Contrasting Fake Reviews in TripAdvisor (discussion paper)

Francesco Buccafurri[1], Michela Fazzolari[2], Gianluca Lax[1], and Marinella Petrocchi[2]

[1] DIIES, University Mediterranea of Reggio Calabria
Via Graziella, Località Feo di Vito
89122 Reggio Calabria, Italy
E-mail:{bucca,lax}@unirc.it
[2] Istituto di Informatica e Telematica - CNR
Via G. Moruzzi, 1
56124 Pisa, Italy
E-mail:{m.fazzolari,m.petrocchi}@iit.cnr.it

Abstract. Fake reviews are a concrete problem still affecting the reliability of systems like TripAdvisor, especially in the case of few reviews, thus in the first, most vulnerable, activity period of operators. In this work-in-progress paper, we present a model aimed to contrast this problem, based on a sort of normalization of scores given by users, to take into account the level of assurance of the reviews. This is done by considering two different dimensions, combined each other, which are the level of assurance of the identity of the review's author and the level of assurance of the occurrence of the evaluated experience. The paper presents a first validation of the approach conducted on real-life data, giving us very encouraging results.

Keywords reputation model, trust management, TripAdvisor.

1 Introduction

Over the last years, online reviews became very important since they reflect the customers' experience about a product or a service and nowadays constitute the basis on which the reputation of an organization is built. Online reviews have a great influence on the purchase decisions of other customers, who are increasingly relying on them [6].

Unfortunately, the confidence in such reviews is often misplaced, due to the fact that scammers are tempted to write fake information in exchange for some reward or to mislead consumers for obtaining business advantages [8]. These reviews are called *opinion spam* or *fake reviews* [4, 5].

SEBD 2018, June 24-27, 2018, Castellaneta Marina, Italy. Copyright held by the author(s).

The identification of fake reviews is not an easy task, since they can be identical to genuine ones. Nevertheless, several automatic techniques have been proposed in recent years. Fake reviews detection involves the identification of a set of features, linked with the content (review centric features) or with the review author (reviewer centric features).

Most of the existing machine learning approaches are not sufficiently effective in spotting fake reviews, nevertheless they are more reliable than manual detection. There exist several studies in the literature that rely on machine learning approaches and consider different set of review features [7, 10]. Further studies consider also reviewer centric features [9], which cannot be extracted from the text of a single review. In addition, graph-theory based approaches have been investigated to find relationships between reviews and their corresponding authors [3]. The spam detection techniques that combine reviews features and reviewers behaviors normally lead to better results [11].

In this work-in-progress paper we propose an approach different from the previous fake review detection approaches. Moreover, we introduce a reputation model and its preliminary experimental validation, designed to contrast the phenomenon of fake reviews in TripAdvisor. TripAdvisor is a very famous travel Website collecting reviews of travel-related contents. On the basis of these reviews, an aggregate score of each content is shown. Due to the economic value related to the effects of this system, and despite the efforts declared by TripAdvisor, the system is not immune from the problem of dishonest reviews, aimed either to fictitiously promote a given operator (i.e., self promoting attack) or to denigrate a competitor (i.e., slandering attack). Therefore, the noise occurring in the reviews derives not only from physiological subjectivity of the users [1]. Starting from a preliminary proposal [2], we define a model and a consequent methodology aimed to (partially) purify the system from the noise coming from fake reviews, in order to obtain more reliable *normalized* scores. This is done by considering two different dimensions, combined each other, which are the level of assurance of the identity of the review's author and the level of assurance of the occurrence of the evaluated experience.

It is worth noting that the approach here proposed is heuristic and any feature used to compute the trust is based on reasonable argumentations and ad hoc observations of the phenomenon, with no a specific validation of any single feature. Anyway, the paper is just aimed to give a general experimental validation of the whole approach and, thus, of the global combination of any feature.

The novel contributions, w.r.t. the initial proposal presented in [2], is that the model does not introduces new features required to TripAdvisor (in favor of the practical relevance of the proposal) and that this paper includes also an experimental validation of real-life TripAdvisor data.

The structure of the paper is the following. In the next section, we define the proposed reputation model. In Section 3, we describe the experiments carried out to validate our proposal. Finally, our conclusions are summarized in Section 4.

2 The Reputation Model

In this section, we describe our reputation model, whose components are listed in the following.

1. A set U of users, corresponding to the set of travelers, potentially covering all Web users.
2. A set S of service providers, corresponding to the set of restaurants, bars, hotels, and other operators registered to TripAdvisor.
3. For each service provider $s \in S$, a list of feedbacks $R(s)$ (which are the reviews), each corresponding to a transaction. A feedback $r_s \in R(S)$ for the service provider s is a tuple $\langle u, d, v, k, t, I \rangle$, where u is the author of the feedback, d is the time of the feedback, v is the time of the transaction, k is the score given by u on s (it is the aggregate score – also detailed into different dimensions), t is a text motivating k (it is the text included by the user to describe the experience) and I is the set of additional resources (it is the images posted by the user). Concerning u, we rely on the set of attributes which can be drawn from the system (and also by using external sources, like social-network sources).

The transaction corresponding to a feedback r_s is denoted by t_r. To implement the notion of certified reputation in this model, for a given feedback r_s, we need two basic measures, which are both numbers ranging in the interval $[0, 1]$: (1) the trustworthiness of the identity of u, denoted by $trust(u)$, and (2) the trustworthiness of the transaction t_r, denoted by $trust(t_r)$.

The first measure $trust(u)$ is related to identity management issues and is a measure of the level of assurance of identity proofing given by the registration phase into the reputation system. Observe that we do not consider the level of assurance of the authentication phase, thus assuming that no attacks on accounts of users occur. This measure may take into account also external information associated with the digital identity of the user in the system (e.g., information coming from different sources as online social networks). We remark that the trustworthiness of the identity of u is directly related to misbehaving users, as any malicious activity is facilitated when the level of assurance of user identity is low. The trustworthiness of the transaction, denoted by $trust(t_r)$ is equally important. Indeed, fake reviews usually correspond to transactions that never occurred.

On the basis of the two measures above, we measure the trustworthiness $trust(r_s)$ of a feedback $r_s = \langle u, d, v, k, t, I \rangle$ associated with a transaction t_r, by the function $trust(r_s) = f(trust(u), trust(t_r))$ such that the higher $trust(u)$ and $trust(t_r)$, the higher $trust(r_s)$. In this paper, we experiment as first attempt a simple function f which is the linear combination of the two contributions. Once $trust(r_s)$ has been computed, the score k of the feedback r_s is *corrected* on the basis of the overall trustworthiness $trust(r_s)$ by means of a function $g(trust(r_s))$. This function is build is such a way that the score k is much closer to the average score obtained by the service provider s as the value $trust(r_s)$ is low.

Specifically, $g(k) = \frac{(\alpha + trust(r_s)) \cdot k}{\sum_{r'_s \in R(S)} (\alpha + trust(r'_s))}$, where α is s suitable (small) offset to avoid null terms. This way, the mean computed over all the scores obtained by the operator s is just the mean weighted by trust values (shifted by α) of each review. Therefore, we obtain the *normalized* feedback r^*_s as $r^*_s = \langle u, d, v, g(k), t, I \rangle$.

Let us explain now how $trust(u)$, representing the level of assurance of identity the user u authoring the review, is computed. Of course, a significant part of the current weakness of the reputation system of TripAdvisor is based on the weakness of its digital identity management system.

Concerning $trust(u)$, we observe that the registration phase of TripAdvisor does not force the user to provide any non-self declared credential. Anyway, the possibility of registering via an existing Facebook profile is also allowed.

To compute the level of assurance of the identity, we consider in our model the following features (which are, in turn, numbers from 0 to 1).

- re (*reviewer experience*): it measures the seniority of the user in the system.
- tc (*text coherence*): it measures the coherence of the known information about the user (for example, the gender) and the text.
- rc (*review count*): it measures the number of reviews of the user. This feature is related to the fact that often fake reviews are done through accounts aimed to a specific goal (for example, for self-promotion or slandering attacks), so they are not reused massively, also to avoid the linkage with de-anonymizing information.
- fi (*Facebook identity*): it measures the level of assurance of the Facebook identity of the user (it is trivially 0 if the TripAdvisor account is not associated with a Facebook profile). Concerning this feature, we argue that a fake reviewer does not have any interest in allowing linkage of her/his TripAdvisor account with other (even fake) accounts, because probably tends to operate as much as possible in an anonymous way.

$trust(u)$ is then obtained as a linear combination of the above components.

To compute the level of assurance of the transaction $trust(t_r)$, we consider in our model the following features (which are, again, numbers from 0 to 1).

- dc (*data coherence*): it measures the coherence between the date of the review and the date of the transaction (this measure is based on the fact, according to our estimation, 50% of the reviews is done within 23 days, 75% within 34 days and 80% within 40).
- ip (*image proof*): it measures the degree of proof given by posted images that the transaction really occurred (trivially, no posted images corresponds to 0, the presence of images recognized as coherent with the other posted by the majority of users corresponds to the maximum value). Indeed, the standard behavior of a faker is to hide as much as possible any information that could be a potential risk for de-anonymization, also the publication of images, which is a typical action done by honest reviewers to give a proof of their claims.

- *rl* (*review locality*): it measures the presence of other reviews by the same user in the same location (city or region) if different from the place of residence corresponding to transactions experienced in the same period.
- *or* (*operator reaction*): it measures the presence of a reaction posted by the operator which is an outlier w.r.t. the standard behavior of the operator.

$trust(t_r)$ is also obtained as a linear combination of the above components.

3 Experiments

In this section, we describe the experiments carried out to validate the proposal: they are based on real-life reviews of restaurants extracted from the TripAdvisor site. First, we describe the data collection procedure and the error metrics used in the experiments. Then, we discuss how a score quantifying the (actual) quality of a restaurant has been computed (it is used as ground truth) and the method adopted to tune the weight of the reputation model parameters. Finally, we show the improvements obtained by our proposal in measuring the score of a restaurant.

3.1 Test Bed

For the study presented in this contribution, we consider data taken from TripAdvisor. Data were collected in January 2017, by developing an ad-hoc scraping software and by using the available API to crawl information. The web scraping process was performed by a Python script that navigated through the restaurants available on the Province of Lucca web page. The metadata related to a review, such as the language of the review, the rating, etc, were obtained by the available APIs. We also stored the reviewers' profiles, which include, when available, the age and country of origin.

At the end of the extraction phase, we obtained a dataset composed of 1.499 restaurants, 60.613 reviewers, and 107.556 reviews. Each review judges the quality of a restaurant by an integer *score* from 1 to 5.

In our experiments, we define the *bias* of a review w as $B^w = \frac{|s^w - GT^r|}{5}$, where s^w is the review score of the restaurant r, the operator $|x|$ denotes the absolute value of the number x, and Q^r (expected quality) is a real number in the interval $[1, 5]$ representing the quality of the restaurant r, which is computed as the average of the scores of that restaurant. B^w represents how much the review score is far from the score of the restaurant and it is normalized w.r.t. the maximum score (i.e., 5). For example, $B^w = 0$ means that the review score is coherent with the quality of the restaurant.

We define *review bias* of a restaurant r as

$$RB^r = \frac{\sum_{i=0}^{n} |B^{W_i}|}{n} \tag{1}$$

https://www.tripadvisor.com/Tourism-g187898-Lucca_Province_of_Lucca_Tuscany-Vacations.html where W_i is the i-th of the n reviews of the restaurant r. In words, it is used to quantify the amount of variation of the reviews of a restaurant w.r.t. the expected quality of that restaurant. For example, $RB^r = 0$ means that all review scores coincide and their value reflects the quality of r.

Given a reputation model t_u, we define its *error %*

$$E^u = \frac{\sum_{i=0}^{m} RB_i^r}{m} \cdot 100 \tag{2}$$

where m is the number of restaurants used in the reputation model and r_i is the i-th one. Clearly, it is the average of the review bias computed for all restaurants.

Finally, to compare the accuracy of two reputation models t_1 and t_2, we define the *improvement %* of t_1 (w.r.t. t_2) as $I_1 = \frac{E^1 - E^2}{E^1} \cdot 100$. Observe that the improvement can be negative in case the reputation model is less accurate than the compared one.

3.2 Parameter Setting

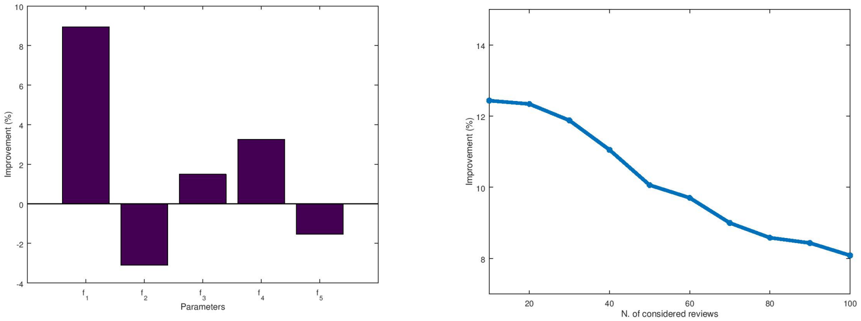
The reputation model presented in this paper uses only five of the parameters defined in Section 2, because they are the only ones evaluable from the collected dataset: two based on the identity of the reviewer and three based on the transaction. They assume a rate from 0 to 1 and how their rate is computed is now discussed.

- f_1 This parameter is related to the number of reviews done by the reviewer. We computed an average of 36 reviews for each reviewer and we assigned 1 to this parameter to reviewers with at least twice the average value. This values is linearly reduced to 0 for reviewers with only 1 review.
- f_2 This parameter is 1 if the reviewer signs in by Facebook, 0 otherwise.
- f_3 This parameter is 1 if the review contains at least an image, 0 otherwise.
- f_4 This parameter is related to length of the reviewer membership. As We have reviewers with activity up to 180 months, we assigned 1 to the rate of ID2 = 1 to reviewers with at least 90 months of activity. This values is linearly reduced to 0 for reviewers registered very recently (clearly, recently w.r.t. the period in which data have been collected)
- f_5 This parameter is related to the coherence between review date and visit date. We set $f_5 = 1$ if *time delta* < 15 days, else $f_5 = .75$ if *time delta* < 30 days, else $f_5 = .25$ if *time delta* < 45 days, 0 otherwise.

In the next experiment, we analyze the performance of our model in which a single parameter is considered. The result of this experiment is reported in Fig. 1.(a) and shows that some parameters are useful to improve the reputation model (namely, f_1, f_3, and f_4), others reduce its performance.

The next task is to combine all parameters and, for this purpose, we need to give a weight to each of them. Such weights are computed by applying a

Observe that this number includes many reviews that are not included in our dataset.



(a) Improvement of the reputation model enabling only one parameter at a time

(b) Performance of the proposed reputation model

Fig. 1: Experiment results

multivariable regression model with the five parameters $f_1 \dots f_5$ which returned the following weights $w_1 = 0.6068$, $w_2 = 0.2520$, $w_3 = 0.0605$, $w_4 = 0.2864$, $w_5 = 0.5778$, where w_x is the weight of the parameter f_x with $1 \leq x \leq 5$. These weights will be used in the next experiment.

3.3 Validation

In this experiment, we measure the performance of the reputation model of TripAdvisor and the performance obtained by using the reputation model proposed in this paper, setting the parameter weight to the values obtained in Section 3.2. We measured the review bias for the first n reviews, with n ranging from 10 to 100: we limited the upper bound to 100 because, for higher values, the review bias is very low. In Fig. 1.(b), we report the improvement % of the proposed reputation model w.r.t. that of TripAdvisor. It is possible to see that our proposal always gives the best results: the improvement is higher when a restaurant has few reviews, that is when it is more vulnerable to fake reviews.

4 Conclusion

TripAdvisor, as well as many other reputation systems, suffers from self-promoting and slandering attacks typically performed by using fake accounts just created for this purpose, which post fake reviews not corresponding to real experiences. In this paper, we defined a new reputation model for the reviews of TripAdvisor. Our proposal aims at evaluate the dependability of a review on the basis of a level of assurance for both identity proofing and truth of the transaction. We validate our proposal by measuring the improvement obtained by enabling our reputation model on a real-life dataset reviews referring to restaurants of the Province of Lucca, Italy. Anyway, the paper is just aimed to give a general experimental validation of the whole approach and, thus, of the global combination of any feature. As a future work, a selective validation of the different features can be also performed.

Acknowledgment

This work has been partially supported by the project "Secure Citizen Remote Identification" (CUP J88C17000150006) and by the project "Stabilimento Virtuale" (CUP J38C17000100006), both funded by Regione Calabria - POR Calabria FESR-FSE 2014-2020 - Asse I and by the European Union under grants 675320 (NECS-Network of Excellence in Cybersecurity) and 700294 (C3ISP-Collaborative and Confidential Information Sharing and Analysis for Cyber Protection) and by Fondazione Cassa di Risparmio di Lucca that partially funds the ReviewLand project.

References

1. J. K. Ayeh, N. Au, and R. Law. Do We Believe in TripAdvisor? Examining Credibility Perceptions and Online Travelers' Attitude toward Using User-Generated Content. *Journal of Travel Research*, 52(4):437–452, 2013.
2. F. Buccafurri, G. Lax, S. Nicolazzo, and A. Nocera. A model implementing certified reputation and its application to tripadvisor. In *Availability, Reliability and Security (ARES), 2015 10th International Conference on*, pages 218–223. IEEE, 2015.
3. E. Choo, T. Yu, and M. Chi. Detecting opinion spammer groups through community discovery and sentiment analysis. In P. Samarati, editor, *Data and Applications Security and Privacy XXIX*, pages 170–187, Cham, 2015. Springer International Publishing.
4. M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada. Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2(1):23, Oct 2015.
5. A. Heydari, M. a. Tavakoli, N. Salim, and Z. Heydari. Detection of review spam. *Expert Syst. Appl.*, 42(7):3634–3642, May 2015.
6. N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 219–230, New York, NY, USA, 2008. ACM.
7. F. Li, M. Huang, Y. Yang, and X. Zhu. Learning to identify review spam. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2488–2493. AAAI Press, 2011.
8. B. Liu. *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers, 2012.
9. A. Mukherjee. Detecting deceptive opinion spam using linguistics, behavioral and statistical modeling. In *ACL*, 2015.
10. A. Mukherjee, B. Liu, and N. Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 191–200, New York, NY, USA, 2012. ACM.
11. S. Rayana and L. Akoglu. Collective opinion spam detection using active inference. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 630–638, 2016.