
Proceedings Paper:
Banerjee, Snehasish orcid.org/0000-0001-6355-0470 (2018) A Methodological Template to Construct Ground Truth of Authentic and Fake Online Reviews. In: 2018 IEEE International Conference on Data Science and Advanced Analytics. IEEE International Conference on Data Science and Advanced Analytics, 01-04 Oct 2018 IEEE , ITA , pp. 641-648.

A Methodological Template to Construct Ground Truth of Authentic and Fake Online Reviews

Snehasish Banerjee
The York Management School
University of York
Freboys Lane, Heslington, York YO10 5GD, UK
Email: snehasish.banerjee@york.ac.uk

***Abstract*—With the emergence of opinion spam, scholars in recent years have been investigating how to distinguish between authentic and fake online reviews. In this research area however, constructing ground truth has been a tricky problem. When labeled datasets of authentic and fake reviews are unavailable, it becomes impossible to systematically investigate differences between the two. In light of this problem, the goal of this paper is three-fold: (1) To review existing approaches of developing ground truth, (2) To present an improved methodological template to construct ground truth, and (3) To conduct a quality-check of the newly constructed ground truth. The existing approaches are dissected to identify several peculiarities. The new approach invests in mitigating pitfalls in the current approaches. In the newly constructed ground truth, authentic reviews were found to be not easily distinguishable from fake reviews. Finally, new research directions are identified with the hope that scholars would be able to stay ahead in their relentless race against spammers.**

***Keywords*—credibility, fake review, ground truth, online review, opinion spam, spam 2.0**

I. Introduction

The credibility of the bewildering array of online reviews, often available now on the Internet for a single product or service, can hardly be taken for granted. This is due to what is referred as spam 2.0, also known as opinion spam [1].

In general, spam refers to the abuse of electronic messaging systems through dissemination of unsolicited messages in bulk [2]. Spam 2.0 specifically refers to the propagation of unsolicited contents to infiltrate social media applications. Since such unsolicited contents are deliberately written to be passed off as authentic, they are difficult to be detected by state-of-the-art countermeasures such as blacklisting and keyword filtering [3, 4].

In the context of online reviews, spam 2.0 includes posting fake reviews by mimicking authentic ones. Review websites are unfortunately peppered with fake reviews. For example, hotel managers have been caught asking employees to post fake positive reviews to maliciously boost the ratings of their properties [5]. Businesses often offer discounts in exchange of fake positive reviews, which could be posted by users without necessarily harboring malicious intentions [6]. Ironically, posting fake positive reviews continuously for about 45 days on Amazon.com elevates a product to the numero uno bestseller rank in its category [7].

To aggravate the problem, the polarity of fake reviews is not always positive. Reviews of negative and moderate polarities could also be fake. Posting negative fake reviews is known to be encouraged by managers as a way to slander competitors [8]. Fake reviews could also express a moderate tone—neither too glowing nor too critical—to make a conscious effort to sound realistic [5, 9].

The undeniable existence of such foul play has expectedly prompted several scholars to look into ways to automatically distinguish between authentic and fake reviews [1, 10]. After all, given that fake reviews are written to resemble authentic ones, differences between the two are inconspicuous to the naked eye. This leaves users clutching at straws to discern review authenticity. When they fail to do so, their perceptions of products and services stand a good chance to be distorted. In consequence, they run the risk of being deceived in making purchase decisions.

In this area of research, constructing ground truth has been recognized as a tricky problem for scholars to tackle [11, 12]. Ground truth here refers to a dataset of online reviews whose authenticity is known a priori with certainty and confidence. When what is authentic and what is fake is not known beforehand, distinguishing between the two conceivably becomes infeasible.

Therefore, the goal of this paper is three-fold: (1) To review existing approaches of developing ground truth, (2) To present an improved methodological template to construct ground truth, and (3) To conduct a quality-check of the newly constructed ground truth. Overall, the paper contributes to the scholarly debate in the growing field of information credibility in social data by dissecting the pitfalls in existing approaches to develop ground truth, and by forwarding a new methodological template. This can serve as a springboard for further research related to spam 2.0.

II. Related Works and Existing Approaches

When it comes to research on authentic and fake online reviews, ground truth is indispensable. However, the difficulty in developing ground truth is heralded as a key problem in this scholarly terrain [11, 12].

To deal with the problem, recent studies have used four major approaches. These include manual annotation, heuristic annotation, crowdsourcing and automation. However, by dissecting these approaches, this paper identifies each of them to suffer from peculiar shortcomings, which can potentially hinder a fair and a meaningful investigation of differences between authentic and fake reviews.

First, manual annotation involves using trained human annotators to label reviews as authentic or fake. For example, [13] trained a few college students to identify review opinion spam. Thereafter, the trained students were asked to annotate reviews as either authentic or fake. Such an approach might appear intuitive. However, the accuracy of the annotations could not be verified because of the lack of access to the contributors of the entries. In other words, there was no way to guarantee that a review annotated by the trained human annotators as authentic (or fake) was really so.

Second, heuristic annotation makes use of some rules of thumb to create ground truth of authentic and fake reviews. For example, [10] relied on the heuristic of text similarity. Unique reviews were treated as authentic, and those that were textually similar to one another were deemed to be fake. However, the validity of such a heuristic could be called into question. For one, unique reviews could be fake because spammers need not always make blatant copies of existing entries. Conversely, textually similar reviews could be authentic. After all, high text similarity between two reviews might be coincidental. In fact, authentic reviews could be occasionally written by drawing ideas from existing entries, thereby resulting in inadvertent text similarity.

Third, crowdsourcing makes use of the online community to create fake reviews, while authentic reviews are collected from the Internet. For example, [1] relied on users of the crowdsourcing platform Amazon Mechanical Turk to collect fake reviews. Again, in [14], users of the same platform were asked to paraphrase authentic reviews to create fake entries. However, in both [1] and [14], authentic reviews were ironically drawn from TripAdvisor.com, an unauthenticated website. It is trivial for anyone to create accounts on the platform to submit fake reviews. Yet, the exact proportion of bogus entries is not possible to estimate. Moreover, no efforts were invested to control for the background of those who contributed authentic reviews against the profile of those who submitted fake entries. Hence, the chance of a systematic bias in the ground truth is not possible to rule out completely.

Fourth, automation involves creating fake reviews using algorithm-generated texts. For example, [15] developed a review synthesizer to create sentences in fake reviews by drawing on those in authentic entries. In a similar vein, [16] used deep neural networks to develop fake reviews by reconstructing input paragraphs from authentic entries. However, such automated approaches are known to result in awkward linguistic phrases and expressions that are not always common in skilfully-crafted fake reviews [14]. Thus, fake reviews created using automation turn out to be too heavily engineered, thus taking a toll on ecological validity.

Overall, it appears that the literature currently lacks a systematic and rigorous approach to create ground of authentic and fake reviews. In the previous works, reviews that were deemed as authentic were not really authentic—at best, they were less likely to be fake. Conversely, reviews that were treated as fake were not fake per se, but were only less likely to be authentic. This in turn seems to dwarf the relevance of prior works that had investigated differences between authentic and fake reviews.

III. Proposed Methodology to Construct GroundTruth

This paper constructed ground truth of authentic and fake online reviews specifically in the context of hotels. For this purpose, 15 hotels uniformly straddling across five popular Asian tourist destinations—Singapore, Hong Kong, Tokyo, Bangkok, and Kuala Lumpur—were identified. The hotels were selected because they were found to attract huge volumes of reviews on websites such as Expedia.com, Hotels.com and Agoda.com.

The ground truth comprised 1,800 reviews altogether. In particular, it included two corpora: one containing 900 authentic reviews, and the other comprising 900 fake reviews. The dataset size of 1,800 reviews was deemed appropriate. It was larger than that used in related works such as [17], [18] and [19]. These used datasets containing some 80, 800, and 160 reviews respectively.

Both authentic and fake reviews were evenly distributed across the 15 identified hotels. Specifically, the corpus of authentic reviews, and that of fake reviews contained 60 entries per hotel (20 positive + 20 negative + 20 moderate). The uniform spread across the three polarities—positive, negative as well as moderate—in the corpora of authentic and fake reviews ensured a well-balanced dataset.

As an improvement to the existing approaches, authentic reviews were obtained from authenticated websites that allow submission of entries only after valid bookings and stays in a given hotel. On the other hand, fake reviews were solicited from participants who had no experience of staying in the hotel. Thus, in the proposed approach, there was no need for any manual annotation as in [13], heuristic annotation as in [10], or automation as in [15].

Nonetheless, one of the problems that supposedly impedes the crowdsourcing approach still pertained: Reviewers who contributed authentic reviews could differ from participants who wrote fake entries [1]. For this reason, care was taken to ensure comparable background between reviewers who contributed authentic reviews, and participants who wrote fake entries as much as possible along five dimensions. These included individuals' country of origin, age, educational profile, travel experiences, and use of review websites.

Reviewers' country of origin could be readily obtained for authentic reviews. Entries for which reviewers had not disclosed their country of origin were not admitted into the corpus of authentic reviews. Reviewers were grouped into four geographical regions that included America, Asia-Pacific, Europe as well as Middle-East and Africa. Such a geographical grouping is widely used in research (e.g., [20, 21]), and in practice by the United Nations World Tourism Organization [22], and the World Economic Forum [23]. Thereafter,

participants were recruited to write fake reviews by keeping in mind the proportions of authentic entries from each of these regions. The target was to have comparable proportions of reviews from the regions in the corpus of authentic reviews as well as in that of fake entries—an approximation of matched sampling. This could help afford a fair comparison between the two corpora.

In terms of age, most reviews are written by young individuals aged 45 years or below [24, 25, 26]. Hence, fake reviews were solicited from participants whose age ranged from 21 to 45 years.

In terms of educational profile, most reviews are written by educated individuals, especially those who have completed secondary/high school [24, 25, 26]. Hence, fake reviews were solicited from participants who were minimally undergraduate students in terms of their educational profiles. In particular, participants included undergraduate students, graduate students, and working adults who minimally had undergraduate degrees. Put differently, all of them had completed secondary/high school education.

In terms of travel experiences, it is conceivable that reviewers writing authentic reviews experienced travelling. That was why they were given access to write reviews in authenticated review websites in the first place [19, 27, 28]. Hence, fake reviews were solicited from participants who had travel experiences in the previous year.

Finally, reviewers are likely to be well-versed with the use of review websites. Hence, fake reviews were solicited from participants who were regular readers or contributors on review websites.

A. *Corpus of Authentic Reviews*

Authentic reviews were collected from Expedia.com, Hotels.com and Agoda.com in evaluation of the 15 identified hotels. Drawing data from multiple authenticated websites—submission of reviews possible only after valid bookings and stays—enhanced representativeness. For each hotel, 60 entries (20 positive + 20 negative + 20 moderate) were collected to yield 900 authentic reviews altogether (60 reviews x 15 hotels). Given that all the hotels had attracted huge volumes of authentic reviews, there were ample submissions to collect 20 entries of each of the three review polarities.

Reviews were admitted into the corpus based on five inclusion criteria. First, they had to be posted as recently as possible. This was necessary to ensure a fair comparison between authentic and fake reviews. Second, the content of reviews must have been written in English. Third, descriptions of reviews had to be at least more than 150 characters in length. Those shorter than 150 characters offer little room for a meaningful analysis [1, 18]. Fourth, reviews must be meaningful. It is possible for reviewers to write unmistakably irrelevant reviews. Such entries would not have facilitated constructing a sound ground truth. Hence, meaningfulness of all reviews were inspected manually to ensure that they contained evaluation of hotels without any nonsensical texts created through random keystrokes. Fifth, reviews had to be accompanied by metadata about the respective reviewers such as country of origin. This provided the basis to solicit comparable number of fake reviews against the volume of authentic entries posted by reviewers across the four regions, namely, America, Asia-Pacific, Europe as well as Middle-East and Africa.

The polarities of reviews—positive, negative and moderate—were determined based on their ratings [29, 30]. Specifically, Expedia.com and Hotels.com require users to rate hotels on a five-point scale. Hence, one- or two-star reviews were treated as negative, three-star reviews were taken as moderate, and four- or five-star reviews were deemed as positive [31]. However, Agoda.com requires users to rate hotels on a 10-point scale. Scales that differ from one another in terms of ranges cannot be linearly interpolated [32, 33]. In other words, a score of one on a five-point scale is not necessarily equivalent to that of two on a 10-point scale. Rating scales with more options generally result in higher scores [34]. Therefore, to make ratings from Agoda.com comparable with those of Expedia.com and Hotels.com, the rescaling approach proposed in [32] was followed. The scheme to assign review polarities is depicted in Table I.

TABLE I. SCHEME TO ASSIGN REVIEW POLARITIES

Website	Scale	Positive	Negative	Moderate
Expedia.com	1-5	4 or 5	1 or 2	3
Hotels.com	1-5	4 or 5	1 or 2	3
Agoda.com	1-10	7.75 or above	5.50 or below	5.51 to 7.74

Collecting 60 authentic reviews for a given hotel involved three steps (Fig. 1). First, reviews for the hotel available in Expedia.com, Hotels.com and Agoda.com were sorted based on their date of posting, most recent entries being at the top. Second, all reviews across the three websites were inspected chronologically to check if they met the inclusion criteria. Third, when a review met all the criteria, it was admitted into the corpus, and its polarity was noted. If any of the criteria were violated, the review was ignored. The last two steps were repeated until the corpus grew to contain 20 positive, 20 negative, and 20 moderate reviews for the hotel.

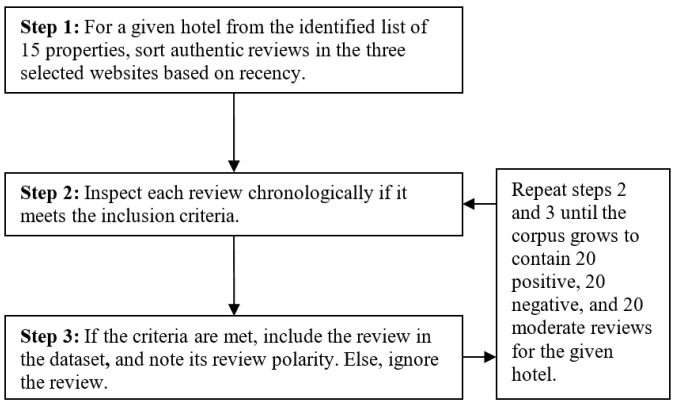

Fig. 1. Collecting authentic reviews for a given hotel.

These steps were iterated for all the 15 identified hotels to yield the corpus of 900 authentic reviews (300 positive + 300 negative + 300 moderate). In particular, 799 reviews were collected from Agoda.com, 45 from Expedia.com, and 56 from Hotels.com. The reviews were contributed by reviewers from more than some 50 countries of origin. In particular, 71 of them were contributed by reviewers from America, 730 by those from Asia Pacific, 88 by those from Europe, and 11 by those from Middle East and Africa. The reviews were posted during the period ranging from April, 2009 to March, 2013. Specifically, four reviews were posted in 2009, 14 in 2010, 24 in 2011, 320 in 2012, and 538 in 2013.

Finally, the validity of the review polarities was verified. For this purpose, three research assistants were recruited. They had graduate degrees in Computer Science or Information Systems with more than one year of professional experience. Moreover, all of them regularly read or contributed reviews, and had travel experience in the previous year.

Each of the three research assistants was randomly assigned one-third of all reviews. They were shown the reviews without their ratings. Hotel names were concealed to avoid biases. They were asked to annotate reviews as either positive, negative or moderate. No strict deadlines were given to prevent fatigue-induced coding errors. Cohen's kappa for the agreement between the polarities annotated by the research assistants, and those inferred from the ratings indicated a non-chance level of agreement: $\kappa = 0.84$ [35].

B. Corpus of Fake Reviews

Fake reviews were solicited from participants via email. For each of the 15 hotels, 60 entries (20 positive + 20 negative + 20 moderate) were obtained to yield 900 fake reviews altogether (60 reviews x 15 hotels).

Participants were instructed to imagine as if they were working for the marketing department of a hotel. Their boss had asked them to write at most six realistic fake reviews in English. Each review had to contain a description of at least 150 characters. These instructions were meant to aid participants in getting into the groove for writing fake reviews, regardless whether malicious intentions were triggered. Moreover, they were asked to submit entry for a hotel only if they had not stayed there earlier.

A pilot study was conducted with six participants who had graduate degrees in Information Systems, and were regular readers or contributors of reviews. Two of them were instructed to write one positive review each for a selection of six hotels. The other two were asked to write one negative review each for the same hotels. The remaining two were instructed to write one moderate review each for the same hotels.

The participants of the pilot study were requested to comment on two aspects: clarity of the instructions, and perceived difficulty in accomplishing the task. With respect to the former, they unanimously agreed that the instructions were clear. With respect to the latter, the consensus was that participants might take substantial amount of time to write realistic fake reviews.

Based on the participants' feedback, two revisions were made. First, the minimum length criterion of "150 characters" in the original instructions was changed to "30 words" in the revised version, assuming five characters per word on average [36]. This revision was necessary because one participant pointed that the criterion of "150 characters" could be easily misinterpreted as "150 words." After all, individuals engaged in writing tasks relate with words more readily than with characters.

Second, an additional line was inserted in the revised instructions to indicate the estimated time the task could take assuming eight minutes per review on average [1]. This revision was necessary because one participant indicated that one would "need more time to think and write [fake reviews]." Hence, this additional line serves to remind participants that writing fake reviews is time-consuming.

Additionally, no deadline was imposed to ensure the quality of fake reviews. If participants are given deadlines, they would write perfunctorily and spontaneously. Writing fake reviews in such a manner can result in greater cognitive load among participants, thereby providing more linguistic cues for detection [37, 38, 39, 40].

Collecting fake reviews involved four steps (Fig. 2). In the first step, invitation for voluntary participation in the study was disseminated using a combination of snowball sampling and maximum variation sampling. The former "identifies cases of interest from people who know people who know what cases are information rich" [41, p. 158], while the latter consists of "determining in advance some criteria that differentiate [the participants]" [41, pp. 156-157]. When differences are maximized before snowballing, better generalization is obtained. Although non-probabilistic, such a purposive sampling approach—selecting participants who serve a specific purpose consistent with the objective of the research [42]—is considered relevant in research on electronic word-of-mouth and online reviews [43, 44].

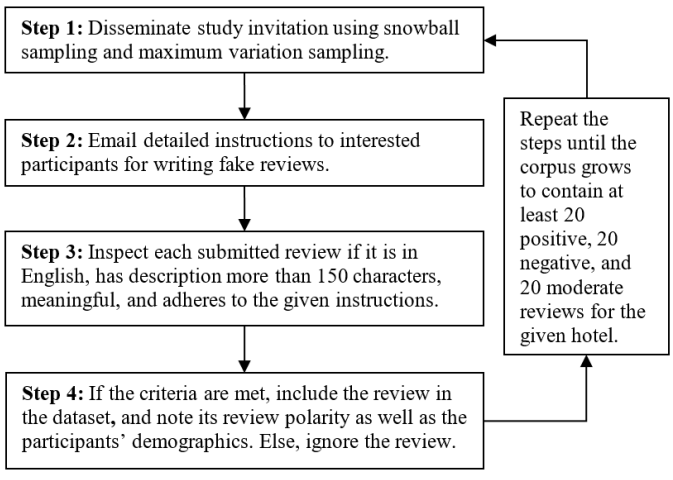

Fig. 2. Collecting fake reviews for a given hotel.

Based on the researcher's personal contacts, individuals who met the selection criteria of age, education, travel experience and review website familiarity were invited for voluntary participation. The invitation was disseminated via email, social networking sites and word-of-mouth. Using email, about 100 seed contacts could be identified. Using social networking sites, about another 150 seed contacts could be reached. More about 50 seed contacts could be contacted through word-of-mouth. All the seed contacts were asked to participate in the study. They were also requested to share the study invitation with as many contacts as possible.

As a part of maximal variation sampling, seed contacts were recruited keeping in mind the proportions of authentic reviews obtained from the four major geographical regions, namely, America, Asia-Pacific, Europe as well as Middle-East and Africa. This helped ensure that the proportions of contributors' country of origin in the corpus of fake reviews were comparable to those in the corpus of authentic entries.

If interested to participate, individuals were asked to send an email to the researcher for further instructions. Over 400 participants indicated interest. The response was generally better from seed contacts who were contacted via email or word-of-mouth vis-à-vis those who were reached via social networking sites.

In the second step of collecting fake reviews, interested participants were sent emails containing detailed instructions. Each interested participant could choose to write at most six reviews for six different hotels. This ensured that the corpus of fake reviews was not skewed with disproportionately more entries from any single participant.

The participants were randomly assigned to hotels and review polarities. However, they were not tasked to write multiple reviews with different polarities to alleviate sudden context switches during the writing task.

After the instructions were sent to participants, no deadline was imposed to return their fake reviews. Nonetheless, call-backs and email follow-ups were used at about weekly intervals up to a maximum of four times as reminders to increase the likelihood of obtaining responses. Not more than four reminders were given because pushing participants excessively might have enticed them to write fake reviews perfunctorily. In that case, their entries could have provided more linguistic cues for detection [37, 38, 39, 40], thereby hindering the construction of a high quality ground truth.

In the third step of collecting fake reviews, entries submitted by participants were manually inspected if they were in English, had descriptions more than 150 characters, were meaningful, and adhered to the given instructions. All submissions were found to be meaningful, and written in English. However, 46 reviews were eliminated as they contained descriptions shorter than 150 characters.

In the fourth step of collecting fake reviews, when a review met all the criteria, it was admitted into the corpus, and its polarity was noted. The demographic information of the respective participant was also archived. If any of the criteria were violated, the review was ignored. The growth of the corpus was closely monitored to ensure that the incoming fake reviews were evenly spread across the 15 hotels.

These steps were iterated until the corpus grew to contain at least 20 positive, 20 negative, and 20 moderate reviews for each hotel. Eventually, 909 fake reviews were obtained from 287 participants. There were two positive and one negative surplus reviews for a hotel in Singapore, two surplus positive reviews for a hotel in Hong Kong, and four surplus positive reviews for a hotel in Tokyo. Nine reviews for these hotels and review polarities were randomly eliminated.

The final corpus of fake reviews contained 900 entries (300 positive + 300 negative + 300 moderate) obtained from 284 participants (aged 21-25 years: 88, aged 26-35 years: 146, aged 36-45 years: 50; educational background: minimally undergraduate students; gender: 128 females, 156 males). The reviews were contributed by participants across some 33 countries of origin. In particular, 69 reviews were contributed by 29 participants from America (13 males, 16 females), 732 by 213 participants from Asia-Pacific (121 males, 92 females), 85 by 37 participants from Europe (20 males, 17 females), and 14 by 5 participants from Middle-East and Africa (2 males, 3 females).

Finally, the validity of the review polarities was verified. The procedure was the same as that employed for verifying the polarity of authentic reviews described earlier. Cohen's kappa for the agreement between the polarities annotated by the research assistants, and those provided by the participants who wrote the fake reviews indicated a non-chance level of agreement: $\kappa = 0.98$ [35].

IV. Quality Check of the New GroundTruth

A. *Manual Annotation*

Initially, a manual quality check of the newly constructed ground truth was performed. For this purpose, the 1,800 reviews in the dataset were equally divided among the three research assistants for another round of annotation. Each of them received a random set of 600 reviews (300 authentic + 300 fake).

The research assistants were told that some reviews were authentic while others were fake. However, taking the cue from prior works [45], they were kept ignorant of the exact proportions. The research assistants were asked to predict if a review was authentic or fake—to the best of their abilities. They were not given any strict deadlines to prevent fatigue-induced prediction errors.

The research assistants were found to accurately identify 415 of the 900 authentic reviews (46.11%), and 469 of the 900 fake reviews (52.11%) in the dataset. Clearly, their performance resembled random guessing. This is consistent with the deception literature on human ability to distinguish between truth and fiction [37, 38, 39, 40]. The sub-par performance assures that the fake reviews in the dataset were not only well written but also difficult to be distinguished from the authentic ones. This inspires confidence in the overall quality of the current dataset.

B. ReviewSkeptic

ReviewSkeptic (http://reviewskeptic.com) is an online tool that allows Internet users to copy and paste any review, and test its authenticity. All reviews in the dataset of [1], one of the most highly cited ground truth datasets thus far, were tested. An accuracy of 100% was obtained. Clearly, the classification algorithm used in ReviewSkeptic is decent.

Therefore, the tool was used to test the newly constructed ground truth. All the 1,800 reviews in the dataset were tested one-by-one. A much lower accuracy of 64.61% was obtained. While 803 of the 900 authentic reviews were accurately identified, only 360 of the 900 fake entries were correctly flagged out by ReviewSkeptic.

The relatively lower accuracy suggests that linguistic nuances between authentic and fake reviews in the new ground truth were blurred to a greater extent compared with those in [1]. This in turn lends support to the proposed methodology employed for the development of ground truth. Table II provides examples of incorrectly identified authentic reviews and incorrectly identified fake reviews from the dataset.

TABLE II. REVIEWS INCORRECTLY IDENTIFIED BY REVIEWSKEPTIC

	Example
Authentic reviews identified as fake	Clean hotel with great prices. The room booking is superior double room, but ended up with a twin bed. was not satified with facilities which got toothbrush, did not have toothpaste. Luckily, did bring it along. Really affordable and mediocre budget hotel.
	everything needs to be paid in this hotel! check emails you want is 1 euro per minute! you want to relax at the pool again pay! Finally, large hotel complex without interest, small room.
	the staff was very friendly and provided excellent customer support. i was happy to be able to walk out of the hotel and easily find transportation via bus or taxi to all my destinations. i definitely want to stay here again and will recommend it to my friends and family.
Fake reviews identified as authentic	Hotel location is good and near to shopping complex! Staffs are very friendly and professional..Facilities are comprehensive and the whole hotel is very elegant!.. Definitely highly recommended if you visit Singapore!
	I am writing this review with so much frustration right after checking out of the USA hotel. The hotel management here seems to be in a frantic mindset of cost cutting, so the bed sheets are laid out without being washed after the previous guests leave, toilets lack tissue rolls, bath towels are stained. Never again to this hotel at any cost.
	Although located at a nice location, my stay in this hotel was not very comfortable. The heater is not working properly most of the time and that affect my health seriously. When I told the staffs to fix it, they said yes but never fix it in the end. Apart from that, I like the fact that the room is clean.

C. Brute-Force Method (Bigrams)

To further validate the quality of the new ground truth, a brute-force method was applied. In particular, [1] showed that a linear support vector machine (SVM)-based classifier using bigrams as features can help classify authentic and fake reviews with an accuracy of nearly 90%. Therefore, another set of experiments was conducted to examine if SVM with a linear kernel was able to classify reviews in the new ground truth with an accuracy of nearly 90%.

A five-fold cross-validation was employed. Given that the dataset comprised 1,800 reviews, a five-fold cross-validation meant splitting it into five subsets, each containing 360 entries (360 reviews x 5 folds = 1800). The five folds were distributed uniformly across the five tourist destinations as well as authenticity. Training was done on four folds containing 1,440 reviews (360 reviews x 4 folds = 1440), and tested on the remaining fold with 360 reviews (360 reviews x 1 fold = 360). This was iterated five times so that every fold was tested exactly once. Performance was assessed by taking the micro-average of the results from each of the k folds.

The accuracy however turned out to be much lower—only 72.61%. One could argue that the lower accuracy was because linear SVM was perhaps not suited for the dataset at hand.

To rule out the argument, the classification was further attempted using a wide variety of algorithms that included logistic regression, C4.5 decision tree, JRip, random forest, and SVM with linear, polynomial as well as radial basis function kernels. The results are shown in Table III.

Even then, the highest accuracy attained using bigrams was only 73.94% particularly when SVM with a radial basis function kernel was used. Overall, these results demonstrate the quality of the newly constructed ground truth. Irrespective of the choice of classification algorithm, bigrams could not classify authentic and fake reviews in the newly constructed ground truth with an accuracy of close to 90%.

TABLE III. ACCURACY USING DIFFERENT CLASSIFICATION ALGORITHMS

Algorithm (Bigrams as Features)	**Accuracy**
Logistic Regression	71.50%
C4.5 Decision Tree	57.61%
JRip	57.22%
Random Forest	64.17%
SVM (linear kernel)	72.61%
SVM (polynomial kernel)	72.78%
SVM (radial basis function kernel)	73.94%

V. DISCUSSION AND CONCLUSIONS

A. Summary of the Proposed Methodological Template

To summarize, this paper presents a three-step solution to construct ground truth of authentic and fake online reviews. The first step involves identifying appropriate data sources that facilitate collecting authentic reviews. After all, ensuring the validity of authenticity is crucial to allow for a meaningful investigation in this research theme. For this purpose, it is essential to rely on authenticated data sources such as Expedia.com, which allow reviews to be posted only after a monetary transaction. Amazon.com too uses the functionality

of 'verified purchase' to indicate if a given review had been posted after payment. Such reviews are generally assured to be authentic [19].

The second step involves collecting authentic reviews from the authenticated sources submitted by contributors who had disclosed maximal information about themselves. Having details about the profile of those who had contributed authentic reviews would be useful to maximize comparability with the background of those who submit fake entries. It should be acknowledged that websites do not always allow contributors to disclose their demographic information such as age, gender and educational profile. For this reason, the extant literature should be relied upon to gain insights into the background of those who are most likely to contribute reviews on the Internet.

The third step involves soliciting fake reviews from humans rather than relying on automation. The selection of participants should be judicious. Adequate efforts need to be invested to ensure comparability in the background of those who contribute authentic reviews against those who submit fake entries. This is necessary to facilitate a fair analysis of differences between authentic and fake reviews. Furthermore, participants should not be pressurized to create fake reviews within a tight deadline. If participants are overly pushed, it can take a toll on the data quality of the ground truth.

That having said, a caveat in this methodology needs to be highlighted. Since the proposed procedure involves labor-intensive human participation, generating a large pool of fake reviews is time-consuming.

B. Future Research Directions

By proposing the above three-step solution to construct ground truth of spam 2.0, this paper paves the way for more rigorous research on differences between authentic and fake reviews vis-à-vis those found in the extant literature. However, readers are cautioned that authentic and fake reviews should not be misinterpreted as truthful and deceitful entries respectively. After all, it is difficult—if not impossible—to ascertain truthfulness of individuals who write reviews.

In addition, this paper carves out two new research directions. First, if participants can be recruited to create fake reviews, it might be useful to interview them after their writing task. Such an endeavor might offer hitherto-unknown insights into the psychology of an opinion spammer. Criminology- and social psychology-related theories could be employed to illumine the findings. A particularly relevant theory is the routine activity theory, which suggests that opinion spammers are not inherently evil but simply have an opportunity to misbehave and deviate from ethical norms [46]. The literature is currently mum on how they make use of the opportunity to create convincing fake reviews.

Second, this paper recognizes that textual reviews could soon lose out to pictorial reviews or video reviews in terms of popularity [47]. Meanwhile, with the advent of sophisticated image-processing software such as Photoshop, it should not be too difficult for opinion spammers to create fake pictorial reviews or fake video reviews. Hence, future works could rely on the methodological template presented in this paper to construct a ground truth of authentic and fake reviews that include pictures and videos too. In this way, data science research on differences between authentic and fake reviews will no longer be restricted to linguistics but will be widened to also encompass image and video analytics.

In addition, the paper invites interested scholars to devise strategies to construct ground truth by combining the proposed methodological template with other novel methods that include hybrid crowd-sensing [48, 49] as well as websites' filtering results [50]. A careful and strategic combination of multiple methods could help alleviate the resource-intensiveness of the proposed approach. By taking a modest step toward gold standard dataset [51] and by identifying the aforementioned research directions, the hope of the paper is that scholars are always able to stay ahead in their relentless race against spammers. It is important that this hope materializes because only then can we expect to make the best use of online reviews.

ACKNOWLEDGMENT

The author thanks Dr Alton Chua for his supervision of the PhD dissertation from which this paper originates.

REFERENCES

[1] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, pp. 309-319, 2011.

[2] B. Whitworth, and E. Whitworth, "Spam and the social-technical gap," Computer, vol. 37, pp. 38-45, 2004.

[3] P. Hayati, V. Potdar, A. Talevski, N. Firoozeh, S. Sarenche, and E. A. Yeganeh, "Definition of spam 2.0: New spamming boom," Proceedings of the Conference on Digital Ecosystems and Technologies. New York, NY: IEEE, pp. 580-584, 2010.

[4] E. Blanzieri, and A. Bryl, "A survey of learning-based techniques of email spam filtering," Artificial Intelligence Review, vol. 29, pp. 63-92, 2008.

[5] E. Allen, "'Dear staff. We could do with some positive comments': Hotel boss is caught telling his workers to post fake reviews on TripAdvisor," MailOnline, 9 October 2012. Retrieved from http://www.dailymail.co.uk/news/article-2214974/Hotel-boss-caught-telling-workers-post-fake-reviews-TripAdvisor.html

[6] E. T. Anderson, and D. I. Simester, "Reviews without a purchase: Low ratings, loyal customers, and deception," Journal of Marketing Research, vol. 51, pp. 249-269, 2014.

[7] The Guardian, "Amazon withdraws ebook explaining how to manipulate its sales rankings," 5 January 2011. Retrieved from http://www.guardian.co.uk/books/2011/jan/05/amazon-ebook-manipulate-kindle-rankings

[8] D. Kerr, "Samsung probed for allegedly bashing rival HTC online," CNET News, 15 April 2013. Retrieved from http://news.cnet.com/8301-1035_3-57579749-94/samsung-probed-for-allegedly-bashing-rival-htc-online/

[9] S. Banerjee, and A. Chua, "Theorizing the textual differences between authentic and fictitious reviews: Validation across positive, negative and moderate polarities," Internet Research, vol. 27, pp. 321-337, 2017.

[10] N. Jindal, and B. Liu, "Opinion spam and analysis," Proceedings of the International Conference on Web search and Web Data Mining. New York, NY: ACM, pp. 219-230, 2008.

[11] S. Gokhman, J. Hancock, P. Prabhu, M. Ott, and C. Cardie, "In search of a gold standard in studies of deception," Proceedings of the Workshop on Computational Approaches to Deception Detection. Stroudsburg, PA: ACL, pp. 23-30, 2012.

[12] E. Fitzpatrick, and J. Bachenko, "Building a data collection for deception research," Proceedings of the Workshop on Computational Approaches to Deception Detection. Stroudsburg, PA: ACL, pp. 31-38, 2012.

[13] F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," Proceedings of the International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI, pp. 2488-2493, 2011.

[14] S. Kim, S. Lee, D. Park, and J. Kang, "Constructing and evaluating a novel crowdsourcing-based paraphrased opinion spam dataset," Proceedings of the International Conference on World Wide Web. New York, NY: ACM, pp. 827-836, 2017.

[15] H. Sun, A. Morales, and X. Yan, "Synthetic review spamming and defense," Proceedings of the International Conference on Knowledge Discovery and Data Mining. New York, NY: ACM, pp. 1088-1096, 2013.

[16] J. Li, M. T. Luong, and D. Jurafsky, "A hierarchical neural autoencoder for paragraphs and documents," arXiv preprint arXiv:1506.01057, 2015.

[17] K. H. Yoo, and U. Gretzel, "Comparison of deceptive and truthful travel reviews," in Information and communication technologies in tourism, W. Höpken, U. Gretzel, and R. Law, Eds. Vienna, Austria: Springer-Verlag, pp. 37-47, 2009.

[18] C. G. Harris, "Detecting deceptive opinion spam using human computation," Proceedings for the Workshops on Artificial Intelligence. Palo Alto, CA: AAAI, pp. 87-93, 2012.

[19] T. Ong, M. Mannino, and D. Gregg, "Linguistic characteristics of shill reviews," Electronic Commerce Research and Applications, vol. 13, pp. 69-78, 2014.

[20] J. Aramberri, "The future of tourism and globalization: Some critical remarks," Futures, vol. 41, pp. 367-376, 2009.

[21] B. Zeng, and R. Gerritsen, "What do we know about social media in tourism? A review," Tourism Management Perspectives, vol. 10, pp. 27-36, 2014.

[22] UNWTO, "International tourism to continue robust growth in 2013," United Nations World Tourism Organization 2013. Retrieved from http://media.unwto.org/en/press-release/2013-01-28/international-tourism-continue-robust-growth-2013

[23] J. Blanke, and T. Chiesa, "The travel & tourism competitiveness index 2011: Assessing industry drivers in the wake of the crisis," in The travel & tourism competitiveness report 2011: Beyond the downturn, J. Blanke, and T. Chiesa, Eds. Geneva, Switzerland: World Economic Forum, pp. 3-34, 2011.

[24] U. Gretzel, K. H. Yoo, and M. Purifoy, "Online travel reviews study: Role and impact of online travel reviews," A&M University, TX: Laboratory for Intelligent Systems in Tourism, 2007.

[25] C. Ip, H. A. Lee, and R. Law, "Profiling the users of travel websites for planning and online experience sharing," Journal of Hospitality & Tourism Research, vol. 36, pp. 418-426, 2012.

[26] B. T. Ratchford, M. S. Lee, and D. Talukdar, "The impact of the Internet on information search for automobiles," Journal of Marketing Research, vol. 40, pp. 193-209, 2003.

[27] E. Bjering, L. J. Havro, and Ø. Moen, "An empirical investigation of self-selection bias and factors influencing review helpfulness," International Journal of Business and Management, vol. 10, pp. 16-30, 2015.

[28] 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, vol. 2, pp. 1-24, 2015.

[29] J. Gerdes, B. B. Stringam, and R. G. Brookshire, "An integrative approach to assess qualitative and quantitative consumer feedback," Electronic Commerce Research, vol. 8, pp. 217-234, 2008.

[30] M. P. O'Mahony, and B. Smyth, "Learning to recommend helpful hotel reviews," Proceedings of the Conference on Recommender systems. New York, NY: ACM, pp. 305-308, 2009.

[31] Z. Chen, and N. H. Lurie, "Temporal contiguity and negativity bias in the impact of online word of mouth," Journal of Marketing Research, vol. 50, pp. 463-476, 2013.

[32] J. Dawes, "Five point vs eleven point scales: Does it make a difference to data characteristics?," Australasian Journal of Market Research, vol. 10, pp. 39-47, 2002.

[33] S. M. Johnson, P. Smith, and S. Tucker, "Response format of the job descriptive index: Assessment of reliability and validity by the multitrait-multimethod matrix," Journal of Applied Psychology, vol. 67, pp. 500-505, 1982.

[34] E. E. Ghiselli, "All or none versus graded response questionnaires," Journal of Applied Psychology, vol. 23, pp. 405-415, 1939.

[35] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, pp. 37-46, 1960.

[36] I. S. MacKenzie, and S. X. Zhang, "The design and evaluation of a high-performance soft keyboard," Proceedings of the Conference on Human Factors in Computing Systems. New York, NY: ACM, pp. 25-31, 1999.

[37] C. F. Bond, and B. M. DePaulo, "Accuracy of deception judgments," Personality and Social Psychology Review, vol. 10, pp. 214-234, 2006.

[38] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," Psychological Bulletin, vol. 129, pp. 74-118, 2003.

[39] S. Porter, and L. ten Brinke, "Reading between the lies identifying concealed and falsified emotions in universal facial expressions," Psychological Science, vol. 19, pp. 508-514, 2008.

[40] S. Porter, L. ten Brinke, and B. Wallace, "Secrets and lies: Involuntary leakage in deceptive facial expressions as a function of emotional intensity," Journal of Nonverbal Behavior, vol. 36, pp. 23-37, 2012.

[41] J. W. Creswell, Qualitative Inquiry and Research Design: Choosing among Five Approaches. California, CA: Sage, 2012.

[42] D. S. Collingridge, and E. E. Gantt, "The quality of qualitative research," American Journal of Medical Quality, vol. 23, pp. 389-395, 2008.

[43] H. J. Jeong, and D. M. Koo, "Combined effects of valence and attributes of e-WOM on consumer judgment for message and product: The moderating effect of brand community type," Internet Research, vol. 25, pp. 2-29, 2015.

[44] E. Keller, and J. Berry, The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and What to Buy. New York, NY: Simon & Schuster, 2003.

[45] N. Du, H. Huang, and L. I. Li, "Can online trading survive bad-mouthing? An experimental investigation," Decision Support Systems, vol. 56, pp. 419-426, 2013.

[46] L. E. Cohen, and M. Felson, "Social change and crime rate trends: A routine activity approach," American Sociological Review, vol. 44, pp. 588-608, 1979.

[47] P. Xu, L. Chen, and R. Santhanam, "Will video be the next generation of e-commerce product reviews? Presentation format and the role of product type," Decision Support Systems, vol. 73, pp. 85-96, 2015.

[48] M. Avvenuti, S. Bellomo, S. Cresci, M. N. La Polla, and M. Tesconi, "Hybrid crowdsensing: A novel paradigm to combine the strengths of opportunistic and participatory crowdsensing," Proceedings of the International Conference on World Wide Web Companion. New York, NY: ACM, pp. 1413-1421, 2017.

[49] S. Cresci, A. Cimino, M. Avvenuti, M. Tesconi, and F. Dell'Orletta, "Real-world witness detection in social media via hybrid crowdsensing," Proceedings of the International Conference on Web and Social Media. Palo Alto, CA: AAAI, pp. 576-579, 2018.

[50] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, "What Yelp fake review filter might be doing?," Proceedings of the International Conference on Web and Social Media. Palo Alto, CA: AAAI, pp. 409-418, 2013.

[51] M. Viviani, and G. Pasi, "Credibility in social media: Opinions, news, and health information—a survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 7, e1209, pp. 1-25, 2017.