# What Yelp Fake Review Filter Might Be Doing?

**Arjun Mukherjee**[†] **Vivek Venkataraman**[†] **Bing Liu**[†] **Natalie Glance**[‡]

[†] University of Illinois at Chicago [‡] Google Inc.

arjun4787@gmail.com; {vvenka6, liub}@uic.edu; nglance@google.com

**Abstract**

Online reviews have become a valuable resource for decision making. However, its usefulness brings forth a curse – *deceptive opinion spam*. In recent years, fake review detection has attracted significant attention. However, most review sites still do not publicly filter fake reviews. Yelp is an exception which has been filtering reviews over the past few years. However, Yelp's algorithm is trade secret. In this work, we attempt to find out what Yelp might be doing by analyzing its filtered reviews. The results will be useful to other review hosting sites in their filtering effort. There are two main approaches to filtering: supervised and unsupervised learning. In terms of features used, there are also roughly two types: linguistic features and behavioral features. In this work, we will take a supervised approach as we can make use of Yelp's filtered reviews for training. Existing approaches based on supervised learning are all based on pseudo fake reviews rather than fake reviews filtered by a commercial Web site. Recently, supervised learning using linguistic n-gram features has been shown to perform extremely well (attaining around 90% accuracy) in detecting crowdsourced fake reviews generated using Amazon Mechanical Turk (AMT). We put these existing research methods to the test and evaluate performance on the real-life Yelp data. To our surprise, the behavioral features perform very well, but the linguistic features are not as effective. To investigate, a novel information theoretic analysis is proposed to uncover the precise psycholinguistic difference between AMT reviews and Yelp reviews (crowdsourced vs. commercial fake reviews). We find something quite interesting. This analysis and experimental results allow us to postulate that Yelp's filtering is reasonable and its filtering algorithm seems to be correlated with abnormal spamming behaviors.

## Introduction

Online reviews are increasingly being used by individuals and organizations in making purchase and other decisions. Positive reviews can render significant financial gains and fame for businesses. Unfortunately, this gives strong incentives for imposters to game the system by posting deceptive fake reviews to promote or to discredit some target products and services. Such individuals are called *opinion spammers* and their activities are called *opinion spamming* (Jindal and Liu 2008).

The problem of opinion spam or fake reviews has become widespread. Several high-profile cases have been reported in the news (Streitfeld, 2012a). Consumer sites have even put together many clues for people to manually spot fake reviews (Popken, 2010). There have also been media investigations where fake reviewers admit to have been paid to write fake reviews (Kost, 2012). In fact the menace has soared to such serious levels that Yelp.com has launched a "sting" operation to publicly shame businesses who buy fake reviews (Streitfeld, 2012b).

Deceptive opinion spam was first studied in (Jindal and Liu, 2008). Since then, several dimensions have been explored: detecting individual (Lim et al., 2010) and group (Mukherjee et al., 2012) spammers, and time-series (Xie et al., 2012) and distributional (Feng et al., 2012a) analysis. The main detection technique has been supervised learning using linguistic and/or behavioral features. Existing works have made important progresses. However, they mostly rely on ad-hoc fake and non-fake labels for model building. For example, in (Jindal and Liu, 2008), duplicate and near duplicate reviews were assumed to be fake reviews in model building. An AUC (*Area Under the ROC Curve*) of 0.78 was reported using logistic regression. The assumption, however, is too restricted for detecting generic fake reviews. Li et al. (2011) applied a co-training method on a manually labeled dataset of fake and non-fake reviews attaining an F1-score of 0.63. This result too may not be reliable as human labeling of fake reviews has been shown to be quite poor (Ott et al., 2011).

In this work, we aim to study how well do the existing research methods work in detecting real-life fake reviews in a commercial website. We choose Yelp.com as it is a well-known large-scale online review site that filters fake or suspicions reviews. However, its filtering algorithm is trade secret. In this study, we experiment with Yelp's filtered and unfiltered reviews to find out what Yelp's filter might be doing. Note that by no means do we claim that Yelp's fake review filtering is perfect. However, Yelp is a

commercial review hosting site that has been performing industrial scale filtering since 2005 to remove suspicious or fake reviews (Stoppelman, 2009). Our focus is to study Yelp using its filtered reviews and to conjecture its review filtering quality and what its review filter might be doing.

Our starting point is the work of Ott et al. (2011) which is a state-of-the-art as it reported an accuracy of 90%. Ott et al. (2011) used Amazon Mechanical Turk (AMT) to crowdsource anonymous online workers (called *Turkers*) to write fake hotel reviews (by paying $1 per review) to portray some hotels in positive light. 400 fake positive reviews were crafted using AMT on 20 popular Chicago hotels. 400 positive reviews from Tripadvisor.com on the same 20 Chicago hotels were used as non-fake reviews. Ott et al. (2011) reported an accuracy of 89.6% using only word bigram features. Feng et al., (2012b) boosted the accuracy to 91.2% using deep syntax features. These results are quite encouraging as they achieved very high accuracy using only linguistic features.

We thus first tried the linguistic n-gram based approach to classify filtered and unfiltered reviews of Yelp. Applying the same n-gram features and the same supervised learning method as in (Ott et al., 2011) on the Yelp data yielded an accuracy of 67.8%, which is significantly lower than 89.6% as reported on the AMT data in (Ott et al., 2011). The significantly lower accuracy on the Yelp data can be due to two reasons: (1) the fake and non-fake labels according to Yelp's filter are very noisy, (2) there are some fundamental differences between the Yelp data and the AMT data which are responsible for the big difference in accuracy.

To investigate the actual cause, we propose a principled information theoretic analysis. Our analysis shows that for the AMT data in (Ott et al. 2011), the word distributions in fake reviews written by Turkers are quite different from the word distributions in non-fake reviews from Tripadvisor. This explains why detecting crowdsourced fake reviews in the AMT data of (Ott et al., 2011) is quite easy, yielding a 89.6% detection accuracy.

However, in the Yelp data, we found that the suspicious reviewers (spammers) according to Yelp's filter used very similar language in their (fake) reviews as other non-fake (unfiltered) reviews. This resulted in fake (filtered) and non-fake (unfiltered) reviews of Yelp to be linguistically similar, which explains why fake review detection using n-grams on Yelp's data is much harder. A plausible reason could be that the spammers according to Yelp's filter made an effort to make their (fake) reviews sound convincing as other non-fake reviews. However, the spammers in the Yelp data left behind some specific psycholinguistic footprints which reveal deception. These were precisely discovered by our information theoretic analysis.

The inefficacy of linguistics in detecting fake reviews filtered by Yelp encouraged us to study reviewer behaviors in Yelp. Our behavioral analysis shows marked distributional divergence between reviewing behaviors of spammers (authors of filtered reviews) and non-spammers (others). This motivated us to examine the effectiveness of behaviors in detecting Yelp's fake (filtered) reviews. To our surprise, we found that the behaviors are highly effective for detecting fake reviews filtered by Yelp. More importantly, the behavioral features significantly outperform linguistic n-grams on detection performance.

Finally, using the results of our experimental study, we conjecture some assertions on the quality of Yelp's filtering and postulate what Yelp's fake review filter might be doing. We summarize our main results below:

1. We found that in the AMT data (Ott et al., 2011), the word distributions of fake and non-fake reviews are very different, which explains the high (90%) detection accuracy using n-grams. However, for the Yelp data, word distributions in fake and non-fake reviews are quite similar, which explains why the method in (Ott et al., 2011) is less effective on Yelp's real-life data.
2. The above point indicates that the linguistic n-gram feature based classification approach in (Ott et al., 2011) does not seem to be the (main) approach used by Yelp.
3. Using abnormal behaviors renders a respectable 86% accuracy in detecting fake (filtered) reviews of Yelp showing that abnormal behaviors based detection results are highly correlated with Yelp's filtering.
4. These results allow us to postulate that Yelp might be using behavioral features/clues in its filtering.

We will describe the detailed investigations in subsequent sections. We believe this study will be useful to both academia and industry and also to other review sites in their fake review filtering efforts. Before proceeding further, we first review the relevant literature below.

## Related Work

Web spam (Castillo et al., 2007; Spirin and Han, 2012) and email spam (Chirita et al., 2005) are most widely studied spam activities. Blog (Kolari et al., 2006), network (Jin et al., 2011), and tagging (Koutrika et al., 2007) spam are also studied. However, opinion spam dynamics differ.

Apart from the works mentioned in the introduction, Jindal et al., (2010) studied unexpected reviewing patterns, Wang et al., (2011) investigated graph-based methods for finding fake store reviewers, and Fei et al., (2013) exploited review burstiness for spammer detection.

Studies on review quality (Liu et al., 2007), distortion (Wu et al., 2010), and helpfulness (Kim et al., 2006) have also been conducted. A study of bias, controversy and summarization of research paper reviews is reported in (Lauw et al., 2006; 2007). However, research paper reviews do not (at least not obviously) involve faking.

Also related is the task of psycholinguistic deception detection which investigates lying words (Hancock et al., 2008; Newman et al. 2003), untrue views (Mihalcea and Strapparava (2009), computer-mediated deception in role-playing games (Zhou et al., 2008), etc.

| Domain | fake | non-fake | % fake | total # reviews | # reviewers |
|---|---|---|---|---|---|
| Hotel | 802 | 4876 | 14.1% | 5678 | 5124 |
| Restaurant | 8368 | 50149 | 14.3% | 58517 | 35593 |

Table 1: Dataset statistics

| Features | P | R | F1 | A |
|---|---|---|---|---|
| Word unigrams (WU) | 62.9 | 76.6 | 68.9 | 65.6 |
| WU + IG (top 1%) | 61.7 | 76.4 | 68.4 | 64.4 |
| WU + IG (top 2%) | 62.4 | 76.7 | 68.8 | 64.9 |
| Word-Bigrams (WB) | 61.1 | 79.9 | 69.2 | 64.4 |
| WB+LIWC | 61.6 | 69.1 | 69.1 | 64.4 |
| POS Unigrams | 56.0 | 69.8 | 62.1 | 57.2 |
| WB + POS Bigrams | 63.2 | 73.4 | 67.9 | 64.6 |
| WB + Deep Syntax | 62.3 | 74.1 | 67.7 | 64.1 |
| WB + POS Seq. Pat. | 63.4 | 74.5 | 68.5 | 64.5 |

(a): Hotel

| P | R | F1 | A |
|---|---|---|---|
| 64.3 | 76.3 | 69.7 | 66.9 |
| 64.0 | 75.9 | 69.4 | 66.2 |
| 64.1 | 76.1 | 69.5 | 66.5 |
| 64.5 | 79.3 | 71.1 | 67.8 |
| 64.6 | 79.4 | 71.0 | 67.8 |
| 59.5 | 70.3 | 64.5 | 55.6 |
| 65.1 | 72.4 | 68.6 | 68.1 |
| 65.8 | 73.8 | 69.6 | 67.6 |
| 66.2 | 74.2 | 69.9 | 67.7 |

(b): Restaurant

Table 2: SVM 5-fold CV results P: Precision, R: Recall, F1: F1-Score on the fake class, A: Accuracy in % across different sets of features. WU means word unigram, WB denotes word bigrams, and POS denotes part-of-speech tags. Top *k*% refers to using top features according to Information Gain (IG).

## Detecting Fake Reviews in Yelp

This section reports a set of classification experiments using the real-life data from Yelp and the AMT data from (Ott et al., 2011).

### The Yelp Review Dataset

To ensure credibility of user opinions posted on Yelp, it uses a filtering algorithm to filter fake/suspicious reviews and puts them in a filtered list. According to its CEO, Jeremy Stoppelman, Yelp's filtering algorithm has evolved over the years (since their launch in 2005) to filter shill and fake reviews (Stoppelman, 2009). Yelp is also confident enough to make its filtered reviews public. Yelp's filter has also been claimed to be highly accurate by a study in BusinessWeek (Weise, 2011)[1].

In this work, we study Yelp's filtering using its filtered (fake) and unfiltered (non-fake) reviews across 85 hotels and 130 restaurants in the Chicago area. To avoid any bias we consider a mixture of popular and unpopular hotels and restaurants (based on the number of reviews) in our dataset in Table 1. We note the class distribution is imbalanced.

### Classification Experiments on Yelp

We now report the classification results on the Yelp data.

**Classification Settings:** All our classification experiments are based on SVM[2] (SVM$^{Light}$ (Joachims, 1999)) using 5-fold Cross Validation (CV), which was also done in (Ott et al., 2011). We report linear kernel SVM results as it outperformed rbf, sigmoid, and polynomial kernels.

From Table 1, we see that the class distribution of the real-life Yelp data is skewed. It is well known that highly imbalanced data often produces poor models (Chawla et al., 2004). Classification results using the natural class distribution in Table 1 yielded an F1-score of 31.9 and 32.2 for the hotel and restaurant domains. For a detailed analysis of detection in the skewed (natural) class distribution, refer to (Mukherjee et al., 2013).

To build a good model for imbalanced data, a well-known technique is to employ under-sampling (Drummond and Holte, 2003) to randomly select a subset of instances from the majority class and combine it with the minority class to form a balanced class distribution data for model building. Since Ott et al., (2011) reported classification on balanced data (50% class distribution): 400 fake reviews from AMT and 400 non-fake reviews from Tripadvisor, for a fair comparison, we also report results on balanced data.

**Results using different feature sets**: We now report classification results using different feature sets. Our first feature set is word unigrams and bigrams[3] (which include unigrams). Our next feature set is LIWC[4] + bigrams. Ott et al., (2011) also used LIWC and word bigrams. Further, we try feature selection using information gain (IG).

Additionally, we also try style and part-of-speech (POS) based features which have been shown to be useful for deception detection (Feng et al., 2012b). We consider two types of features: Deep syntax (Feng et al., 2012b) and POS sequence patterns (Mukherjee and Liu, 2010).

Deep syntax based features in (Feng et al., 2012b) are lexicalized (e.g., PRP → "you") and un-lexicalized (e.g., $NP_2$ → $NP_3$ SBAR) production rules involving immediate or grandparent nodes of Probabilistic Context Free Grammar (PCFG) sentence parse trees. Deep syntax rules (obtained using Stanford Parser) produces a new feature set. We note the following from the results in Table 2:

1. Across both hotel and restaurant domains, word unigrams only yield about 66% accuracy on real-life fake review data. Using feature selection schemes (e.g., Information Gain) does not improve classification much.
2. Adding word bigrams (WB) slightly improves the F1 by 1-2% across both domains and improves accuracy to 67.8% in the restaurant domain. WB+LIWC performs similarly to bigrams alone but slightly reduces F1.
3. Using only POS unigrams deteriorates performance. Thus, POS unigrams are not useful for detection.
4. As word bigrams (WB) performed best on accuracy and F1 metrics, for subsequent feature settings, we add POS feature sets to WB to assess their strength.
5. Word bigrams (WB) + POS bigrams render slight improvements in accuracy over word bigrams but

[1] Yelp accepts that its filter may catch some false positives (Holloway, 2011), and also accepts the cost of filtering such reviews than the infinitely high cost of not having any filter at all which would render it become a *lassez-faire* review site that people stop using (Luther, 2010).
[2] We also tried naïve Bayes, but it resulted in slightly poorer models.
[3] Higher order n-grams did not improve performance.
[4] Linguistic Inquiry and Word Count, LIWC (Pennebaker et al., 2007) groups many keywords into 80 psychologically meaningful dimensions. We construct one feature for each of the 80 LIWC dimensions which was also done in (Ott et al., 2011).

reduces F1.

6. Deep syntax features slightly improve recall over WB + POS bigrams but reduces accuracy. POS sequence patterns also don't help.

Hence, we can say that LIWC and POS features make little difference compared to word unigram and bigram models.

## Comparison with Ott et al. (2011)

The previous experiments on the Yelp data yielded a maximum accuracy of 68.1% which is much less the 90% accuracy reported by Ott et al., (2011) on the AMT data. We now reproduce the setting of Ott et al. for comparison.

Recall that Ott et al., (2011) used AMT to craft fake reviews. Turkers (anonymous online workers) were asked to "synthesize" hotel reviews by assuming that they work for the hotel's marketing department and their boss wants them to write reviews portraying the hotel in positive light. 400 Turkers wrote one such review each across 20 popular Chicago hotels. These 400 reviews were treated as fake. The non-fake class comprised of 400 5-star reviews of the same 20 hotels from TripAdvisor.com.

As Turkers were asked to portray the hotels in positive light, it implies that they wrote 4-5 star reviews. To reproduce the same setting, we also use 4-5 star fake reviews in the real-life Yelp data. Note that our experiments in Table 2 used reviews with all star ratings (1-5★). Further, as in (Ott et al., 2011), we also use reviews of only "popular" Chicago hotels and restaurants from our dataset (Table 1). Applying these restrictions (popularity and 4-5★), we obtained 416 fake and 3090 non-fake reviews from our hotel domain (Table 3). 416 fake reviews are quite close and comparable to 400 in (Ott et al., 2011). Table 4 reports classification results on the real-life fake review data using the setting in (Ott et al., 2011) (i.e., with balanced class distribution on the data in Table 3). We also report results of our implementation on the AMT data of Ott et al., (2011) for a reliable comparison. From Table 4, we note:

1. Our implementation on AMT data of Ott et al. produces almost the same results[5] validating our implementation (Table 4(a)).
2. Reproducing the exact setting in (Ott et al., 2011), the Yelp real-life fake review data in the hotel domain only yields 67.6% accuracy (Table 4(b)) as opposed to 88.8% accuracy using the AMT data. This hints that the linguistic approach in (Ott et al., 2011) is not effective for detecting Yelp's real-life fake (filtered) reviews.

For additional confirmation, we also report results using the 4-5★ reviews from the restaurant domain (data in Table 3). Table 4(c) shows a similar (68.5%) accuracy bolstering our confidence that the linguistic approach is not so effective on Yelp's fake review data in the balanced class distribution. Note that results using n-grams (Ott et al., 2011) yields even poorer results in the natural class distribution. See (Mukherjee et al., 2013) for details.

In summary, we see that detecting real-life fake reviews in the commercial setting of Yelp is significantly harder (attaining only 68% accuracy) than detecting crowdsourced fake reviews which yield about 90% accuracy (Feng et al., 2012b; Ott et al., 2011) using n-grams. However, the actual cause is not obvious. To uncover it, we probe deep and propose a principled information theoretic analysis below.

| | Hotel 4-5★ | Restaurant 4-5 ★ |
|---|---|---|
| Fake | 416 | 5992 |
| Non-fake | 3090 | 44238 |

Table 3: 4-5 ★ reviews from Hotel and Restaurant domains.

| *n* | P | R | F1 | A |
|---|---|---|---|---|
| Uni | 86.7 | 89.4 | 88.0 | 87.8 |
| Bi | 88.9 | 89.9 | 89.3 | 88.8 |

(a) Ott et al., (2011)

| P | R | F1 | A |
|---|---|---|---|
| 65.1 | 78.1 | 71.0 | 67.6 |
| 61.1 | 82.4 | 70.2 | 64.9 |

(b) Hotel

| P | R | F1 | A |
|---|---|---|---|
| 65.1 | 77.4 | 70.7 | 67.9 |
| 64.9 | 81.2 | 72.1 | 68.5 |

(c) Restaurant

Table 4: Comparison with Ott et al. (2011) based on SVM 5-fold CV results. Feature Set: Uni: Word unigrams, Bi: Word bigrams.

| Dataset | # Terms | $KL(F\|\|N)$ | $KL(N\|\|F)$ | $\Delta KL$ | $JS$-Div |
|---|---|---|---|---|---|
| Ott et al. (2001) | 6473 | 1.007 | 1.104 | -0.097 | 0.274 |
| Hotel | 24780 | 2.228 | 0.392 | 1.836 | 0.118 |
| Restaurant | 80067 | 1.228 | 0.196 | 1.032 | 0.096 |
| Hotel 4-5 ★ | 17921 | 2.061 | 0.928 | 1.133 | 0.125 |
| Restaurant 4-5 ★ | 68364 | 1.606 | 0.564 | 1.042 | 0.105 |

Table 5: KL-Divergence of unigram language models.

| % Contr. | Ott. et al. | Ho. | Re. | Ho. 4-5 ★ | Re. 4-5★ |
|---|---|---|---|---|---|
| $\Delta KL$ | 20.1 | 74.9 | 70.1 | 69.8 | 70.3 |

(a) $k = 200$

| Ott. et al. | Ho. | Re. | Ho. 4-5★ | Re. 4-5★ |
|---|---|---|---|---|
| 22.8 | 77.6 | 73.1 | 70.7 | 72.8 |

(b) $k = 300$

Table 6: Percentage (%) of contribution to $\Delta KL$ for top $k$ words across different dataset. Ho: Hotel and Re: Restaurant.

## Information Theoretic Analysis

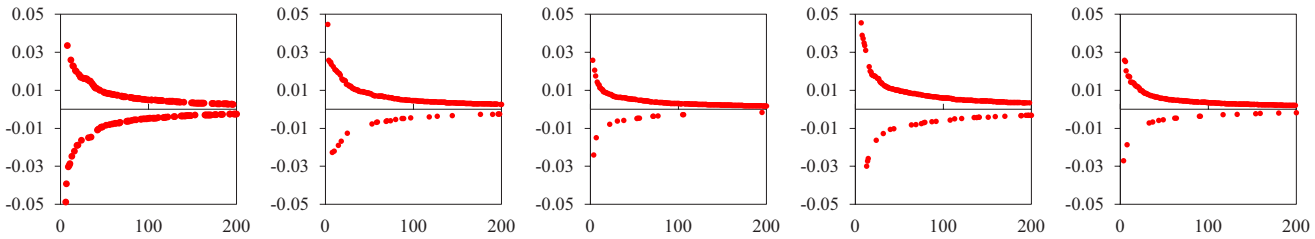
To explain the huge difference in accuracies, we analyze the word distributions in the AMT and Yelp data as text classification relies on the word distributions. We compute word probability distributions in fake and non-fake reviews using Good-Turing smoothed unigram language models[6]. As language models are constructed on documents, we construct a big document of fake reviews (respectively non-fake reviews) by merging all fake (non-fake) reviews.

To compute the word distribution differences among fake and non-fake reviews, we use Kullback–Leibler (KL) divergence: $KL(F||N) = \sum_i F(i) \log_2 \left(\frac{F(i)}{N(i)}\right)$, where $F(i)$ and $N(i)$ are the respective probabilities of word $i$ in fake and non-fake reviews. Note that KL-divergence is commonly used to compare two distributions, which is suitable for us.

In our context, $KL(F||N)$ provides a quantitative estimate of *how much* fake reviews linguistically (according to the

[5] Minor variations of classification results are common due to different binning in cross-validation, tokenization, etc. Ott et al. (2011) did not provide some details about their implementation, e.g., SVM kernel.

[6] Using Kylm - The Kyoto Language Modeling Toolkit




(a) Ott et al. (b) Hotel (c) Restaurant (d) Hotel 4-5 ★ (e) Restaurant 4-5 ★

Figure 1: Word-wise difference of KL-Div ($\Delta KL_{Word}$) across top 200 words (using $|\Delta KL_{Word}|$) for different datasets.

| Word (w) | $\Delta KL_{Word}$ | P(w\|*F*) (in E-4) | P(w\|*N*) (in E-6) |
|---|---|---|---|
| were | -0.147 | 1.165 | 19822.7 |
| we | -0.144 | 1.164 | 19413.0 |
| had | 0.080 | 163.65 | 614.65 |
| night | -0.048 | 0.5824 | 7017.36 |
| out | -0.0392 | 0.5824 | 5839.26 |
| has | 0.0334 | 50.087 | 51.221 |
| view | 0.0229 | 36.691 | 51.221 |
| enjoyed | 0.0225 | 44.845 | 153.664 |
| back | -0.019 | 0.5824 | 3226.96 |
| felt | 0.0168 | 28.352 | 51.222 |

(a) Ott et al.

| Word (w) | $\Delta KL_{Word}$ | P(w\|*F*) (in E-4) | P(w\|*N*) (in E-6) |
|---|---|---|---|
| us | 0.0446 | 74.81 | 128.04 |
| area | 0.0257 | 28.73 | 5.820 |
| price | 0.0249 | 32.80 | 17.46 |
| stay | 0.0246 | 27.64 | 5.820 |
| said | -0.0228 | 0.271 | 3276.8 |
| feel | 0.0224 | 24.48 | 5.820 |
| when | -0.0221 | 55.84 | 12857.1 |
| nice | 0.0204 | 23.58 | 5.820 |
| deal | 0.0199 | 23.04 | 5.820 |
| comfort | 0.0188 | 21.95 | 5.820 |

(b) Hotel

| Word (w) | $\Delta KL_{Word}$ | P(w\|*F*) (in E-4) | P(w\|*N*) (in E-6) |
|---|---|---|---|
| places | 0.0257 | 25.021 | 2.059 |
| options | 0.0130 | 12.077 | 0.686 |
| evening | 0.0102 | 12.893 | 5.4914 |
| went | 0.0092 | 8.867 | 0.6864 |
| seat | 0.0089 | 8.714 | 0.6852 |
| helpful | 0.0088 | 8.561 | 0.6847 |
| overall | 0.0085 | 8.3106 | 0.6864 |
| serve | 0.0081 | 10.345 | 4.8049 |
| itself | -0.0079 | .10192 | 1151.82 |
| amount | 0.0076 | 7.542 | 0.6864 |

(c) Restaurant

Table 7: Top words according to $|\Delta KL_{Word}|$ with their respective *F*ake/*N*on-fake class probabilities P(w|*F*) (in E-4, i.e., $10^{-4}$), P(w|*N*) (in E-6, i.e., $10^{-6}$) for different datasets. As the probabilities can be quite small, we report enough precision. Terms with $\Delta KL_{Word} < 0$ are marked in red.

frequency of word usage) differ from non-fake reviews. $KL(N||F)$ can be interpreted analogously.

An important property of KL-divergence (KL-Div.) is its asymmetry, i.e., $KL(F||N) \neq KL(N||F)$. Its symmetric extension is the Jensen-Shannon divergence (JS-Div.): $JS = \frac{1}{2}KL(F||M) + \frac{1}{2}KL(N||M)$, where $M = \frac{1}{2}(F+N)$. However, the asymmetry of KL-Div. can provide us some crucial information. In Table 5, we report $KL$, $\Delta KL = KL(F||N) - KL(N||F)$, and $JS$ divergence results for various datasets. We note the following observations:

- For the AMT data (Table 5, row 1), we get $KL(F||N) \approx KL(N||F)$ and $\Delta KL \approx 0$. However, for Yelp data (Table 5, rows 2-5), there are major differences, $KL(F||N) > KL(N||F)$ and $\Delta KL > 1$.
- The JS-Div. of fake and non-fake word distributions in the AMT data is much larger (almost double) than Yelp data. This explains why the AMT data is much easier to classify. We will discuss this in further details below.

The definition of $KL(F||N)$ implies that words having very high probability in $F$ and very low probability in $N$ contribute most to KL-Div., $KL(F||N)$. To examine the word-wise contribution to $\Delta KL$, we compute the word KL-Div. difference, $\Delta KL^i_{Word}$ for each word, $i$ as follows: $\Delta KL^i_{Word} = KL_{Word}(F_i||N_i) - KL_{Word}(N_i||F_i)$, where $KL_{Word}(F_i||N_i) = F(i)\log_2\left(\frac{F(i)}{N(i)}\right)$, and analogously $KL_{Word}(F_i||N_i)$.

Fig. 1 shows the largest absolute word KL-div. differences in descending order of $|\Delta KL^i_{Word}|$ of the top words for various datasets. Positive/negative values (of $\Delta KL^i_{Word}$) are above/below the $x$-axis. We further report the contribution of top $k$ words to $\Delta KL$ for $k = 200$ and $k = 300$ in Table 6. Lastly, for qualitative inspection, we also report some top words according to $|\Delta KL^i_{Word}|$ in Table 7. From the results in Fig. 1 and Tables 6 and 7, we draw the following two crucial inferences:

**1. The Turkers' did not do a good job at Faking!**

Fig. 1(a) shows a "symmetric" distribution of $\Delta KL^i_{Word}$ for top words (i.e., the curves above and below $y = 0$ are equally dense) for the AMT data. This implies that, there are two sets of words among the top words: i) set of words, $E$ appearing more in fake (than in non-fake) reviews, i.e., $\forall i \in E$, $F(i) > N(i)$ resulting in $\Delta KL^{i\in E}_{Word} > 0$ and ii) set of words, $G$ appearing more in non-fake (than in fake) reviews, i.e., $\forall i \in G$, $N(i) > F(i)$ resulting in $\Delta KL^{i\in F}_{Word} < 0$. Further, as the upper and lower curves are equally dense, we have $|E| \approx |G|$. Additionally, the top $k = 200$, 300 words (see Table 6(a, b), col. 1) only contribute about 20% to $\Delta KL$ for the AMT data. Thus, there are many more words in the AMT data having higher probabilities in fake than non-fake reviews (like those in set $E$) and also many words having higher probabilities in non-fake than fake reviews (like those in set $G$). This implies that the fake and non-fake reviews in the AMT data consist of words with very different frequencies. This hints that the Turkers did not do a good job at faking probably because they had little knowledge of the domain, or did not put their heart into writing fake reviews as they have little gain in doing so[7].

[7] The Turkers are paid only US$1 per review and they may not have genuine interest to write fake reviews. However, real fake reviewers on Yelp both know the domain/business well and also have genuine interests in writing fake reviews for that business in order to promote/demote.

This explains why the AMT data is easy to classify attaining 90% fake review detection accuracy.

Next, we investigate the *quality* of deception exhibited in AMT fake reviews. We look at some top words (according to largest divergence difference, $|\Delta KL_{Word}|$) in Table 7(a) appearing with higher probabilities in fake reviews (i.e., words with $\Delta KL^i_{Word} > 0$). These are *had*, *has*, *view*, etc. which are general words and do not show much "pretense" or "deception" as we would expect in fake reviews.

**2. Yelp Spammers are Smart but Overdid "Faking"!**

Yelp's fake review data (Table 5, rows 2-5), shows that $KL(F||N)$ is much larger than $KL(N||F)$ and $\Delta KL > 1$. Fig. 1 (b-e) also show that among the top $k$ = 200 words which contribute a major percentage (≈ 70%) to $\Delta KL$ (see Table 6 (a)), most words have $\Delta KL^i_{Word} > 0$ while only a few words have $\Delta KL^i_{Word} < 0$ (as the curves above and below $y = 0$ in Fig. 1 (b-e) are quite dense and sparse respectively). Beyond $k$ = 200 words, we find $\Delta KL^i_{Word} \approx 0$. For top $k$ = 300 words (Table 6(b)) too, we find a similar trend. Thus, in the Yelp data, certain top $k$ words contribute most to $\Delta KL$. Let $A$ denote the set of those top words contributing most to $\Delta KL$. Further, $A$ can be partitioned into sets $A^F = \{i|\Delta KL^i_{Word} > 0\}$ (i.e., $\forall i \in A^F, F(i) > N(i)$) and $A^N = \{i|\Delta KL^i_{Word} < 0\}$ (i.e., $\forall i \in A^N, N(i) > F(i)$) where $A = A^F \cup A^N$ and $A^F \cap A^N = \phi$. Also, as the curve above $y = 0$ is dense while the curve below $y = 0$ sparse, we have $|A^F| \gg |A^N|$.

Further, $\forall i \notin A$, we have $\Delta KL^i_{Word} \approx 0$ which implies that for $\forall i \notin A$, either one or both of the following conditions hold:

1. The word probabilities in fake and non-fake reviews are almost equal, i.e., $F(i) \approx N(i)$. Thus, $\log\left(\frac{F(i)}{N(i)}\right) \approx \log\left(\frac{N(i)}{F(i)}\right) \approx 0$ and $KL_{Word}(F_i||N_i) \approx KL_{Word}(N_i||F_i) \approx 0$, making $\Delta KL^i_{Word} \approx 0$.
2. The word probabilities in fake and non-fake are both very small, i.e., $F(i) \approx N(i) \approx 0$ resulting in very small values for $KL_{Word}(F_i||N_i) \approx 0$ and $KL_{Word}(N_i||F_i) \approx 0$, making $\Delta KL^i_{Word} \approx 0$.

These two conditions and the top words in the set $A$ contributing a large part to $\Delta KL$ for Yelp's data (Table 6) clearly show that for Yelp's fake and non-fake reviews (according to its filter), most words in fake and non-fake reviews have almost the same or low frequencies (i.e., the words $i \notin A$, which have $\Delta KL^i_{Word} \approx 0$). However, $|A^F| \gg |A^N|$ also clearly implies that there exist specific words which contribute most to $\Delta KL$ by appearing in fake reviews with much higher frequencies than in non-fake reviews, (i.e. the words $i \in A^F$, which have $F(i) \gg N(i)$, $\Delta KL^i_{Word} > 0$). This reveals the following key insight.

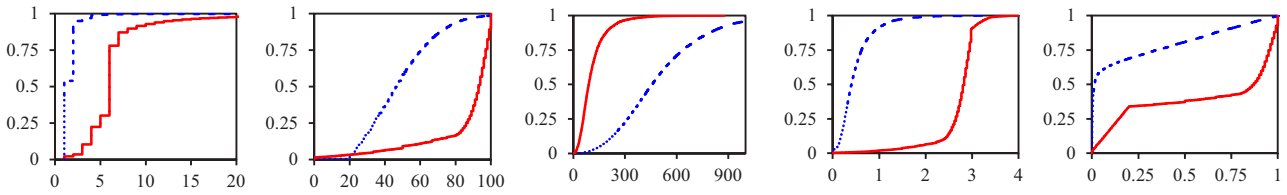
The spammers (authors of filtered reviews) detected by Yelp's filter made an effort (are smart enough) to ensure that their fake reviews have most words that also appear in truthful (non-fake) reviews so as to sound convincing (i.e., the words $i \notin A$ with $\Delta KL^i_{Word} \approx 0$). However, during the process/act of "faking" or inflicting deception, psychologically they happened to *overuse* some words consequently resulting in much higher frequencies of certain words in their fake reviews than other non-fake reviews (i.e., words $i \in A^F$ with $F(i) \gg N(i)$). A quick lookup of these words in $A^F$ with $\Delta KL_{Word} > 0$ in Yelp's data (see Table 7(b, c)) yields the following: *us*, *price*, *stay*, *feel*, *nice*, *deal*, *comfort*, etc. in the hotel domain; and *options*, *went*, *seat*, *helpful*, *overall*, *serve*, *amount*, etc. in the restaurant domain. These words demonstrate marked pretense and deception. Prior works in personality and psychology research (e.g., Newman et al., (2003) and references therein) have shown that deception/lying usually involves more use of personal pronouns (e.g., "us") and associated actions (e.g., "went," "feel") towards specific targets ("area," "options," "price," "stay," etc.) with the objective of incorrect projection (lying or faking) which often involves more use of positive sentiments and emotion words (e.g., "nice," "deal," "comfort," "helpful," etc.).

Thus, the spammers caught by Yelp's filter seem to have "overdone faking" in pretending to be truthful while writing deceptive fake reviews. However, they left behind linguistic footprints which were precisely discovered by our information theoretic analysis.

To summarize, let us discuss again why Yelp's fake review data is harder to classify than the AMT data. Table 6 shows that the symmetric JS-Div. for Yelp data is much lower (almost half) than the AMT data (JS divergence is bounded by 1, $0 \le JS \le 1$, when using $\log_2$). This dovetails with our theoretical analysis which also show that fake and non-fake reviews in the AMT data use very different word distributions (resulting in a higher JS-Div.). Hence, the standard linguistic n-grams could detect fake reviews in AMT data with 90% accuracy. However, for Yelp data, the spammers (captured by Yelp's filter) made an attempt to sound convincing by using those words which appear in non-fake reviews in their fake reviews almost equally frequently (hence a lower JS-Div.). They only overuse a small number of words in fake reviews due to (probably) trying too hard to make them sound real. However, due to the small number of such words, they may not appear in every fake review, which again explains why the fake and non-fake reviews in Yelp are much harder to classify using n-grams. We also tried using $\Delta KL^i_{Word}$ as another feature selection metric. However, using top $k$ = 200, 300, etc. words according to $|\Delta KL^i_{Word}|$ as features did not improve detection performance.

The above analysis shows that reviews filtered by Yelp demonstrate deception as we would expect in fake reviews[8]. Thus, we may conjecture that Yelp's filtering is reasonable and at least to a considerable extent reliable. Our study below will reveal more evidences to back Yelp's filtering quality.

[8] To reflect the real-life setting, we used natural class distribution (Table 1) for our information theoretic analysis. We also did experiments with 50-50 balanced data distribution setting which yielded similar trends.



(a) MNR (b) PR (c) RL (d) RD (e) MCS

Figure 2: CDF (Cumulative Distribution Function) of behavioral features. Cumulative percentage of spammers (in red/solid) and non-spammers (in blue/dotted) vs. behavioral feature value.

## Are AMT Fake Reviews Effective in Real-Life?

An interesting question is: Can we use AMT fake reviews to detect real-life fake reviews in a commercial website? This is important because not every website has filtered reviews that can be used in training. When there are no reliable ground truth fake and non-fake reviews for model building, can we employ crowdsourcing (e.g., Amazon Mechanical Turk) to generate fake reviews to be used in training? We seek an answer by conducting the following classification experiments on a balanced class distribution:

**Setting 1:** Train using the original data of Ott et al. (400 fake reviews from AMT and 400 non-fake reviews from Tripadvisor) and test on 4-5★ fake and non-fake Yelp reviews of the same 20 hotels as those used in Ott et al.

**Setting 2:** Train using 400 AMT fake reviews in Ott et al. and randomly sampled 400 4-5★ unfiltered (non-fake) Yelp reviews from those 20 hotels. Test on fake and non-fake (except those used in training) 4-5★Yelp reviews from the same 20 hotels.

**Setting 3:** Train exactly as in Setting 2, but test on fake and non-fake 4-5★ reviews from all Yelp hotel domain reviews *except* those 20 hotels. This is to see whether the classifier learnt using the reviews from the 20 hotels can be applied to other hotels. After all, it is quite hard and expensive to use AMT to generate fake reviews for every new hotel before a classifier can be applied to that hotel.

Table 8 reports the results. We see that across all settings, detection accuracies are about 52-54%. Thus, models trained using AMT generated (crowdsourced) fake reviews are not effective in detecting real-life fake reviews in a commercial website with detection accuracies near chance. The results worsen when trained models on balanced data are tested on natural distribution. This shows that AMT fake reviews are not representative of fake reviews in a commercial website (e.g., Yelp). Note that we can only experiment with the hotel domain because the AMT fake reviews of Ott et al. (2011) are only for hotels.

| n-gram | P | R | F1 | A |
|---|---|---|---|---|
| Uni | 57.5 | 31.0 | 40.3 | 52.8 |
| Bi | 57.3 | 31.8 | 40.9 | 53.1 |

(a) Setting 1

| P | R | F1 | A |
|---|---|---|---|
| 62.1 | 35.1 | 44.9 | 54.5 |
| 62.8 | 35.3 | 45.2 | 54.9 |

(b) Setting 2

| P | R | F1 | A |
|---|---|---|---|
| 67.3 | 32.3 | 43.7 | 52.7 |
| 67.6 | 32.2 | 43.6 | 53.2 |

(c) Setting 3

Table 8: SVM 5-fold CV classification results using AMT generated 400 fake hotel reviews as the positive class in training.

## Spamming Behavior Analysis

This section studies some abnormal spamming behaviors. For the analysis, we separate reviewers in our data (Table 1) into two groups: i. spammers: authors of fake (filtered) reviews; and ii. non-spammers: authors who didn't write fake (filtered) reviews in our data[9]. We analyze the reviewers' profile on the following behavioral dimensions:

**1. Maximum Number of Reviews (MNR):** Writing many reviews in a day is abnormal. The CDF (cumulative distribution function) of MNR in Fig. 2(a) shows that only 25% of spammers are bounded by 5 reviews per day, i.e., 75% of spammers wrote 6 or more reviews in a day. Non-spammers have a very moderate reviewing rate (50% with 1 review and 90% with no more than 3 reviews per day).

**2. Percentage of Positive Reviews (PR):** Our theoretical analysis showed that the deception words in fake reviews indicate projection in positive light. We plot the CDF of percentage of positive (4-5★) reviews among all reviews for spammers and non-spammers in Fig. 2(b). We see that about 15% of the spammers have less than 80% of their reviews as positive, i.e., a majority (85%) of spammers rated more than 80% of their reviews as 4-5★. Non-spammers show a rather evenly distributed trend where a varied range of reviewers have different percentage of 4-5 ★ reviews. This is reasonable as in real-life, people (genuine reviewers) usually have different rating levels.

**3. Review Length (RL):** As opinion spamming involves writing fake experiences, there is probably not much to write or at least a (paid) spammer probably does not want to invest too much time in writing. We show the CDF of the average number of words per review for all reviewers in Fig. 2(c). We see that a majority (≈ 80%) of spammers are bounded by 135 words in average review length which is quite short as compared to non-spammers where we find only 8% are bounded by 200 words while a majority (92%) have higher average review word length (> 200).

**4. Reviewer deviation (RD):** As spamming refers to incorrect projection, spammers are likely to deviate from the general rating consensus. To measure reviewer's deviation, we first compute the absolute rating deviation of a review from other reviews on the same business. Then,

[9] Our data yielded 8033 spammers and 32684 non-spammers (i.e., ≈20% of reviewers are spammers in our data).

we compute the expected rating deviation of a reviewer over all his reviews. On a 5-star scale, the deviation can range from 0 to 4. The CDF of this behavior in Fig. 2(d) shows that most non-spammers (≈70%) are bounded by an absolute deviation of 0.6 (showing rating consensus). However, only 20% of spammers have deviation less than 2.5 and most spammers deviate a great deal from the rest.

**5. Maximum content similarity (MCS):** To examine whether some posted reviews are similar to previous reviews, we compute the maximum content similarity (using cosine similarity) between any two reviews of a reviewer. The CDF in Fig. 2(e) shows that 70% of non-spammers have very little similarity (bounded by 0.18 cosine similarity) across their reviews showing non-spammers mostly write new content. However only 30% of spammers are bounded by a cosine similarity of 0.3 and the rest 70% have a lot of similarity across their reviews. This is both suspicious and abnormal.

The above analysis shows that the proposed behaviors are quite discriminating. Later, we'll study the efficacy of these features for detecting fake reviews filtered by Yelp.

It is worthwhile to note that there are also various other metadata that can be extracted from Yelp which can be used in detecting fake reviews. These include account activity, review count, friendship and fan relations, usefulness votes, percentage of previous reviews filtered, etc. However, using these features for classification is not fair as they are in some ways directly or indirectly affected by Yelp's filtering, e.g., if a review is filtered, its chance of getting usefulness votes, friend and fan requests reduce automatically. Further, when a spammer sees his reviews are filtered, he is likely to abandon that account which is linked with his account activity and review counts. However, the features MNR, PR, RL, RD, and MCS *record* past behaviors which cannot be undone and has to do with the very reviewing nature per-se than the account's reviews being filtered. Later, we will see that dropping any one feature does not affect detection performance much.

## Statistical Validation

Before applying the behavioral features for fake review detection, we first validate and examine their correlation with filtered reviews. We first normalize all behavioral features to continuous features in [0, 1]. Further to ensure that values close to 1 indicate spamming, we use the flipped version for RL behavior: RL = 1 – RL, as a lower value in this feature indicates spamming. Formally, for a given reviewer behavior $f$, its effectiveness ($Eff(\cdot)$) across fake and non-fake reviews can be defined as follows: $Eff(f) \equiv P(f > \beta|Fake) - P(f > \beta|NonFake)$ where $f > \beta$ is the event that the corresponding behavior exhibits spamming. On a scale of [0, 1] where values close to 1 (respectively 0) indicate spamming (non-spamming), choosing a threshold $\beta$ is somewhat subjective. While $\beta$ = 0.5 is reasonable (as it is the expected value of variables uniformly distributed in [0, 1]), $\beta = 0$ is very strict, and $\beta$ = 0.25 is rather midway. We experiment with all three threshold values for $\beta$. Let the *null hypothesis* be: Reviewers of both fake and non-fake reviews are equally likely to exhibit $f$ (spamming or attaining values > $\beta$), and the *alternate hypothesis*: reviewers of fake reviews are more likely to exhibit $f$ and are correlated with $f$. A Fisher's exact test rejects the null hypothesis with $p<0.01$ across different threshold values $\beta$ for each of the behaviors. Thus, fake (filtered) reviews are strongly correlated with abnormal behaviors of their corresponding reviewers.

Further, since the behaviors are all *anomalous* and not directly linked with filtering, and Fisher's exact test verifies strong correlation of those behaviors with "filtered" reviews, it also renders confidence that the vast majority of the labels (fake: filtered, non-fake: unfiltered) in the Yelp dataset are trustworthy.

## Using Behaviors in Detection

We now study detection performance using behavioral features. For each review we add the behavioral feature of its reviewer. Table 9 reports the results across various settings. We note the following:

1. Using only behavioral features (BF) boosts precision by about 20% and recall by around 7% in both domains resulting in around 14% improvement in F1. Thus, behaviors are stronger than linguistic n-grams for detecting real-life fake reviews filtered by Yelp.
2. N-grams + BF improve F1 by about 3% beyond using only BF showing that for Yelp filtered fake reviews, n-grams do help somewhat in detection. However, the improvement is only 3% because the word distributions in fake reviews in Yelp differ from non-fake reviews only on specific deception words. Thus, the classification result using BF + n-gram dovetails with our information theoretic finding.
3. However, compared to the results in rows 1, 2 and 3, we can see that the major gain is due to BF.

Note that we cannot test the behavioral features on the AMT data of Ott et al., (2011) as its fake reviews are generated by Turkers with no behavior information.

**Performance on Feature Ablation**: We now perform some ablation experiments to investigate the contribution of each behavioral feature. Table 9 shows that Bigrams + BF gives the highest accuracy and F1. Hence, we drop a behavioral feature at a time from the full feature set Bigrams + BF. We report results in Table 10. Note that feature selection metrics, e.g., Information Gain (IG) can also be used to assess the relative strength of each behavior. However, IG of a feature only reports the net reduction in entropy when that feature is used to partition the data. Although reduction in entropy using a feature (i.e., gain obtained using that feature) is correlated with the

discriminating strength of that feature, it does not give any indication on the actual performance loss when that feature is dropped. Here, we want to study the effect of each feature on actual detection performance.

Table 10 shows that dropping individual behavioral features results in a graceful degradation in detection performance. Dropping RL and MCS result in about 4-6% reduction in accuracy and F1-score showing that those features are more useful for detection. Dropping other features also results in 2-3% performance reduction. This shows that all the behavioral features are useful for fake review detection. Furthermore, even with reduced feature set (i.e., dropping one feature at a time), the model significantly outperforms linguistic n-grams. This is quite promising. We believe that our framework should be generic and applicable for fake review detection in other online review websites (as all features can be computed using posting date and star rating which are almost always available). Although the exact *results* obtained on Yelp may not directly apply to other sites, as reviewer activities can be different across different sites (Wang, 2010); the behaviors are general and can be tuned for other sites.

| Feature Setting | P | R | F1 | A |
|---|---|---|---|---|
| Unigrams | 62.9 | 76.6 | 68.9 | 65.6 |
| Bigrams | 61.1 | 79.9 | 69.2 | 64.4 |
| Behavior Feat.(BF) | 81.9 | 84.6 | **83.2** | **83.2** |
| Unigrams + BF | 83.2 | 80.6 | 81.9 | 83.6 |
| Bigrams + BF | 86.7 | 82.5 | **84.5** | **84.8** |

(a): Hotel

| P | R | F1 | A |
|---|---|---|---|
| 64.3 | 76.3 | 69.7 | 66.9 |
| 64.5 | 79.3 | 71.1 | 67.8 |
| 82.1 | 87.9 | **84.9** | **82.8** |
| 83.4 | 87.1 | 85.2 | 84.1 |
| 84.1 | 87.3 | **85.7** | **86.1** |

(b): Restaurant

Table 9: SVM 5-fold CV classification results across behavioral (BF) and *n*-gram features, P: Precision, R: Recall, F1: F1-Score on the fake class, A: Accuracy. Improvements using behavioral features over unigrams and bigrams are statistically significant with p<0.005 based on paired *t*-test.

| Dropped Feature | P | R | F1 | A |
|---|---|---|---|---|
| MNR | 84.9 | 80.6 | 82.7 | 83.3 |
| PR | 82.9 | 78.2 | 80.5 | 80.1 |
| RL | 82.7 | 78.0 | 80.3 | 79.7 |
| RD | 85.2 | 81.6 | 83.4 | 84.0 |
| MCS | 83.9 | 80.1 | 81.9 | 82.9 |

(a): Hotel

| P | R | F1 | A |
|---|---|---|---|
| 82.8 | 86.0 | 84.4 | 84.4 |
| 81.3 | 83.4 | 82.3 | 82.5 |
| 81.8 | 82.9 | 82.3 | 81.8 |
| 83.4 | 86.7 | 85.0 | 85.7 |
| 82.8 | 85.0 | 83.9 | 84.3 |

(b): Restaurant

Table 10: SVM 5-fold CV classification results by dropping behavioral features from the full feature set Bigram+BF (Table 9, last row). Differences in metrics for each dropped feature are statistically significant with $p<0.01$ based on paired *t*-test.

## Final Words about Yelp's Filtering

Grounding on the results of our experimental study, it is safe to assert the following.

**Yelp is at Least Doing a Reasonable Job at Filtering:** This assertion can be supported by four evidences from our study: i. Classification under balanced class distribution gives an accuracy of 67.8%, which is significantly higher than random guessing of 50%. Thus, linguistically there is a difference between filtered and unfiltered reviews in Yelp. Our information theoretic analysis characterizes the precise linguistic difference. Filtered reviews in Yelp exhibit noticeable deception and pretense implying filtered reviews to be fake. ii. Using abnormal behavioral features render even higher accuracy. It is not likely for a genuine reviewer to exhibit these behaviors. iii. Fisher's exact test verifies strong correlation of abnormal behaviors with filtered reviews. iv. Yelp has been doing industrial scale filtering since 2005 to provide the most trustworthy content (Stoppelman, 2009). If Yelp's filtering is ineffective, it would not have used it for the past 7 years. Although there are some complaints about filtering genuine reviews, considering the huge number of filtered reviews in Yelp some false positives are possible. Although these may not be hard evidences and we do not claim that Yelp's filtering is perfect, they are strong enough to render confidence that Yelp is doing a reasonable job at filtering.

**How does Yelp Filter?** Although an interesting question, it is hard to know the exact clues that Yelp uses as they are trade secrets. However, from our results, we can speculate that Yelp might be using a behavioral based approach for filtering. But we are not saying that Yelp only uses features similar to what we have proposed. In fact, Yelp can employ many other behavioral features extracted from its internal data, e.g., features based on IP addresses, geo-locations, network/session logs, mouse gestures, click behaviors, social network interactions (friend and fan relations) of reviewers at its website, etc. (Wang, 2010). Unfortunately, such information is not publicly available for us to experiment with. This may explain why our behavioral features although effective, are not achieving even higher detection accuracy. Nevertheless, our study and its results are worthwhile and give an insight into various types of clues that may be used by Yelp's filter.

## Conclusions

This paper performed an in-depth investigation of the nature fake reviews in the commercial setting of Yelp.com. Our study shows that although linguistic methods in (Ott et al., 2011; Feng et al., 2012b) reported very high (90%) detection accuracy on crowdsourced fake reviews, linguistic features do not work well on real-life fake reviews in the commercial setting of Yelp.com. Behavioral features yielded a respectable 86% accuracy indicating that Yelp's filter might be using a behavioral based approach. Finally, the results of our experimental study also demonstrated to a great extent that Yelp's filtering is reliable. Lastly, we showed that fake review detection using linguistic features (n-grams, POS, etc.) is not so effective in the real-life setting, and crowdsourced fake reviews may not be representative of real-life fake reviews. To the best of our knowledge, this is the first work that investigates the nature of real-life fake reviews filtered by

a commercial website and is but a nascent effort towards an escalating arms race to combat Web 2.0 abuse.

## Acknowledgements

This project is supported in part by a Google Faculty Award and by a NSF grant under no. IIS-1111092.

## References

Castillo, C., Donato, D., Gionis, A., Murdock, V., and Silvestri, F. 2007. Know your neighbors: web spam detection using the web topology. SIGIR.

Chawla, N., Japkowicz, N. and Kolcz, A. 2004. Editorial: Special issue on learning from imbalanced data sets. SIGKDD Explorations.

Chirita, P. A., Diederich, J., and Nejdl, W. 2005. MailRank: using ranking for spam detection. CIKM.

Drummond, C. and Holte, R. 2003. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats oversampling. In Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets.

Feng, S., Xing, L., Gogar, A., and Choi, Y. 2012a. Distributional Footprints of Deceptive Product Reviews. ICWSM.

Feng, S., Banerjee, R., and Choi, Y. 2012b. Syntactic Stylometry for Deception Detection. ACL (short paper).

Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R. 2013. Exploiting Burstiness in Reviews for Review Spammer Detection. ICWSM.

Hancock, J. T., Curry, L.E., Goorha, S., and Woodworth, M. 2008. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. Discourse Processes, 45(1):1–23.

Holloway, D. 2011. Just another reason why we have a review filter. http://officialblog.yelp.com/2011/10/just-another-reason-why-we-have-a-review-filter.html. Yelp Official Blog.

Jindal, N. and Liu, B. 2008. Opinion spam and analysis. WSDM.

Jindal, N., Liu, B. and Lim, E. P. 2010. Finding Unusual Review Patterns Using Unexpected Rules. CIKM.

Jin, X., Lin, C.X., Luo, J., and Han, J. 2011. SocialSpamGuard: A Data Mining-Based Spam Detection System for Social Media Networks. PVLDB.

Joachims, T. 1999. Making large-scale support vector machine learning practical. Advances in Kernel Methods. MIT Press.

Kim, S. M., Pantel, P., Chklovski, T. and Pennacchiotti, M. 2006. Automatically assessing review helpfulness. EMNLP.

Kolari, P., Java, A., Finin, T., Oates, T. and Joshi, A. 2006. Detecting Spam Blogs: A Machine Learning Approach. AAAI.

Kost, A. 2012. Woman Paid to Post Five-Star Google Feedback. ABC 7 News. http://www.thedenverchannel.com/news/ woman-paid-to-post-five-star-google-feedback.

Koutrika, G., Effendi, F. A., Gyöngyi, Z., Heymann, P. and H. Garcia-Molina. 2007. Combating spam in tagging systems. AIRWeb.

Lauw, H.W., Lim, E.P., Wang, K. 2006. Bias and Controversy: Beyond the Statistical Deviation. KDD.

Lauw, H. W., Lim, E.P. and Wang, K. 2007. Summarizing Review Scores of Unequal Reviewers. SIAM SDM.

Li, F., Huang, M., Yang, Y. and Zhu, X. 2011. Learning to identify review Spam. IJCAI.

Lim, E., Nguyen, V., Jindal, N., Liu, B. Lauw, H. 2010. Detecting product review spammers using rating behavior. CIKM.

Liu, J., Cao, Y., Lin, C., Huang, Zhou, M. 2007. Low-quality product review detection in opinion summarization. EMNLP.

Luther, 2010. Yelp's Review Filter Explained. http://officialblog.yelp.com/2010/03/yelp-review-filter-explained.html.

Mihalcea, R. and Strapparava, C. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. ACL-IJCNLP (short paper).

Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. 2013. Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews. UIC-CS-03-2013. Technical Report.

Mukherjee, A., Liu, B. and Glance, N. 2012. Spotting fake reviewer groups in consumer reviews. WWW.

Mukherjee, A, and Liu, B. 2010. Improving Gender Classification of Blog Authors. EMNLP.

Newman, M.L., Pennebaker, J.W., Berry, D.S., Richards, J.M. 2003. Lying words: predicting deception from linguistic styles, Personality and Social Psychology Bulletin 29, 665–675.

Ott, M., Choi, Y., Cardie, C. Hancock, J. 2011. Finding deceptive opinion spam by any stretch of the imagination. ACL.

Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., and Booth, R.J. 2007. The development and psychometric properties of LIWC2007. Austin, TX, LIWC.Net.

Popken, B. 2010. 30 Ways You Can Spot Fake Online Reviews. The Consumerist. http://consumerist.com/2010/04/14/how-you-spot-fake-online-reviews/

Spirin, N. and Han, J. 2012. Survey on web spam detection: principles and algorithms. ACM SIGKDD Explorations.

Stoppelman, J. 2009. Why Yelp has a Review Filter. Yelp Official Blog. http://officialblog.yelp.com/2009/10/why-yelp-has-a-review-filter.html

Streitfeld, D. 2012a. Fake Reviews Real Problem. New York Times. http://query.nytimes.com/gst/fullpage.html?res=9903E6DA1E3CF933A2575AC0A9649D8B63

Streitfeld, D. 2012b. Buy Reviews on Yelp, Get Black Mark. New York Times. http://www.nytimes.com/2012/10/18/technology/yelp-tries-to-halt-deceptive-reviews.html.

Wang, G., Xie, S., Liu, B., and Yu, P. S. 2011. Review Graph based Online Store Review Spammer Detection. ICDM.

Wang, Z. 2010. Anonymity, social image, and the competition for volunteers: a case study of the online market for reviews. The BE Journal of Economic Analysis & Policy.

Weise, K. 2011. A Lie Detector Test for Online Reviewers. BusinessWeek. http://www.businessweek.com/magazine/a-lie-detector-test-for-online-reviewers-09292011.html

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

Xie, S., Wang, G., Lin, S., and Yu, P.S. 2012. Review spam detection via temporal pattern discovery. KDD.

Zhou, L., Shi, Y., and Zhang, D. 2008. A Statistical Language Modeling Approach to Online Deception Detection. IEEE Transactions on Knowledge and Data Engineering.