Analyzing and Detecting Opinion Spam on a Large-Scale Dataset via Temporal and Spatial Patterns

Huayi Li[†], Zhiyuan Chen[†], Arjun Mukherjee[‡], Bing Liu[†] and Jidong Shao[§]
[†]Department of Computer Science, University of Illinois at Chicago, IL, USA
[‡]Department of Computer Science, University of Houston, TX, USA
[§]Dianping Inc., Shanghai, China
{lhymvp, czyuanacm}@gmail.com, arjun@cs.uh.edu, liub@cs.uic.edu, jidong.shao@dianping.com

Abstract

Although opinion spam (or fake review) detection has attracted significant research attention in recent years, the problem is far from solved. One key reason is that there is no large-scale ground truth labeled dataset available for model building. Some review hosting sites such as Yelp.com and Dianping.com have built fake review filtering systems to ensure the quality of their reviews, but their algorithms are trade secrets. Working with Dianping, we present the first large-scale analysis of restaurant reviews filtered by Dianping's fake review filtering system. Along with the analysis, we also propose some novel temporal and spatial features for supervised opinion spam detection. Our results show that these features significantly outperform existing state-of-art features.

1 Introduction

Despite the prevalence of opinion spam, existing methods are not keeping pace due to the unavailability of large-scale ground truth datasets in the real world commercial setting which impedes research of opinion spam detection. Existing work typically relies on pseudo fake reviews rather than real fake ones. For example, Jindal and Liu (2008) treated duplicate and near-duplicate Amazon product reviews as fake reviews. Li et al. (2011) manually labeled fake reviews by reading the reviews and comments, which are unreliable. Ott et al. (2011) used Amazon Mechanical Turk (AMT) to crowdsource anonymous online workers to write fake hotel reviews. The review dataset that they compiled had only 800 reviews which is too small to support reliable statistical analysis. In addition to that, the motivations and the psychological states of mind of hired Turkers and the professional spammers in the real world can be quite different as the results shown in (Mukherjee et al. 2013).

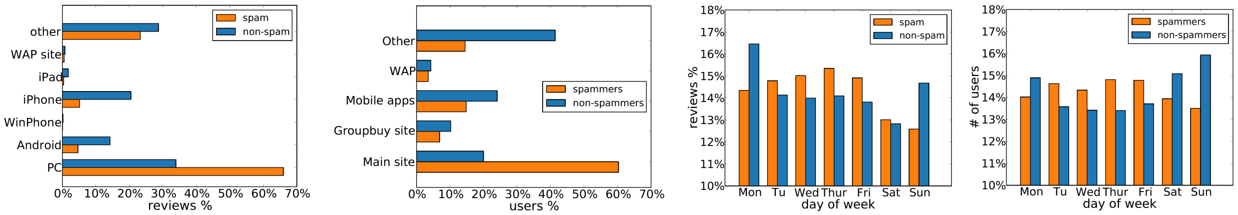
Companies such as Dianping and Yelp have developed effective fake review filtering systems against opinion spam. Mukherjee et al. (2013) reported the first analysis of Yelp's filter based on reviews of a small number of hotels and restaurants in Chicago. Their work showed that behavioral features of reviewers and their reviews are strong indicators of spamming. However, the reviews they used were not provided by Yelp but crawled from Yelp's business pages. Due to the difficulty of crawling and Yelp's crawling rate limit, they only obtained a small set of (about 64,000) reviews.

In this work, we study a large scale real-life restaurant review dataset with fake review labels provided by Dianping's spam detection system. Our work is demarcated from all previous works in the following dimensions:

- Data Volume: Our dataset is shared by Dianping with users' identity anonymized. It contains over 6 million reviews of all restaurants in Shanghai, China (Section 3). To the best of our knowledge, no existing study has been performed on such a large scale.
- Data Richness: Compared with other datasets, reviews in our dataset come with a much richer context, including users' IP addresses, users' profile. These additional data allow us to create more useful features for building machine learning models to spot review spammers.
- Feature Novelty: This paper is the first to give comprehensive insights of temporal and spatial features at various levels (reviews, users, IPs). Our experimental results show that the features and patterns that we propose in this paper build markedly more accurate classification models.

2 Related Work

Supervised learning is the most commonly used technique in opinion spam detection. Jindal and Liu (2008) built a logistic regression classifier with review feedback features, title and content characteristics and rating related features. Other researchers (Li et al. 2011; Ott et al. 2011; Feng, Banerjee, and Choi 2012) focused solely on the textual features, for instance, unigrams and bigrams. Mukherjee et al. (2013) further boosted the performance by appending users' behavioral features. Network-based approaches are exploited in (Wang et al. 2011; Akoglu, Chandy, and Faloutsos 2013; Li et al. 2014b) using various relational classifiers or graph propagation algorithms. Besides, with only a small portion of labeled reviews, researchers pointed out that using Positive-Unlabeled Learning (PU learning) (Li et al. 2014a; Ren, Ji, and Zhang 2014) outperforms traditional supervised learning. Since PU learning is not the focus of this work, we treat filtered reviews as positive and unfiltered reviews as negative.



(a) Review client pattern (b) Reg. portal pattern (c) Review temporal pattern (d) Reg. temporal pattern

Figure 1: Bar charts of patterns for users and reviews

Other interesting findings include rating behaviors (Günnemann, Günnemann, and Faloutsos 2014), spam topic models (Li, Cardie, and Li 2013), review burstiness and time-series analysis (Fei et al. 2013; Xie et al. 2012), spammer groups (Mukherjee, Liu, and Glance 2012) and reviewers with multiple userids or accounts (Qian and Liu 2013). Although the above works have made progresses, they rely on pseudo fake reviews (e.g., manually labeled, crowdsourced) which are noisy as opposed to real-world fake reviews.

3 Dataset

Our reviews shared by Dianping[1] consist of all reviews of all restaurants in Shanghai from November 1st, 2011 to April 18th, 2014. Users' IDs and IPs are anonymized. The dataset is not only much larger than those review datasets used in existing studies but also contains class labels produced by Dianping's fake review filter. Note that those fake reviews detected by Dianping's system were removed from business web pages. We further infer the class label of users and IP addresses by considering the majority class of all their reviews. That is, users/IPs are considered as spam users/IPs if more than 50% of their reviews are filtered by Dianping (i.e., fake). Table 1 shows the statistics of our data. Due to the confidentiality agreement with Dianping, we are unable to disclose more detailed information about their system. The large number of reviews, users, IPs and restaurants enables us to conduct a wide range of analyses that have never been done before. For simplicity, we use "spam" to represent fake reviews, "spammers" to refer to users who write fake reviews and "spam IPs" to represent IPs with a majority of fake reviews and "non-spam" to represent truthful reviews, and authentic users and organic IPs to represent users and IPs with less than one half of their reviews that are fake, respectively. Dianping further provide us with the city level locations associated with the IPs in our dataset. Locations of IPs encourage spatial analysis of spammers' behaviors.

Table 1: Statistics of restaurant review dataset in Shanghai

# of reviews	# of users	# of IPs	# of restaurants
6,126,113	1,074,604	1,331,471	108,787

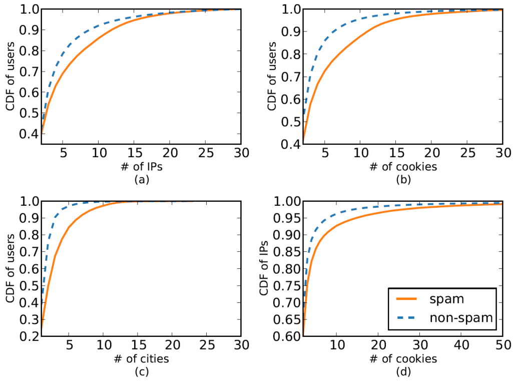


Figure 2: CDF of users and IPs v.s. number of other entities

4 Opinion Spam Analysis

4.1 Meta-data Patterns

Reviewers can use various devices. In the main site of Dianping, Spammers can quickly start writing fake reviews once registered. Consequently, the percentage of reviews with fake review labels posted from Dianping's main site is higher than all other clients in Figure 1(a). Dianping uses different site names to categorize the portals through which users register. Figure 1(b) shows the registration distribution of spammers and non-spammers in percentage based on each class. Since registering accounts through Dianping's main site is the fastest and most convenient way, spammers show a preference in registering on the main site. To show that spammers switch IP addresses/cities and even browser cookies more often than ordinary/genuine users, we thus plot the cumulative distribution (Figure 2) of the number of users as a function of the number of IPs, cookies and so on.

4.2 Temporal Patterns

Now we would like to show some longitudinal studies along the time dimension. In Figure 1(c), spammers are more active in weekdays except Mondays and less active in weekends comparing to organic users. This demonstrates that many spammers may be part-time workers who are usually busy on Monday with their own work and write more fake reviews in other weekdays. On the contrary, non-spammers who write authentic reviews based on their real personal experiences are more likely to post reviews on Sundays and

[1] http://www.dianping.com

Mondays after returning from dinner parties or hangouts that happen over weekends. A similar study of the number of users registered on days of week in Figure 1(d) reveals similar patterns.

4.3 Spatial Patterns

An interesting question we investigate here is that in order to maximize the profits from writing fake reviews, do spammers work for restaurants in the other cities beyond where they reside? First, we would like to see how spammers and non-spammers distribute across the major cities in China assuming the city where a user is registered is the city where the user lives. We map the IPs that users used in registration to a city level coordinate and we use geo-tagged pie charts for visualization as illustrated in Figure 3. The size of a pie chart represents the total number of users mapped to the city and its color portion reflects the ratio of spammers to non-spammers. Due to the fact that most users are registered in Shanghai, we exclude them from the study as we want to focus on the users outside Shanghai. There are two observations from this chart: (1) People in large cities (a few biggest charts) are dominated by non-spammers. This makes sense because people in large cities have higher salary and are likely to travel to Shanghai for vacation or business purposes. So their reviews are more likely to be authentic because they write reviews given their own experiences; (2) the further the cities are from Shanghai, the higher the ratios of spammers. A possible explanation is the travel cost. As the distance increases, the chance of people traveling to Shanghai drops and this is especially true when it comes to the underdeveloped cities in the western part of China where the profits of writing spam is reasonably attractive given the local average income. We can say that opinion spamming exhibits geographical outsourcing. We use side by side histogram in Figure 4 to represent the malicious IP ratio as a function of distance to Shanghai. It can be easily seen from the chart that IPs in cities that are 200+ miles away from Shanghai are mostly malicious.

4.4 Temporal and Spatial Patterns

In addition to the success of finding individual temporal and spatial patterns, there are more novel dynamics to explore when we combine spatial and temporal dimensions. Following the hypothesis that some professional spammers frequently change IP addresses to register many accounts in a short period of time, we postulate such spammers would also change IP addresses frequently when posting reviews to fool the Dianping's fake review filtering system.

We thus propose a novel metric to quantify the abnormal behaviors of such spammers and we call it the Average Travel Speed (ATS) measure. We define $S_u =< r_1, r_2, ..., r_{|S_u|} >$ as the sequence of reviews ordered by posting time-stamp of a reviewer u. Each review r_i consists of two primary attributes, IP address and time-stamp. As mentioned earlier, Dianping tagged each IP with a pair of coordinates of the city where it locates. Only 3.2% of IPs are not found in their IP database, we thus remove the reviews pertaining to those IPs for each user. The ATS measure aims to simulate the traveling sequence of a user. It averages the speed (miles per second) of a user from one location to the next in the sequence of movement. The rationale is that users who frequently and randomly "move" all over China with unusual speeds are highly likely to be spammers. The formal definition of ATS is in Equation 1 where $r_i = (t, IP, loc)$ and $r_i.t < r_j.t$ for $i < j$. The function $distance$ takes in two geo-coordinates and returns the Vincenty distance of the two points on earth. ATS of users with only one review is set to zero since ATS requires at least two reviews. Note that the IPs that spammers use can be IPs of proxy rather than the actual IPs of their end devices. Thus the ATS measure can also spot abnormal behaviors of frequent switching between IPs that are far apart.

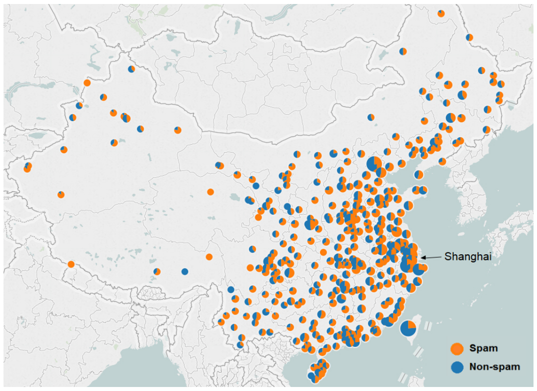


Figure 3: Distribution of spammers v.s. non-spammers registered in different cities. Shanghai City is excluded from the chart as its pie is too big. We also manually enlarge the spammers portion a little bit for the ease of demonstration.

$$ATS_u = \frac{\sum_{i=2}^{|S_u|} distance(r_i.loc, r_{i-1}.loc)}{|S_u| - 1} \quad (1)$$

There are two caveats for this analysis. First, we only have a complete set of reviews of restaurants in Shanghai between November 1st, 2011 and April 18th, 2014. Therefore, reviews to restaurants outside Shanghai are not counted. Second, city locations of IPs that we retrieved from Dianping IP database may not reflect the correct city locations of IPs as of the time when reviews were posted. In spite of these issues, we found many users whose ATS are exceptionally high which we can see from Figure 5. Most users are stationary who barely change IPs or city locations. It is also noteworthy that the majority of the users with unusually fast mobility rate are filtered by Dianping showing that novel spatio-temporal dynamics such as the average travel speed can be useful in spam detection.

5 Opinion Spam Detection

Now, we want to test the efficacy of our discovered patterns using a supervised learning approach to classify spammers and non-spammers (users). We propose a set of user level features that are strong indicators of opinion spammers. These new features are listed in Table 2. Some of the features/measures show in the tables are not at the user

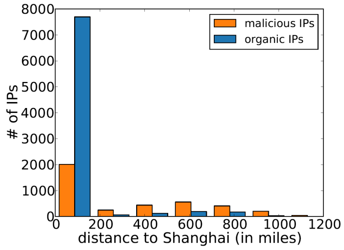


Figure 4: Histogram of IPs for various distance to Shanghai

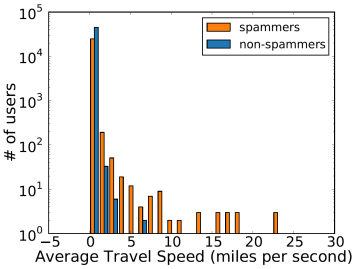


Figure 5: Abnormal patterns measured by ATS

level, so we define the features with respect to users. The spammers class is a very skewed class so we under-sample a subset of non-spammers and combine it with spammers to form a balanced set of users following the existing work in (Mukherjee et al. 2013). Experimental results are averaged based on 10-fold cross validation.

Table 2: Proposed Features

Feature Name	***Description***
regMainsite	Whether the user is registered on main site of Dianping
regTu2Tr	Whether the user is registered between Tue. and Thur.
regDist2SH	Distance from the city where a user registered to Shanghai
ATS	Average Travel Speed (Equation 1)
weekendPcnt	% of reviews written at weekends
pcPcnt	% of reviews posted through PC
avgDist2SH	Average distance from user city to Shanghai
AARD	Average absolute rating deviation of users' reviews
uIPs	# of unique IPs used by the user
ucookies	# of unique cookies used by the user
ucities	# of unique cities where users write reviews

Our compared state-of-the-art baselines are Support Vector Machines (SVM) with n-gram features (Ott et al. 2011) and behavioral features (Mukherjee et al. 2013). Table 3 shows the performances of the baselines, our proposed features, and the combination of all features respectively. From the results we can see that the proposed new features markedly outperforms the state-of-the-art baselines because it captures more subtle characteristics of opinion spammers. The combination of them achieves slightly better results.

6 Conclusions

This paper performed opinion spam analysis using a large-scale real-life dataset with high accuracy fake review labels shared by Dianping.com. To our knowledge, no such a large scale investigation has been done before. The rich content and the large scale data enabled us to broadly and deeply investigate the differences between spammers and non-spammers along many dimensions and to classify them.

Table 3: Results based on 10-fold cross validation

Method	**Accuracy**	**Precision**	**Recall**	**F1**
Unigram and Bigram	0.68	0.71	0.63	0.67
Behavioral Features	0.74	0.71	0.78	0.73
Proposed New Features	0.84	0.81	0.86	0.83
Combined	**0.85**	**0.83**	**0.87**	**0.85**

7 Acknowledgment

We would like to thank Dianping.com for allowing us to use their review dataset in this analysis. This work was supported in part by grants from National Science Foundation (NSF) under grant no. IIS-1111092 and IIS-1407927.

References

Akoglu, L.; Chandy, R.; and Faloutsos, C. 2013. Opinion fraud detection in online reviews by network effects. In *ICWSM*.

Fei, G.; Mukherjee, A.; Liu, B.; Hsu, M.; Castellanos, M.; and Ghosh, R. 2013. Exploiting burstiness in reviews for review spammer detection. In *ICWSM*.

Feng, S.; Banerjee, R.; and Choi, Y. 2012. Syntactic stylometry for deception detection. In *ACL*, 171–175.

Günnemann, S.; Günnemann, N.; and Faloutsos, C. 2014. Detecting anomalies in dynamic rating data: A robust probabilistic model for rating evolution. In *KDD*, 841–850.

Jindal, N., and Liu, B. 2008. Opinion spam and analysis. In *WSDM*, 219–230.

Li, F.; Huang, M.; Yang, Y.; and Zhu, X. 2011. Learning to identify review spam. In *IJCAI*, 2488–2493.

Li, H.; Chen, Z.; Liu, B.; Wei, X.; and Shao, J. 2014a. Spotting fake reviews via collective positive-unlabeled learning. In *ICDM*, 899–904.

Li, H.; Mukherjee, A.; Liu, B.; Kornfield, R.; and Emery, S. 2014b. Detecting campaign promoters on twitter using markov random fields. In *ICDM*, 290–299.

Li, J.; Cardie, C.; and Li, S. 2013. Topicspam: a topic-model based approach for spam detection. In *ACL*, 217–221.

Mukherjee, A.; Venkataraman, V.; Liu, B.; and Glance, N. S. 2013. What yelp fake review filter might be doing? In *ICWSM*.

Mukherjee, A.; Liu, B.; and Glance, N. S. 2012. Spotting fake reviewer groups in consumer reviews. In *WWW*, 191–200.

Ott, M.; Choi, Y.; Cardie, C.; and Hancock, J. T. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *ACL*, 309–319.

Qian, T., and Liu, B. 2013. Identifying multiple userids of the same author. In *EMNLP*, 1124–1135.

Ren, Y.; Ji, D.; and Zhang, H. 2014. Positive unlabeled learning for deceptive reviews detection. In *EMNLP*, 488–498.

Wang, G.; Xie, S.; Liu, B.; and Yu, P. S. 2011. Review graph based online store review spammer detection. In *ICDM*, 1242–1247.

Xie, S.; Wang, G.; Lin, S.; and Yu, P. S. 2012. Review spam detection via temporal pattern discovery. In *KDD*, 823–831.