CHAPTER 8

SUMMARY AND CONCLUSION

Increase in the user-generated content has motivated millions of persons to search and use opinions to make decision. Positive opinions have huge impact on sales of products and services, while negative opinions can reverse this impact. The huge positive impact also gives good incentives for reviewers to write spam reviews, which are written with the aim of misleading customers by giving undeserving positive or negative ratings and opinions. Spam detection systems are designed as a classification problem to group the reviews into two categories, namely, ham or genuine reviews and spam or fake reviews. The main focus of this research work is the development of a classification system that can identify ham and spam online reviews.

The primary objective of this research work is to propose classification system whose individual steps are enhanced using algorithms, so that decision of an online review being spam or ham, can be identified in an accurate manner. For this an online spam review detection system was designed with two steps, namely, feature engineering and classification. This objective was achieved using a research methodology that was designed with three phases, with each phase focusing on improving the steps of the detection system. Phase I of the research methodology designs a feature engineering algorithm that performs two tasks, namely, feature extraction and optimal feature vector construction using feature selection algorithm. Phases II and III dedicates itself to improve the second step of the online spam review detection, that is, the classification.

In order to improve the accuracy of the classification step, the research work proposes the use of multiple feature sets extracted from the three main components of an online review platform, namely, review, reviewer and product. Eight groups of review features, namely, textual features, metadata features, content similarity features (bag of words features, POS features and n grams), rating features, sentiment score feature and burst / peak pattern features, were extracted. A total of 36 features were extracted in this group. Twelve reviewer-based features, namely, reviewer activities, maximum number of reviews, review length, reviewer deviation, burst review ratio (brr), ratio of verified

purchases, reviewer burstiness, extreme ratings, reviewer average proliferation, reviewer spamicity, % of positive reviews, % of negative reviews, were extracted. Two product-based features, namely, rank in sales and average rating, were extracted. Thus a total of 50 features were extracted, which resulted in curse of dimensionality.

This problem was solved through the use of a feature selection algorithm. The proposed feature selection algorithm is a two-step algorithm, where the first step produces an optimal candidate feature set that consists only of non-redundant and relevant features. This was selected using maximum relevant minimum redundant algorithm combined with information gain and mutual information. This candidate feature sets were then combined using a fusion algorithm to form a super feature vector. In the second step, an enhanced ant colony optimization algorithm that is combined with genetic algorithm was proposed and used. The SVM classifier was used as an fitness evaluation method to decide on the optimal feature set.

The second step of the proposed online spam review detection step is classification, which was performed using hybrid systems that combined clustering, classification and ensembling. Phase II of the research methodology is dedicated to enhancing the ensemble classification system. The ensemble system was enhanced through the use of enhanced SVM classifier. The SVM classifier was enhanced by first improving hyperplane construction through the use of Mahalanobis distance measure. Then, the irrelevant support vectors that have no impact on the performance of the classifier were removed. Both optimization approaches intended to increase hybrid systems' classification accuracy while reducing their time complexity. The ensemble classifiers were constructed by varying the kernel function used by the SVM classifier.

In the final Phase of the research, the above enhanced ensemble system was further improved by using a preprocessing step. This preprocessing step used either a clustering or classification or classification and clustering algorithm to improve the quality of the training set, thus improving the performance of the detection system. Thus, the hybrid systems were designed to have two steps. The first step performed preprocessing, while the second used the enhanced ensemble SVM classifier to detect ham and spam reviews. Three clustering algorithms, namely, K-means, Mean shift and expectation maximization,

were selected. The classification algorithms used were SVM, KNN and NB. Using these algorithms, a total of seven systems were designed by varying the algorithms used in step 1. The results from the base classifiers of the ensemble system were aggregated using a weighted majority voting algorithm.

To assess the impact of the algorithms proposed in each phase, several investigations were performed out. Two review datasets were used, one from Amazon and the other from Yelp. During the evaluation, five performance indicators were used: precision, recall, F-Measure, accuracy, and speed. All the proposed algorithms were compared with their conventional and/or existing counterparts. From the experiments results the following findings were discovered.

- From Phase I experimental results, it was understood that the usage of any feature selection algorithm has a positive impact on the performance of online spam review detection. In particular, this impact was maximum, while using the proposed algorithm. The proposed feature selection algorithm combining MRMR with MI and MRMR with IG with Ant Colony Optimization enhanced with GA showed 9.11% (Amazon) and 9.08% (Yelp) efficiency gain with respect to accuracy when compared with classifier that used no feature selection algorithm.

- Phase II experimental results proved that all the optimization methods incorporated in SVM classifier were successful. The performance of the enhanced ensemble system was correlated with the conventional SVM and conventional ensemble SVM classifier. This observation divulged that the proposed ensemble system that used enhanced SVM as base classifier was more successful and was able to achieve a high accuracy of 86.79% with amazon dataset and 83.20% with Yelp dataset. Moreover, the usage of the proposed optimization procedures with the ensemble system also reduced the time complexity of review spam detection system considerably. The conventional SVM classifier took 22.03 seconds with Amazon dataset and 17.37 seconds with Yelp dataset to classify a review as either spam or ham. This reduced to 18.04 seconds (Amazon) and 13.53 seconds (Yelp) when the ensemble system incorporated with enhanced SVM was used.

- From Phase III experiments, it could be deciphered that the hybridization of classifier and clustering was successful in improving spam detection. Among Type 1 hybrid systems proposed, the system that used KM for step 1 and proposed ensemble classifier with enhanced SVM in step 2 was more efficient. Comparison of Type 2 hybrid systems revealed that the system that used SVM in step 1 and proposed ensemble classifier in step 2 produced maximum efficiency. While comparing type 3 hybrid system, the system that used SVM, followed by KM in step 1 and proposed ensemble classifier in step 2 improved spam detection impressively. Comparison of the three winning systems showed that the type 3 system improved the review spam identification more efficiently and achieved a high spam detection accuracy of 98.54% with Amazon dataset and 97.05% with Yelp dataset. This algorithm was also able to reduce the time complexity of ensemble classifier by 3.5 seconds on average.

The presence of spam online reviews results with customer dissatisfaction and reduce business. Currently, online e-commerce sites use several systems to detect and remove such damaging reviews. These systems identify spam reviews and help to improve the truthfulness of the business. But, these systems are not always adequate and may not lead to removal of all spam reviews. Therefore, systems that advocate for better detection systems are required. This research work proposed an online review spam detection system based on multiple features, enhanced feature selection and hybrid systems that combine clustering with classification algorithms, which can help both individuals and businesses to provide a safe and efficient environment during an online purchase.

FUTURE RESEARCH DIRECTIONS

The points given below can be considered in future to enhance the proposed spam review identification system.

- The proposed systems can be further improved by including an outlier detection algorithm, that can detect abnormal behaviors in reviews. This can further improve the detection of spam reviews.

- The time complexity of the proposed hybrid ensemble model can be reduced by removing irrelevant base classifiers using algorithms that can identify unwanted base classifier during spam review detection.

- The system can also be improved by parallel processing algorithms. This is feasible, by way of figuring out operations that are unbiased to every different and recommend a parallel structure to enhance the performance.

- Different linguistic constructs such as modifiers, negations, emojis, and ironic words can all be used to improve the effectiveness of the proposed Spam Identification framework's classification module.

- The points of mobile-enabled social networking, such as IP address, mac address, and geotagging, can also be explored to improve the system's performance.

- Furthermore, combining user debts on current review sites, like as Amazon, with social media websites (Facebook, Twitter, etc.) can help to reduce spam in user reviews.