International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 6 | ISSN : 2456-3307

Semi-supervised Learning with Ensemble Method for Online Deceptive Review Detection

Miss. Priyanka Shinde, Prof. Hemlata Channe
Pune Institute of Computer Technology, Pune, Maharashtra, India

ABSTRACT

Now-a-days not only organizers but users also prefer to give opinion after using any kind of resource. Opinion of user is very important for business. Because of opinion of actual user further consumers should think to use that resource. In Business, opinion review has great impact to economical bottom line. Unsurprisingly, opportunistic individuals or groups have attempted to abuse or manipulate online opinion reviews (e.g., spam reviews) so that they credit or degrade the target product. Because of this detecting deceptive and fake opinion reviews is a topic of ongoing research interest. In this paper semi-supervised learning approach with ensemble learning methods is used for finding out these spam reviews. Utility is demonstrated using a data set of online hotel booking websites.

Keywords : Opinion Spam, Multilabel and Multiclass, Ensemble of classifiers , Co-training, PU learning, EM algorithm

I. INTRODUCTION

As with more end users are using online opinion reviews to inform their service decision making,opinion reviews have an economical impact on the bottom line of business. Opinion spamming is becoming more sophisticated and, in some cases, organized, due to the potential to profit from such activities. For example, some businesses reportedly recruited online users such as professional fake review writers to post fake opinions.[1] These opinions can be used to market and promote a particular business, spread rumors and damage the reputation of a competing business, or influence online users opinions and views about a particular topic[2]. While supervised learning has been traditionally used to detect fake reviews, supervised learning approaches suffer from several limitations. For example, unless one can be assured of the "quality" of the reviews used in the training dataset, we will have a garbage-in-garbage-out situation. In addition, the amount of labeled data points used to train the classifier can be difficult to obtain and update, given the dynamic nature of online reviews Some limitations in supervised learning methods could be addressed using automatic labeling, a process known as semisupervised learning[1][3]. In the latter, a large number of unlabeled data points are used, instead of labeled data points. As such, labeled data points can be sparsely present and using those points, labels of the unknown instances are automatically generated first, which can then be used to train a classifier and generated the review [4].

In other domains, it has been found that using unlabeled data in conjunction with a small amount of labeled data can considerably improve learner accuracy compared to completely supervised methods. a two-view semi-supervised method for review spam detection was created by employing the framework

of a co-training algorithm to make use of the large amount of unlabeled reviews available. The cotraining algorithm is a bootstrapping method that uses a set of labeled data to incrementally apply labels to unlabeled data.[1] It trains 2 classifiers on 2 distinct sets of features and adds the instances most confidently labeled by each classifier to the training set. This effectively allows large datasets to be generated and used for classification, reducing the demand to manually produce labeled training instances.[3][5] A modified version of the co-training algorithm that only adds instances that were assigned the same label by both classifiers was also proposed. Their dataset was generated with the assistance of students who manually labeled 6000 reviews collected from Epinions.com, 1394 of which were labeled as review spam. Four groups of review centric features were created: content, sentiment, product and metadata. Another two groups of reviewer centric features were created: profile and behavioral[6].

In order to use the two-view method for adding unlabeled instances to the training set, classifiers were trained on each set of features (i.e., one with review centric features and another with reviewer centric ones). Note that these 2 classifiers are only used to add instances to the labeled data and the final classifier is trained using all available features, both review centric and reviewer centric. Experiments were conducted using Nave Bayes, Logistic Regression and SVM with 10-fold cross validation, and it was found that Naive Bayes was the best performer, so all additional work was performed with Nave Bayes.[3] They observed that using the co-training semi-supervised method, they were able to obtain an F-Score of .609, which was higher than the 0.583 they obtained when not including any unlabeled data. Further, it was observed that by using their co-training with agreement modification, they were able to raise this value to 0.631. While these F-Scores appear low, it is hard to compare them with the performance from other studies as they used their own dataset. The results do seem to indicate that this type of semi-supervised learning may indeed help in the area of review spam detection and demands further study with additional datasets[3][7].

PU-Learning is a second type of semi-supervised learning approach, this is used to learn from a few positive examples and a set of unlabeled data. Montes-yGmez and Rosso adapt this approach for review spam detection in their work Using PU-Learning to Detect Deceptive Opinion Spam [3]. PU-learning is an iterative method which tries to identify a set of reliably negative instances in the unlabeled data. The model is trained and evaluated using all of the unlabeled data as the negative class and any instances that are classified as positive are removed. The process is repeated until some stop criterion is reached. For evaluation purposes, the dataset generated by was used and the performance was evaluated using F-Measure. Classifiers were trained using both Nave Bayes and SVM as learners. PU-learning achieved an F-measure of 83.7 percentage with NB, using only 100 positive examples. While this is better than the results achieved using 6000 labeled instances and co-training it is difficult to make a conclusive statement as the methods use different datasets and, as previously discussed, the dataset created by Ott et al. may not provide an accurate indication of real world performance[3][7][8].

There are various approaches that can be used for semi-supervised learning. These include Expectation Maximization, Graph Based Mixture Models, Self-Training and Co-Training methods.In the similar system of semisupervised learning for online deceptive review detection some new features are extracted in which author shows that by incorporating new dimension in feature vector gives better results. These extracted features are as follows: Part Of Speech Tags (POS tags),Linguistic Inquiry

and Word Count , Sentiment polarity and Bigram frequency count [1].

Proposed Method: In our dissertation, we will be focusing on applying the Self-Training approach to Yelps reviews[1][3][7]. In self-training, the learning process employs its own predictions to teach itself. An advantage of self-training is that it can be easily combined with any supervised learning algorithm as base learner [6][18] . We will be using three different supervised learning methods - Nave Bayes,Decision Trees and Logistic Regression as base learners. We would then be comparing the accuracy of each of the semi-supervised learning methods with its respective base learner. The base learners would be using both behavioral and linguistic features as mentioned above.[9]So, here the dissertation idea is that we are going use ensemble method for these learned classifiers such as twin SVM with naive bayes to solve multilabel and multiclass in text categorization. For learning we are going to use three semi-supervised learning algorithms that are cotraining ,expectation maximization and PU learning algorithms.

II. SEMI-SUPERVISED LEARNING WITH ENSEMBLE METHOD

In semi-supervised learning there is a small set of labeled data and a large pool of unlabeled data. We assume that labeled and unlabeled data are drawn independently from the same data distribution. In our project, we consider datasets for which $n_l << n_u$ where n_l and n_u are the number of labeled and unlabeled data respectively. First, we use Nave Bayes as a base learner to train a small number of labelled data.The classifier is then used to predict labels for unlabeled data based on the classification confidence. Then, we take a subset of the unlabeled data, together with their prediction labels and train a new classifier. The subset usually consists of unlabeled examples with high-confidence predictions above a specific threshold value .

In addition to using Naive Bayes, we are also planning to use Decision Trees and Logistic Regression as base learners. The performance of each of the semi-supervised learning models would then be compared with its respective base learner here we use naive bayes classifier.[9][2][10]

Naive Bayes is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class. This is also known as Maximum A Posteriori (MAP)[18][20].In this equation E is the evidence while H is the prior probablity.

1) Bayes Theorem:

$$P(H/E) = \frac{(P(E|H) * P(H))}{(P(E))}$$

2) The MAP for a hypothesis is:

$$MAP(H) = max(P(H/E))$$
$$= max((P(E/H)*P(H)) / P(E))$$
$$= max(P(E/H)*P(H)$$

A. Co-Training Algorithm

This algorithm used for a large unlabele sample (U) to boost the performance of a learning algorithm when only a small set of labeled (L)examples is available. In particular, we consider a setting in which the description of each example can be partitioned into two distinct views.

Initially, a collection of data points is chosen, of which some are labeled (L) and the others are unlabeled (U). The U set is then iteratively exhausted by incrementally. earning and classifying member instances to the L set. First, u instances are considered at random from U and inserted into a set

U. Each instance is a composition of two views, x_1 and x_2. The algorithm then runs for k iterations or until the set U is exhausted. In each iteration, a classifier h1 is trained on only the x_1's view of the instances in L , and another classifier h_2 on only the x_2's view of the instances in L ,here h_1 and h_2 are naive bayes classifiers .Each classifier is allowed to label p positive and n negative instances, which are added to the set L . Finally, 2(p + n) examples are randomly sampled from U and are used to replenish U'[21].

The co-training algorithm is described in Algorithm 1 [1].

Algorithm 1: Co-training Algorithm

Input : Labeled instance set L and unlabeled instance set U

Output: Deployable classifier , C

1 Create set of unlabeled examples, U' ,by randomly sampling u examples from U';
2 **for** *each feature vector x in L U U* **do**
3 Partition x to tuple of views, (x_1, x_2) ;
4 **for** *k iterations* **do**
5 $h_1 \leftarrow train(x_1) \forall (x_1, x_2)$;
6 $h_2 \leftarrow train(x_2) \forall (x_1, x_2)$;
7 Let h_1 label p positive and n negative examples from U';
8 Let h_2 label p positive and n negative examples from U';
9 Add labeled examples to L;
10 Randomly sample $2(p+n)$ examples from U to U';

Figure1. Co-training algorithm

B. EM Algorithm

Expectation-Maximization (EM) algorithm to learn a classification model from a small set of labeled examples and a large set of unlabeled examples .This Data can be used in any supervised learning classification algorithm like In Base paper classifier used are KNN, Logistic Regression, Random Forest. In this dissertation same classifier used in EM for training and testing of multiple classifiers.

Here, the learning of the algorithm with the conjunction of the labeled and predicted labeled sets is the Expectation step (E-step) and the prediction of the labels of the unlabeled set is the Maximization step (M-step) .The pseudocode for EM learning is described in Algorithm 2 [1].

Algorithm 2: Expectation-Maximization Algorithm

Input : Labeled instance set L and unlabeled instance set U

Output: Deployable classifier , C

1 $C \leftarrow trainL$;
2 **while** *true* **do**
3 PU = Predict (C,U);
4 **if** *PU same as in previous iteration* **then**
5 return C;
6 **else**
7 $C \leftarrow train(LUPU)$;

Figure 2. EM algorithm

C. PU Learning

In order to clarify the construction of the opinion spam classifier, Algorithm 3 presents the formal description of the proposed method. In this algorithm P is the set of positive instances and U_i represents the unlabeled set at iteration i; U_1 is the original unlabeled set. C_i is used to represent the classifier that was built at iteration i, and W_i indicates the set of unlabeled instances classified as positive by the classifier C_i. These instances have to be removed from the training set for the next iteration. Therefore, the negative class for next iteration is defined as U_i,W_i. Line 4 of the algorithm shows the stop criterion that we used in our experiments, $|W_i| <= |W_{i-1}|$ [22].

The idea of this criterion is to allow a continue but gradual reduction of the negative instances [8].Pseudocode for PU learning is described in Algorithm 3

Algorithm 3: PU Algorithm

Input : Labeled instance set L and unlabeled instance set U

Output: Deployable classifier , C

$i \leftarrow 1$;

$|W_0| \leftarrow |U|$;

$|W_1| \leftarrow |U|$;

while $|W_i| \leq |W_{i-1}|$ **do**

  $C_i \leftarrow train(P, U_i)$;

  $U_i^L \leftarrow predict(C_i, U_i)$;

  $U_i^L \leftarrow extract\,positives(U_L^i)$;

  $U_{i+1} \leftarrow U_i - W_i$;

  $i \leftarrow i+1$;

return C_i ;

Figure 3. PU learning

D. Ensemble Method

Ensemble learning algorithms train multiple classifiers and then combine their predictions. Since the generalization ability of an ensemble classifier can be much better than a single learner, the algorithms and applications of ensemble learning have been widely studied in recent years. To solve the multi-class and multilabel problem of text categorization by binary SVM, a novel text categorization method based on twin-SVM with nave Bayes ensemble is proposed.[20]

Twin-SVM for multi-label:

1. To solve the multi-label problem, for each pair of training sets c_1 andc_2 sharing common training samples, we proposed a twin-SVM method which respectively trains two binary classifiers SVM_1 to distinguish c_1 against c_1-c_2 and SVM_2 to distinguish c_2 against c_2 -c_1.

2.The combination the two SVM called a twin-SVM may predict a sample to be classified to both classes, that is, a twin-SVM may give votes to both the two parties it wants to differentiate. Thus the combination by the decomposition-based strategies of all the twin-SVMs may classify a testing sample to more than one class[20].

Twin-SVM with nave Bayes ensemble for multi-class :

1. The main idea of the twin-SVM with Naive Bayes ensemble is not to distinguish among all the classes but the most likely classes a testing sample may belong to.

2. In the training phase, we firstly train a nave Bayes classifier for all the classes. Secondly, like onevs-one, we train twin-SVM classifiers for every pair of all the classes.

3. In the testing phase, we select the top ranked classes of Naive Bayes by the principle that the sum of their posterior probabilities is bigger than a threshold of . The label decision strategy for twin-SVMs is based on the validation results of the naive Bayes classifier so that possible labels with lower posterior probabilities of the selected classes are refined by the twin-SVM classifiers. The proposed method takes advantages both of the fast speed of the naive Bayes and the high precision of the SVM[19][20].

III. RESULTS AND DISCUSSION

In this paper we studied semisupervised learning approaches and ensemble method for online deceptive review detection system,proposed method gives the better performance in terms of classification accuracy. the available dataset was partitioned into subsets with sizes in the ratios of a a: (100 - a), where a assumes values in (75: 80: 90). In each process described, (0.2 * a)% instances were taken as labeled training dataset and the rest as unlabeled training dataset. Also, four variations of classifiers were used across all evaluations, namely thek-Nearest Neighbor classifier (k-NN), the Logistic Regression classifier, the Random Forest classifier and the Stochastic Gradient Descent classifier. For the k-NN classifier, the value of `k' was chosen as 4. Also, for the Random Forest classifier, 100 worker instances were used for evaluations.The algorithms implemented and their results are presented in Sections A to D.

A. ***Co-Training Algorithm***:

In [24] author considers two dimensions in feature vector for classification of web spam data. In this paper the dataset used is more richer in sense that it considers 15 dimensions in feature vector. So as per the algorithm the feature vector is randomly partitioned in two views and then algorithm applies on it.For the evaluation, the best score obtained is 73.25% while cross-vaildation accuracy obtained for cotraining algorithm on k=10 is 70.48% In this particular evaluation, the dataset was divided in a 75:25 partition for training and test dataset. Of the training dataset, 20% of the instances were chosen as labeled and the rest as unlabeled. The k-NN classifier was used for the evaluations.

B. ***Expectation Maximization Algorithm***

In previous cotraining algorithm the dataset is divided for training phase and for testing phase same ratio is considered for the EM as well. Classifier is first derived from form divided labled dataset and is then predicts the lables for remaining unlabled data.The process continues until algorithm stops. The best score for EM algorithm are,the classification accuracyis83% and cross-validation accuracy is 81.86%.we have used k-NN classifier for final evaluation.

C. ***Positive Unlabeled Learning***

The PU learning algorithm used by D. Hernández [8] for classifying web spam data in spam and non-spam categories. The same approach in this paper is used for classifying hotel reviews in deceptive and truthful class.For this classification dataset [8] is as input for PU learning algorithm. The best results were obtained when the dataset was partitioned 80% for training and 20% for testing. Out of the 80% training data comprising 1280 reviews, 256 positively labeled reviews were chosen as labeled instances and the remaining instances were treated as unlabeled.A balanced mix of 320 data points was chosen for testing purposes as compared to 160 used in [] and the same in [38], [45] which reported a maximum F-Score of 0.837 when applied only on the set of deceptive reviews with the mentioned dataset partitioning scheme and an undisclosed accuracy of classification.

In this paper the best results obtained by using PU learning algorithm for classification accuracy is 81.25% and cross-validation accuracy is 79.98%.

The proposed method obtains better results than the previous results, classification accuracy obtained for semi-supervised learning with ensemble method is 92.75%.Where,cross-validation accuracy obtained here is 89.46%.Results shows that the proposed method gives the best results .Fig shows the graph of the classification accuracy of Co-training algorithm, EM algorithm , PU learning algorithm and Ensemble method ,which shows that ensemble method outperforms than individual semi-suppervised learning approach.

D. ***Ensemble Method***

Ensemble method basically used in machine learning to improve the overall results.So as the results obtained for above three semi-supervised learning algorithms are compared with our proposed method. In semi-supervised learning with ensemble method the main goal is to run the three basic leaning algorithms and apply the ensemble method on these semi-supervised learing approach. Obeservation shows that twin-SVM with Naïve Bayes ensemble method outperforms the other two methods in multi-label and multi-class classification. Although for single label and multiclass classification, one-vs-one SVM is very effective, it can rarely make precise perditions for multi-label samples because there will be only one label getting the most votes from all the binary classifiers. This results the lower performance of one-vs-one than that the combination of twin-SVM classifiers where likely labels may get the same number of votes.

In fig. 4 classification accuracy analysis is shown by using graph, it shows that proposed method outperforms than previous similar system which uses semi-supervised learning approach.

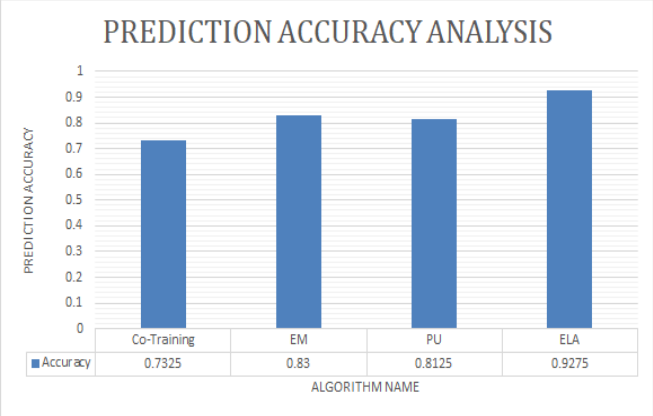

Figure 4. Prediction accuracy analysis

In fig. 5 cross-validation analysis is shown,where proposed method have CV score 0.92 which is better than semi-supervised algorithms. The CV analysis is generated by using k = 10, 20, 50.

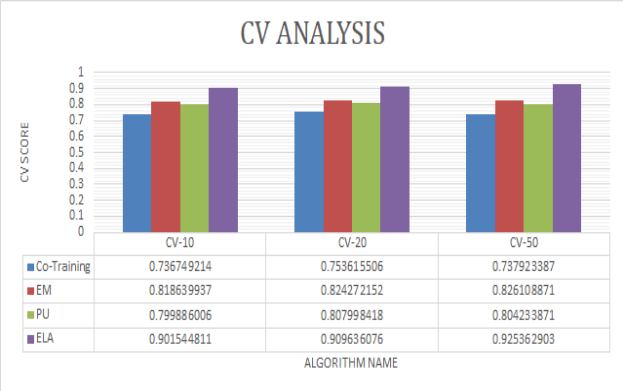

Fig.5 Cross-validation analysis

IV. CONCLUSION

In today's era everyone is so concern to use the best service in less expenditure. For this they cheak the reviews of that product or service on respective website. Opinions about that particular service if positive it gives lots of profit and fame to the business. Unfortunately this makes some imposters to post some fake review to make credit or discredit the target product or service.

In this paper we propose that semi-supervised learning approach with ensemble method gives the best accuracy in online deceptive review detection. Classification accuracy achieved for proposed method is 0.92 which is better than the similar system. System works on the text reviews for future interest multimedia reviews should be taken in consideration for online deceptive review detection.

V. REFERENCES

[1] J. Rout, A. Dalmia, S. Bakshi, and S. Jena, "Revisiting semi-supervised learning for online deceptive review detection,"in Proc. 15th IEEE Int. Conf. Trust,Secur. Privacy Comput. and Internet of things, vol. 2, no. 1, pp. 15–25, 2017

[2] O. chappel," Semi-supervised learning," https:// www. molgen. mpg. de / 3659531/ MITPress SemiSupervised- Learning.pdf, 2006

[3] F. Li, M. Huang, and Y. Y., "Learning to identify review spam," Proc. 22nd International Joint Conference Artif. Intell. (IJAI), pp. 24–84, 2011.

[4] L.Zhou, Y. Sh, and D. Zhang, "A statistical language modeling approach to online deception detection,"IEEE Transaction on Knowledge and Data Engineering, vol. 20, no. 8, pp. 11–28, 2009.

[5] P. Mallapragada, R. Jin, A. Jain, and Y. Liu, "Semiboost: Boosting for semisupervised learning," Tech. Rep. MSU-CSE, Michigan State University, pp. 7–197, 2007.

[6] H. Li, B. Liu, and A. S. J. Mukherjee, "Spotting fake reviews using positiveunlabeled learning," Comput. Sistemas , IEEE International Conference,vol. 18, no. 3, pp. 46–74, 2011.

[7] A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal, "Detecting group review spam,"

Proceedings 20th International Conference in Companion World Wide Web, pp. 93–94, 2011.

[8] D. Hernndez and G. R., "Using PU-learning to detect deceptive opinion spam,"Proc. 4th Workshop Computer Approaches Subjectivity, Sentiment Social Media Anal., vol. 1, pp. 38–45, 2013.

[9] Y. Ren and Y. Zhang, "Deceptive opinion spam detection using neural network," Singapore University of Technology and Design, Singapore, pp. 11–28,2016.

[10] Kaggle, "Kaggle," http://www.kaggle.com

[11] Morgan and Claypool, "Sentiment analysis: mining opinions, sentiments, and emotions," Cambridge University Press, pp. 1–8, 2015

[12] Z. Gyngyi, G. H., Y. Molina, and P. J., "Combating web spam with trustrank,"Proc. 13th Int. Conf. Very Large Data Bases (VLDB), IEEE International Conference, vol. 30, pp. 57–87, 2004.

[13] A. Ntoulas, M. Najork, M. Manasse, and F. D., "Detecting spam web pages through content analysis," Proceedings in 15th International Conference World Wide Web (WWW), pp. 83–92, 2006

[14] H. Drucker, D. Wu, and V. Vapnik, "Support vector machines for spam categorization," IEEE Transaction Neural Network, vol. 10, no. 5, pp. 10–54, 1999.

[15] W. Feng and G. Hirst, "Detecting deceptive opinions with prole compatibility," Proceedings 6th IEEE International Conference on Natural Lang. Process(IJCNLP), pp. 33–83, 2013

[16] J. Lee and S. Yoo, "An elliptical boundary model for skin color detection," Proc. of the 2002 International Conference on Imaging Science, Systems, and Technology, 2002

[17] K. Lau, Y. Li, and Y. Jing, "Toward a language modeling approach for consumer review spam detection," Proc. IEEE 7th Int. Conf. e-Bus. Eng. (ICEBE),pp. 1–8, 2010.

[18] A. Mukherjee and V. Venkataraman, "What yelp fake review filter might be doing?" in Proc. 7th Int. AAAI Conf. Weblogs Social Media, pp. 409–418,2013.

[19] B. Liu and W. Lee, "Building text classifiers using positive and unlabeled examples," ICDM-03, Melbourne Florida, pp. 19–22, 2003.

[20] Y. Dai and Y. Philip, "Partially supervised classification of text documents,"Proceedings of the Nineteenth International Conference on Machine Learning (ICML-2002),Sydney, pp. 387–394, 2002.

[21] W. Liu, Y. Li, D. Tao, and Y. Wang, "A general framework for co-training and its applications,"IEEE transaction on Neurocomputing, vol. 167, no. 10, pp.112–121, 2015.

[22] D. Fusilier, M. Montes-y Gmez, P. Rosso, and R. Cabrera, "Detecting positive and negative deceptive opinions using PU-learning,"Inf. Process. Manage.,vol. 51, no. 4, pp. 433–443, 2015.

[23] J. Peng, R. Choo, and H. Ashman, "Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles,"J.Netw. Comput. Appl., vol. 70, pp. 171–182, 2016.