Comparing Human Computation, Machine, and Hybrid Methods for Detecting Hotel Review Spam

Christopher G. Harris(✉)

School of Mathematical Sciences, University of Northern Colorado, Greeley, CO 80639, USA
christopher.harris@unco.edu

Abstract. Most adults in industrialized countries now routinely check online reviews before selecting a product or service such as lodging. This reliance on online reviews can entice some hotel managers to pay for fraudulent reviews – either to boost their own property or to disparage their competitors. The detection of fraudulent reviews has been addressed by humans and by machine learning approaches yet remains a challenge. We conduct an empirical study in which we create fake reviews, merge them with verified reviews and then employ four methods (Naïve Bayes, SVMs, human computation and hybrid human-machine approaches) to discriminate the genuine reviews from the false ones. We find that overall a hybrid human-machine method works better than either human or machine-based methods for detecting fraud – provided the most salient features are chosen. Our process has implications for fraud detection across numerous domains, such as financial statements, insurance claims, and reporting clinical trials.

Keywords: Crowdsourcing · Human computation · Word of mouth · Web 2.0 · Machine learning · TripAdvisor · Review spam

1 Introduction

Consumers today have a vast amount of information at their fingertips when making a purchase decision. Despite the availability of a variety of resources, customers place a significant emphasis on the advice and recommendations of their peers; 4 of every 5 adults in the U.S. adults indicate they use online customer reviews before purchasing an item, with half of these (2 in 5) indicating they nearly always do [1]. Other industrialized nations also rely heavily on peer-generated online reviews (also called electronic word of mouth, or eWOM) before purchases [2–5]. This translates into a competitive advantage for retailers and service providers that maintain higher ratings and better reviews than their competitors; indeed, a one-star increase in a restaurant's Yelp review score translates into a 5 to 9 percent increase in revenue [6].

These high stakes create opportunity; some unscrupulous retailers have recognized an advantage to boost their own business or disparage their competitors, creating a market for generating fraudulent reviews. As many as a third of online reviews may be

Published by Springer Nature Switzerland AG 2019
I. O. Pappas et al. (Eds.): I3E 2019, LNCS 11701, pp. 75–86, 2019.
https://doi.org/10.1007/978-3-030-29374-1_7

fraudulent [7, 8], with an estimate of 16% for Yelp [9] and a similar percentage estimated for unverified hotel review websites such as TripAdvisor [10].

In this paper, we focus on evaluating fraud in lodging reviews (also called *opinion spam* or *review spam*) on websites with unverified reviews. As with restaurant reviews, hotel reviews represent a complex mix of a product-related and a service-related good. Some websites contain only verified reviews; for example, Priceline and Booking only allow customers that purchased lodging through their website to contribute a review within a specified period (typically 28 days after the stay). Others, such as TripAdvisor, do not verify identities or stays. However, TripAdvisor branded sites make up the largest travel community in the world, reaching 350 million unique monthly visitors, with more than 570 million reviews and opinions covering more than 7.3 million accommodations, airlines, attractions, and restaurants [11].

A variety of methods have been employed in review spam detection. TripAdvisor claims to use a machine approach with 50 filters in its vetting process [12], but several recent, high-profile review spamming campaigns have demonstrated that their approach is not infallible. Humans are well-established judges in online fraud detection (e.g., [13, 14]), although they are considered poor at spotting deception [15]. Can a hybrid human-machine interface can outperform either of these models? We address this question in this paper.

The remainder of this article is organized as follows. In Sect. 2, we discuss related work in review spam detection. We describe our experiment methodology in Sect. 3, results and analysis in Sect. 4. We conclude and describe future research in Sect. 5.

2 Related Work

Efforts to detect fraudulent advertising claims has existed for centuries, with humans serving as the primary arbiters. The juries of many court systems worldwide are designed around this paradigm. In 2006 crowdsourcing gained prominence as a mechanism to perform small focused tasks in which humans outperformed machines; detecting fraudulent or misleading information using crowdworkers appeared to be a natural extension. Few studies to date, however, have used crowdworkers to detect online review spam (e.g. [16, 17]). Review spam detection provides an unusual scenario in the assessment of human-created data, since machine-based methods have been shown to outperform human judges. Review spam is created with the specific intent of misleading customers and is therefore difficult for humans to detect [18].

With the advent of natural language processing (NLP), machine-based techniques have been the primary focus in detecting review spam. These techniques can be divided into three basic forms: supervised learning, unsupervised learning, and semi-supervised learning. A comprehensive review of the various machine learning techniques applied to review spam can be found in [19].

Supervised learning is a popular technique in which the machine uses labeled training data to learn the class label (i.e., either “fake” or “genuine” review). Primarily using three types of learners – Logistic Regression (LR), Naïve Bayes (NB) and Support Vector Machine (SVM) – they make use of linguistic features in the review title and text, such as parts of speech (POS), Linguistic Inquiry and Word Count

(LIWC), and sentiment polarity. Ott et al. conducted a study of deceptive opinion spam limiting their scope to n-gram based features and achieved an accuracy with an SVM of 88% using unigram and bigram term frequency features for reviews on 1- and 2-star hotels [20] and 89% for bigrams for reviews on 4- and 5-star rated hotels [16]. Mukherjee et al. was only able to achieve an accuracy of 68% on Yelp data using the same approach [21]. Human judges were not able to outperform these classifiers on these same datasets, with the best judge achieving an accuracy of only 65%.

Unsupervised learning occurs when learning is from a set of unlabeled data and is often represented as clustering. It involves finding unseen relationships in the data that are not dependent on the class label. Few researchers to date have applied an unsupervised approach; Lau et al. achieved a true positive rate of 95% using an unsupervised probabilistic language model to detect overlapping semantic content among untruthful reviews on an Amazon review dataset [22], but their methods depend on having a large sample of fake reviews from which to build a language model.

Semi-supervised learning is a hybrid approach, in which learning occurs from both labeled and unlabeled data. It makes use of very little labeled data and a large amount of unlabeled data to determine the class label. This is ideal for online review spam because most data are unlabeled – in other words, there is rarely an oracle to tell if a review is genuine or fake. Although little research has applied the use of semi-supervised learning for review spam detection, results may yield better performance than supervised learning while reducing the need to generate large labeled datasets. To date, the best performer on review spam has been Li et al., who used a co-training algorithm and a two-view semi-supervised method to learn from a few positive examples and a set of unlabeled data [23]. They obtain a precision of 0.517, recall of 0.669 and an F-score of 0.583.

Little research to date in review spam has examined hybrid methods, in which the output of machine learning methods is then evaluated by humans before a final decision is made. Harris looked at bodybuilding supplement reviews in [16], first by examining the linguistic qualities identified by Yoo and Gretzel in [24] and then asking human evaluators to identify fake reviews. He found that human evaluators significantly improved their decision making by comparing each review against the dataset's linguistic features. In this study, we take a comparable approach – provide human evaluators with the linguistic qualities of the dataset, the machine recommendation, and then asking the evaluators to classify the data as either a genuine or fake review.

3 Detecting Hotel Review Spam

We seek to compare three different methods of identifying review spam – by non-expert human evaluation, by applying machine learning techniques, and by using a hybrid approach. We begin by constructing the dataset, describing the metrics, and then discussing the various methods and features from which review spam is assessed.

3.1 Dataset Construction

We wish to create a dataset containing a mix of genuine and fake reviews that appear to be drawn from TripAdvisor. We construct the dataset by selecting hotels on TripAdvisor from three markets: New York, London, and Hong Kong. We select these three markets as they have many international visitors which helps minimize cultural differences in language usage.

We create two pools of hotels in each market: those with high TripAdvisor ratings (a rating of four- and five- stars on TripAdvisor) and those with low TripAdvisor ratings (one- and two-star ratings). We filter out reviews in languages other than English and hotels that do not also appear on Booking.com. We eliminate those properties that have fewer than 300 Booking.com reviews.

From each of our 3 markets, we select five properties from the low-rated property pool and five from the high-rated property pool, comprising 30 properties in total. We randomly sample 90 Booking.com reviews from each property. The distribution of reviewer ratings from high-rated and low-rated properties differ as do the ratings in our samples. Booking.com verifies the reviewers have stayed at the property, therefore we assume that these reviews are genuine.

TripAdvisor scores hotels on a scale of 1 to 5 while Booking.com scores hotels in the range from 2.5 to 10; however, according to [25] a linear transformation can be made between the two. We transform the Booking.com score to a TripAdvisor score, rounded to the nearest half-star.

Using a conservative 10% estimate of fake reviews on travel websites mentioned in [12] as a guide, we then asked three non-experts to create 10 fake reviews for each of the 30 properties: 5 four- and five-star reviews *(boosting spam*) and 5 one- and two-star reviews (*vandalism spam*), which represent fake reviews used to either boost a given property or disparage a competing property, respectively. None of our fake review writers have stayed at any of the properties but are permitted to perform searches. They are asked to make the review "as convincing as possible" with respect to the type of review being asked (either high or low rating) and are asked to pay careful attention to the language used in all reviews for that property on the internet. For each property, the 10 fake reviews are then comingled with the 90 genuine reviews.

3.2 Metrics

We calculate accuracy, which is the number of correctly classified reviews divided by all reviews evaluated. We also calculate the precision and recall and the corresponding F-score. These are reported separately for the 15 high-rated hotels and the 15 low-rated hotels. We separately examine these metrics for boosting spam and vandalism spam.

3.3 Feature Extraction and Engineering

Identifying the correct features is essential for the review spam identification task. We apply the output obtained from the LIWC software [26] to derive a classifier, similar to the approach made by Ott et al. [17]. We constructed features for each of the 80 LIWC dimensions, which fall into four categories: linguistic (the average number of words per

sentence, the rate of misspelling, use of exclamation marks, etc.), psychological (the use of language representing social, emotional, cognitive, perceptual and biological processes, as well as temporal and/or spatially-related terms), personal (references to work, leisure, money, religion, etc.) and construction (filler, connection, and agreement words). Additional details about LIWC and the LIWC categories are available at http://liwc.net.

In addition to LIWC, we also examine POS, term frequency and use bigram feature sets, with their corresponding language models, since bigrams performed best in a comparison made in [17]. We apply the Kneser-Ney smoothing method to provide absolute-discounting interpolation of n-gram terms [27].

3.4 Machine Approach

We use a supervised learning approach for our machine learning task, since we have the labels for all reviews. Using this dataset, we design a fully-supervised method using various features in the language. We use both Naïve Bayes (NB) and SVMs as our supervised methods.

NB assumes the features are conditionally independent given the review's category. Despite its inherent simplicity and the lack of applicability of the conditional independence assumption to the real world, NB-based categorization models still tend to perform surprisingly well [28].

$$P_{NB}(c|d) = \frac{P(c)\prod_{i=0}^{m} P(f_i|c)}{P(d)} \tag{1}$$

We use the Natural Language Toolkit (NLTK) [29] to estimate individual language models, for truthful and deceptive opinions.

SVMs [30] can make use of certain kernels to transform the problem to allow linear classification techniques to be applied to non-linear data. Applying the kernel equations arranges the data instances within the multi-dimensional space so that there is a hyperplane that separates the data into separate classes. We restrict our evaluation to linear kernels since these performed best in preliminary evaluations using our features. We use Scikit-learn [31] to train our linear SVM models on the POS, LIWC, and bigram feature sets. We normalize each LIWC and bigram feature to unit length before combining them.

To ensure all hotel reviews are learned using the same language model, we evaluated using a 5-fold nested cross validation (CV) procedure [32]. Each fold contains all reviews (boosting and vandalism, genuine and fake) from 12 hotels; thus, our model applies its learning on reviews from the remaining 3 hotels. This avoids some pitfalls of learning from an incomplete set of data, as is described in [33].

3.5 Human Computation Based Approach

For each of our 30 hotels, we randomly allocated our 100 reviews into 4 batches of 25 reviews. We hired 360 human assessors from Amazon Mechanical Turk (MTurk) to examine reviews from each batch of 25 presented in random order. To allow us to

assess inter-annotator agreement, each hotel review was examined by three separate assessors. We created a simple web-based interface that displayed the title and text for each review, along with a prompt for the assessor to determine if the displayed review is genuine or fake. Assessors were paid $0.50 per batch; to provide an incentive for careful assessment, they were told that if they correctly classified all 25 hotels, they would be compensated an additional $0.50. We take the majority label for the 3 assessments for each review.

3.6 Hybrid Approach

To create a hybrid evaluation, we provide human assessors with the information ascertained by the machine approach. Along with each review, we provide the LIWC output for each feature for each review, the average LIWC output for all reviews for the 100 reviews, as well as the SVM and NB-determined classes using the best SVM and NB models. Human assessors recruited from MTurk were provided with the SVM- and NB-determined class and the LIWC output. They were asked to decide on whether each review was genuine or fake and had the opportunity to go along with the machine assessment or override it. They were provided the same payment and incentive as those in the human computation approach. As with the human computation approach, we take the majority label for the 3 assessments for each review.

4 Results and Analysis

Table 1 illustrates the results (accuracy, precision, recall, and F-score) obtained for both machine learning-based approaches, the human computation approach, and the hybrid approach.

From Table 1 we can observe that the hybrid approach does better (t-test) than the $SVM_{POS+LIWC+Bigrams}$ approach in accuracy, and for F-score of fake reviews. There was no difference for detection of genuine reviews for the hybrid method and the $SVM_{POS+LIWC+Bigrams}$ approach. The hybrid approach performs significantly better than the human computation approach (two-tailed t-test: $t(718) = 13.414$, $p < 0.001$ for F-score, $t(718) = 3.6116$, $p = 0.003$ for accuracy)

Initially this appears unsurprising; the hybrid approach provides the human assessor with the class decision (either fake or genuine) from the Naïve Bayes and SVM approaches and provides the LIWC feature information for the review that is being evaluated and the average for all reviews in the collection. With all this information, certainly a human decision maker's answer would have greater accuracy, precision and recall scores than the decision tool providing information. After all, one would expect that the information provided, the greater the confidence in the decision-making process.

Table 1. Classifier performance for our approaches. Machine learning approaches use nested 5-fold cross-validation. Reported precision, recall and F-score are computed using a micro-average, i.e., from the aggregate true positive, false positive and false negative rates.

Approach	Features used			Accuracy	P	R	F
	POS	LIWC	Bigrams				
NB	*		*	89.7%	48.8	70.0	57.5
NB		*	*	90.5%	52.0	71.0	60.0
NB	*	*	*	91.0%	53.5	73.3	61.9
SVM	*		*	93.4%	64.1	76.3	69.7
SVM		*	*	94.4%	68.8	81.7	74.7
SVM	*	*	*	94.9%	71.0	84.0	76.9
Human Comp				90.2%	50.6	74.7	60.3
Hybrid		*		95.1%	69.6	90.0	78.5

Upon closer examination, we find that this is true – to a point. Of the 3000 reviews evaluated, 271 were incorrectly classified according to the best Naïve Bayes approach ($NB_{POS+LIWC+Bigrams}$), 151 according to the best SVM approach ($SVM_{POS+LIWC+Bigrams}$), and 148 according to the Hybrid approach. However, we see in Fig. 1(a) the Hybrid approach misclassified 17 reviews (11%) in which it differed from the class label given by both SVM and NB but got correct 55 reviews (37%) in which both NB and SVM misclassified the review type. Therefore, the Hybrid approach was three times as likely to override the class decision from both machine learning approaches and make a correct decision as it was to override their decision and get the class label incorrect.

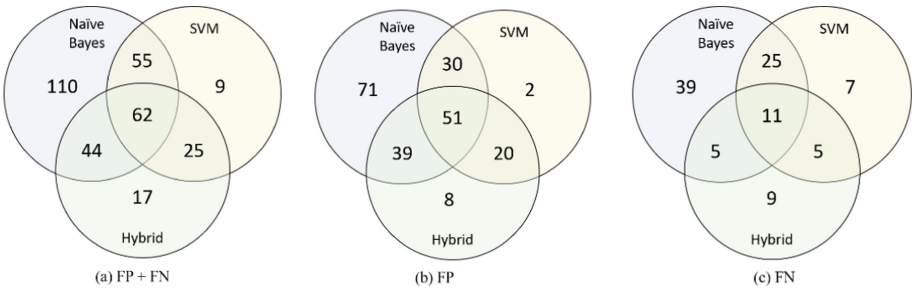

Fig. 1. Venn diagrams showing counts of (a) all incorrectly labeled answers (False Positive and False Negative), (b) False Negative answers only and (c) False Negative answers for $NB_{POS+LIWC+Bigrams}$, $SVM_{POS+LIWC+Bigrams}$, and the Hybrid approaches.

The number of false positives (Fig. 1(b)) is considerably larger than the number of false negatives (Fig. 1(c)) for all three approaches. Comparing Fig. 1(b) and (c), we see that the best Naïve Bayes approach obtains a greater percentage of false positive decisions (i.e., it classifies more genuine reviews as fake than the converse) than the

best SVM approach, indicating a slightly more "aggressive" approach towards classifying reviews as fake (a ratio of 2.4:1) than the best SVM approach (2.1:1). This, however, pales in comparison to the Human Computation approach (2.9:1) and the Hybrid approach (3.9:1). The relative aggressiveness of the human-based approaches may have to do with the incentives offered for correctly classifying reviews.

When the human decision-maker using the Hybrid model had to decide between the Naïve Bayes and SVM classes (i.e., when the two machine methods did not provided the same class label), they chose the Naïve Bayes 64% of the time and SVM 36% of the time. Had only the best SVM class labels have been offered and the humans classified the labels according to the SVM output, the 44 misclassified answers would have boosted Hybrid accuracy to 96.5% and obtained an F-score of 84.1 – a significant increase (two-tailed t-test: $t(718) = 9.156$, $p < 0.001$ for F-score, $t(718) = 2.289$, $p = 0.0224$ for accuracy)

We saw no distinguishable patterns between the 3 hotel markets we examined. However, we discovered that there was a discernable difference between detecting boosting spam and vandalism spam for high-rated and low-rated hotels. Table 2 illustrates the number of false negative classification errors (fake reviews classified as genuine) by hotel type and by review spam type. Figure 2 illustrates the relative proportion of false positive to false negative errors by approach. Overall, we observe that vandalism review spam on low-rated hotels were most difficult to detect (an average of $\frac{18}{75}$ or 24%, were not detected as review spam) whereas vandalism review spam for high-rated hotels was least difficult, with an average of $\frac{12}{75}$, or 16%, not detected.

Table 2. Number of false negative classification errors for each approach, broken down by hotel type and review spam type.

Hotel type	Review spam type	NB	SVM	Human Comp	Hybrid	Average	Total # fake reviews
High-rated	Boosting	17	16	13	12	14.5	75
High-rated	Vandalism	15	6	24	3	12.0	75
Low-rated	Boosting	26	7	18	5	14.0	75
Low-rated	Vandalism	22	19	21	10	18.0	75
All deceptive reviews		80	48	76	30	58.5	300

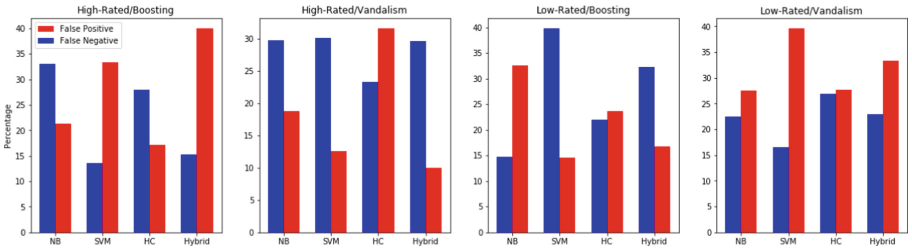

Fig. 2. Relative percentage of classification errors, comparing false positive and false negative values, by hotel type, review type, and approach.

In general, the best Naïve Bayes approach had a more difficult time with the low-rated hotels (i.e., those with one- and two-star ratings) while the best SVM and the Hybrid approaches had a more challenging time with boosting (positive) reviews on high-rated properties and vandalism (negative) reviews on low-rated properties. The human computation approach had a harder problem with vandalism on the low-rated properties and boosting on the high-rated properties. In part, this shows the influence of the best SVM approach on the Hybrid approach, but it also shows how language may also be a factor.

Next, we examine the output provided by the LIWC classifiers as this information is also provided to the assessors in the Hybrid approach. Spatial details were considerably more prominent in genuine reviews than in fake reviews, which supports the reality monitoring (RM) theory of Johnson and Raye [34]. This type of information provides more details about the room layout, bathroom configuration, etc. that can be verified by other guests. This also backs up other work (e.g. [35]) indicating that encoding spatial information into lies is challenging.

Emotion-laden terms, such as a description of the front desk staff's attitude, were more prominent in fake reviews – claims containing these experiences cannot be easily corroborated by other guests. We also noticed that fake reviews contained more external terms – providing background on their vacation, for instance – and less focus on terms that could be verified by others who stay in the same hotel.

Several other researchers have found that while deception is often associated with negative emotional terms (e.g., [36, 37]), the fake reviews were more extreme – the boosting review spam our study used contained more positive and fewer negative emotion terms whereas the vandalism review spam was just the opposite. This exaggeration of emotional terms was most readily picked up by the best SVM approach and least readily picked up by the Human Computation approach, showing how challenging it is for humans to associate strong emotional terms with fake reviews without guidance from machine techniques.

Regarding parts of speech, Gunther et al. [38] indicates that deceptive communications were characterized by fewer first-person singular pronouns, fewer third-person pronouns, more negative emotion words, fewer exclusive words, and more motion verbs. Our findings generally concur with this earlier work with one notable exception: in our study, deceptive reviews contained more first-person singular pronouns. This echoes the findings of Ott et al. [17], who speculated that the use of more first-person pronouns was an attempt to boost the writer's credibility by emphasizing that they were a participant and thus able correctly observe the situation.

One of the biggest indicators of a deceptive review was the generous use of punctuation, particularly exclamation marks. Both human and machine approaches detected this association. An examination of reviews of the 15 hotels on TripAdvisor and Booking.com indicates a much more prolific use of exclamation marks on the former.

It was challenging to analyze the performance of our human assessors, primarily because there were 360 used for each approach (720 total). Figure 3 illustrates the distribution of misclassification errors (false positive and false negatives) for each batch of 25 reviews for human computation and for hybrid approaches. Comparing these two bar graphs illustrates the value of the hybrid approach, as humans in both approaches

were provided the same incentives. We note that false positive reviews were more prominent than false negative reviews because only 10% of reviews were fake. It is worth noting that many reviewers did not have a fake review in their batch, and therefore did not have a point of reference to know what constituted a fake review. This unfortunately raises a potential risk of confirmation bias [39].

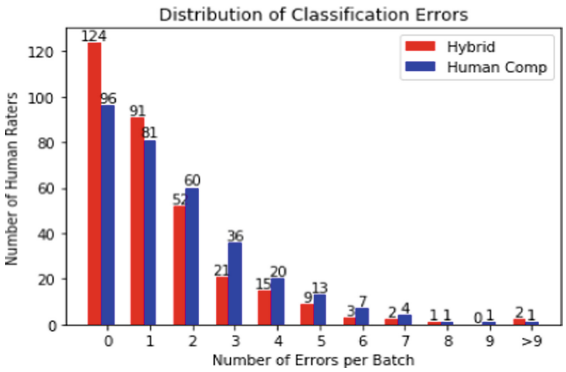

Fig. 3. Distribution of classification errors per batch, comparing human computation and hybrid methods.

5 Conclusion

We have conducted an empirical experiment in which we merged verified reviews for 15 hotel properties in 3 markets (New York, London, and Hong Kong), extracted 90 reviews for each hotel from Booking.com, and then merged them with 10 reviews we had non-experts write. Half of the review spam was boosting, or trying to positively influence the hotel's rating, while the other half was vandalism, or written to negatively influence a competitor's hotel – a growing area of review spam.

From these 3000 reviews, we employed four different methods to determine if each review was genuine or fake: two supervised machine learning methods (Naïve Bayes and SVM), human computation (using MTurk), and a hybrid method (also using MTurk) that allowed humans to make decisions using the class labels from the machine learning methods as well as LIWC output. While it is not surprising that the hybrid method outperformed either of the other methods, it was surprising that when humans were presented with too much information – particularly information that presented more than one possible decision – it negatively impacted human decision-making. Only when the additional information was presented in a non-conflicting manner did humans excel.

This study provided us with considerable data to evaluate, and we intend to do this with an extension of this study. We would also like to evaluate temporal aspects of reviews – the order in which they are posted – since the burstiness of reviews for a property can provide additional evidence of possible review spam.

Although this study was limited in scope to English-language reviews for five hotels, we believe that the overall findings of hybrid man-machine decision making can be extended to other situations outside of validating reviews, such as evaluating financial statements, investigating insurance claims, and evaluating the validity of clinical trials.

References

1. Smith, A., Anderson, M.: Online shopping and e-commerce, online reviews. Pew Research Center report (2016). https://www.pewinternet.org/2016/12/19/online-reviews/
2. Floh, A., Koller, M., Zauner, A.: Taking a deeper look at online reviews: the asymmetric effect of valence intensity on shopping behaviour. J. Mark. Manag. **29**(5–6), 646–670 (2013)
3. Burton, J., Khammash, M.: Why do people read reviews posted on consumer-opinion portals? J. Mark. Manag. **26**(3–4), 230–255 (2010)
4. Mikalef, P., Pappas, I.O., Giannakos, M.N.: Value co-creation and purchase intention in social commerce: the enabling role of word-of-mouth and trust. In: Americas Conference on Information Systems AMCIS (2017)
5. Mikalef, P., Giannakos, M.N., Pappas, I.O.: Designing social commerce platforms based on consumers' intentions. Behav. Inf. Technol. **36**(12), 1308–1327 (2017)
6. Luca, M.: Reviews, reputation, and revenue: the case of Yelp.com. In: Harvard Business School NOM Unit Working Paper (12–016) (2016)
7. Streitfeld, D.: The best book reviews money can buy. The New York Times **25** (2012)
8. Mukherjee, A., Liu, B., Glance, N.: Spotting fake reviewer groups in consumer reviews. In: Proceedings of the 21st International Conference on World Wide Web, pp. 191–200. ACM (2012)
9. Luka, M., Zervas, G.: Fake it till you make it: reputation, competition, and Yelp review fraud. Manage. Sci. **62**(12), 3412–3427 (2016)
10. Harris, C.: Decomposing TripAdvisor: detecting potentially fraudulent hotel reviews in the era of big data. In: 2018 IEEE International Conference on Big Knowledge (ICBK), pp. 243–251. IEEE (2018)
11. Comscore: Top 50 Multi-Platform Properties (Desktop and Mobile) April 2019. https://www.comscore.com/Insights/Rankings. Accessed 06 Nov 2019
12. Belton, P.: Navigating the potentially murky world of online reviews (2015). http://www.bbc.com/news/business-33205905. Accessed 06 Nov 2019
13. Vrij, A.: Detecting lies and deceit: pitfalls and opportunities in nonverbal and verbal lie detection. Interpersonal Commun. **6**, 321 (2014)
14. Lim, E.P., Nguyen, V.A., Jindal, N., Liu, B., Lauw, H.W.: Detecting product review spammers using rating behaviors. In: CIKM 2010 (2010)
15. Jindal, N., Liu, B.: Opinion spam and analysis. In: WSDM 2008 (2008)
16. Harris, C.G.: Detecting deceptive opinion spam using human computation. In: Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence (2012)
17. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 309–319. Association for Computational Linguistics (2011)
18. Bond Jr., C.F., DePaulo, B.M.: Accuracy of deception judgments. Pers. Soc. Psychol. Rev. **10**(3), 214–234 (2006)

19. Crawford, M., Khoshgoftaar, T.M., Prusa, J.D., Richter, A.N., Al Najada, H.: Survey of review spam detection using machine learning techniques. J. Big Data **2**(1), 23 (2015)
20. Ott, M., Cardie, C., Hancock, J.T.: Negative deceptive opinion spam. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 497–501 (2013)
21. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: What yelp fake review filter might be doing? In: Seventh International AAAI Conference on Weblogs and Social Media (2013)
22. Lau, R.Y., Liao, S.Y., Kwok, R.C.W., Xu, K., Xia, Y., Li, Y.: Text mining and probabilistic language modeling for online review spam detecting. ACM Trans. Manag. Inf. Syst. **2**(4), 1–30 (2011)
23. Li, F.H., Huang, M., Yang, Y., Zhu, X.: Learning to identify review spam. In: Twenty-Second International Joint Conference on Artificial Intelligence (2011)
24. Yoo, K.H., Gretzel, U.: Comparison of deceptive and truthful travel reviews. Inf. Commun. Technol. Tourism **2009**, 37–47 (2009)
25. Martin-Fuentes, E., Mateu, C., Fernandez, C.: Does verifying uses influence rankings? Analyzing Booking.com and TripAdvisor. Tourism Anal. **23**(1), 1–15 (2018)
26. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of LIWC2015 (2015)
27. Ney, H., Essen, U., Kneser, R.: On structuring probabilistic dependences in stochastic language modelling. Comput. Speech Lang. **8**(1), 1–38 (1994)
28. Coe, J.: Performance comparison of Naïve Bayes and J48 classification algorithms. Int. J. Appl. Eng. Res. **7**(11), 2012 (2012)
29. Loper, E., Bird, S.: NLTK: the natural language toolkit. arXiv preprint cs/0205028 (2002)
30. Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: a review of classification techniques. Emerg. Artif. Intell. Appl. Comput. Eng. **160**, 3–24 (2007)
31. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**(Oct), 2825–2830 (2011)
32. Huang, C.M., Lee, Y.J., Lin, D.K., Huang, S.Y.: Model selection for support vector machines via uniform design. Comput. Stat. Data Anal. **52**(1), 335–346 (2007)
33. Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S.: Cross-validation pitfalls when selecting and assessing regression and classification models. J. Cheminform. **6**(1), 10 (2014)
34. Johnson, M.K., Raye, C.L.: Reality monitoring. Psychol. Rev. **88**(1), 67 (1981)
35. Vrij, A., et al.: Outsmarting the liars: the benefit of asking unanticipated questions. Law Hum Behav. **33**(2), 159–166 (2009)
36. Newman, M.L., Pennebaker, J.W., Berry, D.S., Richards, J.M.: Lying words: predicting deception from linguistic styles. Pers. Soc. Psychol. Bull. **29**(5), 665–675 (2003)
37. Toma, C.L., Hancock, J.T.: What lies beneath: the linguistic traces of deception in online dating profiles. J. Commun. **62**(1), 78–97 (2012)
38. Gunther, A.C., Perloff, R.M., Tsfati, Y.: Public opinion and the third-person effect. In: The SAGE Handbook of Public Opinion Research, pp. 184–191 (2008)
39. Yin, D., Mitra, S., Zhang, H.: Research note—when do consumers value positive vs negative reviews? An empirical investigation of confirmation bias in online word of mouth. Inf. Syst. Res. **27**(1), 131–144 (2016)