Distributional Footprints of Deceptive Product Reviews

Song Feng **Longfei Xing** **Anupam Gogar** **Yejin Choi**
Department of Computer Science
Stony Brook University
NY 11794, USA
songfeng, loxing, agogar, ychoi@cs.stonybrook.edu

Abstract

This paper postulates that there are natural distributions of opinions in product reviews. In particular, we hypothesize that for a given domain, there is a set of representative distributions of review rating scores. A deceptive business entity that hires people to write fake reviews will necessarily distort its distribution of review scores, leaving distributional footprints behind. In order to validate this hypothesis, we introduce strategies to create dataset with pseudo-gold standard that is labeled automatically based on different types of distributional footprints. A range of experiments confirm the hypothesized connection between the distributional anomaly and deceptive reviews. This study also provides novel quantitative insights into the characteristics of natural distributions of opinions in the TripAdvisor hotel review and the Amazon product review domains.

Introduction

There has been a lot of speculation and anecdotal evidence about the prevalence of deceptive product reviews, i.e., fictitious customer reviews that are written to sound authentic in order to promote the business (e.g., Dellarocas (2006), Yoo and Gretzel (2009), Mukherjee et al. (2011)). There are a small number of cases where it is possible to identify the deceptive reviewers with high confidence. For instance, some deceptive reviewers mistakenly leave trails of their misconducts, e.g., account names that can link to their employment with the company they were writing fake reviews for.[1]. Unrealistically prolific reviewers who write reviews for several instances of the same type of products within short period of time would be another clear-cut case to raise suspicion (e.g., multiple simultaneous reviews for high-end electronic gadgets or dentists in several locations across the country). However, it is unrealistic to expect most deceptive reviewers will leave such obvious traces behind. In fact, it has been shown that recognizing the fake reviews is a very daunting task (e.g., O'Connor (2008)), and human can perform only slightly better than chance (Ott et al. 2011).

Computers are surprisingly better than human in detecting deceptive reviewers based on shallow lexico-syntactic patterns, achieving accuracy close to 90% in the work of Ott et al. (2011). However, such high performance is attainable only when the in-domain training data with true gold standard is available. Because it is not possible to accurately annotate existing reviews as fake or genuine, it is necessary to hire people to write fake reviews (Ott et al. 2011), which limits the scalability across many different domains.

In this study, we explore an alternative direction that does not require supervised training data in detecting suspicious business entities and reviewers. The premise of our approach is that there are *natural distributions of opinions* in product reviews. In particular, for a given domain, we hypothesize that there is a set of representative distributions of review rating scores. A deceptive business entity that hires people to write fake reviews will necessarily distort its distribution of review scores, leaving distributional footprints behind.

The existence of the prominent *shape* of the distribution of product reviews has been first recognized in the recent work of Hu, Zhang, and Pavlou (2009), which found that the typical shape of Amazon review distribution is asymmetric bimodal (J-shaped), rather than uni-modal. However, no prior work has directly linked the representative distribution of review scores to deceptive reviewing activities.

In order to validate the hypothesized connection between the distributional anomaly and deceptive reviews, we explore strategies to create dataset with pseudo-gold standard that is labeled automatically based on different types of distributional footprints. We show that a statistical classifier trained on such dataset can detect fake product reviews with accuracy as high as 72% on previously unseen data with true gold-standard. The three contributions of this study are highlighted below:

- We introduce the notion of *natural distribution of opinions*, and present the first quantitative studies characterizing the representative distributions of opinions in the TripAdvisor hotel review and the Amazon product review domains.

[1] http://blogs.wsj.com/wallet/2009/07/09/delonghis-strange-brew-tracking-down-fake-amazon-raves/

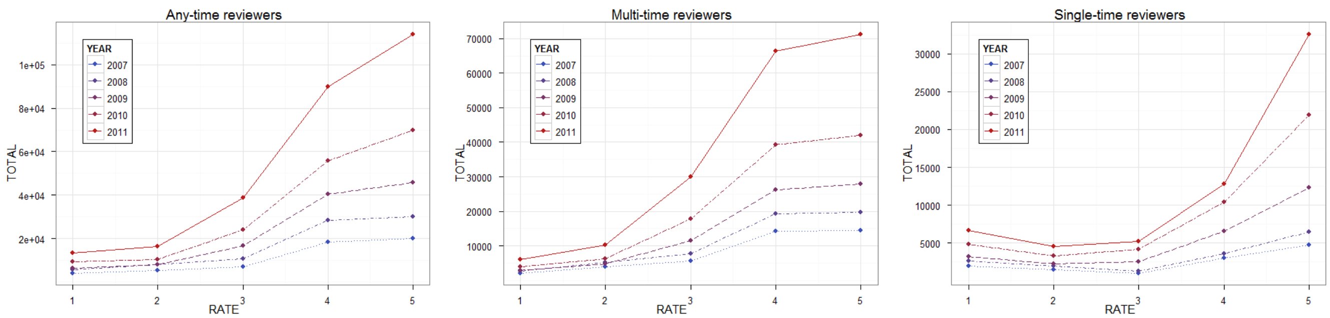

Figure 1: Representative distributions of review-ratings for year $y \in [2007, 2011]$ (Data: TripAdvisor)

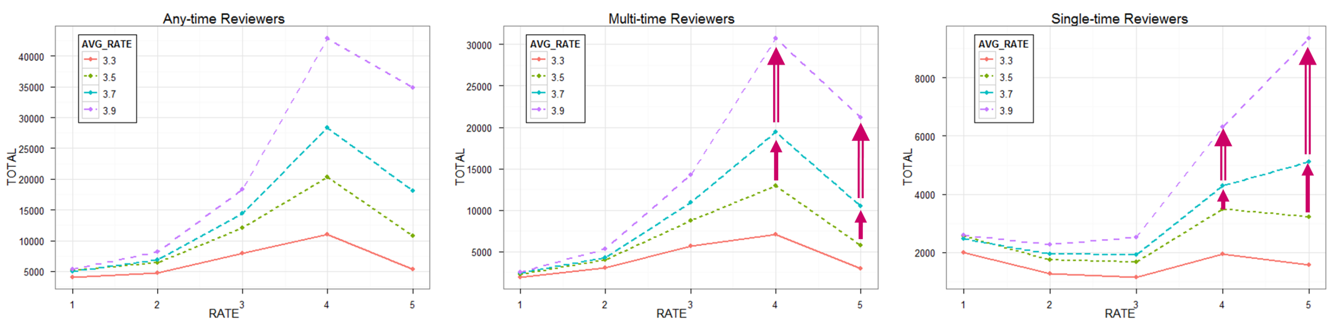

Figure 2: Representative distributions of review-ratings for products with average rating $\bar{r} \in [3.2, 3.9]$ (Data: TripAdvisor)

- We examine different types of distributional footprints of deceptive reviews, and evaluate them *directly* and *statistically* using NLP techniques, rather than relying on human judgments that are known to be not so reliable for deception detection.
- We introduce data collection strategies to obtain (noisy) gold-standard automatically, which can be readily applied to new domains.

The fake review detection strategies introduced in this paper can be employed together with the supervised classification approach of Ott et al. (2011). The distinct strengths of our approach over supervised one are (1) it can be applied to other domain with little cost as it does not require hiring people to write fake reviews, and (2) it is not susceptible to deceptive reviewers who are trained to avoid certain lexical cues that are highly indicative of fake reviews, since our detection strategies are content-independent.[2] Only for the evaluation and validation purposes, we employ content-based classification techniques based on lexical cues.

Distributional Anomaly in TripAdvisor.com

We crawled hotel reviews from `www.tripadvisor.com` for nearly 4000 hotels located in 21 big cities such as London, New York, and Chicago. The crawled data amounts to 839,442 reviews over the period of 2007 – 2011. The number of reviewers increased from approximately 53,000 in 2007 to 170,000 in 2011, while the percentage of anonymous reviewers dropped from over 70% in 2003 to 10% in 2011. Among the reviewers who are not anonymous, about 25% reviewers are *one-time* reviewers, i.e., reviewers who have written only one review under their accounts. We found that this ratio between *one-time* reviewers to *multi-time* reviewers has been more or less stable since 2007.

Ever More Exceedingly Positive Reviews!

Using the data described above, we plot the *representative distributions* of review ratings, as shown in Figure 1. On the x-axis, rating score 5 corresponds to the highest (positive) value, and 1 corresponds to the lowest (negative) rating. Y-axis shows the count of reviews corresponding to the given rating. The right-most graph is plotted for only those reviews written by one-time reviewers, the middle graph for multi-time reviewers, and the left-most for all reviewers. We have two interesting observations: first, every year, the number of positive reviews (rating = 4 & 5) increases much more substantially then the number of negative reviews (rating = 1 & 2). It is as if all these hotels are dramatically enhancing their service each year, impressing the reviewers ever more positively than the preceding years!

Second, notice that the distribution of ratings by multi-time reviewers corresponds to a monotonically increasing line, while the distribution of ratings by one-time reviewers corresponds to a J-shaped (bi-modal) line, such that the

[2] E.g., overusing self-references ("I", "me", "my"), and lacking spatial information. Refer to Ott et al. (2011) for a sample set of lexical cues.

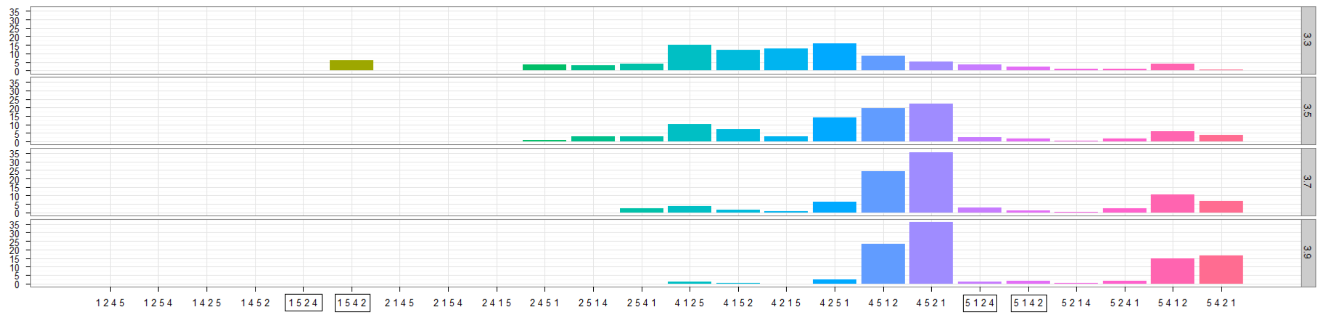

Figure 3: *Distribution of distribution* of review-ratings by *any-time* reviewers (Data: TripAdvisor). The row indexes the average rating of the corresponding products, and the column indexes a particular ordering of ratings sorted by corresponding review counts (i.e., each column represents a particular shape of the distribution of review-ratings). The length of each bar is proportionate to the number of products with the corresponding shape of the review distribution.

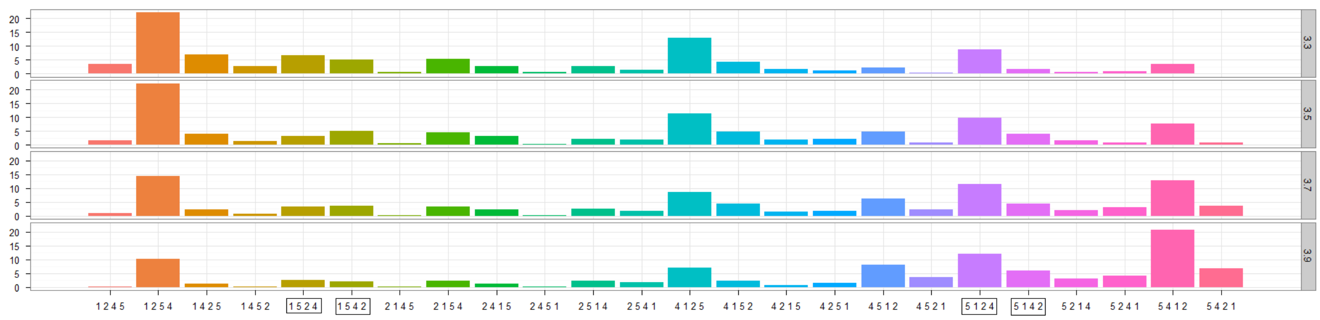

Figure 4: *Distribution of distribution* of review-ratings by *single-time* reviewers (Data: TripAdvisor).

count of rating = 1 is higher than the count of rating = 2 or 3. In contrast, the distribution of ratings of multi-time reviewers has relatively more mass in rating $r = \{2, 3, 4\}$. In other words, one-time reviewers are more likely to have *extreme* opinions, i.e., they are more biased towards the most positive (5-star) and the most negative (1-star) reviews in comparison to multi-time reviewers.[3]

Unimodal V.S. J-shaped (bi-modal) Distributions

We postulate that for a set of hotels of the same average star rating $\bar{r}$, there exists a *natural* distribution of the *truthful* customer ratings. We cannot measure this distribution directly and exactly, because deceptive reviews distort this natural distribution, and it is not possible to identify all of the deceptive reviews. Nonetheless, as will be shown, the notion of the natural distribution helps us identifying the distributional footprints of deceptive reviews.

Figure 2 shows the representative distributions of the review ratings of the given average star rating $\bar{r}$ in the range of $[3.2, 3.9]$.[4] As before, we see that the review ratings of single time reviewers are relatively more skewed toward extreme opinions: 5-star and 1-star ratings. Similarly as in Figure 1, the distribution of single-time reviewers forms a J-shaped, bi-modal line. However, the distribution of multi-time and any-time reviewers are different, i.e., here we see unimodal graphs with the highest point at rating = 4.[5]

Also notice that if we compare the distribution of reviews written by *single-time* reviewers across different $\bar{r} \in [3.2, 3.9]$, then we see that the number of 5-star reviews increases faster than the number of 4-star reviews as the average rating goes up, as highlighted by red arrows in Figure 2.[6] In contrast, if we compare the distribution of reviews written by *multi-time* reviewers, then the increase in the number of 4-star and 5-star reviews across different $\bar{r}$ is generally comparable.

This indicates that hotels that are maintaining an average rating as high as 3.9, are substantially supported by an *unnaturally* higher portion of single-time reviewers giving the 5-star reviews, a bulk of which might as well be fakes. With-

[3] One possible conjecture to this phenomenon is that much of strongly positive one-time reviewers are deceptive reviewers who are paid to write positive reviews, while much of the strongly negative one-time reviewers are truthful reviewers who rarely participate in online reviews, except for that one time when they became upset enough to vent their dissatisfaction. Or it could be also that much of the strongly negative one-time reviewers are also deceptive ones, who are paid to write negative reviews for competitors.

[4] For brevity, we omit the distribution of review ratings corresponding to hotels whose average rating is outside this range.

[5] This implies that the monotonically increasing graphs in Figure 1 are due to reviews for (hotel, year) pairs whose average rating is higher than 3.9.

[6] Notice the delta difference in the length of arrows between multi-time and single-time reviewers.

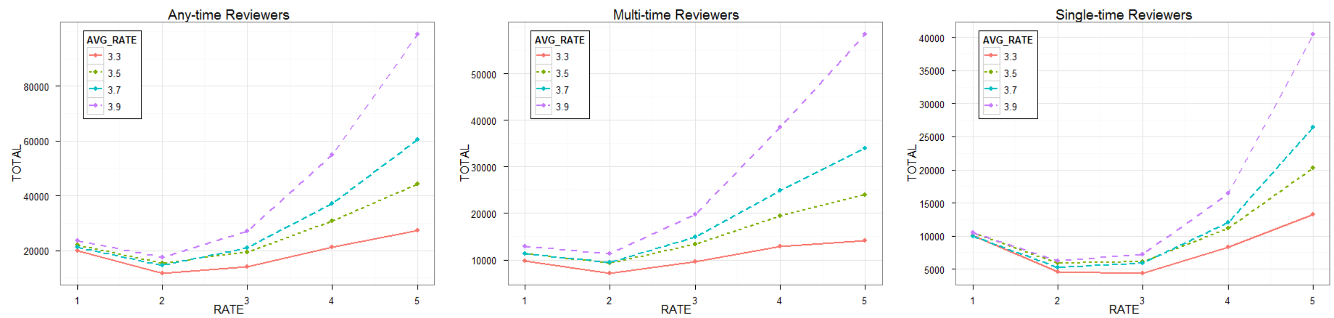

Figure 5: Representative distributions of review-ratings for products with average rating $\bar{r} \in [3.2, 3.9]$ (Data: Amazon)

out solid evidence however, such hotels might insist that all those single-time reviewers are genuinely happy customers, who were impressed enough to write a single strongly positive review just for them, just once in their lives. The evaluation presented later in this paper will provide the first quantitative proof to fundamentally challenge such arguments.

Distribution of Distribution

For any hotel that became active in soliciting (fake) positive reviews today, there must have been a point in time when the hotel got engaged in the solicitation for the first time. That is, some of the deceptive hotels of 2011 might have not been deceptive in e.g., 2010. We therefore consider each year as a coarse time unit, and regard the pair of (hotel, year) as a separate entity. After filtering out hotels that do not have sufficient reviews (20 reviews per year), we obtain 7389 combinations of (hotel, year) pairs over 2165 hotels for the duration of 2007 – 2011.

For each (hotel, year) pair p, let $\bar{r}_p$ be the average review rating based on all reviewers' rating. Let n_i be the count of reviews with i-star rating. Then $D_p := \{n_i, 1 \leq i \leq 5\}$ is the (unnormalized) distribution of the review ratings of the given (hotel, year) pair p. Let D_p^S and D_p^M denote D_p computed only based on single-time reviewers and multi-time reviewers respectively.

Figure 3 and 4 provide deeper insights into the distributional anomaly. To proceed, let us first define the *shape* of the distribution of review ratings as follows. Let $\hat{D}_p$ be the sorted list of indices of D_p, such that index $i \in \{1, 2, 4, 5\}$ is sorted ahead of index $j \in \{1, 2, 4, 5\}, i \neq j$ in $\hat{D}_p$ if $n_i >= n_j$ in D_p, breaking the tie at random. For instance, for $D_p = \{n_1, n_2, n_4, n_5\}$ such that $n_5 \geq n_1 \geq n_2 \geq n_4$, the shape of D_p can be characterized as $\hat{D}_p = (5 \succ 1 \succ 2 \succ 4)$.[7]

The columns (bars) in Figure 3 and 4, correspond to these shape definitions, sorted by the numeric order of the sorted list of indices, i.e., from $\hat{D}_p = (1 \succ 2 \succ 4 \succ 5)$ to $\hat{D}_p = (5 \succ 4 \succ 2 \succ 1)$. The rows correspond to the bin of different average review ratings $\bar{r}$, ranging from 3.2 to 3.9. The y-axis within each row corresponds to the # of hotels (in %) that belong to the bin defined by the average review rating $\bar{r}$ and the shape of review distribution $\hat{D}_p$. In a nutshell, these figures provide the visualization of the *distribution of the distribution*, i.e., the distribution of $\hat{D}_p$, which defines different shapes of the distribution D_p.

In Figure 3, we see that the mass of the distribution generally shifts from left to right, as the average rating goes up, which is only as expected. For $\bar{r} \in [3.5, 3.9]$, notice that the most prominent shape of the distribution is $\hat{D}_p = (4 \succ 5 \succ 2 \succ 1)$. We see a similar shifting trend in Figure 4, where the mass of the distribution is gradually moving from left to right as the average rating increases, but there are subtle, yet distinctive differences:

- First, if we examine the mass focused on the shape of distribution indexed by $\hat{D}_p = (5 \succ 1 \succ 2 \succ 4)$, we see that there is a lot more concentration in Figure 4 than in Figure 3. In fact, this particular shape of distribution, which indicates $n_5 \geq n_1 \geq n_2 \geq n_4$, is a highly suspicious one: how could it be that for a hotel for which 5-star reviews are the most dominant, there are more number of 1 & 2-star reviews than 4-star reviews?
- Second, also notice that the distribution of single-time reviewers (Figure 4) is much more *divergent* than that of all reviewers (Figure 3), suggesting *distributional perturbation* caused by various single-time reviewers.

Distributional Anomaly in Amazon.com

For comparative analysis, we examine the representative distributions of review ratings in another popular review website, `www.amazon.com`. We use the Amazon review dataset of Jindal and Liu (2008), which consists of reviews for the duration of June 2006, over 700,000 products.

Figure 5 shows the representative distributions of review ratings for products whose average rating $\bar{r}$ is in the range of $[3.2, 3.9]$, computed with respect to all reviewers (leftmost), multi-time reviewers (middle), and single-time reviewers (right-most) respectively. In contrast to Figure 2 of TripAdvisor, here we see all distributions are in the shape of J (bi-modal), where the J-shape of single-time reviewers shows relatively more extreme opinions (5-star & 1-star

[7] Since there are $4!$ possible permutations of indices, this definition will categorize various (unnormalized) distributions D_p of various (hotel, year) pairs into $4!$ different categories. We omit the index $i = 3$ for brevity.

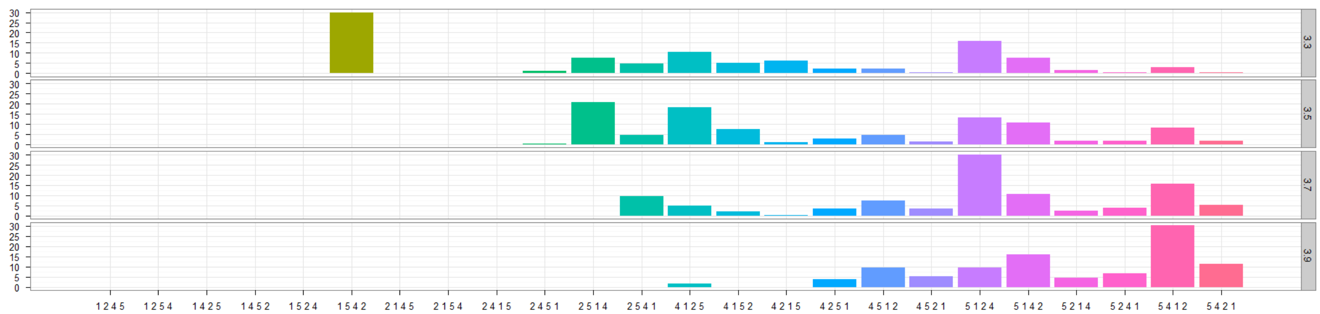

Figure 6: *Distribution of distribution* of review-ratings by *any-time* reviewers (Data: Amazon).

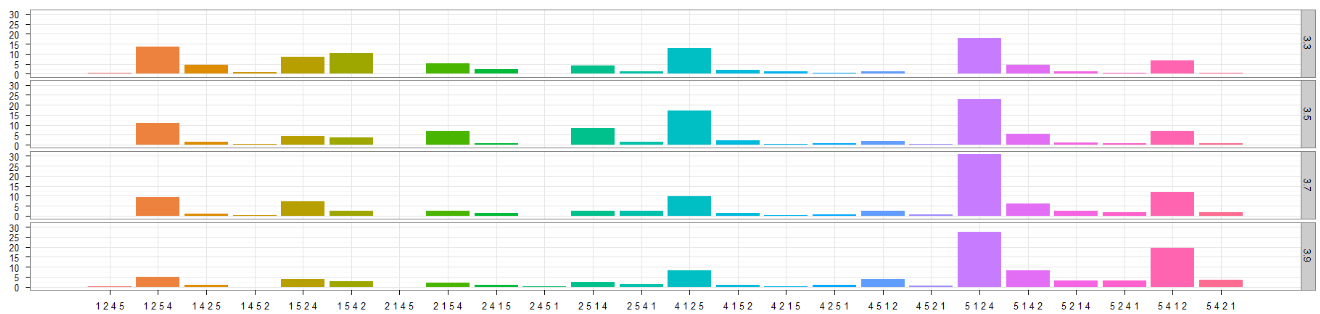

Figure 7: *Distribution of distribution* of review-ratings by *single-time* reviewers (Data: Amazon).

ratings) than that of multi-time reviewers. Similarly in Figure 2, the slope towards 5-star reviews grows steeper as the average review rating increases.

Figure 6 and 7 show the *distribution of distribution* of review-ratings by all reviewers and single-time reviewers respectively, similarly as Figure 3 and 4 of TripAdvisor.

Here we see similar trends that we found in TripAdvisor. First, in both Figures, we see the mass of the distribution gradually shifts from left to right as the average rating increases. Second, the distribution of the single-time reviewers is much more divergent than that of all reviewers. Third, the suspicious shape of the distribution $\hat{D}_p = (5 \succ 1 \succ 2 \succ 4)$ stands out again among the single-time reviewers. In fact, even more so in the Amazon data than it was in the TripAdvisor data. It is interesting to see that in Figure 7, the most dominant shape for any average rating is $\hat{D}_p = (5 \succ 1 \succ 2 \succ 4)$.

Deception Detection Strategies

In this section, we introduce deception detection strategies guided by statistics that are suggestive of distributional anomaly. Our detection strategies are *content independent*, in that it will rely only on the meta data, such as, the rating distribution of a hotel, or the historic rating distribution of a reviewer.

Committee of Truthful Reviewers $\mathcal{T}$

We first begin by collecting the *"committee of truthful reviewers"*, which will become handy in some of the deception detection strategies, as well as evaluation setup. We conjecture that reviewers with a long history of reviews are more likely to be trustworthy. We collect a set of reviewers who have written more than 10 reviews. One thing regular reviewers hardly do is to post several reviews in a very short time interval (Lim et al. 2010). We therefore discard any reviewer who has written more than 1 review within 2 consecutive days, as such reviewers might be engaged in deceptive activities. Finally, we only keep those reviewers whose rating trends are not outrageous. For instance, we discard reviewers whose ratings are always far away ($\delta = r(h) - \bar{r}_h, |\delta| \geq 1$) from the the average ratings of *all* the reviewees (i.e., hotels).[8] The resulting committee has 42766 reviewers as its trustworthy member, which we denote as $\mathcal{T}$.

Identifying Deceptive Business Entities

Next we present three different strategies for identifying deceptive hotels.

[1] STRATEGY-$avg\Delta$

This strategy is based on the insights we gained from Figure 2. For a hotel h, we calculate the discrepancy between the average rating by the committee of truthful reviewers ($\mathcal{T}$) and the average rating by single-time reviewers $\mathcal{S}$:

$$\delta_h = \bar{r}_h^{\mathcal{S}} - \bar{r}_h^{\mathcal{T}}$$

[8]Such reviewers who are consistently far off from the average might not be necessarily deceptive, but nonetheless do not reflect the general sentiment of the crowd.

$\mathcal{S}$	Set of single-time reviewers.
$\mathcal{M}$	Set of multiple-time reviewer.
$\mathcal{T}$	Set of regular reviewers .
$\mathcal{R}^*(h)$	Set of $*$ type reviewers that reviewed h.
$\overline{r}_h$	average rate of hotel h .
$\overline{r}_h^{\mathcal{R}}$	average rate of hotel h based on reviews by $\mathcal{R}$ type of reviewers.
$rv_\lambda^{\mathcal{R}}(h)$	a review with rate λ of hotel h by a reviewer in $\mathcal{R}$.

Table 1: Notational Definitions.

After sorting the hotels by δ in a *descending* order, hotels ranked at top are assumed to be more suspicious (in Table 3), and hotels ranked at bottom are assumed to be credible (in Table 6).

[2] STRATEGY-$dist\Phi$

This strategy is based on the insights we gained from Figure 3 and 4. Remind that the percentage of the distribution ($5 \succ 1 \succ 2 \succ 4$) with respect to single-time reviewers in Figure 4 is substantially higher than that of any-time reviewers in Figure 3. Therefore, we first calculate the ratio of the number of strongly positive reviews to the number of strongly negative reviews among different groups of reviewers, i.e. $\mathcal{S}$ and $\mathcal{M}$.

$$\tau_h^{\mathcal{R}} = \frac{|rv_\lambda^{\mathcal{R}}(h), \lambda \geq \lambda_{high}|}{|rv_\lambda^{\mathcal{R}}(h), \lambda \leq \lambda_{low}|}$$

For suspicious hotels, we pick those with bigger r_h:[9]

$$r_h = \frac{\tau_h^{\mathcal{S}}}{\tau_h^{\mathcal{M}}}$$

For trustful hotels, we pick those with the smaller r'_h:

$$r'_h = \frac{max(\tau_h^{\mathcal{S}}, \tau_h^{\mathcal{M}})}{min(\tau_h^{\mathcal{S}}, \tau_h^{\mathcal{M}})} - 1$$

[3] STRATEGY-$peak \uparrow$

A sudden burst in the reviewing activity can be a sign for deceptive activities (e.g., Jindal, Liu, and Lim (2010)). We therefore translate this idea into a strategy so that we can compare it against other strategies. Specifically, if $\overline{r}(h, M)$ among reviews posted in month M for h is greater than the average rating among reviews posted within the two months before and after M, then we assume the corresponding hotel is suspicious.

Evaluation

Evaluation Strategy

We want to measure the quality of deception detection strategies introduced earlier, but there is no direct and straightforward method to do so. One might wonder whether we could perform human judgment study on our proposed strategies, but there are two major problems: first, it has been shown in prior literature that human are not good at detecting deceptions (Vrij et al. 2007), including detecting fake reviews (Ott et al. 2011). Second, because our strategies are essentially developed based on our own human judgment guided by relevant statistics, human judgment study guided by the same set of statistics is likely to lead to the conclusion that might be overly favorable for this study.

Therefore, we introduce an alternative approach to evaluation that can directly measure the utility of deception detection strategies. More specifically, we exploit the gold standard dataset created by Ott et al. (2011), which includes 400 deceptive reviews that are written by hired people, and contrastive 400 truthful reviews that are gathered from TripAdvisor, modulo filtering rules to reduce incidental inclusion of deceptive reviews. Henceforth, we refer to this dataset as the *gold standard data*, as this is the only dataset publicly available with true gold standard in the product review domain.

For all our strategies, we mix and match the gold standard data and the pseudo-gold standard data in three different combinations as follows:

(C1) $rule, gold$: Train on the dataset with pseudo gold standard determined by one of the strategies, and test on gold standard dataset of Ott et al. (2011).

(C2) $gold, rule$: Train on gold standard dataset and test on pseudo gold standard dataset.

(C3) $rule, rule$: Train and test on the pseudo gold standard dataset (of different split).

The purpose of the above variations is in order to probe whether a high performance in (C1) and/or (C2) correlate with (C3) empirically. If it does, then it would be suggestive that one could resort to the experiment in the (C3) configuration alone, when the gold standard dataset is not readily available.

Experimental Configuration

Whenever possible, the dataset with the pseudo-gold standard determined by one of our strategies will include 400 reviews per class, where 80 % is used for training, and 20% is used for testing for 5-fold cross validation. Note that for certain variations of strategies, it might be impossible to find as many as 400 reviews for each class. In those cases, the number of training and test instances are given in the parenthesis in Table 6 and 4. [10] We use the LIBSVM (Chang and Lin 2011) classifier and feature values are term frequencies scaled with respect to the document length.

Notational Definitions

In Table 2 – 6, the pseudo gold standard dataset is defined using notations of the following format: $(H, \mathcal{R})$, where H corresponds to the set of hotels, and $\mathcal{R}$ corresponds to the

[9] We set $\lambda_{high} = 5$ and $\lambda_{low} = 2$.

[10] To avoid overlap between the pseudo-gold standard determined by our strategies and the gold standard data, we exclude all those reviews for the 20 hotels that are selected by Ott et al. (2011). We also truncate each review at 150 tokens, to balance the length with the gold standard data. We exclude hotels with less than 20 reviews per year, assuming deceptive hotels are likely to be much more productive than generating only a handful reviews per year.

DECEP	TRUTH	TRAIN	TEST	ACC. (%)
$*, *$	$*, *$	$rule$	$gold$	43.5
		$gold$	$rule$	42.0
		$rule$	$rule$	48.4
$H^*, \mathcal{S}$	$H^*, \mathcal{T}$	$rule$	$gold$	50.0
		$gold$	$rule$	58.1
		$rule$	$rule$	61.3
$H^*, \mathcal{S}$	$H^*, \mathcal{M}$	$rule$	$gold$	38.5
		$gold$	$rule$	44.0
		$rule$	$rule$	55.0

Table 2: Classification on 5-star reviews: BASELINES

DECEP	TRUTH	TRAIN	TEST	ACC. (%)
$H_S, \mathcal{S}$	$H'_S, \mathcal{T}$	$rule$	$gold$	65.7
		$gold$	$rule$	65.1
		$rule$	$rule$	67.1
$H_S, \mathcal{S}$	$H_S, \mathcal{T}$	$rule$	$gold$	**70.0**
		$gold$	$rule$	66.3
		$rule$	$rule$	65.0
$H_S, \mathcal{S}$	$H_S, \mathcal{M}$	$rule$	$gold$	58.3
		$gold$	$rule$	45.6
		$rule$	$rule$	43.1

Table 3: Classification on 5-star reviews: STRATEGY-$avg\Delta$

DECEP	TRUTH	TRAIN	TEST	ACC. (%)
$H_S, \mathcal{S}$	$H'_S, \mathcal{T}$	$rule$	$gold$	**72.5**
		$gold$	$rule$	**73.8**
		$rule$	$rule$	**74.4**
$H_S, \mathcal{S}$	$H_S, \mathcal{T}$	$rule$	$gold$	60.3 (160/40)
		$gold$	$rule$	62.0
		$rule$	$rule$	63.2 (160/40)
$H_S, \mathcal{S}$	$H_S, \mathcal{M}$	$rule$	$gold$	36.9
		$gold$	$rule$	45.6
		$rule$	$rule$	58.0

Table 4: Classification on 5-star reviews: STRATEGY-$dist\Phi$.

DECEP	TRUTH	TRAIN	TEST	ACC. (%)
$H_S, \mathcal{S}$	$H'_S, \mathcal{T}$	$rule$	$gold$	54.1 (200/50)
		$gold$	$rule$	64.4
		$rule$	$rule$	60.4 (200/50)
$H_S, \mathcal{S}$	$H_S, \mathcal{T}$	$rule$	$gold$	53.8 (200/50)
		$gold$	$rule$	**72.0**
		$rule$	$rule$	61.0 (200/50)
$H_S, \mathcal{S}$	$H_S, \mathcal{M}$	$rule$	$gold$	40.2 (200/50)
		$gold$	$rule$	40.5
		$rule$	$rule$	56.6 (200/50)

Table 5: Classification on 5-star reviews: STRATEGY-$peak \uparrow$.

set of reviewers. $\mathcal{R}$ can be any of the top three notations in Table 1. H can be one of the following three options:

- H_S denotes the set of hotels selected by strategy S.
- H'_S denotes the set of hotels randomly selected from the complement set of H_S, so that $H_S \cap H'_S = \emptyset$.
- H^* stands for a set of randomly selected hotels.

The first column in Table 2 – 6 defines how the instances in 'DECEP'tive and 'TRUTH'ful classes are created using above notations.

Baselines

Next we define three different pseudo gold standard datasets that correspond to baselines, using notations defined above. These baseline datasets will contrast the quality of other pseudo gold standard dataset created by deception detection strategies discussed earlier.

- BASELINE-1: (DECEP = $*, *$ TRUTH = $*, *$)

  Both hotels and reviews are randomly selected.
- BASELINE-2: (DECEP = $H^*, \mathcal{S}$ TRUTH = $H^*, \mathcal{M}$)

  First a set of hotels are randomly selected, then reviews written by $\mathcal{S}$ for the corresponding set of hotels H^* are considered as deceptive reviews, and reviews written by $\mathcal{M}$ are considered as truthful reviews. Note that the same set of hotels are used by both deceptive and truthful class.
- BASELINE-3: (DECEP = $H^*, \mathcal{S}$ TRUTH = $H^*, \mathcal{T}$)

  First randomly select a set hotels, then reviews by $\mathcal{S}$ are considered as deceptive, and reviews by $\mathcal{T}$ are considered as truthful. Again, the same set of hotels are used by both deceptive and truthful class.

Experimental Results

Baselines: First consider the baseline results in Table 2. As can be seen, none of the three baselines could perform consistently better than chance (50%). This clearly demonstrates that *not all* single-time reviewers are deceptive.

Three strategies on positive reviews: Table 3, 4, and 5 show the classification performance based on the pseudo gold standard determined by the three strategies defined earlier: STRATEGY-$avg\Delta$, STRATEGY-$dist\Phi$, and STRATEGY-$peak \uparrow$ respectively. In Table 4, we see that choosing the complement set of hotels (H'_S) for truthful reviewers yields better performance than sharing the same set of hotels as the deceptive reviewers.[11]

It is quite astonishing to see that the classifier trained *only* on the pseudo gold standard data, which consists of reviews written for the set of hotels that are completely disjoint from those in the gold standard data, achieves deception detection accuracy as high as 72.5%. Recall that Ott et al. (2011) report the human judges could determine deceptive reviews only slightly better than chance. This is a highly encouraging and exciting result for two reasons: first, it demonstrates an effective strategy for automatic data collection with (noisy) gold standard. Second it validates the long-standing suspicions in the community regarding the existence of deceptive

[11]The best performing construction of DECEP and TRUTH class labels differs across different strategies. We conjecture this is due to uneven size of training and test data. Note that some of these strategies can be highly selective when they are combined with a particular construction rule of class labels.

DECEP	TRUTH	TRAIN	TEST	ACC. (%)
$H_S, \mathcal{S}$	$H'_S, \mathcal{T}$	$rule$	$rule$	63.8 (160/40)
$H_S, \mathcal{S}$	$H_S, \mathcal{T}$	$rule$	$rule$	56.3 (320/80)
$H_S, \mathcal{S}$	$H_S, \mathcal{M}$	$rule$	$rule$	65.5 (100/25)

Table 6: Classification on 1-star reviews: STRATEGY-$avg\Delta$

DECEP	TRUTH	TRAIN	TEST	ACC. (%)
$H_S, \mathcal{S}$	$H'_S, \mathcal{T}$	$rule$	$rule$	60.4 (160/40)
$H_S, \mathcal{S}$	$H_S, \mathcal{T}$	$rule$	$rule$	64.0 (320/80)
$H_S, \mathcal{S}$	$H_S, \mathcal{M}$	$rule$	$rule$	58.8 (160/40)

Table 7: Classification on 1-star reviews: STRATEGY-$dist\Phi$.

reviews, and provides a technique to pin-point the dishonest business entities.

Another important observation to make from Table 3 is, simply trusting multi-time reviewers (third row) is dangerous, as the classification accuracy turns out to be very bad, especially in comparison to the second row, where the definition of "truthful reviewers" $\mathcal{T}$ is much more restrictive than that of $\mathcal{M}$ for the identical set of hotels H_S. This indicates that the deception is prevalent even in the multi-time reviewers, at least with respect to those who have written reviews for highly suspicious hotels.

Three strategies on negative reviews: We also extend our strategies to negative reviews, as shown in Table 6 and 7. Because we do not have gold standard dataset available (none is publicly available), we resort to the TRAIN=$rule$ and TEST=$rule$ configuration, which we have seen to correlate reasonably well with TRAIN=$rule$ and TEST=$gold$ in Table 3, 4, and 5. The best accuracy achieved is 65.5%, which is substantially lower than what we could achieve for the positive reviews. We conjecture that detecting fake negative reviews is much harder, as many of them can be truthful negative reviews.

Related Work & Discussion

There has been a number of previous work that investigated deception detection strategies on product reviews (e.g., Yoo and Gretzel (2009), Mukherjee et al. (2011)). The evaluation has been always a challenge, as it is nearly impossible to manually determine whether a review is truthful or not. Prior work therefore resorted to various alternatives. Some researchers relied on human judgments that can be imperfect and biased (e.g., G. Wu and Cunningham (2010), Mukherjee et al. (2011)). Others focused on slightly different problems, e.g., detecting duplicate reviews or review spammers (e.g., Jindal and Liu (2008), Lim et al. (2010), Jindal, Liu, and Lim (2010)). A very recent work of Ott et al. (2011) performed a more direct and explicit evaluation by creating a gold standard data, in particular, by hiring Amazon turkers to write fake reviews. One limitation however, is that it is not cost efficient when exploring different domains. In this work, we have presented a novel evaluation strategy that exploits existing gold standard, and empirically validated the connection between the performance evaluated using the gold standard and the performance evaluated using only the peudo gold standard data.

Some previous work has recognized the notion of anomaly in the review activities (e.g., G. Wu and Cunningham (2010)), however, our work is the first to provide a comprehensive, direct, and large-scale analysis on representative distribution of product reviews, accompanying quantitative evaluations that are not based on human judgments that can be imperfect and biased.

References

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Dellarocas, C. 2006. Strategic manipulation of internet opinion forums: Implications for consumers and firms. In *Management Science, Vol. 52, No. 10*.

G. Wu, D. Greene, B. S., and Cunningham, P. 2010. Distortion as a validation criterion in the identification of suspicious reviews. In *Technical report, UCD-CSI-2010-04, University College Dublin*. University College Dublin.

Hu, N.; Zhang, J.; and Pavlou, P. A. 2009. Overcoming the j-shaped distribution of product reviews. *Commun. ACM* 52:144–147.

Jindal, N., and Liu, B. 2008. Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, 219–230. New York, NY, USA: ACM.

Jindal, N.; Liu, B.; and Lim, E.-P. 2010. Finding unusual review patterns using unexpected rules. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, 1549–1552.

Lim, E.-P.; Nguyen, V.-A.; Jindal, N.; Liu, B.; and Lauw, H. W. 2010. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, 939–948. New York, NY, USA: ACM.

Mukherjee, A.; Liu, B.; Wang, J.; Glance, N. S.; and Jindal, N. 2011. Detecting group review spam. In *Proceedings of the 20th International Conference on World Wide Web (Companion Volume)*, 93–94.

O'Connor, P. 2008. User-generated content and travel: A case study on tripadvisor.com. In *Information and Communication Technologies in Tourism*. Springer Vienna. 47–58.

Ott, M.; Choi, Y.; Cardie, C.; and Hancock, J. T. 2011. 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*, 309–319. Portland, Oregon, USA: Association for Computational Linguistics.

Vrij, A.; Mann, S.; Kristen, S.; and Fisher, R. 2007. Cues to deception and ability to detect lies as a function of police interview styles. *Law and human behavior* 31(5):499–518.

Yoo, K.-H., and Gretzel, U. 2009. Comparison of deceptive and truthful travel reviews. In *Information and Communication Technologies in Tourism*, 37–47. Springer Vienna.