Let's vote to classify authentic and manipulative online reviews : the role of comprehensibility, informativeness and writing style

Banerjee, Snehasish; Chua, Alton Y. K.; Kim, Jung-Jae

2015

Banerjee, S., Chua, A. Y. K., & Kim, J. J. (2015). Let's vote to classify authentic and manipulative online reviews: The role of comprehensibility, informativeness and writing style. 2015 Science and Information Conference (SAI), 77-83 .

<https://hdl.handle.net/10356/80816>

<https://doi.org/10.1109/SAI.2015.7237129>

© 2015 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at:
[<http://www.dx.doi.org/10.1109/SAI.2015.7237129>].

Downloaded on 26 Apr 2023 19:28:25 SGT

Let's Vote to Classify Authentic and Manipulative Online Reviews: The Role of Comprehensibility, Informativeness and Writing Style

Snehasish Banerjee, Alton Y. K. Chua

Wee Kim Wee School of Communication & Information
Nanyang Technological University
Singapore
{snehasis002, altonchua}@ntu.edu.sg

Jung-Jae Kim

School of Computer Engineering
Nanyang Technological University
Singapore
jungjae.kim@ntu.edu.sg

Abstract—Scholars increasingly seek to investigate differences between authentic and manipulative online reviews. A common line of research argues that authentic and manipulative reviews are distinguishable based on three textual characteristics, namely, comprehensibility, informativeness and writing style. Although recent studies have analyzed differences between authentic and manipulative reviews in terms of these textual characteristics, they often lack in terms of methodological rigor. For one, datasets used for analysis are not always representative. Moreover, only few machine learning algorithms are used to classify authentic and manipulative reviews. Recognizing the value of methodological rigor, this paper extends prior studies by examining textual differences between authentic and manipulative reviews using a more representative dataset. Moreover, authentic and manipulative reviews were classified using a voting among multiple classifiers that had been used in recent literature. The implications of the results are discussed.

Keywords—online reviews; authentic; manipulative; voting; classification; comprehensibility; informativeness; writing style

I. INTRODUCTION

The advent of Web 2.0 has empowered users by allowing them to freely contribute content on the Internet. Users are now able to share content about themselves in social networking sites, upload multimedia content in image sharing and video sharing sites, as well as express opinion about products and services in review websites. The abundance of user-generated content available on the Internet serves as valuable information sources for the online community.

A popular form of user-generated content comprises online reviews posted in review websites such as Amazon and TripAdvisor. Reviews enable users to share their post-purchase experiences of products and services with other potential consumers in the online community. The growing popularity of reviews however serves as a double-edged sword. On the one hand, the plethora of reviews allows consumers to obtain a glimpse into others' post-purchase experiences prior to purchase. This in turn has the potential to assist consumers in making better informed purchase decisions. On the other hand, not all reviews posted in review websites are guaranteed to be authentic accounts of post-purchase experiences. Some users could write manipulative reviews out of imagination merely to

pass time or to poke fun [1]. Even businesses could at times persuade consumers or employees to contribute manipulative reviews [2]. Such entries could potentially mislead consumers in their purchase decisions.

This problem has propelled a number of scholarly investigations in recent years to distinguish between authentic and manipulative reviews. A common line of research argues that authentic and manipulative reviews could be distinguished based on their textual differences [3], [4], [5]. In particular, three textual characteristics, namely, comprehensibility, informativeness and writing style, could help sift the grain from the chaff [6], [7]. Comprehensibility is a measure of the understandability of reviews. Informativeness refers to the degree to which reviews are specific and detailed. Writing style indicates the ways words are used in reviews to express opinions.

Even though recent studies have analyzed differences between authentic and manipulative reviews in terms of such textual characteristics, two research gaps exist in terms of their methodological rigor. One, most studies are constrained by the nature of their datasets. While some are impaired by the lack of valid ground truth – for example, [6] and [7], others use datasets that are not representative – for example, [3] and [8]. Two, most related studies such as [9] and [10] limit the scope of their analyses to few machine learning techniques. However, combining multiple techniques through meta-classification approaches such as voting has thus far been seldom used.

Guided by the above methodological caveats, the objective of this paper is to examine differences between authentic and manipulative reviews in terms of the textual characteristics of comprehensibility, informativeness, and writing style. The key contribution of this paper lies in its methodological rigor. It uses a more representative ground truth dataset comprising authentic and manipulative hotel reviews, as well as a more comprehensive analytical approach compared with those used in prior studies.

The remainder of the paper is organized as follows. The following section offers an overview of the related work. This is followed by the methods and the results. The key findings of the paper are presented next. Finally, the paper concludes by highlighting its contributions and limitations.

II. RELATED WORK

The increasing popularity of reviews along with the rising concerns on their manipulation have given rise to adequate scholarly efforts in recent years. A common research theme involves distinguishing between authentic and manipulative reviews based on their textual differences. In one of the earliest studies [7], a dataset of 82 positive reviews (40 authentic + 42 manipulative) was analyzed. It was found that authentic reviews were less complex compared with manipulative entries. The former was found to contain fewer pronouns, fewer positive emotion words, but more negative emotion words than the latter.

In another related study [3], a dataset of 800 positive reviews (400 authentic + 400 manipulative) was analyzed. It was found that authentic and manipulative reviews could be distinguished with substantial accuracy by examining their part-of-speech (POS) tags, psycholinguistic dimensions, as well as unigrams, bigrams and trigrams. A similar investigation was replicated later to obtain largely comparable findings on a dataset of 800 negative reviews (400 authentic + 400 manipulative) [8]. The main difference was that the use of pronouns in negative manipulative reviews was found to be less abundant than that in positive manipulative reviews, albeit higher than their respective authentic counterparts. Even though the results of these two studies were promising, a limitation is that they relied on a bewildering array of textual characteristics for analysis. In particular, taking into account all possible unigrams, bigrams and trigrams to distinguish between authentic and manipulative reviews not only lacks a sound theoretical basis, but could also result in coincidental findings.

Taking the cue, some scholars have recently started being more discerning in the choice of textual characteristics to distinguish between authentic and manipulative reviews. For example, examining a dataset of 800 reviews (400 authentic + 400 manipulative) distributed across both positive and negative comments, [4] relied on the textual characteristics of review complexity, and sentiment polarity. Contrary to [7], authentic reviews were found more complex compared with manipulative entries. No difference between the two was however detected in terms of sentiment polarity.

Given the fractured and rather inconclusive nature of this research area, [6] recently proposed a theoretical framework to identify authentic reviews amidst manipulative ones. It posited that authentic and manipulative reviews are distinguishable in terms of their comprehensibility, informativeness and writing style. Comprehensibility was conceptualized as the length and the readability of reviews. Informativeness comprised the proportion of several POS tags in reviews. Writing style entailed the use of emotion words, perceptual words, tenses and punctuations in reviews.

However, to the best of our knowledge, the extent to which comprehensibility, informativeness and writing style helps ascertain review authenticity is yet to be rigorously examined using a sound dataset. This stems from two reasons. One, related studies are often impaired by the lack of valid ground

truth. Obtaining labelled datasets of authentic and manipulative reviews is certainly challenging. To mitigate this challenge, prior studies such as [6] and [7] have often assumed hotel reviews posted in TripAdvisor – an unauthenticated website that allows anyone with Internet access to post entries – as being authentic. It could at best be argued that reviews posted in TripAdvisor are less likely to be manipulative. However, they are certainly not guaranteed to be authentic. Given the questionable validity of ground truth, the findings of such prior studies might not have painted a fair picture of textual differences between authentic and manipulative reviews.

Two, most studies confine the scope of their datasets to positive reviews that are meant to applaud products and services, for example, [3] and [5]. Some recent studies have started to look into negative reviews that are intended to chastise products and services, for example, [6] and [8]. However, no study has thus far taken into consideration moderate reviews that shed light on both pros and cons of offerings. It appears as if scholars implicitly assume all moderate reviews to be authentic. As a result, they have not sought to investigate review authenticity for such entries. However, extant literature has yet to explicitly demonstrate that all moderate reviews, which are conceivably abundant on the Internet, are guaranteed to be authentic. The datasets used in previous studies thus fail to closely simulate real world settings.

Therefore, it is pertinent to analyze textual differences between authentic and manipulative reviews using a dataset that would be sound in terms of its validity as well as its representativeness. For the purpose of this study, the validity of the ground truth was ascertained by drawing authentic reviews from authenticated hotel review websites such as Agoda.com, Expedia.com, and Hotels.com, which allow entries to be posted only after an actual payment and stay in a given hotel. The manipulative reviews were solicited from participants in an experimental research setting. Unlike studies that confine the scope of datasets to either positive reviews [7] or only popular hotels in a single location [3], the dataset of authentic and manipulative reviews used in this study straddles across positive, negative as well as moderate entries for different types of hotels located in multiple tourist destinations. This was done to enhance the representativeness of the dataset in order to simulate real world settings as closely as possible.

Furthermore, related studies have mostly limited the scope of analyses to few machine learning techniques. For example, [11] used techniques such as C4.5 Decision Tree and Logistic Regression, while [12] used Naïve Bayes and JRip. Yet again, [9] used Support Vector Machine, while [10] used Random Forest. In contrast, this paper investigates textual differences between authentic and manipulative reviews using classification by voting among six commonly used machine learning techniques used in recent literature. These include C4.5 Decision Tree, JRip, Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine. Such a meta-classification approach of combining multiple techniques has rarely been used heretofore in research on authentic and manipulative reviews.

III. METHODS

A. Data Collection

The dataset created for this paper comprises 900 authentic reviews, and 900 manipulative reviews written in evaluation of hotels. Authentic reviews refer to those that are written after a stay in a given hotel. On the other hand, manipulative reviews refer to those that are written without any stay in the hotel.

Authentic reviews were collected from three authenticated review websites, namely, Agoda.com, Expedia.com and Hotels.com. These review websites allow reviews to be posted for a given hotel only after a payment and a stay in the hotel. It is therefore not possible to post manipulative entries in these websites by creating fictitious user accounts. This assures that all reviews obtained from the three websites are authentic. Moreover, drawing authentic reviews from multiple authenticated review websites facilitated greater scope for triangulation.

Collecting authentic reviews involved two steps. In the first step, a set of 15 hotels that attract large volumes of reviews in Agoda.com, Expedia.com and Hotels.com were identified. In particular, three hotels were identified from each of the five tourist destinations, namely, Hong Kong, Bangkok, Kuala Lumpur, Singapore and Tokyo. The list of three hotels included one luxury hotel, one budget hotel, and one mid-range hotel. The hotels were categorized into luxury, budget or mid-range properties by checking their star-ratings in the three selected websites. Hotels with a consistent star-rating of five were deemed as luxury properties, those with a consistent star-rating of one were deemed as budget properties, and those with a consistent star-rating of three were deemed as mid-range properties. All the identified hotels had attracted more than some 1,000 reviews altogether across Agoda.com, Expedia.com and Hotels.com. This allowed for a wide scope of data collection.

In the second step, a total of 60 authentic reviews were randomly collected for each of the 15 hotels (15 hotels x 60 reviews = 900). Reviews were admitted into the dataset only if they were in English, and contained meaningful descriptions of at least 150 characters. Reviews shorter than 150 characters do not allow for a meaningful analysis [3]. The set of 60 reviews collected for every hotel included 20 positive, 20 moderate, and 20 negative entries. The sentiment polarity of reviews was determined based on the review ratings [13]. These two steps yielded 900 authentic reviews. Specifically, it included 300 positive, 300 moderate, and 300 negative reviews distributed uniformly across the 15 identified hotels.

To ensure comparability with the authentic reviews, 60 manipulative reviews were collected for each of the 15 hotels. These entries were solicited from participants through email instructions. Informed by studies such as [3] and [7], participants were asked to imagine as if they work for the marketing department of a hotel. Their boss had asked them to write some realistic manipulative reviews in English for hotels. Each review had to contain a meaningful description of at least 150 characters.

Participants were recruited for the study based on three criteria. One, they had to be in the age group of 21 to 45 years.

After all, most reviews are posted by users in this age group [14]. Two, their educational profile had to be minimally undergraduate students. This is because reviews are generally posted by educated users [15]. Three, they had to be familiar with the use of reviews in review websites. These criteria ensure that the participants selected to write manipulative reviews were realistically capable of doing the task.

Collecting manipulative reviews from participants involved two steps. In the first step, the invitation to participate in the study was disseminated using snowballing. The researchers' personal contacts were leveraged through social networking sites and word-of-mouth. When a participant volunteered to participate in the study, they were sent the detailed email instructions to write manipulative reviews. Participants were asked not to submit entries if they had stayed in a hotel earlier. This assures that the manipulative reviews obtained from participants were not unduly biased.

In the second step, when a participant sent the manipulative reviews, the entries were examined if they could be admitted into the dataset. If a review was in English, and contained a meaningful description of at least 150 characters long, it was included into the dataset. Otherwise, it was ignored. These two steps were iterated for the 15 hotels concurrently to yield 900 manipulative reviews. Specifically, it included 300 positive, 300 moderate, and 300 negative reviews distributed uniformly across the 15 identified hotels. The entries were submitted by a total of 284 participants.

To summarize, the overall dataset contained a total of 1,800 reviews (900 authentic + 900 manipulative). The reviews were uniformly distributed across the 15 identified hotels (5 luxury + 5 budget + 5 mid-range) spread across five tourist destinations, namely, Hong Kong, Bangkok, Kuala Lumpur, Singapore and Tokyo. For each hotel, the dataset comprised 60 authentic reviews (20 positive + 20 moderate + 20 negative), and 60 manipulative reviews (20 positive + 20 moderate + 20 negative).

B. Measures

The measures for the three textual characteristics of comprehensibility, informativeness and writing style were largely inspired from [6]. Specifically, comprehensibility entails the length and the readability of reviews. Length was measured as the number of words in a given review. For readability, some metrics widely used in prior research include Gunning-Fog Index, Coleman-Liau Index, Automated Readability Index, and Flesch-Kincaid Grade Level [10], [16]. These rely on factors such as average sentence length in words, and average word length in characters or syllables. A Java program was written to compute these metrics. Thereafter, readability of a given review was measured as the arithmetic mean of all the metrics. Thus, the comprehensibility-related features for classification include number of words and mean readability.

Informativeness comprises the proportion of several POS tags in reviews. Specifically, eight POS tags could differ between authentic and manipulative texts. These include nouns, adjectives, articles, prepositions, verbs, adverbs, pronouns, and personal pronouns [3], [17]. The first four could

be higher in authentic reviews, which are supposed to be informative. In contrast, the rest could be higher in manipulative reviews, which could be found wanting in terms of informativeness. To offer greater nuances, the use of personal pronouns was further measured separately as the fraction of singular self-references such as “I” as well as that of plural self-references such as “we”. This was necessary because there is evidence from prior research that manipulative reviews could contain fewer singular self-references but more plural self-references to dissociate oneself, thus transferring accountability to some imaginary others [18]. For every review, Stanford parser’s POS tagger was used to measure the fraction of the POS tags [19]. Thus, the informativeness-related features for classification include a total of 10 POS tags.

Writing style encompasses the use of emotion words, perceptual words, tenses and punctuations in reviews. The use of emotion words was measured as the fraction of both positive emotion words and negative emotion words in reviews. The use of perceptual words was operationalized as the fraction of words related to the senses of sight, hearing, and touch. The use of tense was operationalized as the fraction of past, present, and future tense. The use of punctuations entails the fraction of question marks and exclamation points in reviews. These were measured using the Linguistic Inquiry and Word Count toolkit [20]. Thus, there were a total of 10 writing style-related features for classification: the two types of emotion words, the three types of perceptual words, the three tenses, and the two forms of punctuations.

Overall, there were a total of 22 features: two related to comprehensibility, 10 pertaining to informativeness, and 10 for writing style. The values corresponding to each of these features were measured for all the 1,800 reviews (900 authentic + 900 manipulative) in the dataset. The resulting feature matrix (1,800 reviews x 22 features) was used for subsequent data analyses.

C. Data Analyses

Data were analyzed using classification by voting among six machine learning techniques, namely, C4.5 Decision Tree (C4.5), JRip, Logistic Regression (Logistic), Naïve Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM). Related studies have often restricted the analysis to only some of these machine learning algorithms. For instance, [11] used techniques such as C4.5 and Logistic, while [12] used NB and JRip. Yet again, [9] used SVM, while [10] used RF. Given that no single algorithm has been consistently superior in performance, all of these were combined using voting through average probability. For SVM, a polynomial kernel was chosen as it was found to result in superior performance compared with either linear or radial basis function kernels.

Classification performance was evaluated using a five-fold cross-validation. For the purpose of this paper, if a manipulative review was correctly classified, it is termed as a true negative. If incorrectly classified, it is called a false

positive. Likewise, if an authentic review was correctly classified, it is termed as a true positive. If incorrectly classified, it is called a false negative. Based on these definitions, performance was evaluated using six metrics, namely, precision, sensitivity, specificity, accuracy, F_1 -measure, and area under the receiver operating characteristic curve (AUC).

The machine learning algorithms were implemented using the Weka toolkit [21]. All parameters were set to their default values. Furthermore, a cost-sensitive version of the classifiers was also implemented. For this purpose, the cost of classifying manipulative reviews as authentic was treated twice as expensive as that of classifying authentic reviews as manipulative [22]. The results of the cost-sensitive classification in terms of the performance metrics did not differ adequately up to two decimal places from those of the original classification. Hence, they are not reported for the sake of brevity.

After examining the classification performance, the top five features that helped to distinguish between authentic and manipulative reviews were identified using Information Gain (IG), and Chi-squared (χ^2) feature selection techniques. In particular, IG ranks features used in a classification problem in decreasing order of their relative importance based on entropy [12]. Besides, χ^2 ranks features by measuring divergence from the expected distribution when feature occurrence is independent of the class value [23]. For classification with less than 100 features, IG and χ^2 feature selection techniques are known to perform well [23], [24]. Moreover, related studies have widely used these feature selection techniques [12], [25].

To further delve deeper, differences between authentic and manipulative reviews in terms of these top five features were examined using independent samples t-tests. This statistical procedure is appropriate when comparing differences in means between two samples [26]. For the purpose of this paper, t-tests helped to disinter the differences in textual characteristics between authentic and manipulative reviews pertaining to the top five features.

IV. RESULTS

Results of the classification analysis are summarized in Table I. Using voting among the other machine learning algorithms, 587 of the 900 authentic reviews as well as 581 of the 900 manipulative entries could be correctly identified. This yielded a classification accuracy of about 65%. Thus, it seems that comprehensibility, informativeness and writing style could help ascertain review authenticity with a reasonable accuracy.

As shown in Table II, the top five features that aid distinguishing between authentic and manipulative reviews in terms of IG and χ^2 feature selection include the use of (1) past tense, (2) singular self-references, (3) pronouns, (4) nouns, and (5) articles. Independent samples t-tests indicated that authentic and manipulative reviews differed from one another significantly across all the five features.

TABLE I. CLASSIFICATION PERFORMANCE

Analysis	Precision	Sensitivity	Specificity	Accuracy	F ₁ -measure	AUC
C4.5	0.59	0.56	0.61	0.58	0.57	0.59
JRip	0.57	0.64	0.52	0.58	0.60	0.60
Logistic	0.63	0.64	0.63	0.64	0.63	0.69
NB	0.63	0.64	0.63	0.63	0.63	0.68
RF	0.64	0.54	0.69	0.62	0.58	0.66
SVM	0.63	0.64	0.63	0.63	0.63	0.63
Voting	0.65	0.65	0.65	0.65	0.65	0.71

TABLE II. TOP FIVE FEATURE SELECTION

Feature	Information Gain	χ^2	t-test
Past tense	0.036	81.973	t(1798) = -7.00***
Singular self-references	0.035	86.309	t(1751.55) = -9.07***
Pronouns	0.026	64.405	t(1798) = -7.58***
Nouns	0.019	48.134	t(1705.71) = 6.96***
Articles	0.014	35.736	t(1761.26) = -4.23***

*** Statistically significant results at $p < 0.001$

In particular, past tense was significantly less used in authentic reviews (3.90 ± 3.65) vis-à-vis manipulative entries (5.13 ± 3.78) [t(1798) = -7.00, $p < 0.001$]. Singular self-references were significantly fewer in authentic reviews (1.87 ± 2.42) vis-à-vis manipulative entries (3.00 ± 2.85) [t(1751.55) = -9.07, $p < 0.001$]. As a result, it is not surprising that the fraction of pronouns in general was significantly lower in authentic reviews (8.20 ± 4.62) compared with manipulative ones (9.83 ± 4.47) [t(1798) = -7.58, $p < 0.001$]. The proportion of nouns was significantly higher in authentic reviews (28.05 ± 7.03) than manipulative entries (25.98 ± 5.55) [t(1705.71) = 6.96, $p < 0.001$]. The fraction of articles was significantly lower in authentic reviews (8.28 ± 3.87) compared with manipulative entries (9.00 ± 3.35) [t(1761.26) = -4.23, $p < 0.001$].

To offer a qualitative treatment of the dataset in terms of the top five features, a word-level analysis technique known as the ‘Difference Between Proportions Method’ (DBPM) was adopted [13]. It helps identifying words that differ significantly in their occurrences across two types of documents. For the purpose of this paper, the DBPM was used to identify examples of words related to the top five features that differ significantly in their occurrences between authentic and manipulative reviews in the dataset.

Specifically, differences in word count frequency of specific words between authentic and manipulative reviews were computed. Thereafter, the differences were standardized into z-scores such that a positive z-score for a given word would indicate that the word is more abundant in authentic reviews vis-à-vis manipulative entries. Likewise, a negative z-score would mean that the word is less frequent in the former. An absolute z-score greater than 1.96 for a given word would suggest that the occurrence of the word among authentic

reviews significantly differs from its occurrence among manipulative entries ($p < 0.05$).

With respect to past tense, the DBPM analysis revealed that words such as ‘was’ ($z = -20.91$), ‘were’ ($z = -12.51$), and ‘had’ ($z = -7.55$) appeared significantly fewer in authentic reviews vis-à-vis manipulative entries. Among pronouns, singular self-references such as ‘I’ (-32.91) and ‘my’ ($z = -21.83$) were significantly fewer in the former. Moreover, nouns such as ‘floor’ ($z = 7.12$), location ($z = 5.06$), and ‘airport’ ($z = 2.54$) were significantly more abundant in authentic reviews compared with manipulative entries. However, articles such as ‘a’ ($z = -20.15$) and ‘an’ ($z = -4.72$) were significantly fewer in the former. Overall, it appears that authentic reviews are generally richer in nouns but scantier in terms of past tense, pronouns and articles vis-à-vis manipulative entries.

V. DISCUSSION

Three key findings are gleaned from this paper. First, it offers a set of heuristics to distinguish between authentic and manipulative reviews. On the one hand, reviews that are rich in nouns are likely to be authentic. On the other hand, reviews that are rich in past tense, pronouns and articles are likely to be manipulative.

The finding that authentic reviews contain more nouns but fewer pronouns and articles vis-à-vis manipulative entries is also echoed in related prior studies such as [6] and [7]. However, to the best of our knowledge, the variation in past tense between authentic and manipulative reviews has not been reported in extant literature. Manipulative reviews were richer in past tense perhaps because they were deliberately written to mimic authentic entries, which are supposed to be written after experience with products and services. In contrast, authentic reviews seemed to be written without any attempt to connote a

past experience. Consistent with the deception research of [27], authentic reviews appear more imperfect compared with manipulative entries. That is why, they were not overly rich in past tense despite supposedly being recollections of past experiences.

Second, classification by voting represents a viable and robust method to distinguish reviews in terms of their authenticity. In terms of five performance metrics, namely, precision, sensitivity, accuracy, F_1 -measure, and AUC, voting emerged superior over the individual classification algorithms. Prior studies have also suggested that combining multiple classification algorithms simultaneously has the potential to enhance overall performance [28].

It is conceivable that different classification algorithms strive to optimize different energy functions. In consequence, some could turn out to be better in accurately identifying authentic reviews, while others could emerge as being better in accurately identifying manipulative entries. Moreover, their performance could also be dependent on the nature of the dataset. Therefore, it is not surprising that most studies generally stick to few algorithms that provide the best performance [10], [11]. Even though such a selective strategy might yield compelling results, it might not be methodologically rigorous. Therefore, it is important to use meta-classification approaches such as voting. Interestingly, classification by voting has long been used in domains such as pattern recognition [29]. It is perhaps time for scholars investigating differences between authentic and manipulative reviews to join the bandwagon.

Third, the classification accuracy found in this paper was not as high as those reported in some of the earlier studies such as [4] and [6]. In particular, both these studies found accuracy ranging from some 70% to 76%, whereas this paper obtained an accuracy of 65%. The lower accuracy could be attributed to the nature of the dataset, which straddles across different types of reviews, namely, positive, negative and moderate, submitted in evaluation of various types of hotels – luxury, budget and mid-range. In contrast, as indicated earlier, the datasets used in the previous studies were largely confined in their scope, which perhaps resulted in over-estimation of accuracy.

Nonetheless, an error analysis was conducted to study the nature of incorrectly classified instances. Specifically, 632 of the 1,800 reviews (313 authentic + 319 manipulative) in the dataset were found to be incorrectly classified. In other words, 313 authentic reviews were incorrectly classified as manipulative (false negatives). Conversely, 319 manipulative reviews were incorrectly classified as authentic (false positives). The volumes of false positives and false negatives were generally comparable for all three types of reviews, as well as all three types of hotels. Interestingly, across review types, the number of incorrectly classified instances was the highest among negative reviews. Across hotel types, the number of incorrectly classified instances was the highest among reviews for budget hotels. Overall, these results suggest that extant literature on differences between authentic and manipulative reviews is still not comprehensive. Future scholarly inquiry needs to take into account nuances in the contexts, such as the type of products and services being

reviewed as well as the nature of reviews, to further enrich understanding of this research theme.

VI. CONCLUSION

Recognizing the value of methodological rigor, this paper extends the literature by examining textual differences between authentic and manipulative reviews using a more representative dataset, and a more rigorous analytical approach. For this purpose, a ground truth dataset of authentic and manipulative reviews was created. The validity of the ground truth was ascertained by drawing authentic reviews from authenticated review websites. On the other hand, manipulative reviews were solicited from participants in an experimental research setting. The differences between authentic and manipulative reviews were examined in terms of three textual characteristics, namely, comprehensibility, informativeness and writing style. Analysis was done by combining multiple classification algorithms through voting. Results indicate that reviews that are rich in nouns are likely to be authentic, whereas those rich in past tense, pronouns and articles are likely to be manipulative.

The key significance of this paper is that it analyzed differences between authentic and manipulative hotel reviews using a more representative dataset than those used in extant literature. In particular, the dataset could be considered representative in at least three ways. One, it comprised positive, negative as well as moderate reviews. However, prior studies had confined datasets to either positive and/or negative entries [7], [8]. Two, the dataset included reviews for three types of hotels, namely, luxury, budget, and mid-range. In contrast, prior studies had mostly studied reviews for popular hotels [6], [8]. Three, the dataset contained reviews for hotels from five tourist destinations: Hong Kong, Bangkok, Kuala Lumpur, Singapore and Tokyo. However, prior studies mostly relied on reviews for hotels located in a single destination [7], [3]. Therefore, the dataset used in this paper could more closely simulate real world settings compared with previous studies, thereby yielding more realistic results compared with previous studies.

Moreover, related studies had mostly relied on few machine learning algorithms to classify authentic and manipulative reviews. In contrast to such piecemeal analytical approach, this study combined multiple techniques, which had been found to perform well in recent literature, using the meta-classification approach of voting. This study therefore allows for a more comprehensive and rigorous analysis of textual differences between authentic and manipulative reviews compared with previous related works.

However, a limitation of this paper is that it did not propose any new algorithm to distinguish between authentic and manipulative reviews. Rather, its contribution lies in its methodological rigor. Interestingly, the accuracy of classifying authentic and manipulative reviews found in this paper was not as high as reported in some of the earlier studies such as [4] and [6]. Perhaps, the higher accuracy in previous studies could be attributed to the nature of the datasets, which appear far from being representative. This suggests that the current understanding of textual differences between authentic and manipulative reviews is only the tip of the iceberg. Much more

needs to be done to obtain a more comprehensive and a clearer view of differences between the two. Hence, this paper calls for further scholarly inquiry to shed greater light on the research theme on authentic and manipulative online reviews.

REFERENCES

- [1] X. Chen, and S. C. J. Sin, "Misinformation? What of it? Motivations and individual differences in misinformation sharing on social media," *Proceedings of the American Society for Information Science and Technology*, 2013, pp. 1-4.
- [2] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review spam detection via temporal pattern discovery," *Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: ACM, 2012, pp. 823-831.
- [3] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," *Annual Meeting of the Association for Computational Linguistics*, 2011, pp. 309-319.
- [4] C. Harris, "Detecting deceptive opinion spam using human computation," *AAAI Workshop on Artificial Intelligence*, 2012, pp. 87-93.
- [5] T. Ong, M. Mannino, and D. Gregg, "Linguistic characteristics of shill reviews," *Electronic Commerce Research and Applications*, vol. 13, pp. 69-78, 2014.
- [6] S. Banerjee, and A. Y. K. Chua, "A theoretical framework to identify authentic online reviews," *Online Information Review*, vol. 38, pp. 634-649, 2014.
- [7] 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: Springer, 2009, pp. 37-47.
- [8] M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam," *North American Chapter of the Association for Computational Linguistics*, 2013, pp. 497-501.
- [9] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing - style features and classification techniques," *Journal of the American Society for Information Science and Technology*, vol. 57, pp. 378-393, 2006.
- [10] A. Ghose, and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, pp. 1498-1512, 2011.
- [11] L. Zhou, J. K. Burgoon, D. P. Twitchell, T. Qin, and J. F. Nunamaker Jr, "A comparison of classification methods for predicting deception in computer-mediated communication," *Journal of Management Information Systems*, vol. 20, pp. 139-165, 2004.
- [12] M. P. O'Mahony, and B. Smyth, "Learning to recommend helpful hotel reviews," *Proceedings of the Conference on Recommender systems*, New York: ACM, 2009, pp. 305-308.
- [13] 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.
- [14] 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.
- [15] J. Rong, H. Q. Vu, R. Law, and G. Li, "A behavioral analysis of web sharers and browsers in Hong Kong using targeted association rule mining," *Tourism Management*, vol. 33, pp. 731-740, 2012.
- [16] N. Korfiatis, E. Garcia-Bariocanal, and S. Sánchez-Alonso, "Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content," *Electronic Commerce Research and Applications*, vol. 11, pp. 205-217, 2012.
- [17] P. Rayson, A. Wilson, and G. Leech, "Grammatical word class variation within the British National Corpus sampler," *Language and Computers*, vol. 36, pp. 295-306, 2001.
- [18] L. Zhou, J. K. Burgoon, and D. P. Twitchell, "A longitudinal analysis of language behavior of deception in e-mail," *Proceedings of the Conference on Intelligence and Security Informatics*, Berlin: Springer, 2003, pp. 102-110.
- [19] D. Klein, and C. D. Manning, "Accurate unlexicalized parsing," *Proceedings of the Association for Computational Linguistics*, 2003, pp. 423-430.
- [20] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, "The development and psychometric properties of LIWC 2007," Austin, TX: LIWC.net, 2007.
- [21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, pp. 10-18, 2009.
- [22] L. S. Chen, and J. Y. Lin, "A study on review manipulation classification using decision tree," *Proceedings of the IEEE International Conference on Service Systems and Service Management*, 2013, pp. 680-685.
- [23] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289-1305, 2003.
- [24] Y. Yang, and J. O. Pedersen, "A comparative study on feature selection in text categorization," *Proceedings of the International Conference on Machine Learning*, California: AAAI, 1997, pp. 412-420.
- [25] F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," *Proceedings of the International Joint Conference on Artificial Intelligence*, California: AAAI, 2011, pp. 2488-2493.
- [26] J. E. McEnroe, and S. C. Martens, "Auditors' and investors' perceptions of the 'expectation gap'," *Accounting Horizons*, vol. 15, pp. 345-358, 2001.
- [27] 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.
- [28] F. Kimura, and M. Shridhar, "Handwritten numerical recognition based on multiple algorithms," *Pattern recognition*, vol. 24, pp. 969-983, 1991.
- [29] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 66-75, 1994.