Engineering Science
2021; 6(3): 33-38
http://www.sciencepublishinggroup.com/j/es
doi: 10.11648/j.es.20210603.11
ISSN: 2578-9260 (Print); ISSN: 2578-9279 (Online)

Ensemble Classifiers Employed for Spam Review Detection

Alhassan Jamilu Ibrahim, Maheyzah Siraj, Usman Abubakar Jauro

Information Assurance and Security Research Group (IASRG), Faculty of Computing, Universiti Teknologi Malaysia, Skudai Johor, Malaysia

Email address:
jamilualhassan@yahoo.com (A. J. Ibrahim), maheyzah@utm.my (M. Siraj), ausamn63@gmail.com (U. A. Jauro)

To cite this article:
Alhassan Jamilu Ibrahim, Maheyzah Siraj, Usman Abubakar Jauro. Ensemble Classifiers Employed for Spam Review Detection. *Engineering Science*. Vol. 6, No. 3, 2021, pp. 33-38. doi: 10.11648/j.es.20210603.11

Received: June 12, 2021; **Accepted**: July 7, 2021; **Published**: August 11, 2021

Abstract: The advancement of technology and the use of internet have changed many aspects of human culture over the years. Today, consumers take confidence in e-commerce platforms like amazon and eBay for comprehensive understanding of products and services when making a purchase decision. Here the web or user-generated content from consumers of such products and services, known as reviews, are exploited by spam reviewers to falsely promote or downgrade some targeted products. Despite potential solutions, Identifying and preventing review spam are still one of the top challenges faced by web search engines today. Therefore, in the quest to provide a more improved and efficient classification of review spam, this research probed different techniques in order to find most effective solution to spam detection. The research employed three base classifiers, Naïve Bayes, Support Vector Machines and Logistic Regression to form ensemble classifiers complimented with Arching classifier. The Arching classifier performs the weighted voting that produces the final class label with performance and accuracy higher than either of the individual base classifiers. Cross-validation is used as evaluation metrics to measure the performance or effectiveness of the ensemble classifiers while the experimental results shows that the ensemble classifiers achieve the best results compared to the single based classifier in terms of Precision, Recall, F1-measure and Accuracy.

Keywords: Spam Review, Detection, Ensemble Classifier, Arching Classifier, Weighted Voting

1. Introduction

The world is witnessing more and more participation in modern electronic commerce, where online review is playing a vital role. Customers now engage in reading reviews on products and stores when they are making decisions on what to buy or where to buy it. Spam reviewers seized this opportunity to write malicious reviews to discredit decent stores or use fake reviews to deceive customers on low quality products. This is often regarded as spam review. These spam reviews had posed a serious threat to e-commerce, with individuals, companies, cooperates and organizations loosing huge sum of fortune in the process.

Now, not only do potential consumers search these reviews to make purchase decisions but are also used by manufacturers to identify defects in their products as well as competitive information on their potential competitors [1]. Thus, opinion spamming or review spam are terms used in identifying fake reviews which were deliberately concocted to deceive potential users or opinion mining systems by providing them with undeserving positive opinions or false negative opinions to promote or downgrade some targeted products. Identifying and preventing similar spam are one of the top challenges faced by web search engines today. Amit Singhal, principal scientist of Google Inc. estimated that the search engine spam industry had a revenue potential of $4.5 billion in year 2004 if they had been able to completely fool all search engines on all commercially viable queries [2]. Google also pointed out concerns of fake reviews in an official report and clearly directed the innovators and users to not purchase and receive payments from firms that make available false reviews [15].

However, different companies or organization have different method of collecting reviews from customers. According to [12], Reviews at Xianyu, the largest second-hand goods app in China, differ from reviews in other e-commerce websites in several aspects. users have no idea about the quality and possible lowest price of the second-hand goods. Thus, reviews at Xianyu act as a communication tool for buyers and sellers (e.g., query for details and discounts) and review action usually happens before purchase.

It is generally believed that manual distinguishing of spam and truthful reviews is so formidable or even impossible for human. Therefore, there is no spam or non-spam gold standard dataset to be used as training data in machine learning algorithms to exploration spam reviews. Existing approaches in spam review detection mainly focused on exacting linguistic features and behavioral features. However, according to [13], linguistic features are ineffective when they are used to detect the real-life fake reviews and it usually requires a large number of samples to make the observations on behavior features.

However, some spam detection methods are proposed using diverse aspects of reviews as features. Some other researchers used both of the unsupervised and supervised as well as ensemble methods in their approaches and at the end, they have compared the results. Though previous work was promising, there still remains a precarious research gap in the fight against review spam. This gap is largely due to the efficiency exhibited by this logic machines to carry out accurate predictions during detection. Therefore, this research is focused on developing a review spam identification engine that first build an ensemble of classifiers and then employed to identify spam reviews from real ones.

2. Related Work

Ensemble techniques involve the analysis of reviews, mostly at content level, and employ classification algorithms, such as Bayesian, Support Vector Machines, and others to segregate spam from legitimate review. These approaches have been extensively applied in spam filtering and exhibit different capabilities. Recently, researchers in [3] designed a classifier ensemble using Naïve Bayes (NB), Support Vector Machine (SVM) and Genetic Algorithm (GA). The ensemble model shows higher percentage of classification accuracy than the base classifiers and enhances the testing time due to data dimensions reduction and significant improvement over the single classifiers.

In the meanwhile, a discriminant model has been proposed in [4] that combined logistic regression with Naïve Bayes to form ensemble classifiers. The authors observed that Naïve Bayes approaches its asymptotic error without the need for a large number of training examples, and it does so very quickly. Logistic regression, on the other hand, is capable of outperforming naive Bayes, given the number of training examples is large enough [11]. The overall classification result was also observed to be superior to that of the base classifiers due to their respective diversity.

An ensemble model of Naïve Bayes and Logistic Regression was also designed by [5], which employs supervised learning that is partly generative and partly discriminant. In order to exploit the generative, joint probability from the inputs and outputs of the supervised learning task, large portion of the subset are trained. Furthermore, to further exploit the conditional probability of the outputs for given inputs, another much smaller subset of the parameters are discriminatory trained. Over the years, various researchers have combined these classifiers in different form of ensemble classifications.

3. The Proposed Methodology

The proposed methodology consists of four phases as illustrated in Figure 1: data collection, pre-processing, implementation and performance measurements. The detail for each phase is explained in the following sections.

3.1. Data Collection

The dataset employed for the purpose of this research are customer reviews from various categories that are originally collected from amazon.com by [6]. To prepare the dataset in a practicable and suppler manner, DOM Parser was used to parse the XML version of the dataset in Java before the other categories are discarded. To eliminate review duplicates, researcher in [7] presented four methods for finding duplicates in SAS dataset using SAS versions 6 and 8. The first three utilize various combinations of the SORT procedure, the FREQ procedure, and the DATA step, while the fourth is a SAS macro that allows greater flexibility for dealing with duplicates. A current study in [8] implemented the third method, which is PROC SORT THEN DATA STEP to eliminate reviewers with multiple reviews or duplicates from the dataset. The final library dataset is in a directory of 2000 reviews with 2 sub directories of 1000 positive and 1000 negative reviews respectively.

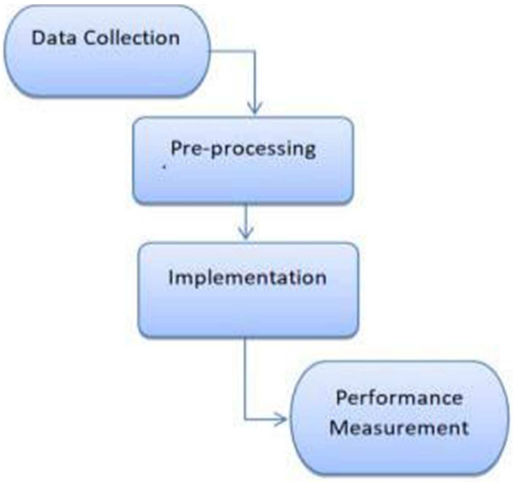

Figure 1. *The proposed methodology.*

3.2. Pre-processing

Before the dataset can be employed for classification purposes, it is essential that it is processed into a format suited for the classifiers and one which can increase their accuracy and speed of processing. Processing resources like stop-word elimination and stemming are applied to process the dataset into a reduced corpus. English words that are functional like ‘a’, ‘and’, ‘of’, ‘the’ etc., often occur frequently but are not of any use in classification are referred to as Stop-words. Stemming on the other hand, is the process of plummeting words by shortening and taking them to their root or base form. The final output of this phase is a training text that is reduced to more than one-third of its original length and after the features identified, the corpus is converted into vector equivalent.

Many data analysis software packages provide for feature extraction and dimension reduction. Common numerical programming environments such as MATLAB, SciLab, NumPy and the R language provide some of the simpler feature extraction techniques (e.g., principal component analysis) via built-in commands. Other popular suites in machine learning software that provide for numerous phases of complex data mining and analysis are WEKA application and GATE application. More specific algorithms are often available as publicly available scripts or third-party add-ons. For the purpose of this research, WEKA application, a collection of machine learning algorithms for data mining tasks is used for both pre-processing, ensemble as well as classification.

3.3. Implementation

The proposed architecture is illustrated in Figure 2. The core processes which are feature extraction and classification are retained. But, when considering ensemble of classifiers, the combination of the output of a number of classifiers is only valuable if they conflict on some inputs. This study refers to the measure of disagreement as the diversity in classifier ensemble (i.e., voting). There have been several methods proposed to measure ensemble diversity, usually dependent on the measure of accuracy. Therefore, it is pertinent to highlight how the ensemble ended up with Naïve Bayes, SVM and logistic Regression or how these base classifiers where chosen.

1) Base Classifier Selection: When considering an ensemble of classifiers, the combination of the output of a number of classifiers is only valuable if they conflict on some inputs. This study refers to the measure of disagreement as the diversity in classifier ensemble [9]. Bagging is a concept that implements this same philosophy with a reasonable amount of precision. Build-in classifiers in the WEKA application where used to create the bagging models. It creates separate samples of the training dataset and creates a classifier for each sample. The results of these multiple classifiers are then combined (such as averaged or majority voting). The trick is that each sample of the training dataset is different, giving each classifier that is trained, a subtly different focus and perspective on the problem.

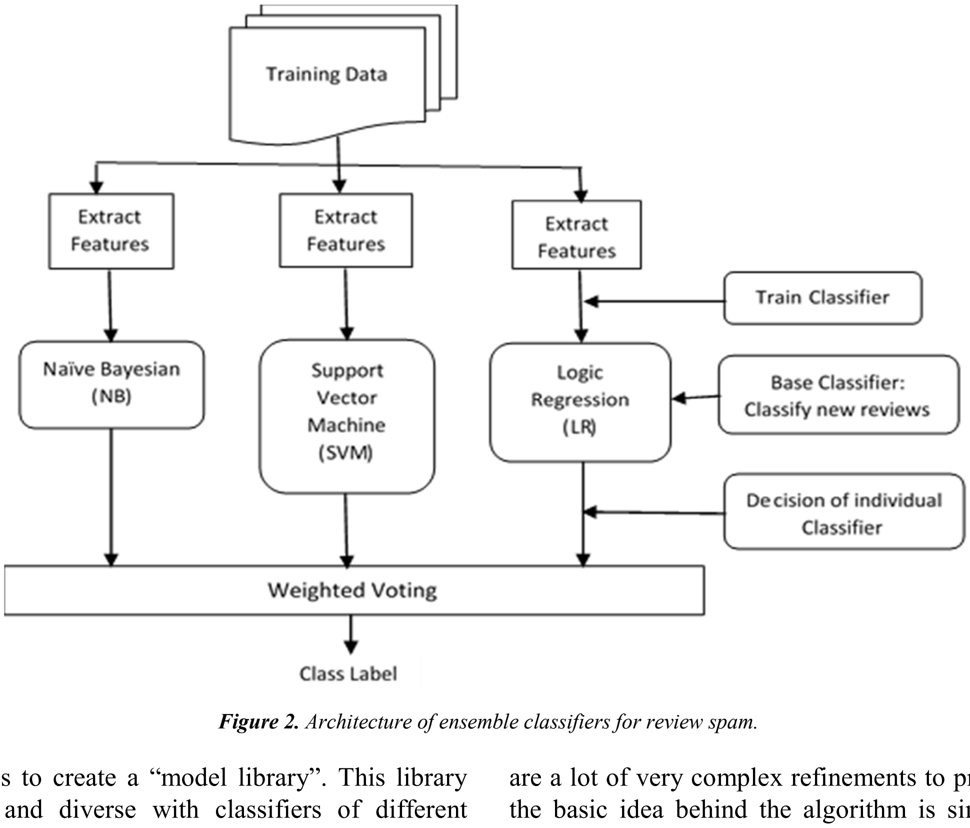

Figure 2. *Architecture of ensemble classifiers for review spam.*

The first step is to create a “model library”. This library should be large and diverse with classifiers of different parameters. Keeping in mind that it should not really hurt performance to train ‘bad’ models as they will simply not be chosen by the ensemble selection algorithm if they hurt performance. The second step is to combine models from the library with the ensemble selection algorithm. Although there are a lot of very complex refinements to prevent over fitting, the basic idea behind the algorithm is simple. Basically, it starts with best model from the whole library. At this point the ensemble has only one model in it. Then models are added one at a time. This is done for some number of times and the weights determined by ensemble selection for each bag are added until the final ensemble of three base classifiers: Naïve Bayes (NB), Support Vector Machine (SVM) and Logistic Regression (LR) were produced.

2) Ensemble Classification: After the base classifiers are identified the next step is to train and test all the three base classifiers using the same training and test sets. WEKA is used to implement the ensemble classification phase which was possible through the choice voting algorithm provided. Generally, voting in machine learning counts the number of times a given data instance was predicted for each class. The aggregate prediction favors the class that received the most votes. In situations where multiple classes received the same number of votes, the predicted class will be selected at random from those classes. In an attempt to place most emphasis on individual predictions that are most informative, weighted voting assigns weights to individual predictions based on the AUC attained via nested cross validation for the relevant combination of data category and algorithms.

Since all the built-in classifiers Naïve Bayes, Support Vector Machines, Logistic Regression in WEKA are identified, they are configured for the ensemble. The CLI can be used to declare these classifier algorithms for ensemble process or simply from the class of combining classifiers where the three classifiers are added into the Generic Array Editor one after the other. The voting itinerary provides a set of choice for combination rule where majority voting was used in this current research. In the test option, cross validation was utilized where the folds are set to 10 through the entire process. Once command is ushered for processing, ensemble cog runs the three classifiers algorithm using the same training and test set and then use majority voting to predict classes for all instances.

3) Weighted Voting: When a vote deals with only two alternatives, all reasonable voting methods have the same outcome as majority rule. A weighted voting system is one in which the preferences of some voters carry more weight than the preferences of other voters. Dynamic weighted voting was used for each of the last classification fold where the weights changes with each input vector in the operational phase. Mainly the classifiers were assigned weighs during the training phase of this current research. A conceptually similar technique is the mixture-of-experts model, where the set of classifiers Logistic, Naïve Bayes that constitute the ensemble, and followed by a second level classifier, arching classifier, was used for assigning weights for the consecutive combiner. The gating network is trained through the Expectation Maximization (EM) algorithm [16]. The inputs to the gating network are the actual training data instances themselves (unlike outputs of first level classifiers for stacked generalization), hence the weights used in combination rule are instance specific, creating a dynamic combination rule. The combination rule selects the most appropriate classifier, or classifiers weighted with respect to their expertise, for each instance *x*.

For a given set *D*, of *d* tuples, arcing can be demonstrated using the following steps; For iteration *i*=3 and the set of training data, D_i, of *d* tuples is sampled using the original tuple sets, *D*, in place. The training dataset *D* will witness the presence of examples from dataset *D* more than once. The rest of examples that could not make it into the training dataset are used as test data. Classifier model, M_i, is then learned using training dataset D_i for each training examples *d*. A classifier model, M_i, is learned using training set, D_i. In order to carry out classification of an unknown tuple, *X*, each classifier, M_i, will then return its class prediction, where vote counts depends on classifier weighs. The hybrid classifier (NB, SVM and LR), *M**, compute the votes and then assign the class with the most votes to *X.*

3.4. Performance Measurements

In this section the evaluation metrics used in measuring the classifier performances are highlighted and mathematically represented. The computed results from implementing these matrices are summarized. Both table and graphical comparison as well as result discussion was used to explain the classifier performances. This study employs Cross Validation Technique to measure the performance of the base and new ensemble classifier. The matrices for the evaluating the performance of the classifiers is thus represented by (1) until (4); Spam Precision (*SP*), Spam Recall (*SR*), Spam F1-measure (*F1*) and Accuracy (*A*) to be calculated. Let *TN*=number of legitimate reviews classified as legitimate (true negatives), *TP*=number of spam reviews classified as spam (true positives), *FP*=number of legitimate reviews classified as spam (false positives), *FN*=number of spam reviews classified as legitimate (false negative).

$$SP = TP/(TP + FP) \qquad (1)$$

$$SR = TP \ / \ (TP + FN) \qquad (2)$$

$$F1 = (2 \times SP \times SR) \ / \ (SP + SR) \qquad (3)$$

$$A = (TP + TN) \ / \ (TP + FP + TN + FN) \qquad (4)$$

Weighted Accuracy (WA) of the ensemble classifier was also calculated using (5);

$$A = (\lambda TN + TP) \ / \ (\lambda(TN + FP) + TP + FN) \qquad (5)$$

For a stratified K-fold cross-validation, folds are chosen in such a way that the mean response value is approximated to the rest of the folds equal [10]. Implementation of the current research was carried out with a *k*-fold cross validation where *k* 10. This means that the dataset used was separated into 10 different learning sets. The testing set is 40% while the remaining was used as training set.

The key metric which is used in the evaluation of classifier performance is widely believed to be classification accuracy which accounts for the percentage of test samples where it has been correctly classified. Accuracy of a given classifier is referred to as the capacity for that classifier to make correct predictions on the label of new or previously unseen data (i.e., tuples without class label information). Likewise, the accuracy of a given predictor is how well a given predictor can guess the value of the predicted attribute for new or previously unseen data.

4. Experimental Results and Analysis

Table 1 summarizes some of the evaluation metrics which are Precision, Recall and F1-measure for both base and ensemble classifiers. Figure 3 illustrates the comparison of the performance metrics in a graphical form. The precision, which is the fraction of retrieved instances that are relevant, can be seen to have jumped higher than that of all the base classifiers in the ensemble. It can be concluded that the positive predictive value had improved with the ensemble. Spam Recall, which is the fraction of relevant instances that are retrieved, has also appreciated and thus the sensitivity of ensemble classifiers, is superior to that of the base classifiers. The F1-measure is a measure of a test's accuracy. It considers both the Spam Precision and the Spam Recall of the test to compute the score, which is why it can be seen closely to both as they increase.

Table 1. *Evaluation results for base and ensemble classifiers.*

Classifier	Precision	Recall	F1-measure
SVM	0.790	0.790	0.790
Naïve Bayes	0.809	0.808	0.808
Logic Regression	0.639	0.639	0.638
Ensemble	0.882	0.881	0.881

Table 2. *Accuracy for base and ensemble classifiers.*

Classifier	Accuracy (%)
SVM	79.00
Naïve Bayes	80.80
Logistic Regression	63.85
Ensemble	88.09

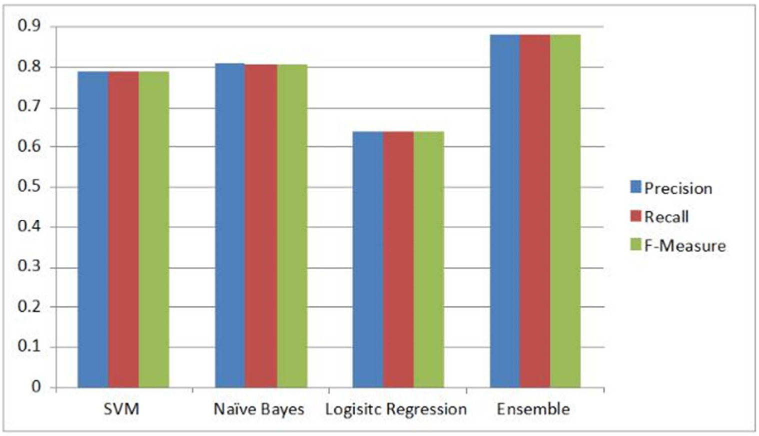

Figure 3. *Comparison on Precision, Recall and F1-measure for base and ensemble classifiers.*

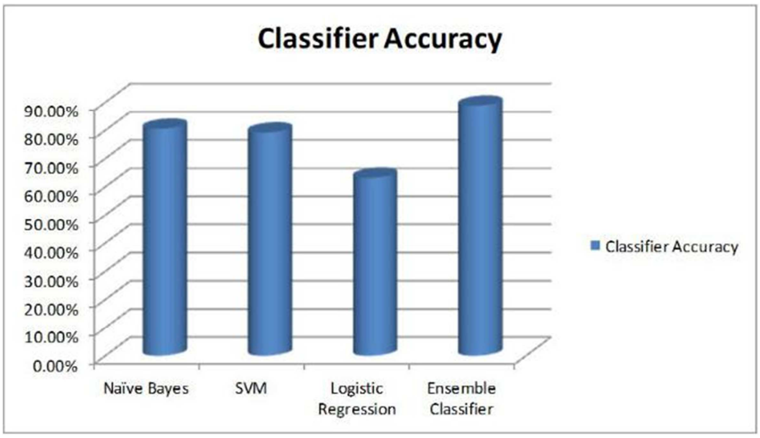

Figure 4. *Comparison on Accuracy for base and ensemble classifiers.*

The model of the current research has indicated better and larger improvement in classification accuracy when these classifiers are tested. The approach displayed better accuracy than the performance of the individual base classifiers with statistically significant results. For instance, the F1 measure in [12] for even the embedding smoothing on the subset of training samples is 0.8448 with 71.02% precision. Also, researchers in [13] where able to achieve 78.1% and 75.4% accuracy for hotels and restaurants respectively, using unlabeled dataset. Researchers from [15] in an attempt draw classification results using top 10 features selected by Chi-square., where able to achieve 89.26% accuracy using ensemble classifier.

The approach in the current research of combining Naïve Bayes, Support Vector Machine (SVM) and Logistic Regression classifiers are found to be superior to individual approaches to spam review detection in terms of classification accuracy. Table 2 highlights the performance of each of the three base classifiers and that of the ensemble classifier, while Figure 4 depicted the classification accuracy in graphical format.

From the graph, Logistic Regression performed the least followed by SVM that is very closely followed by Naïve Bayes and finally the ensemble classifier with the greatest performance. The difference between the least performed classifier, Logistic Regression and Support Vector Machine stood at an average of 15%. Support Vector Machine and Naïve Bayes exhibited a near performance, with Naïve Bayes leading with just an average of 1.8%. On the other hand, the ensemble classifier surpasses all the base classifiers with an average difference of 8.1% to Naïve Bayes itself.

5. Conclusion and Future Work

This research presented the model, operational framework and the methodology of designing and investigating the architecture of an ensemble classifier to perform the function of spam review detection. The model was designed using Naïve Bayes, Support Vector Machine and Logistic Regression as base classifiers. These base classifiers are justifiably selected through the process of 'bagging', where each was assigned a weight depending on accuracy of prediction. The base classifiers in the ensemble perform classification in parallel which can be implemented using WEKA application, MATLAB or other effective tools.

Arching classifier is used to perform weighted voting for overall class label. Lastly, cross-validation is used as evaluation metrics to measure the performance or effectiveness of the ensemble classifiers. The experimental results show that the ensemble classifiers (SVM, Naïve Bayes and Logistic Regression) achieve the best results compared to single based classifier in terms of Precision, Recall, F1-measure and Accuracy.

Therefore, it can be concluded that the aim of the research was achieved. Nevertheless, there are possibilities this performance can be improved by future researchers when different dataset, feature combinations and/or when other parameter settings are explored. Some of the future work that worth to be explored are combination of different classifier in the ensemble architecture. Investigation on the effectiveness of the ensemble classifiers on real and bigger dataset is also an open research issue.

References

[1] N. Jindal, and B. Liu, "Analyzing and detecting review spam," Seventh IEEE International Conference on Data Mining, 2007, pp. 547-552.

[2] A. A. Benczur, K. Csalogany, T. Sarlos, and M. Uher, "SpamRank– fully automatic link spam detection," Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web, 2005.

[3] G. Muthukumarasamy, "Spam review detection using a hybrid classification method," International Journal of Advances in Engineering Sciences, vol. 4, 2014, pp. 22-27.

[4] J. Halloran, "Classification: naive bayes vs logistic regression," Technical report, University of Hawaii, 2009.

[5] R. Rajat, S. Yirong, Y. N. Andrew, and M. Andrew, "Classification with hybrid generative/discriminative models," The Annual Conference on Neural Information Processing Systems (NIPS), 2003.

[6] J. Mcauley, and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text. ACM Proceedings of the 7th ACM Conference on Recommender Systems, 2013, pp. 165-172.

[7] B. J. Peterson, "Finding a duplicate in a haystack," Proceedings of the Thirty-first Annual SAS® Users Group, 2006.

[8] Y. Bi, "The Impact of diversity on the accuracy of evidential classifier ensembles," International Journal of Approximate Reasoning, vol. 53 (4), 2012, pp. 584-607.

[9] P. Melville, and R. J. Mooney, "Constructing diverse classifier ensembles using artificial training examples," IJCAI Citeseer, 2003, pp. 505-510.

[10] F. Barigou, N. Barigou, and B. Atmani, "Spam detection system combining cellular automata and naïve bayes classifier," ICWIT Citeseer, 2012, pp. 250-260.

[11] Hussain N, Turab Mirza H, Rasool G, Hussain I, Kaleem M. Spam Review Detection Techniques: A Systematic Literature Review. Applied Sciences, 2019; 9 (5): 987.

[12] Li A, Qin Z, Liu R, Yang Y, Li D. Spam review detection with graph convolutional networks. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 2703-2711.

[13] You Z, Qian T, Liu B. An attribute enhanced domain adaptive model for cold-start spam review detection. In Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1884-1895.

[14] Hussain N, Mirza HT, Hussain I, Iqbal F, Memon I. Spam review detection using the linguistic and spammer Behavioral methods. IEEE Access, 2020, 8: 53801-16.

[15] Fayaz M, Khan A, Rahman JU, Alharbi A, Uddin MI, Alouffi B. Ensemble Machine Learning Model for Classification of Spam Product Reviews. Complexity, 2020.

[16] Shahariar GM, Biswas S, Omar F, Shah FM, Hassan SB. Spam review detection using deep learning. In 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2019, (pp. 0027-0033). IEEE.