ABSTRACT

Online reviews are often contemplated as the primary factor in beholding customer's decision to purchase a product or service. It is observed as a valuable source of information in determining public opinion of these products and services. On considering this key aspect and their impact over the products, manufacturers and retailers are highly concerned and are fretful about the customer feedback and reviews. The increase in the reliance on online reviews has a direct effect on the sale of products since it may also induce the fake reviews to promote or devalue the products and their services. This practice is known as Opinion (Review) Spam, where the spammers manipulate and produce poisonous and fake reviews (i.e, making fake, untruthful, or deceptive reviews) to incur profit or gain. In view of the fact that not all online reviews are truthful and trustworthy, it is essential to develop techniques for detecting review spam. By means of extracting meaningful features from the text using machine learning techniques for classification and evolutionary algorithms for feature selection. In addition, the reviewer information, apart from the text, can be used to aid in spam review classification process. Moreover, majority of current, contemporary research has focused on supervised learning methods, which requires labelled data. It is a scarcity, while dealing with the online review spam. The web contains large possessions and assets of opinions about products, politicians and more, which are expressed in different means including news group posts, review sites. The number of customer reviews received for a product is growing exponentially at a rapid rate. An important issue associated to the trustworthiness of the online opinions has been neglected and left uncared for more often. Hence, it is exceedingly essential to have a mechanism which is proficient and capable of assessing the trustworthiness of reviews for pertinent and proper decision making or for marketing and business

intelligence. Trusted customer reviews are useful and valuable for both potential buyers and product manufacturers. Further, it is more convenient and less time consuming for buyer to see at a glance, feature by feature comparison of the reviews written by a good number of the customers in taking buying decisions without getting biased. The product manufacturer gets acquainted with the strengths and weaknesses of his/her own products and also that of the competitors, consumer preferences and interests which could help them decide to maximize profits. Thus, it is indispensable to identity the trusted customer reviews. Hence the bio inspired algorithms based feature selection is proposed to solve the problem of spam review classification.

In general, the spam reviews are characterized by a vector space model in which every review is considered as a vector of terms. Since there exists many different terms in the spam review, not all feature selection techniques such as Principal component analysis, information gain and Latent semantic indexing can handle or deal with the high dimensional data. Certain definite terms could be powerful discriminatory terms which are required for classification; others might not be influential and could carry redundant information, confusing and mystifying the classifier. Moreover, the problem of curse of dimensionality further reduces the performance of the classifier. Therefore, the dimensionality reduction techniques are important to obtain the best constructed features. Feature selection and feature extraction are significant research problems in the applications allied to the classification problem including spam review classification. Feature selection techniques can be applied to produce optimal subset. Among the feature selection approaches either a filter method or a wrapper method can be employed to do so. The evaluation is made according to the statistics computed from the data in the filter method and the evaluation of the feature

subset is done based on the predictive performance of the classifier in the case of wrapper method.

The thesis investigates the use of bio inspired algorithms for feature selection in which the performance of a classifier is used to assess and evaluate the quality of a feature subset. Research contribution deals with feature extraction based on improved binary particle swarm optimization (iBPSO). The proposed methods include iBPSO based Naive Bayes (NB) classifier and iBPSO based K Nearest Neighbour (kNN) classifier. The performance of the proposed methods has been compared with the existing algorithms using evaluation measures such as precision, recall and accuracy.

The second contribution makes use of hybrid algorithms of improved Binary Particle Swarm optimization (iBPSO) with Shuffled Frog Leaping Algorithm (SFLA), for feature selection. These bio-inspired computational algorithms are employed in feature subset selection and the selected significant attributes are given to the NB classifier in order to classify the review as spam or ham. Amongst these proposed methods iBPSO_SFLA algorithm showed better performance in classifying the reviews. The proposed evolutionary computational technique suffers from exploitation inability and slow convergence. The experimental results indicate that the proposed iBPSO_SFLA method is more effective than the existing methods.

A hybrid technique based on improved Binary Particle Swarm optimization (iBPSO) with Cuckoo Search (CS), for solving optimization problems is proposed to overcome the limitations of the previous contribution. A new meta-heuristic algorithm, called Cuckoo Search for solving optimization problem is proposed. It is based on the obligate brood parasitic behaviour of some cuckoo species in combination with the Lévy flight behaviour of some birds. The simulation results prove that the proposed iBPSO_CS method provides better performance.

The fourth contribution includes a technique that involves a hybrid improved Binary Particle Swarm optimization (iBPSO) with Binary Flower Pollination Algorithm (BFPA) for feature selection. Global pollination process and Local pollination are the two key steps the Flower pollination algorithm. The positions of each particle in the iBPSO are updated using Levy Flight distribution in the Flower pollination to attain optimal solution. The results show that the proposed iBPSO_BFPA method yields a high quality of feature subset and also overcomes the problem of local optima, thus increasing the accuracy of the NB classifier.

Hybrid Cuckoo Search with Harmony Search (HCS-HS) algorithm is proposed. The search space is traversed using the CS algorithm to identify the global optimal value at a faster rate. However, the exploitation of the required solutions is found to be poor due to large iteration steps. To resolve and determine this, a standard HS algorithm is utilised to explore the solution by well adjusting the pitch adjustment rate (PAR). For this reason, the hybridization of HCS_HS would help in providing overall improvements by rectifying their own disadvantages by other algorithms and retaining their advantages. The performance of the proposed HCS_HS for spam review classification has been evaluated considering various metrics including accuracy, precision and recall and accuracy. The simulation and comparisons made the results indicate that the proposed HCS_HS is better than other techniques. The hotel review dataset is used for evaluating the proposed methods. The results show that the proposed HCS_HS method shows improved performance for spam review classification on comparison with other traditional, existing methods.