CHAPTER 3
METHODOLOGY

Online e-commerce industry is continuously experiencing rapid changes in terms of tools being used and infrastructure that focus on customer satisfaction. This attracts several users to use them, which in turn, has attracted developers to research on areas that can be improved to further improve the experience of the user and make the environment safe to use. One such area is the field of spam review detection and the challenges and complexities involved have motivated businesses to adapt and use sophisticated methods that can improve the accuracy of spam detection and removal. All these methods aim to provide truthful information by increasing the speed and detection rate of spam reviews.

From the literature study conducted (Chapter 2, Review of Literature), it is evident that several solutions have been proposed for detecting online spam reviews. However, as mentioned in Chapter 1 (Introduction), the cunningness and cleverness of spammers in creating spam reviews, has made it a difficult to achieve 100% performance accuracy. Thus, the quest for methods that can improve the detection accuracy is still ongoing. For this, the developers have to continuously monitor the posted spam reviews, in order to

(i) keep track of latest trick used by spammers while creating spam reviews
(ii) identify and understand the ever-changing common characteristics of spam reviews

Apart from developing solutions that counteract the above, they should also be user friendly and highly accurate. The performance of these solutions has to be proved individually, before it could be implemented by the Internet and e-commerce websites. This constant vigilance is required in order to keep track of latest tricks used by spammers to create fake reviews.

According to researchers (Bar-Ilan University, 2019; Greengrad, 2019), the performance of a spam detection system can be improved through the implementation of new innovative algorithms and ideas. Alternatively, according to Chavolla, *et al.* (2018) and Clune (2019)instead of looking for new ideas, it will be more beneficial to enhance

the working of the existing systems. Several suggested solutions belong to the second category and this research work also works towards enhancing the working of the existing systems.

Improving the performance of existing solutions can be done using two types of methodology. The first methodology identifies the issues that may exist in the algorithms used currently in the existing system and enhances the performance by providing solutions to solve these issues. However, the solution has to be designed and implemented carefully, so that, apart from helping to meet the demands on performance, they do not add additional complexities or create new issues. This kind of methodology has been used and proved to be successful Thang and Pashchenko (2017) and Mosavi *et al.*(2019). The second methodology, travels in a different path, and is based on the hybridization concept. Hybridization is defined as the task that amalgamates two or more existing algorithms in order to take advantage of both its advantages. This method has also been proved successful (O'Driscoll *et al.*, 2019;Ren *et al.*, 2019).

This research works uses the second type of methodology and proposes algorithms that enhance the working of each step of spam review detection using hybridization concept. Each algorithm is designed to combine advantages of different algorithms so as to provide a solution that can result with a detection system that is stable, versatile and accurate. The proposed system is designed as a binary classification system, which classifies an online review as ham or spam.

3.1. DEVELOPMENT METHODOLOGY, PHASES AND INTERACTIONS

Online Spam Review Detection (OSRD)systems perform two main steps to identify spam reviews from ham (honest or genuine) reviews. They are, feature engineering and spam filtering (Bajaj *et al.*, 2017). During feature engineering important characteristics are extracted from reviews to form a feature vector, which is then used in the second step to filter spams. Feature engineering is defined as a task that constructs or extracts features from the source of information (Nabi *et al.*, 2020). In this research, the sources of information are the online reviews posted in a website.

The feature engineering is performed to construct a feature vector that encompasses methods which fully utilize all available information pertaining to a review, a reviewer or a product and works with the aim of accurately detecting the spam activities in the customer review section of e-commerce portals. The available information can belong to three main perspectives, namely, the content of the review (Review Centric Features), the author of the review (Reviewer Centric Features) and the by-product being reviewed (Product Centric Features)(Kumar, 2018). Each category provides full information regarding the three main entities of the online reviews. This research work proposes the use of all three information to obtain a more complete form of information and proposes methods to construct an optimal feature vector.

The second step, spam filtering, is dedicated to the separation of spam reviews and ham reviews. Existing methods used for review spam filtering can be formatted into two main lists, namely, machine learning approaches and lexicon based avenues. Among these two approaches, machine learning algorithms are more popular and can be implemented using either a supervised (classification) or unsupervised (clustering) algorithm (Krithiga and Ilavarasan, 2020).

Researchers are always in search of methods to improve the working of these two machine learning algorithms. The most popular manner is to hybridize two or more classification or clustering algorithms that belong to the same domain (Roy *et al.*, 2018). Another novel way of hybridization is to systematically combine a single instance of classification algorithm and a single instance of clustering algorithm to produce an effective spam filter (Gupta *et al.*, 2019). In this research work, the hybrid system combining the supervised and unsupervised algorithms is enhanced and used to identify spam reviews.

To build the envisaged feature engineering and hybrid spam filtering algorithm, the research methodology was constructed with three phases (Figure 3.1). The first phase is focused on feature engineering and constructing an optimal feature vector. The second phase is focused on improving the performance of the classifier that will be used during the construction of the contemplated hybrid system. The third phase is dedicated to the construction of the hybrid system, that merges clustering and classification algorithms.

Each of these phases were designed independently and were combined using a simple I/O (Input/Output) interface, where the output of the previous phase is used as input to current phase. The interaction of the algorithmsproposed in each phase is presented Figure 3.2. .

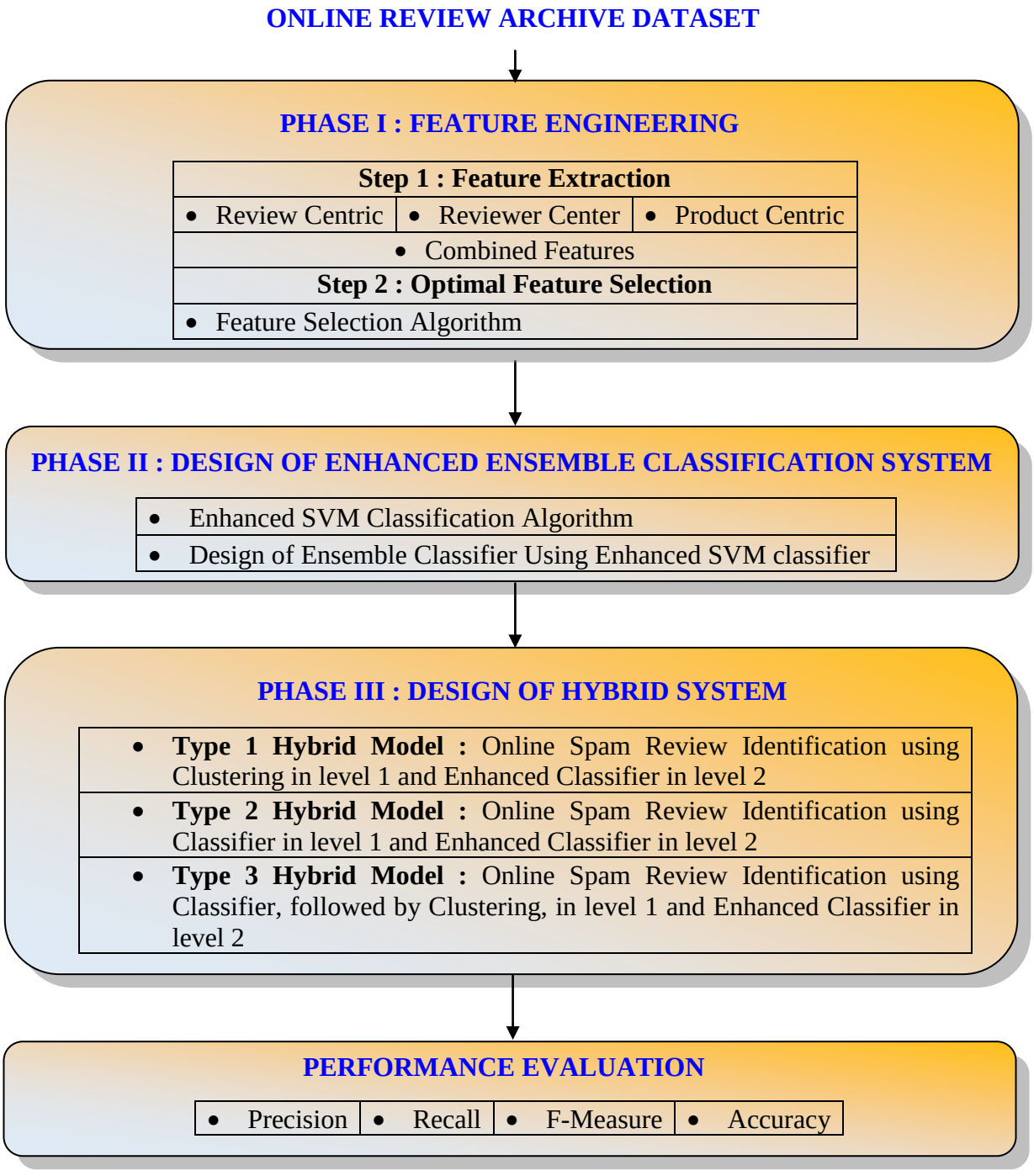

Figure 3.1 : PROPOSED RESEARCH METHODOLOGY

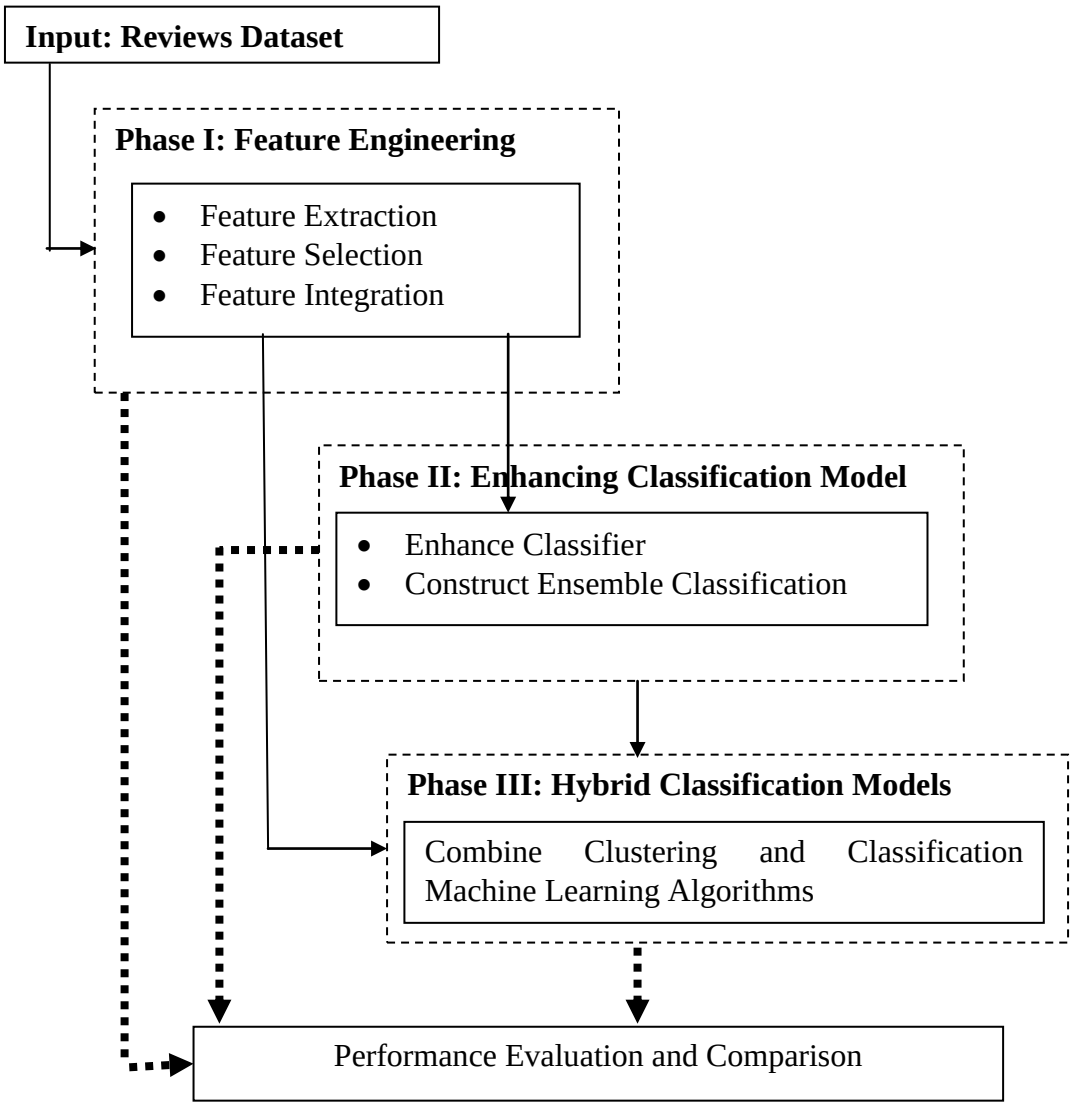

Figure 3.2 : Interaction of Algorithms and Research Phases

3.2. PHASE I : FEATURE ENGINEERING

The methods proposed in Phase I perform feature engineering to construct a feature vector that has only optimal features and have positive impact on review spam detection. Phase I achieves this using two stages, namely, feature extraction and feature selection.

3.2.1. Feature Extraction

The influential portion of review spam detection system is the Feature Extraction and is defined as the method that revolutionizes the raw data into a reduced form. This reduced data is termed as feature vector or feature space. It has proven that compared to

the usage of single set of feature vectors, multiple features can advance the accomplishment of classification (Arora *et al.*, 2008). This research work also extracts multiple features extracted from the three participants of an online purchase, namely, review content, reviewer and product. In the first stage of Phase I, a total of 53 sets of features are extracted (Table 3.1).

TABLE 3.1

FEATURES EXTRACTED

Review Centric Features (38)	Textual features (9), meta data (8), content similarity using Bag of Words (5), PoS tags (9), n-grams (4), rating (1), sentiment (1), burst patterns (1)
Reviewer Centric Features (13)	Reviewer activities, Maximum Number of Reviews, Percentage of positive reviews, Review length, Reviewer deviation, Burst Review Ratio, Ratio of Verified Purchases, Reviewer Burstiness, Extreme ratings, Reviewer average proliferation, Reviewer spamicity, % of positive reviews, % of negative reviews
Product Centric Features (2)	Rank in Sale, Average Rating

3.2.2. Feature Selection

The features extracted in stage 1 may have irrelevant and noisy or redundant features. The presence of these two unwanted features raises two issues during classification. They are increased computations and thus increased time complexity and over fitting. Moreover, the usage of multiple feature sets, while the performance of classifier is raised as regards training time and accuracy, also suffers from the problem of high dimensionality. Feature selection works towards providing solutions to these issues. A feature selection algorithm works on the principle that not all features extracted are important for review spam detection and if these insignificant features could be identified correctly, can be removed on the outside influencing the performance of the spam detection system (Reddy and Reddy, 2019). Moreover, the feature selection algorithms also aim to choose features that have

(i) maximum class discriminating power and

(ii) maximum influence on the accuracy of the classification output.

Thus, usage of feature selection algorithm helps to cut down the final feature vector size, reduce irrelevant and redundant features and improve the accuracy of spam detection.

The second stage of Phase I, thus focuses on feature selection algorithm and a method to combine all the three sets of features extracted from Stage 1 into a single feature vector. For this purpose, a two-step algorithm is proposed. In the first step, a simple feature selection algorithm is used to generate a set of candidate feature set. This feature set is then used in the second step, which further identifies the optimal and accurate features that are important for spam detection.

In the first step, an enhanced Maximum Relevance Minimum Redundant (MRMR) feature selection algorithm is accomplished to identify unique and correspondent features. The conventional MRMR algorithm considers the relationship between features to identify relevant and non-redundant features, but ignores the relationship between features and its target class label. Moreover, the performance is directly proportional to the dataset size, that is, the number of redundant terms retained grows in magnitude with respect to relevancy. To solve this issue, two variants of MRMR are designed, whose results are then combined to form the candidate feature set. The two variants proposed differ in the feature analysis method used. The first method uses the information gain, while the second method uses mutual information. Both the methods are modified to consider relationships that exist between features and between features and class label. The result of the two MRMR algorithms are then combined using score-based method.

To further improve the quality of the candidate feature set obtained from step 1, in the second step an hybrid Ant Colony Optimization (ACO) combined with Genetic Algorithm (GA) based feature selection algorithm is came up with. The main advantage of ACO is that it can perform local searching, while GA considers a global panorama through functioning on the comprehensive populace from the commencement. Thus, amalgamating ACO and GA can take advantage of each and work towards producing a reduced feature space having vital features.

Phase I results with a feature vector that has only optimal features, that is, most relevant features without much redundancy and which exhibits high discrimination

between target classes. Complete portrayal of the working of the algorithms introduced in this section is presented in Chapter 4, Feature Engineering. A large number of experiments were carried out to demonstrate the benefits of the suggested feature engineering techniques. In Chapter 7, Findings and Discussion, the results of these tests are reported.

3.3. PHASE II: DESIGN OF ENHANCED ENSEMBLE CLASSIFICATION SYSTEM

An online review detection system as mentioned earlier, frequently use the clustering or classification algorithms. In this research task, a hybrid system that integrates clustering algorithm and classification algorithm for ensemble learning is proposed. The proposed system is designed in three steps.

Step 1 : Enhance a classification algorithm

Step 2 : Design ensemble systems using the classifier enhanced in step 1.

Step 3 : Design hybrid systems using algorithms from step 2.

Phase II of the research methodology is focused on the first two steps. The classifier used is Support Vector Machine (SVM), which was chosen as it is most frequently used algorithm for classification and prediction and moreover has a great proven success record for achieving high performance when compared to several other classifiers like K-Nearest Neighbour (KNN), Naïve Bayes (NB), Decision trees and Back Propagation Neural Network (Edwin and Bogdan, 2017; Guosheng and Guohong, 2008). However, the SVM classifier also has the issues of high training complexity while working with very high dimensional dataset, which leads to speed limitation. Phase II aims to solve this issue by using speed optimization and ensembling.

3.3.1. Enhanced SVM Classification System

Optimization of SVM classifier is done in two manners, in this research work. The first is to remove irrelevant support vectors which do not have any relevancy during classification. This lessens the quantity of computations and thus solves the high training time required. The second is to replace the conventionally used Euclidean distance with Mahalanobis distance measure.

The algorithm begins by mapping all training features into the vector space of SVM. The optimization method first computes the margin of the training feature set for each category. The obtained margins are arranged in ascending order and the first M smallest margins are selected as relevant Support Vectors (SVs). The rest of SVs are treated as noise or irrelevant SVs from the training set are eliminated. In the later step, the identified SVs are outlined into the prototypal vector space. These new features are mapped with SVs to the same original vector spaces. Now Mahalanobis distance measure is used to estimate the average distance bounded by the brand-new features and each set of SVs from various leagues. The category whose new feature is closest to the set of SVs (or having minimum distance) is identified and is taken as the category to which the new feature will be grouped.

A major decision while using SVM is the preference among choices related to kernel function. The kernel functions perform the task of mapping the features into high dimensional feature space. The most common and well-performing kernel functions used with SVM are linear, sigmoid, polynomial and Radial Basis Function(RDF) (Nanda *et al.*, 2018). When dealing with data that is divergent in nature, each of these functions has its own fundamental traits and responses. Thus, an important aspect of SVM is the correct select of kernel function. Correct kernel function derivation and execution in SVM aids in scaling effectively with high-dimensional datasets, and the balance between order intricacy and mistake rate may be precisely regulated. This issue is solved in this research work by using the concept of ensembling.

3.3.2. SVM Ensemble Classification System

Ensembling is defined as a method that pools multiple machine learning classifiers whose data are pooled to produce a more precise classification result. Ensemble classifiers (Noor et al., 2020) can address a wide range of application concerns and add to the typical inclination/fluctuation tradeoff by offering for setups that would be insufficient to reach with a single classifier. Moreover, it reduces the likelihood of giving high weightage to a decision made by a poorly chosen model. All these advantages have made Ensemble of Classifier(EoC) as a sought after solution to many real-world applications (Zhang*et al.*, 2020a; Pisula, 2020) and therefore is analyzed in this research.

The ensemble SVM classification system is designed to have four base classifiers, each created with the same training data but with different kernels. The weighted majority voting scheme is used to total the consequences of these four classifiers. The proposed ensemble SVM classifier is designed as a binary classifier, as it needs to classify the input review as either spam review or ham review.

Phase II of this research work proposes an enhanced ensemble-based SVM classification system to improve the process of ham/spam reviews. The enhancement methods incorporated along with the method used during ensembling are presented in Chapter 5, Design of Enhanced SVM Ensemble Classification System. A few tests were carried out to show the benefits of the proposed feature engineering methods. In Chapter 7, Results and Discussion, the results of these tests are reported.

3.4. PHASE III: DESIGN OF HYBRID SYSTEMS

The final phase of the research methodology is to design hybrid systems using the algorithms proposed in Phases I and II. As stated previously, classification and clustering machine learning algorithms are the two frequently used methods for identifying spam reviews. The systems proposed in Phase III, differs in the manner these two algorithms are combined and aims to increase the spam detection accuracy. The hybrid systems perform spam detection in three steps, as detailed below.

- Step 1 : Construct feature vector using methods proposed in Phase I. Partition this vector into training and testing sets.
- Step 2 : Improve quality of training data using single classifier or single classification algorithm.
- Step 3 : Categorize online reviews as spam and ham by combining Step 1 with ensemble classifier proposed in Phase II

Using the above three steps, three types of hybrid systems are proposed, as listed below.

- Type 1 : Hybrid System using clustering algorithm in step 1 and ensemble classification algorithm in step 2
- Type 2 : Hybrid System using classification algorithm in step 1 and ensemble classification algorithm in step 2

- Type 3 : Hybrid System using classification followed by clustering in step 1 and ensemble classification algorithm in step 2

In this research work, three classifiers and three clustering algorithms were analyzed. The classifiers used are Naïve Bayes (NB), Support Vector Machine (SVM) and K-Nearest Neighbour (KNN). The clustering algorithms used are K-Means (KM), Mean Shift (MS) and Expectation-Maximization (EM). By varying the algorithms used in level 1, a total of seven systems were proposed for identifying spam and ham online reviews (Table 3.2).

TABLE 3.2

PROPOSED HYBRID SYSTEMS

Type 1		**Type 2**		**Type 3**		
Step 1	**Step 2**	**Step 1**	**Step 2**	**Step 1**	**Step 2**	**Step 3**
KM	ESM	SVM	ESM	KM	SVM	ESM
MS		KNN				
EM		NB				

In Type 1 systems, the clustering algorithm is smeared to obtain a set of clustered data and it centroids. The nearest neighbour method is used to identify clumps of features having a feature and its neighbour. Using this information, a new feature vector is created, which is one dimensional and contains only the optimal features. Using this, the nearest cluster to the input review is estimated and classification is performed using the proposed enhanced SVM-based ensemble system.

In Type 2 systems, the classification algorithm is first used on the training data and only those features that are correctly classified (true positives and true negatives)are selected to form a new training set. This new training set is again size reduced and quality improved. This training batch is then exercised to train the enhanced SVM-based ensemble system (ESM).

Encouraged by the promising results obtained by Type 1 and Type 2 system, another hybrid system, Type 3, was also designed. This system begins by obtaining optimal feature set using the concept proposed in Type 2 system. This new feature set is

then retuned using concepts proposed in Type 2 system. The final feature vector is then used to train the enhanced SVM-based ensemble system.

The final phase of the research work designs three types of hybrid systems and the detailed description of the steps involved are presented in Chapter 6, Design of Hybrid Classification Systems. Considerable amount of experiments were super intendenced to prove the advantages of the proposed hybrid system over conventional and enhanced counterparts. The results of these experiments are presented in Chapter 7, Results and Discussion.

3.5. EXPERIMENTAL RESULTS

All the algorithms proposed in the three phases of the research methodology were analyzed vigorously using several experiments. During performance evaluation, two datasets, namely Amazon (http://liu.cs.uic.edu/download/data) and Yelp (http://www.yelp.com) were used. Five performance metrics, namely, speed, , accuracy, precision, recall and f-measure were used during analysis. The results obtained by the proposed algorithms were compared with its conventional counterparts to determine the performance gain obtained.

The performance evaluation showed that the algorithms proposed in each phase of the study was successful in improving its working and when combined to form the hybrid systems produced maximum efficiency during the detection of spam reviews. A maximum of 98.54% with Amazon dataset and 97.05% with Yelp dataset was obtained while using the proposed hybrid system that combined clustering with enhanced ensemble classification system. Description of the Amazon and Yelp datasets, metrics used to evaluate the proposed algorithms inclusive of the results of experiments are put forwarded in Chapter 7, Results and Discussion.

3.6. CHAPTER SUMMARY

The numerous algorithms offered to enhance the feature engineering process and classification steps of online spam review detection system were introduced in this chapter. This chapter also presented the research methodology along with the flow of the proposed ham/spam detection system. The working of the feature engineering is elaborately discussed in the following chapter, Chapter 4, Feature Engineering.