Can We Identify Manipulative Behavior and the Corresponding Suspects on Review Websites using Supervised Learning?

Huiying Duan[1] and Cäcilia Zirn[2]

[1] Heidelberg Institute for Theoretical Studies gGmbH
Heidelberg, Germany
Huiying.Duan@h-its.org
[2] KR & KM Research Group, University of Mannheim
Mannheim, Germany
caecilia@informatik.uni-mannheim.de

Abstract. Identification of manipulative behavior and the corresponding suspects is an essential task for maintaining robustness of reputation systems integrated by review websites. However, this task constitutes a great challenge. In this paper, we present an approach based on supervised learning to automatically detect suspicious behavior on travel websites. We distinguish between two types of manipulation, treating them as separate tasks: promoting manipulation, which is performed in order to push the reputation of a hotel, and demoting manipulation, which is used to demote competitors. Both tasks consist of three separate levels: detecting suspicious reviews (review level), suspicious reviewers (reviewer level) and suspicious objects of the reviews, i.e. hotels (object level). A separate classifier for each of the levels is trained on various sets of textual and non-textual features. We apply state-of-the-art machine learning algorithms like Support Vector Machines. The performance of our approach is evaluated on a new dataset that we created based on reviews taken from the platform TripAdvisor and which was carefully annotated by human judges. The results show that it is possible to identify manipulating reviewers and objects of manipulation with over 90% accuracy. Identifying suspicious reviews, however, seems to be a much harder task, for which our classifier achieves an accuracy of 68% detecting promoting manipulation and 84% detecting demoting manipulation. We argue that there is the need to identify more efficient features for the classification on review level. Finally, we analyze and discuss statistical characteristics of manipulative behavior based on the predictions of the reviewer and object level classifiers.

Keywords: reputation system, trust management, manipulative behavior identification and analysis, opinion mining, supervised learning, TripAdvisor

1 Introduction

Recently, a large number of review websites like TripAdvisor[3] or Yelp[4] have gained great success by integrating reputation into their systems. Reputation refers to public opinions regarding trust on a certain object, e.g. a restaurant or a hotel. Trust is defined as a subjective probability by which an individual A expects that another individual B performs a given action on which its welfare depends [3]. The reason behind the magnificent achievement of these websites is that reputation acts as a social catalyst which aids travelers to decrease the degree of uncertainty and the risk of decision making in virtual environments, where much information of concern is missing. For instance, via a review website, a traveler can check the reputation of hotels which are around his travel location before booking.

Usually, once these websites become popular, manipulative behavior emerges[5] [6] [7]. In this paper, manipulative behavior is defined as an operation of injecting fraudulent reviews. A service provider (e.g. a hotelier or a restaurateur) "hires" people either to give fraudulent positive reviews or to give negative ones on the service provider's competitors. We call the former one promoting manipulation and the latter one demoting manipulation. Generally, a review refers to a personal and subjective evaluation in terms of quality of service. Its content consists of both a numerical and a textual value. For instance, a hotel review on TripAdvisor contains a total score, sub-scores on particular aspects of a hotel (such as the room, value, cleanliness etc.) and a piece of text describing the experience of the service consumption. In the perspective of trust and reputation management, the robustness of reputation systems is largely threatened by these "attacks". For instance, the reputation value of a hotel could be miscalculated by considering fraudulent reviews, and this miscalculation indirectly influences the ranking of hotels in terms of their reputation value.

The research on the robustness of reputation systems [5–13] can be categorized in terms of different criteria. Regarding the choice of the dataset, [5–8] study Amazon[8] product reviews; [9–13] use TripAdvisor as a case study. There is a large difference between the two datasets. On Amazon, reviews are about products, while on TripAdvisor, most of the reviews are related to hotel or restaurant services. Service is a more complicated concept for trust management than products due to the subjectivity and the variation of quality. Hence, manipulative behavior detection on TripAdvisor is supposed to be more difficult than on Amazon. For instance, the quality of a product usually does not

[3] www.tripadvisor.com

[4] www.yelp.com

[5] http://www.dailymail.co.uk/travel/article-2013391/Tripadvisor-Hotel-owners-bribe-guests-return-good-reviews.html

[6] http://www.rhinocarhire.com/Car-Hire-Blog/May-2012/Hotel-Reviews-Faulty-Towers-or-The-Ritz.aspx

[7] http://www.sfgate.com/technology/article/Yelp-s-trust-at-risk-from-phony-reviews-3708962.php

[8] http://www.amazon.com/

change, whereas the quality of hotel service might vary over time. Regarding types of features for machine learning, textual features are used in [5, 7, 9–12]; non-textual features in [5, 6, 8, 9, 12, 13]. Considering detection as a supervised learning approach, [9] labels reviews by the proportion of positive feedback given to a review. We adopt some of the representative features in our work, e.g. the proportion of “Positive Singletons“, which refers to positive reviews written by users who wrote only this one review, and “Reactive Positive Singletons“, which refer to positive reviews written by a hotel as reaction to negative reviews [12]. Textual features such as unigrams and bigrams are commonly used in suspicious review identification [7, 10, 11] to capture the textual content of a review.

The main contribution of this paper is to identify manipulative behavior - both promoting and demoting - on three different levels: the object level, the reviewer level and the review level. Considering all types of information, i.e. non-textual and textual features, six classifiers (three levels, for each two types of manipulative behavior) are trained using Support Vector Machines. Second, we annotate a corpus of hotel reviews. Considering one of the main results in [11], which shows that human performance is low for annotating fake reviews[9], we create the gold-standard annotation for our dataset in a careful manner. Experienced annotators are selected and trained to perform the annotation task using the relation of information between reviews, reviewers and hotels, and we finally only choose those reviews on which all annotators reach a consensus. We evaluate the performance of the classifiers on the newly created gold standard.

The rest of this paper is organized as follows. The next section introduces the dataset. In section 3, the basic idea of suspect identification is represented and features for learning are proposed. Section 4 introduces the process of annotation generation. Section 5 shows the main results about learned classifiers and the corresponding discussion.

2 Dataset Review

We selected 167,909 reviews about New York City’s hotels from TripAdvisor for our dataset. New York City (NYC) is considered as one of the most ideal cities for traveling all around the world, and there are a large number of hotels. We assume it is more likely to find manipulative behavior in NYC than in any other region due to keen competition. A basic statistics of the dataset is listed in Table 1. We collect hotel information, reviewers who provide reviews about the hotels and the corresponding reviews. Note that the amount of reviews which have the singleton feature in [12] is only 3.24%. The fact indicates that the most representative features might be different from one dataset to another. Therefore, it is necessary to specify the most representative features in terms of a concrete dataset.

[9] In their work, only textual information is considered by human annotators to make a decision whether it is fraudulent or not.

Table 1. Basic Statistics of TripAdvisor's Dataset

Dataset Statistics	New York City (NYC)
Duration	January 1999 - June 2011
Hotels	404
Reviewers	110,128
Reviews	167,909
Singletons	5,446 (3.24%)

3 Suspect Detection

Obviously, there are three types of objects that are involved in manipulative behavior: the review, the reviewer and the hotel. A service provider (e.g. the hotelier) "hires" reviewers either to give positive fraudulent reviews, or to give negative ones on the service provider's competitors. We call the former one promoting manipulation and the latter one demoting manipulation. For each level and each type of manipulation we build a separate classifier. In the following, we introduce the features we use for classification.

Advanced Positive Singleton (AdvPositiveSingleton), formalized by formula (1), is the improved version of Positive Singleton [12, 13]. It is defined on the review level for promoting manipulation. In [12, 13], a positive rating[10] is one assigned 4 or 5 points, and a rating with less than 4 points is negative. This definition could be inaccurate. For instance, a newly posted 4-point rating should be considered as negative, if the previous 100 ratings are all 5-point. Therefore, we improve the feature by adding a new condition which estimates the distance between a new rating and the current reputation of the hotel, i.e. the reputation evaluated at the moment when the rating is created. If the distance is larger than a threshold TH_p, we then consider the rating as positive. In the experiment, we empirically set the threshold to 1. Likewise, **Advanced Negative Singleton (AdvNegativeSingleton)** is specified for demoting manipulation.

$$AdvPositiveSingleton(r_i^t) = \begin{cases} 1 & \text{if } r_i^t \text{ is PS and } (r_i^t - TV^t) > TH_p \\ 0 & \text{Otherwise} \end{cases} \tag{1}$$

Time Interval between Posted Date and Stayed Date (TimePosted-Stayed) refers to the difference between the date a review is posted and the date the reviewer stayed in this hotel. It is defined on the review level for both promoting and demoting cases.

Time Interval between Consecutive Contributions (TimeConsecContributions). Contributions of a reviewer are ordered by the time a review is posted. Then the time interval between two consecutive reviews can be regarded as a random variable. This feature contains two subfeatures, mean (TimeConsecContributions_MEAN) and variance (TimeConsecContributions_VAR) of the

[10] In this section, "rating" refers to the numerical value and "review" refers to the textual value.

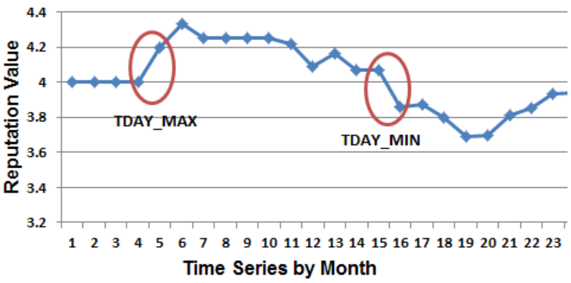


Fig. 1. An example for TurningDay

time interval variable. They are defined on the review level for both promoting and demoting cases.

Rating Preference (RatingPreference), formalized by formula (2), is an indicator for describing a reviewer's attitude towards rating provision. In formula (2), $SUBR()$ denotes a function whose inputs are the overall rating and the index of the sub-rating, and the output is the value of the corresponding sub-rating. When writing a review, a reviewer does not only give an overall rating r_i^t, but sub-ratings $SUBR(r_i^t, k)$ for value, rooms, location, cleanliness, service, etc. It is defined on the review level for both promoting and demoting cases.

$$RatingPreference(r_i^t) = r_i^t - \frac{\sum_{k=1}^{N} SUBR(r_i^t, k)}{N} \tag{2}$$

Turning Day (TurningDay), demonstrated by Fig. 1, indicates the maximal reputation variation of a hotel. Each point represents an evaluation of reputation on a certain time stamp. The circle of evaluation is one month. Then we develop a simple algorithm to identify the intervals which have the largest and smallest slopes TurningDay_MAX and TurningDay_MIN. These are the places where the reputation value has the largest variation during a hotel's life time. We specify TurningDay_MAX as a feature for promoting manipulation, and TurningDay_MIN for demoting manipulation. Furthermore, the logical relationship among a hotel, a reviewer and a review is also taken into consideration, since the variation results from reviews and reviewers who provide them. Therefore, the corresponding reviews and reviewers are also covered by this feature. TurningDay is defined on all levels for both promoting and demoting cases.

Inactive Duration (InactiveDuration) refers to the duration from the last post to the time when data is collected. It is defined on the reviewer level for both promoting and demoting cases.

Contribution Statistics (ContributionStatistics) contains the number of contributions (ContributionNum), mean (Contribution_MEAN) and variance (Contribution_VAR) of contributions which are generated by a reviewer. A unit of contribution refers to a review. All the three subfeatures are defined on the reviewer level for both promoting and demoting cases.

Consistency of Ratings (ConsistencyRating), contains variance of mode (VAR_ MODE) and variance of mean (VAR_MEAN) with respect to different types of ratings for a hotel. First, we categorize ratings of a hotel by the type of traveler, such as "business", "couples" etc., then we calculate mode and mean of these variables respectively. Finally, variance of each mode and mean are calculated. Formula (3) shows the calculation of (VAR_MEAN), (VAR_ MODE) is calculated respectively. R_j denotes the set of ratings for a hotel. $SUBS()$ is a function which returns the subset of ratings in terms of type index k. $MEAN$ and VAR are defined to evaluate mean and variance respectively. The idea behind this feature is to measure to what degree different types of ratings are consistent with each other, and it is defined on the hotel level for both promoting and demoting cases.

$$VAR\_MEAN(R_j) = VAR(MEAN(SUBS(R_j, k))) \tag{3}$$

Average Number of Reviews per Month (AverageNumPerMonth) refers to the mean of the amount of reviews posted on a hotel in one month. It is defined on the hotel level for both promoting and demoting cases.

Proportion of Advanced Positive Singleton (PropAdvPositiveSingleton) refers to the proportion of AdvPositiveSingleton and it is defined on the hotel level for promoting manipulation. The feature is adopted from [12]. We only replace Positive Singleton by AdvPositiveSingleton. Parallel to this, **Proportion of Advanced Negative Singleton (PropAdvNegativeSingleton)** is defined for demoting manipulation.

Reactive Advanced Positive Singletons (ReactiveAdvPositiveSingleton), is also adopted from Reactive Positive Singletons [12]. In order to recover from negative ratings, the management may react by posting some positive shill reviews. The strength of evidence can be quantified as $\frac{T-t_i}{T}$ where T is the length of the entire time period, and t_i is the reaction time associated with shill i. It is formalized by formula (5), where T_h is a normalization factor for hotel h. It is defined on the hotel level for promoting manipulation.

$$ReactiveAdvPositiveSingleton(h) = \frac{1}{|T_h|}(1 - \prod_{i=1}^{n}(1 - \frac{T - t_i}{T})) \tag{4}$$

Truncated Positive Rating (TruncPositiveRating) is adopted from [12], in which it is called Truncated Rating. The idea is to remove a portion of the most positive ratings for a hotel and recalculate the average to see if it deviates much from the original value. It is formalized by formula (4), where R_h^{tr} is the truncated rating set. It is defined on the hotel level for promoting manipulation. Parallel to this, **Truncated Negative Rating (TruncNegativeRating)** is defined for demoting manipulation.

$$TruncPositiveRating(h) = \frac{1}{|R_h|}\sum_{r \in R_h} r - \frac{1}{|R_h^{tr}|}\sum_{r \in R_h^{tr}} r \tag{5}$$

Rating Mean (Rating_MEAN) refers to the mean of the overall ratings on a hotel. It is defined on the hotel level for both promoting and demoting cases.

Rating Variance (Rating_VAR) refers to the variance of the overall ratings on a hotel. It is defined on the hotel level for both promoting and demoting cases.

Ratio of Room Number to Review Number (RatioRoomReview) refers to the ratio of the amount of rooms a hotel owns to the amount of reviews for it. The intuition is that it is suspicious for a hotel who owns only few rooms to have a large number of reviews. It is defined on the hotel level for promoting behavior.

Hotel Reviews Contradiction Degree (ContradictionDegree), formalized by formula (6), refers to the maximum variance of sub-ratings for a hotel. There are N sub-ratings for items as value, rooms, location, cleanliness, service, etc.. MAX is a function to find the maximum of the ratings. It is defined on the hotel level for both promoting and demoting cases.

$$ContradictionDegree(h) = MAX(\{VAR(SUBS(r_i^h, k)), \text{where } i = 1...N\}) \tag{6}$$

Textual Features (UniBigram) refer to the textual features extracted from the review content. Like in [7, 10, 11], we use unigrams and bigrams, representing the review text by the amount of its words and consecutive word pairs.

4 Gold Standard Annotation

Since we are going to apply classic supervised learning approaches, having properly labeled data is the most significant part in our work. Before describing the annotation process, we have some comments on the experiment in [11], who used Amazon Mechanical Turk[11] to purposely create fake reviews. They mix them with real reviews from TripAdvisor which they consider to be written by honest reviewers, and they ask human annotators to spot the malicious ones. One of the main findings of their experiment is that humans are bad at identifying fraudulent reviews. We agree with that, yet we argue that generating fraudulent reviews using Amazon Mechanical Turk is a valid way which has its own limits. It is still unclear whether the character of fraudulent reviews written by virtual workers is matchable to that in TripAdvisor[12]. Furthermore, the annotators in [11] make their decision based on the review text only. In our opinion, a better solution is to identify fraudulent reviews which are extracted from a dataset using all complimentary information given, i.e. checking various reviews of the same reviewer or the date they were posted. We therefore assume that if the annotation process is carefully handled, an appropriate gold standard can be manually generated.

[11] https://www.mturk.com/mturk/welcome

[12] http://tripadvisorwatch.wordpress.com/2010/10/10/tripadvisor-pay-review-fake/

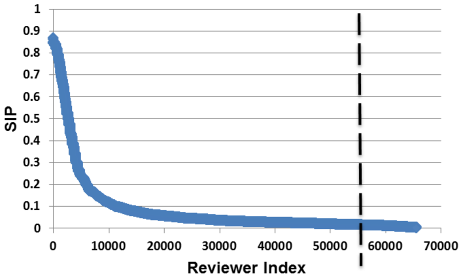


Fig. 2. Distribution of Suspicious Index for Promoting Manipulation

We select three well-trained and independent annotators. Well-trained means every annotator has at least a basic notion of manipulative behavior. They are encouraged to evaluate each review by identifying logical inconsistency within the information related to a review. The information does not only refer to the numerical and textual value of a review per se, but all types of information of the corresponding reviewer, such as uploaded pictures, reviewer profile etc. Interestingly, the annotators based their decisions on facts like one of the uploaded pictures was the only one looking quite different from the pictures uploaded by other reviewers. We randomly pick 1000 reviews from the dataset whose total rating score is either 1 or 5, and let all of the annotators evaluate the same 1000 reviews separately. We believe that a review with a score of 1 or 5 is most likely to be suspicious. Like we expected, the annotation process is a very time-consuming procedure, since an annotator has to check a lot of information in order to make a decision. In addition, we calculate the inter-annotation agreement using Fleiss Kappa, which is κ= 0.18. This indicates only slight agreement, which is consistent with the findings in [11]. To provide a reliability of the labels, we chose only those reviews for our final gold standard that were unanimously labeled by all three annotators. Thus having a complete agreement level and considering the fact that our annotators made use of all information provided about the review, the reviewer and the hotel, we assume the labels in the gold standard to constitute the truth.

So far, only reviews are labeled, but we still need suspiciousness labels for the reviewer and the hotel level. Considering logical relations among different levels, a set of labeled suspicious reviewers and hotels can be generated from labeled reviews. There are two logical implications we use. If a review is suspicious, the corresponding reviewer is also suspicious; if a number of reviews posted on a hotel are all suspicious, the hotel is also suspicious. Following this idea, the sets of suspicious reviewers and hotels are generated. In addition, in our previous work [1], we succeeded in assigning a Suspicious Index (SI) to the objects on different levels. Fig. 2 demonstrates the distribution of the SI on the reviewer level with respect to promoting manipulation. The data can be fitted by an exponential function. In this case, we simply set a threshold for SI (e.g. 0.01) to choose a set of genuine reviewers with respect to promoting manipulation. Parallel to that, we are able to find a set of least suspicious reviewers and hotels by choosing the set of objects which have the lowest SI. The statistics of annotated objects is listed in Table 2.

Table 2. Annotations Statistics

Annotated Object	**Number**
Number of Genuine Reviews	180
Number of Promoting Reviews	139
Number of Demoting Reviews	24
Number of Genuine Reviewers	390
Number of Promoting Reviewers	131
Number of Demoting Reviewers	20
Number of Genuine Hotels	43
Number of Promoting Hotels	26
Number of Demoting Hotels	2

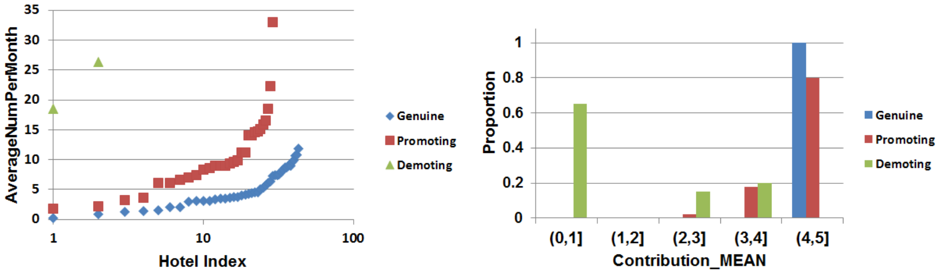
5 Experimental Results and Discussion

In this section, in order to evaluate the effectiveness of the proposed features, we compare the feature value distribution of the different groups (i.e. genuine, promoting manipulation and demoting manipulation) with respect to the annotations. We list the classification results and present a ranking of the most effective features based on the training data. Using the predictions of the classifiers for hotels and reviewers, we explore statistical characteristics of the suspects.

5.1 Feature Evaluation

To illustrate the effectiveness of the proposed features, we plot the distribution of feature values with respect to genuine and suspicious objects considering the gold standard annotations. In this section, we sample the most representative features only due to the paper limitation.

Average number of reviews per month (AverageNumPerMonth) is one of the key features specified on the hotel level. The value distribution of AverageNumPerMonth is plotted in Fig. 3(a). All the hotels are ordered by their AverageNumPerMonth value, which is represented on the y-axis. The x-axis corresponds to the indexes of the hotels. There are three groups of hotels: those with genuine reviews (genuine group), those with promoting reviews (promoting group) and those with demoting reviews (demoting group). The values of the demoting group clearly differ from those of the genuine group. Comparing the promoting group to the genuine group, all of the hotels whose AverageNumPerMonth is larger than 15 are suspicious. This numerical difference can be captured by machine learning approaches.



(a) AverageNumPerMonth

(b) Contribution Means

Fig. 3. Some Results for Feature Evaluation

Contribution Mean (Contribution_MEAN) is a feature specified on the reviewer level. Its value distribution is plotted in Fig. 3(b). On the x-axis, there are 5 points and each represents a range of values. The y-axis denotes the percentage of reviewers whose feature value falls into this range. Fig. 3(b) shows that the range of Contribution_MEAN of the genuine group is between 4 and 5[13]. Contribution_MEAN of the promoting group is mostly distributed between 4 and 5, whereas Contribution_MEAN of the demoting group is distributed between 0 and 4. Again, boundaries among the different groups can be learned.

5.2 Learning Results

Table 3. Classification Results, where A for Accuracy, P for Precision, R for Recall and F for F-Score [2] in %. UniBigram denotes that both Unigrams and Bigrams are considered during learning process. Non-textual denotes all the corresponding features described in section 3.

			Genuine			**Fraudulent**		
Types	**Features**	**A**	**P**	**R**	**F**	**P**	**R**	**F**
Hotel$_{\text{promoting}}$	Non-Textual	**91.3**	100.0	87.8	93.5	76.9	100	87.0
Hotel$_{\text{demoting}}$	Non-Textual	**100.0**	100.0	100.0	100.0	100.0	100.0	100.0
Reviewer$_{\text{promoting}}$	Non-Textual	**96.4**	100	95.4	97.6	85.5	100	92.2
Reviewer$_{\text{demoting}}$	Non-Textual	**100.0**	100.0	100.0	100.0	100.0	100.0	100.0
Review$_{\text{promoting}}$	Non-Textual	65.2	71.1	68.4	69.8	57.6	60.6	59.0
Review$_{\text{promoting}}$	UniBigram	68.3	76.7	70.1	73.2	57.6	65.6	61.3
Review$_{\text{demoting}}$	Non-Textual	80.4	89.4	88.5	89	14.3	13.6	13.0
Review$_{\text{demoting}}$	UniBigram	84.3	90.0	92.0	91.0	41.7	35.7	38.5

[13] This is determined by the way we generate the labels for genuine reviewers within the annotation process.

The main learning results are shown in Table 3. For machine learning, we use the toolkit Weka[14]. Due to the experience of previous work [5, 11], several classic supervised learning approaches are applied, such as linear logistic regression, SVMs and Naive Bayes. Since SVMs clearly outperform other classifiers, we only show those classification results. Achieving accuracies above 90%, identifying manipulative behavior on hotel and reviewer level seems to work quite well. Especially demoting manipulation could be detected correctly in all cases. However, the classification results on review level are not what we had expected. All the scores are much lower than those on the reviewer and hotel level. Although the accuracies ranging between 65% and 84% do not seem to be that low, the actual performance for detecting fraudulent reviews has an f-measure as low as 13% for detecting demoting behavior. Comparing non-textual features and textual features, the latter ones clearly outperform non-textual features. We draw the conclusion that it is extremely difficult to identify fraudulent reviews. More representative features for identifying suspicious reviews need to be developed. In the following part, we will focus on the results for reviewer and hotel level only.

5.3 Feature Selection

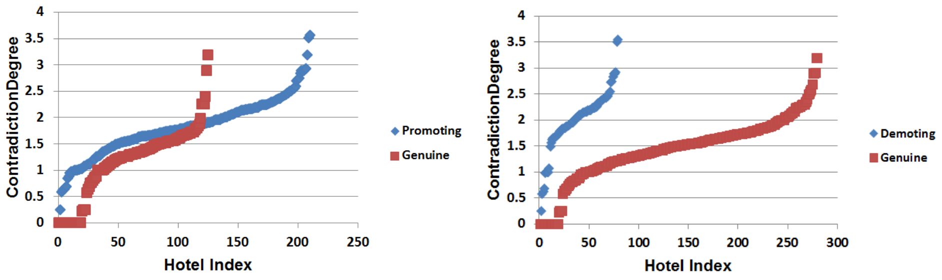
In this section, we explore the performance of the single features. Given the human annotations, features are ranked by the weight assigned by the SVMs [4]. Table 4 shows the top five features for suspicious hotel classification. As we expected, Average Number of Reviews per Month (AverageNumPerMonth) is the best feature for detecting promoting manipulation, and second best for detecting demoting manipulation. A hotel suffering from demoting manipulation usually has a large value for AverageNumPerMonth, since in order to recover from slander, the hotels "hire" reviewers to give fraudulent positive reviews. The singleton related feature is shown in the list as well.

Table 4. Top 5 Features in the Hotel Level

Ranking	**Features$_{\mathbf{PM}}$**	**Features$_{\mathbf{DM}}$**
1	AverageNumPerMonth	Rating_VAR
2	Rating_VAR	AverageNumPerMonth
3	RatioRoomReview	PropAdvNegativeSingleton
4	TurningDay	Rating_MEAN
5	PropAdvPositiveSingleton	VAR_MODE

Table 5 shows the top five features for suspicious reviewer classification. As we expected, Contribution Mean (Contribution_MEAN) is the top one for both promoting and demoting manipulation detection. Interestingly, Inactive Duration (InactiveDuration) is ranked second for promoting manipulation detection,

[14] www.cs.waikato.ac.nz/ml/weka/



(a) Promoting Manipulation

(b) Demoting Manipulation

Fig. 4. ContradictionDegree Evaluation Results

since providing a singleton review usually implies a large value of InactiveDuration. Contribution Variation (ContributionVAR) is ranked third for promotion detection and second for demoting detection.

Table 5. Top 5 Features on the Reviewer Level

Ranking	**Features$_{\mathbf{PM}}$**	**Features$_{\mathbf{DM}}$**
1	Contribution_MEAN	Contribution_MEAN
2	InactiveDuration	ContributionVAR
3	ContributionVAR	ContributionNum
4	TurningDay	InactiveDuration
5	ContributionNum	TimeConsecContributions_MEAN

5.4 Statistical Characteristics of Suspects

In this section, we investigate uncertain assumptions and explore statistical characteristics of suspects by considering the predictions made by our trained classifiers.

In section 3, we specify Hotel Reviews Contradiction Degree (ContradictionDegree) with the expectation that the larger the ContradictionDegree of a hotel is, the more suspicious is the hotel. Applying the same technology for feature evaluation, we plot the ContradictionDegree value distribution for both promoting and demoting cases in Fig. 4. Hotels are ranked by their ContradictionDegree value. Both cases show that the ContradictionDegree ranges of the suspicious and the genuine group completely overlap. This result completely rejects the validity of ContradictionDegree, which is not very useful for suspect identification.

[9] considers the helpfulness of a review as a representative feature for evaluating the trustworthiness of a review. We can not evaluate this hypothesis on the review level since we do not have a good classifier. However, we can still learn some similar notion on the reviewer level where we have qualified classifiers. The

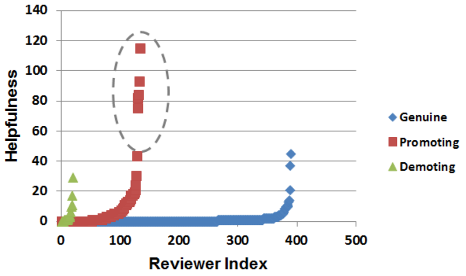
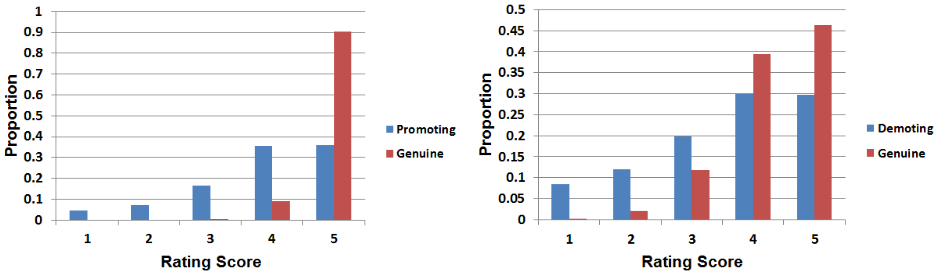


Fig. 5. Distribution of Reviewer's Helpfulness

helpfulness of a reviewer is equal to the sum of helpfulness of all the reviews which are provided by the reviewer. Fig. 5 shows the distribution of helpfulness of reviewers with respect to the different groups. In the dotted circle area, the value for the promoting group is much larger than that for the genuine group. It is an indirect evidence to reject the hypothesis of [9] that the more helpfulness, the less suspicious.

It is not enough to restrict the observations to the human annotations, we would like to explore the statistical character of the whole dataset. One of the most important questions is what the rating distribution looks like with respect to different groups of reviewers. Do reviewers who try to promote a hotel always give 5 points? Similarly, do reviewers who try to demote a hotel always give the lowest rating? The prediction for the whole population is done using the trained classifiers. The results are shown in Fig. 6. In Fig. 6(a) we can see, regarding the predictions made by the classifier for promoting manipulation detection, that genuine reviewers provide mostly 4 or 5 points, whereas suspicious reviewers provide all from 1 to 5 points. This is a surprising insight. The proportion of 1 or 2 points given by suspicious reviewers is much larger than that given by genuine reviewers. It is shown that suspects who intend to promote a hotel provide more negative ratings than honest reviewers, which is a very counterintuitive result. A reasonable explanation for that is that it is a strategy to avoid being identified by TripAdvisor's detection algorithm. An alternative explanation is that in order to maximize the profit per account, a reviewer provides both positive and negative fraudulent reviews. For the case of demoting, which is plotted in Fig. 6(b), suspicious reviewers do not only provide negative fraudulent reviews but positive ones as well. The reason for that is similar than before. Another result we can derive is that most of the negative reviews are fraudulent. Note that, as we mentioned before, the results are subject to the particular dataset. We might draw quite different conclusions in different areas.

Regarding the ranking of hotels in terms of their reputation value, the ranking distribution of different groups is shown in Fig. 7. Three groups are extracted from the prediction which is generated by the classifiers. Promotion group refers to the set of suspicious hotels which are predicted to be related to promoting manipulative behavior; demotion group refers to the set of suspicious hotels



(a) Promoting Manipulation (b) Demoting Manipulation

Fig. 6. Rating Distribution of Subpopulations

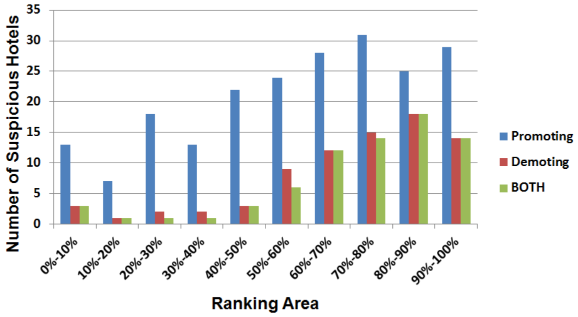


Fig. 7. Reputation Ranking Distribution of Subpopulations

which are predicted to be related to demoting manipulative behavior; BOTH group is the intersection of the first two groups. The x-axis represents 10 intervals that hotels fall into in terms of their ranking. For instance, the top 10% ranked hotels fall into the first interval and so on. The y-axis denotes the number of suspicious hotels which fall into an interval. Fig. 7 shows that manipulation appears in all intervals and promotion is much more popular than demotion. Note that this result is derived from TripAdvisor, which probably already applies detection mechanisms. Even considering TripAdvisor applies manipulation prevention, there is still a number of suspects existing in the system. Another fact is that most suspicious hotels suffering from demotion are also related to promotion. It seems that promotion behavior is triggered by demotion behavior, since in order to recover from slander, the hotels “hire” reviewers to provide fraudulent positive reviews.

6 Conclusion

This paper shows the results for learning classifiers to identify manipulative behavior on the levels of reviewers and hotels, and explores the statistical characteristics of manipulative groups. The experiments are conducted on review data from TripAdvisor about NYC's hotel scene. Manipulative behavior annotations regarding the review, reviewer and hotel level are generated by taking the unanimous votes of three human annotators considering the logical relationship among the levels. Annotations for genuine behaviour are generated by the clustering approach presented in [1]. Sets of features are specified regarding different levels and types of manipulative behavior, promoting and demoting manipulation. Using the annotations and SVMs, several classifiers are learned. The results show that it is possible to learn highly accurate classifiers on the levels of reviewers and hotels, but not on the level of reviews, even considering both non-textual and textual features. Regarding the levels of reviewer and hotel, the specified features are ranked using the weight assigned by the SVMs, such as the Average Number of Reviews per Month (AverageNumPerMonth), the Contribution Mean (Contribution_MEAN) and the Hotel Reviews Contradiction Degree (ContradictionDegree). The value distributions show that different groups of suspects can be distinguished using features such as AverageNumPerMonth and Contribution_MEAN, but not using ContradictionDegree. Characteristics of the data based on the predictions of the classifiers are shown as well. The rating distribution with respect to the different groups (genuine, promoting and demoting) indicates that suspicious reviewers provide reviews with a large variation. The reason could be either that a suspicious reviewer provides both fraudulent positive and negative reviews in order to maximize the profit, or that he does this to avoid being detected by TripAdvisor's manipulation detection mechanism. Even though TripAdvisor applies prevention of manipulative behavior, there is still manipulative behavior going on attempting to take advantage of the reputation system.

The practical significance of the methodology proposed in this paper deserves to be discussed here. Supervised learning is an expensive methodology to apply in general, if the labels for the training have to be created first. However, considering the current situation of review websites like e.g. TripAdvisor and Yelp, some of them already possess labeled data regarding different levels (i.e., review, reviewer and hotel) from earlier manual manipulation detection approaches. As far as we know, some of the websites filter reviews manually. They can make accurate judgments based on the complete information they have, such as the IP address attached to a review, previous reviews written by the user, etc.. It is not as difficult to generate the annotations as we encounter in this work. A problem is that the character of manipulation might differ from region to region, e.g. it might be different in NYS compared to big cities in China due to cultural and social factors. Once obtained sufficient labeled data, there are two ways of applying supervised learning. Either different local classifiers capturing the characteristics of manipulation in a certain region could be learned, or a general classifier capturing universal characteristics could be applied. It is the freedom of analysts to make a choice depending on the dataset and the particular goal.

Regarding future work, there are several ways to improve the learning quality with respect to fraudulent review identification. The logical relationship among the different levels is not explored enough yet. This type of relationship could be used as an advanced feature for identifying fraudulent reviews. Whether a reviewer or a hotel is suspicious can be treated as a new feature for learning on the review level. Furthermore, since we have two different types of features, textual and non-textual ones, semi-supervised learning might be suitable in this case. Using semi-supervised learning we can learn from both types of information and combine them in order to achieve a better result.

References

1. Duan, H., Yang, P.: Building robust reputation systems for travel-related services. In: In Proceedings of the 10th Annual Conference on Privacy, Security and Trust (PST 2012). Paris, France (2012), http://sites.google.com/site/duanhuiying/publications
2. Forman, G., Scholz, M.: Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. SIGKDD Explorations 12(1), 49–57 (2010)
3. Gambetta, D.: Can we trust trust? In: Trust: Making and Breaking Cooperative Relations. pp. 213–237. Basil Blackwell (1988)
4. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. 46(1-3), 389–422 (Mar 2002)
5. Jindal, N., Liu, B.: Opinion spam and analysis. In: Proceedings of the international conference on Web search and web data mining. pp. 219–230. WSDM '08, ACM, New York, NY, USA (2008)
6. Jindal, N., Liu, B., Lim, E.P.: Finding unusual review patterns using unexpected rules. In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 1549–1552. CIKM '10, ACM, New York, NY, USA (2010)
7. Lau, R.Y.K., Liao, S.Y., Kwok, R.C.W., Xu, K., Xia, Y., Li, Y.: Text mining and probabilistic language modeling for online review spam detection. ACM Trans. Manage. Inf. Syst. 2, 25:1–25:30 (Jan 2012)
8. Lim, E.P., Nguyen, V.A., Jindal, N., Liu, B., Lauw, H.W.: Detecting product review spammers using rating behaviors. In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 939–948. CIKM '10, ACM, New York, NY, USA (2010)
9. O'Mahony, M.P., Smyth, B.: Learning to recommend helpful hotel reviews. In: Proceedings of the third ACM conference on Recommender systems. pp. 305–308. RecSys '09, ACM, New York, NY, USA (2009)
10. Ott, M., Cardie, C., Hancock, J.: Estimating the prevalence of deception in online review communities. In: Proceedings of the 21st international conference on World Wide Web. pp. 201–210. WWW '12, ACM, New York, NY, USA (2012)
11. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 309–319. HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
12. Wu, G., Greene, D., Cunningham, P.: Merging multiple criteria to identify suspicious reviews. In: Proc. 4th ACM Conference on Recommender Systems (RecSys'10) (2010)
13. Wu, G., Greene, D., Smyth, B., Cunningham, P.: Distortion as a validation criterion in the identification of suspicious reviews. In: Proceedings of the First Workshop on Social Media Analytics. pp. 10–13. SOMA '10, ACM, New York, NY, USA (2010)