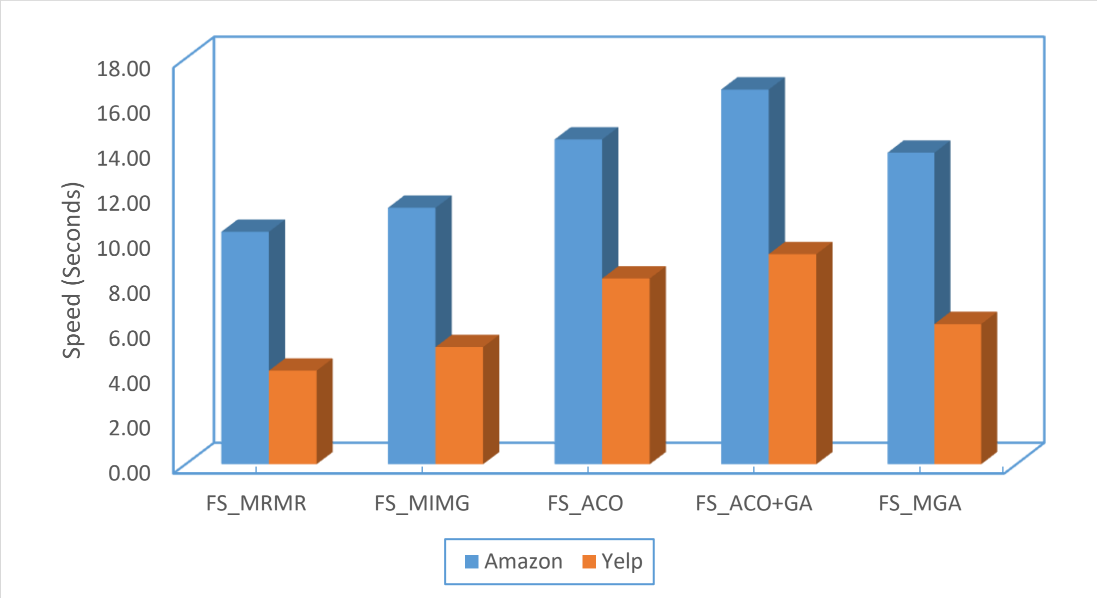

Figure 7.6 :Speed (Seconds) of Feature Selection Algorithms

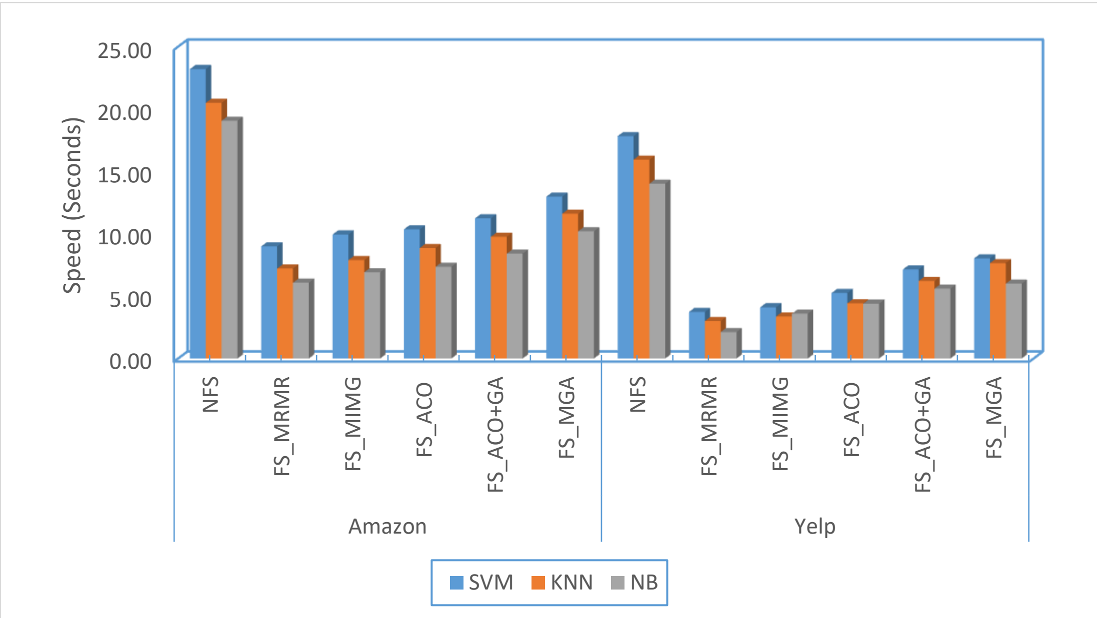

Figure 7.7 : Effect of Feature Selection Algorithms on Ham/Spam Detection System in Terms of Speed (Seconds)

Thus, from Stage 1 experiments, it could be concluded that the inclusion of feature selection algorithm during ham/spam detection is beneficial, with respect to all the four selected performance metrics. Further, it was also understood that the proposed MGA feature selection algorithm that used the combined versions of maximum relevant minimum redundant algorithms for candidate selection and ACO + GA algorithm for optimal feature vector generation had maximum positive effect on spam identification. Using this algorithm, several advantages, as listed below, were obtained.

(i) Improve the quality of the feature vector by reducing redundancy and maximizing relevancy
(ii) Improve prediction accuracy
(iii) Reduce the size of the feature vector

Further, it was also easy to see that the performance of SVM classifier was high when compared to KNN and NB classifiers. The high values obtained indicate that the objective framed for Phase I (Chapter 1, Introduction) is met by the proposed MGA feature selection algorithm and therefore was used in the design of the enhanced ensemble classification system using enhanced SVM classifier proposed in Phase II of the research methodology.

7.4 EVALUATION OF PHASE II ALGORITHMS

Stage 2 experiments are involved in evaluating the OSRD systems that use enhanced versions of SVM classifier proposed in Phase II of the research methodology. The coding scheme used during discussion is presented in Table 7.4.The optimal feature vector produced by MGA algorithm is used to train and test all the classifiers.

Figures 7.8 to 7.12 shows the precision, recall, F-Measure, accuracy and speed of the ham/spam classification systems respectively.

According to Mienye *et al.* (2020), an ensemble system always improve the process of classification. In accordance with this fact, the ES (Ensemble SVM) has improved the performance of OSRD system with respect to all selected performance metrics. This trend was the same with both the datasets.

TABLE 7.4

CODING SCHEME USED – ENHANCED CLASSIFICATION SYSTEMS

Code	Algorithm
S	SVM Classification System
ES	Ensemble SVM Classification System
ES_SO	Enhanced Ensemble SVM Classification System with Speed Optimizers
ES_SO+ED	Enhanced Ensemble SVM Classification System with Speed Optimizers and Euclidean Distance Measure
ES_SO+MD	Enhanced Ensemble SVM Classification System with Speed Optimizers and MahalanobisDistance Measure

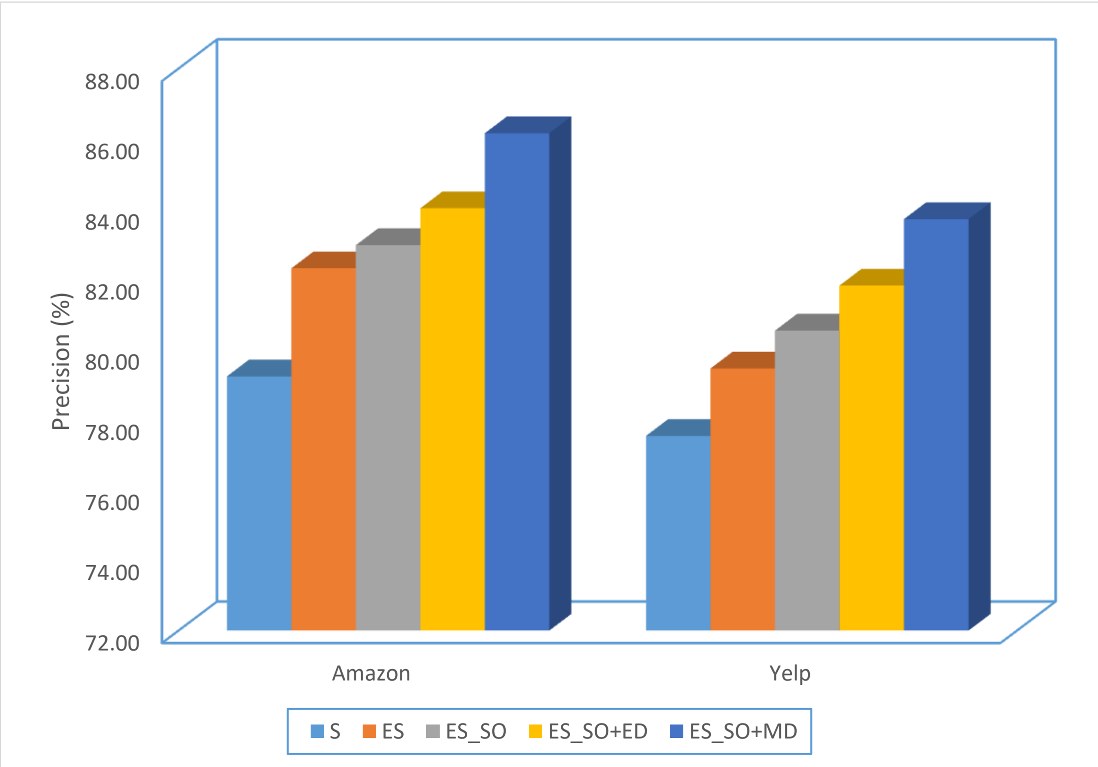

Figure 7.8 : Precision (%) of OSRD Systems

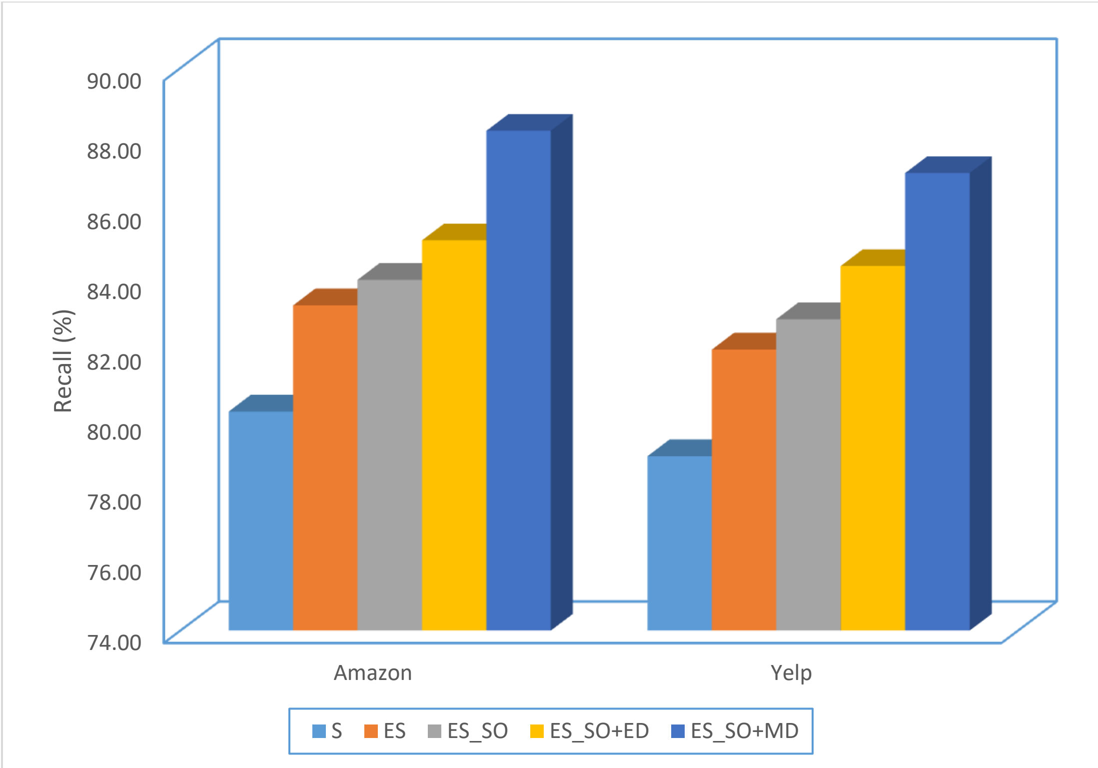

Figure 7.9 : Recall (%) of OSRD Systems

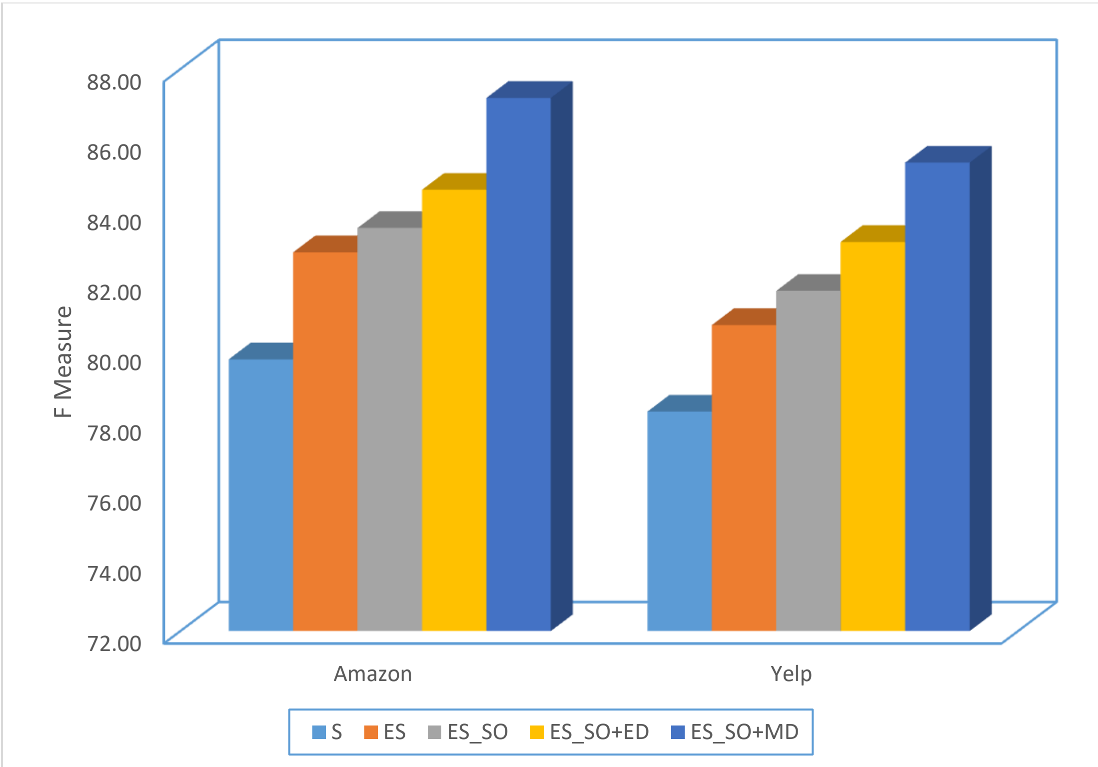

Figure 7.10 : F Measure(%) of OSRD Systems

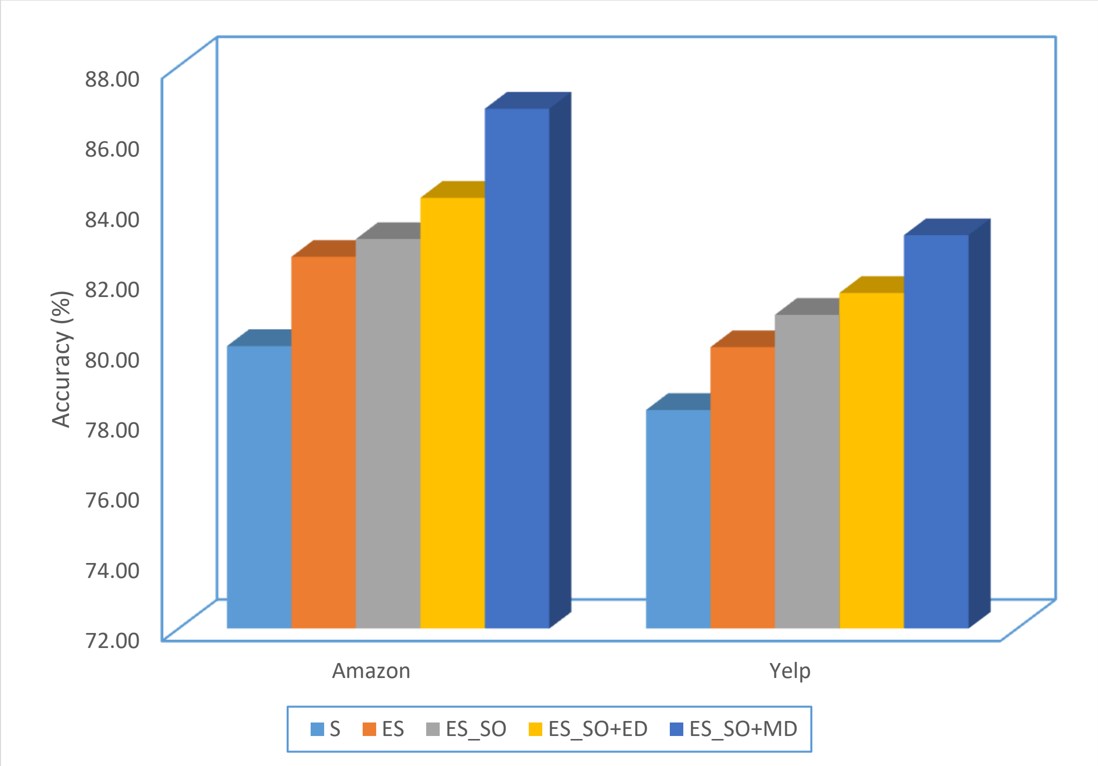

Figure 7.11 : Accuracy (%) of OSRD Systems

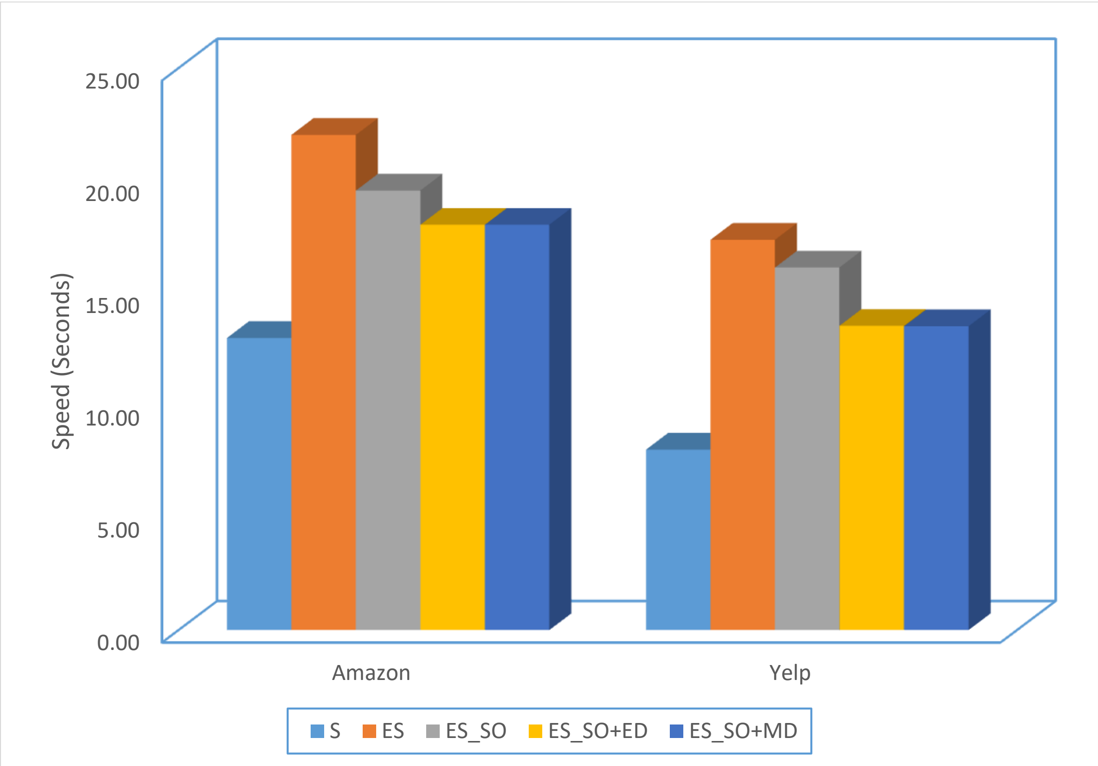

Figure 7.12 : Speed (Seconds) of OSRD Systems

The above results present the evidence that the all the optimization procedures have improved the process of ham/spam review classification in an effective manner, when compared to the single SVM and ES classifier system.

In order to enhance the ensemble SVM (ES) system, the base classifier used was enhanced in two manners. The first was to include a speed optimization procedure, that removed irrelevant SVs, thus reducing the number of computations involved. The second optimization procedure was to use a distance measure to improve the construction of hyperplane. The effect of these two optimization procedures is discussed in this section.

The inclusion of speed optimization procedure to the single SVM classifier improved its performance by 4.5% (precision and recall), 4.47% (F-Measure) and 3.66% (accuracy) with Amazon dataset. When tested with Yelp dataset, the efficiency gain obtained was 3.71% (precision), 4.69% (recall), 4.20% (F-Measure) and 3.34% (accuracy). When the optimization procedure was implemented with the ensemble system, the performance was increased by 0.79% (precision), 0.86% (recall), 0.83% (F-Measure) and 0.61% (accuracy), when tested with Amazon dataset. The same with Yelp dataset showed the efficiency gain as 1.34% (precision), 1.04% (recall), 1.19% (F-Measure) and 1.14 seconds (accuracy). However, the real benefit obtained was with the speed parameter. The single classifier system is always less complex than the ensemble systems. The ES took 22.03 and 17.37 seconds while using Amazon and Yelp datasets respectively. This reduced to 19.56 and 16.14 seconds after the incorporation of the speed optimization procedure. This shows that the incorporation of the speed optimization is successful in improving the quality of SVs and thus has improved the process of ham and spam identification.

The optimization achieved through the use of the distance measure revealed that the use of Mahalanobis distance measure increase classification performance than Euclidean distance. The system ES_SO+MD, when compared with ES and while using Amazon dataset, showed an average efficiency gain of 4.67% (precision), 5.95% (recall), 5.30% (F-Measure) and 5.09% (Accuracy) in terms of precision, recall, F-Measure and accuracy respectively. The same algorithm while using Yelp dataset showed an efficiency gain of 5.34%, 6.11%, 5.72% and 3.97% in terms of precision, recall, F-Measure and accuracy respectively. Analysis of speed parameter shows that the there is no (or very little difference) between ES_SO+ED and ES_SO+MD. But, when compared with conventional ES and ES+SO, the speed gain was still high. The ES_SO+ED and ES_SO+MD took on average less than 18.04 seconds (Amazon) and 13.53 seconds (Yelp) to classify a review to its appropriate target class. This shows that both these algorithms have reduced the time complexity to a great extent when compared to ES. However, the high accuracy (or F-Measure) shown by ES_SO+MD stresses the fact that, this system has more capacity than ES_SO+ED.

Thus, from the various results it can be seen that the combination of speed optimization and the hyperplane construction while using Mahalanobis distance measure has a high impact on the performance of OSRD system both in terms of classification accuracy and speed, thus meeting the objective of Phase II of the research work.

7.5 EVALUATION OF PHASE III ALGORITHMS

Phase III of the research proposed seven OSRD systems, grouped into three types. They are re-listed below.

- Type 1 Hybrid Systems
  - KM_ESVM, MS_ESVM and EM_ESVM
- Type 2 Hybrid Systems
  - SVM-ESVM, KNN-ESVM and NB-ESVM
- Type 3 Hybrid System
  - SVM_KM_ESVM

The main aim of Stage 3 experiments is to analyze these seven system and identify the winner among them as the safe system for identifying fake reviews. The coding scheme used is given in Table 7.5.

TABLE 7.5

CODING SCHEME – HYBRID SYSTEMS

Code	Description
KM_ESVM	Hybrid Model Using KM in Step 1 and ESVM in Step 2
MS_ESVM	Hybrid Model Using MS in Step 1 and ESVM in Step 2
EM_ESVM	Hybrid Model Using EM in Step 1 and ESVM in Step 2
SVM_ESVM	Hybrid Model Using SVM in Step 1 and ESVM in Step 2
KNN_ESVM	Hybrid Model Using KNN in Step 1 and ESVM in Step 2
NB_ESVM	Hybrid Model Using NB in Step 1 and ESVM in Step 2
SVM_KM_ESVM	Hybrid Model Using SVM, KM in Step 1 and ESVM in Step 2

Figures 7.13 to 7.17 show the comparison of the seven hybrid systems when tested with the two datasets with respect to the five selected performance metrics respectively and the results are compared with ES_SO+MD, to understand the efficiency of the proposed hybrid systems.

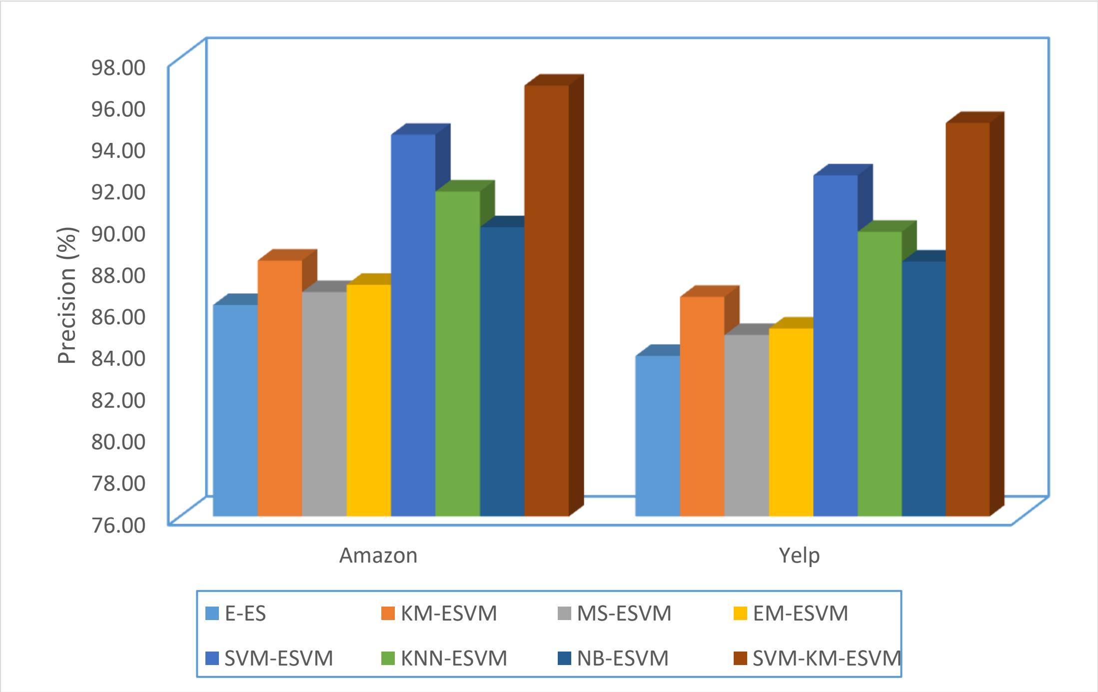

Figure 7.13 : Precision (%) of Hybrid Systems

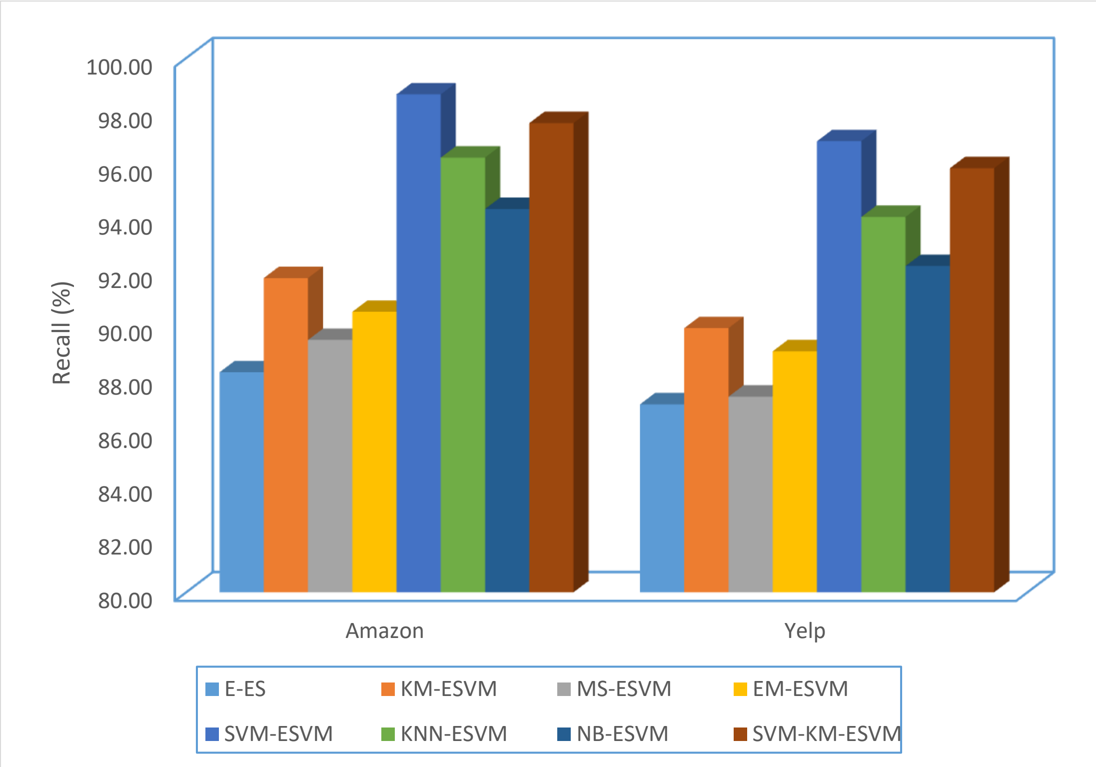

Figure 7.14 :Recall (%) of Hybrid Systems

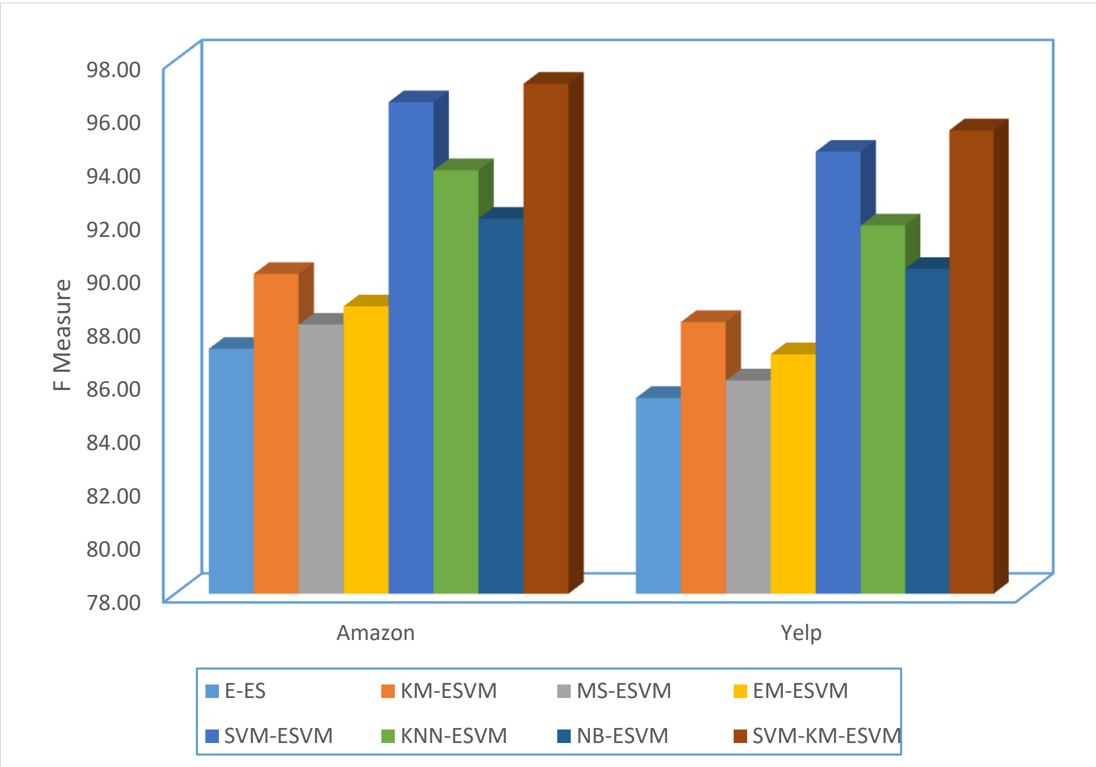

Figure 7.15 :F-Measure (%) of Hybrid Systems

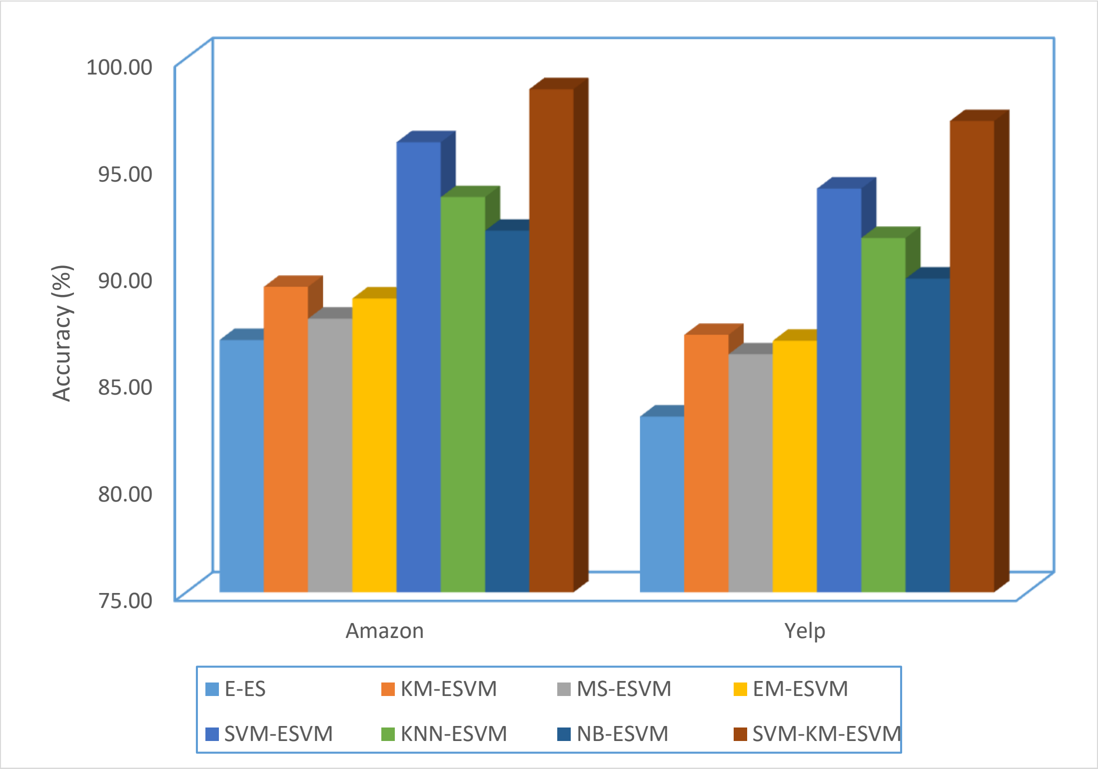

Figure 7.16: Accuracy (%) of Hybrid Systems

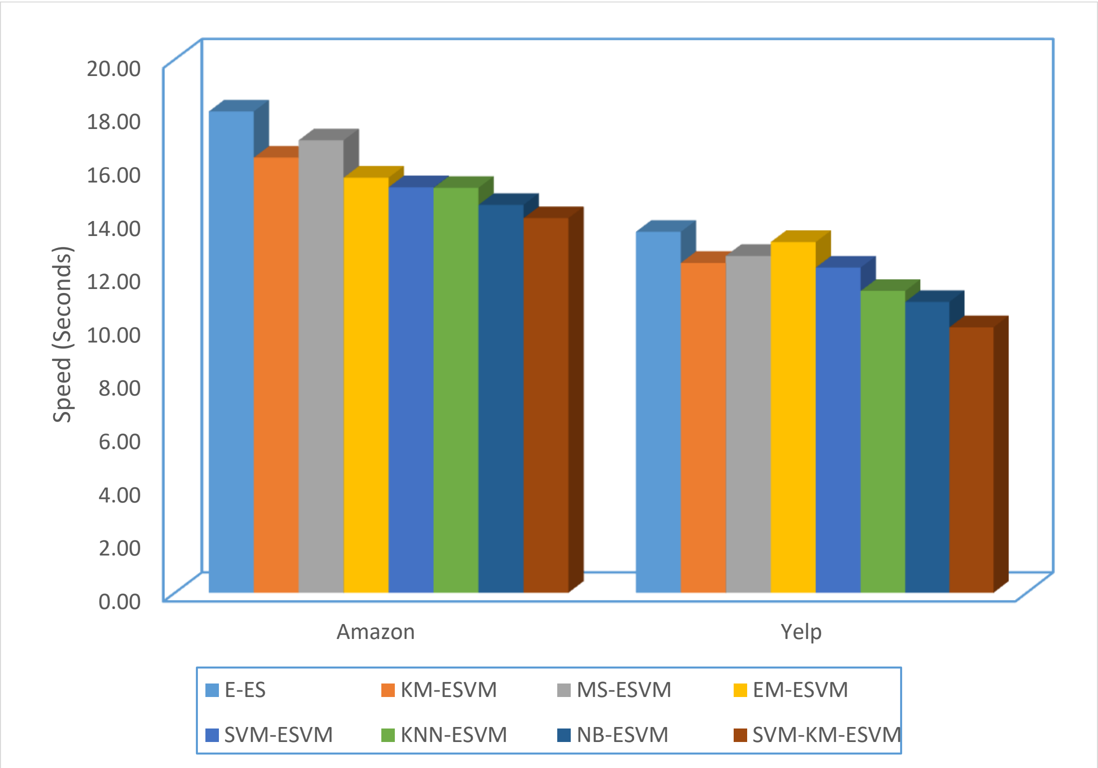

Figure 7.17 : Speed (Seconds) of Hybrid Systems

Comparison of type 1 systems showed that the system that used KM clustering in step 1 and enhanced ensemble SVM classifier in step 2 produced maximum performance. Comparison of type 2 systems showed that system that used SVM for step 1 and enhanced ensemble SVM classifier. However, maximum efficiency was provided by the system that used SVM classifier first to remove outliers, followed by the usage of KM clustering algorithm and then classifying using enhanced ensemble SVM classifier. This system showed a maximum accuracy of 98.54% with Amazon dataset and 97.05% with Yelp dataset. This trend obtained was the same with all the performance metrics. The high values obtained indicate that the objective framed for Phase III is met by the proposed SVM_KM_ESVM system is met.

7.6. CHAPTER SUMMARY

From the various results presented in this chapter, it is clear that the proposed OSRD system that used the optimal feature vector from MGA algorithm and used SVM and KM in step 1 and enhanced ensemble SVM classifier in step 2, can perform spam detection in a highly accurate and fast manner. Moreover, this system is also proved to be stable, as proved by the similar performance trend obtained while varying datasets. Thus, the proposed system has increased both accuracy and speed of the detection system, thus achieving the research problem. The research work is summarized and concluded in the next chapter, Chapter 8, Summary and Conclusion.