A Quantitative Study of Software Reviews Using Content Analysis Methods

HONGYU ZHANG¹, ANDREW YANG², ALLEN PENG², LAVINIA F. PIEPTEA³, JIAN YANG³, AND JUNHUA DING⁴

¹School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
²Hamilton High School, Hamilton, AZ 13346, USA
³Department of Information Science, University of North Texas, Denton, TX 76203, USA

Corresponding author: Junhua Ding (junhua.ding@unt.edu)

This work was supported in part by the National Security Agency (NSA) under Grant H98230-20-1-0417, and in part by the NSF under Grant 2225229 and Grant 1852249.

ABSTRACT Online product reviews play a critical role for consumers to make a decision of product purchasing, and become an important data source for vendors to build recommendation systems. Some consumers won’t even buy a product without reading online reviews first, and some vendors invite reviewers to write product reviews before a product is released to the market. However, the review quality could greatly impact the use of the reviews for supporting the purchase decision and recommendation. In order to produce enough high-quality reviews, it is a common practice that vendors provide incentives for writing product reviews. Current research is divided. Some research showed that incentive reviews could be more biased compared to organic reviews, but other scholars showed incentive reviews contain more useful information. Furthermore, other academic publications showed very different results regarding the differences between the incentivized and organic reviews. One of the reasons explaining the observation differences could be due to the quality of the reviewers and reviews. Therefore, it is necessary to control the quality of the reviewers and their reviews for a more objective comparison. In this research, we first discuss an approach for ensuring the quality of the data collection and processing to ensure the quality of the comparison study. Then, we explore the differences between incentivized reviews and organic reviews collected from a website that provides reviews for enterprise software systems. Several parameters of the reviews, such as the overall score, sentiment, and subjectivity of the reviews, were analyzed and compared. Our results could provide a reference for appropriately using reviews and managing the reviewing process.

INDEX TERMS Content analysis, information quality, quantitative study, sentiment analysis, text mining.

I. INTRODUCTION

Online reviews or online consumer reviews (OCR), also referred to as electronic word-of-mouth (e-WOM), have been recognized as a valuable information source for the user perception of goods and services. Pricing, performance, and customer satisfaction are some of the most common aspects evaluated by users through text, pictures, or videos [2]. The information is then used to determine customer interests, needs, and opinions regarding businesses’ products and services alike. Given the increased usage of internet, which became indispensable in our modern lifestyle, large amounts of reviews are now available for analysis [15], providing consumers with references, or businesses with resources to build product recommendation systems. Online reviews have been explored throughout variate research, providing progress-driven insights to both individuals and businesses.

Researchers reported that individuals’ benefits as recipients gravitate around trust and usefulness gained from the reviews, regardless of the domain, for example in sales [1] [17] or in the hospitality and tourism industry [31], [36], just to name a few. According to data from a survey, 77% of consumers read online reviews, and 75% trust online reviews more than personal recommendations [24].
Moreover, statistics from the same article show consumers’ dependence on posted opinions while making decisions, online reviews being read and trusted, with over 80% of internet shoppers acknowledging reviews’ helpfulness. Academic findings show that online reviews have a significant impact on consumer attitudes [34] towards brand image, and consequently on the purchase intent [28], [29]. On average, consumers rely on 112 reviews before deciding on a product online [9]. Similar for businesses, online represents a convenient source of information that captures the interests on the market. Consumers’ critique brings awareness of possible improvements in their activities, as shown by [10], who analyzed the connection between clients’ expectations and business sales achievements. Moreover, the feedback from buyers and sellers showed strong correlation, indicating that both parties would respond and react back [27] to each other in similar ways [39]. Clearly, analyzing the client feedback, companies can develop new product or services, thereby further advancing their market presence. In addition, recommendation systems could be built on reviews to provide better product recommendations for customers.

Although the online reviews are important to individuals and businesses, it may take long time to cumulate sufficient number of reviews for a product. Therefore, some businesses provide incentives for reviewers to write reviews. Incentivized reviews are expected to be posted by consumers who have used the product and share their honest and unbiased opinions. Some studies showed that incentivized reviews may be susceptible to manipulation, but other studies showed that incentivized reviews are more useful than organic reviews, which were written by users voluntarily without receiving any incentives [7], [8], [18], [19]. Current research on the impact of incentivized reviews in comparison to the organic reviews is limited. Kim et al. [19] explored the differences between incentivized and organic reviews though comparing the review attributes such as sentiment, rating, and review length of the two types of reviews, which included 666 reviews on 52 skincare products. The analysis results showed that incentivized reviews were more positive and gave higher rating compared to organic reviews. Costa et al. [7] used a machine learning method to analyze the review dataset from the University of California San Diego and showed the incentivized reviews are longer and more positive compared to organic reviews. However, the analysis results could not be generated due to the limitations of the quality of the reviews and the reviewers. For example, the review dataset used by Kim et al. [19] is very small and it may contain many fake reviews, or reviews without much information, such as only giving a rate but no review content. The review dataset used by Costa et al. [7] did not explicitly classify the reviews into incentivized or organic, but were rather grouped by the authors, which could be misleading.

In this research, the reviews were collected from a commercial website that provides online reviews of enterprise software systems. The quality of the online review is highly assured. For example, the website filters out suspicious fake reviews, the reviews without much content, or those where the reviewer cannot be verified. Therefore, the quality of the reviews is guaranteed, and the background of the reviewers is similar for both invited and organic reviewers, which mitigates some confounding problems, such as the fact that the invited reviewers may have a much better understanding of the product than the organic ones. The analysis and comparison were conducted on the overall rating a reviewer gave to a product, and the linguistic attributes of the reviews, such as the length of a review, sentiment, and subjectivity that are important to define the quality of reviews. Text mining methods were used for analyzing the reviews and A/B testing was conducted to understand the statistical significance of the differences between the two types of reviews.

The paper is organized as follows: Section II describes the related work, Section III discusses the methodology, Section IV discusses the experimental results, and Section V is conclusion.

II. RELATED WORK

Considering the important information carried by the online reviews, online reviews have been investigated by researchers and largely divided into five categories of features [40], where each addresses an important component of the review system: reviewer, review, recipient, channel, and response.

Reviewer related features include source credibility [38], reviewer’s expertise [35], reviewer’s information [4], [13] and reviewer’s reputation [16]. At the opposite pole, the recipients provide feedback with their response [6]. Recipient’s familiarity and identification with the product [26] represent deterministic features triggering influence in the final decision. The platform hosting the reviews influence in the customer decision. Zheng [40] identifies the types of channels to be non-commercially linked third-party channels, commercially linked third-party channels, and seller-owned channels. Published literature research indicates that the type of communication channel can influence the recipient’s assessment of a review [3], such as their arousal, valence, and trust.

Most of the review content analyzed in literature is adopting the dual-process theory Elaboration Likelihood Model (ELM) to distinguish two types of influential factors in individuals’ perception [11]: informational (or central route) and normative (or peripheral route) influences. Normative influence regards the individual pressure of fitting into society norms and trends. Evaluating dimensions of peripheral route are information quantity and product ranking. Thus, users seek easily accessible overall opinion [6], considering the volume of available data at the same time. Many websites address this by using summarized score systems for each product or service to represent the average rating provided by all consumers who have reviewed it. Informational influence, on the other hand, refers to the receiver’s perception on the content quality of the message. Attempts have been made by scholars to develop criteria by which the value of an argument can be evaluated. Park et al. [23] considered credibility, objectivity, clarity, and logic. Cheung et al. [6] measured
review quality only in terms of relevance, timeliness, accuracy, and understandability, while Filieri and McLeay [12] considers these four dimensions and further investigates value-added, and completeness. Relevance [37] is how useful a review is for a certain activity and depends on consumer needs. Nelson et al. [22] refers to timeliness as up-to-date, contemporary, and state-of-the-art product/service information. Same study defines information accuracy as the proper mapping of recorded data to its real-world representation. Wang and Strong [37] present readability, interpretability, and ease of understanding as aspects in information understandability, and usefulness and beneficial information as value-added characteristics. Complete information, according to them, has enough breadth, depth, and scope for the task at hand.

Review content data has been extensively analyzed. Some of the most popular features found in literature are the review valence, length, review volume, rating score, message quality, and subjectivity. The valence of a review has been extracted by Ghaseimaghai et al. [14] using two-way ANOVA analysis that shows a link between high ratings and short reviews with a positive attitude. These results are further confirmed through sentiment analysis methods that show review length negative correlation to the review rating, whereas the sentiment correlates positively.

Overall review helpfulness, according to [37], builds on review depth, extremity, readability, and rating. However, research conducted by Almansour et al. [2] has found evidence of text content and score rating discrepancy. Often biased and challenging to determine the exact meaning of a text, approaches such as Latent Semantic Analysis (LSA) have been applied to mitigate subjectiveness. LSA is a statistical method used in natural language processing for analyzing relationships between large texts and the words contained in them, yielding a set of relevant notions within paragraphs. Text preparation, parsing, term reduction, singular value decomposition (SVD), and factor analysis [5] are the main steps included in this process. The review helpfulness has been studied in literature mainly through logistic and regression models, with primary focus on the content quality (e.g., semantic characteristics, online rating) [4], [5], [15].

Message quality has been quantified through several indexes that represent readability, such as the Automated Readability Index (ARI), Gunning’s Fog Index (FOG), Coleman–Liau Index (CLI), and the Flesch–Kincaid Reading Ease index (FK) [15], [21], [30]. Liang et al. [20] implemented a multilevel approach to control variables related to the features of a hotel business (room, pictures, rank) and the existence of nested structure in their dataset by implementing FK and constructing a multiple regression model. Authors select the Gunning’s Fog Index (FOG) as an additional analysis to prove their method validity.

Research results on the differences of incentivized reviews and organic reviews were reported recently [7] and [19] as we discussed in last section. The focus of our research is on the assurance of the quality of the reviews as well as the reviewers so that the analysis results can be much more reliable.

III. METHODOLOGY

In this section, we are going to elaborate the process of the data analysis and the methods we developed to conduct the data analysis and comparative study. We collected the review data from commercial website which provides the software purchasing consultant services for customers. The website collected millions of reviews for hundreds of software systems. One of the examples of the website is G2.com (at https://www.g2.com). The reviews included incentivized reviews and organic reviews. These reviews could be collected directly by the website itself or copied from other websites that provide the similar service. The reviewers shall be verified via the contact information, such as LinkedIn account, the job title, and the experience of reviewer, which is also listed along with the review. Low quality reviews, such as a review that is too short or has no useful information, would be removed. The website includes more incentivized reviews than organic reviews. Further, we pre-process the review data. We referred to other articles [19] to select the review parameters to be analyzed, such as ARI index [30] and sentiment, so that our results could be compared to other similar work. In order to analyze the difference between Organic and Incentivized reviews in one product, we conduct A/B testing to assert the difference in the mean value of each parameter in a statistical way, and the LDA [32] method was used to analyze textual difference between the two types of reviews.

A. DATA COLLECTION AND PRE-PROCESSING

The review website contains reviews of many different enterprise products, but we only collect reviews of project management software products. The forms of the reviews on this website are mainly divided into text and video, including Incentivized reviews and Organic reviews.

We developed a Python program to crawl the website to collect the text reviews. We first crawled all project management software product links from the website, store the links and cookies in prepared for crawling the specific product reviews. Then, we loaded the product links and traverse them. For each product, we crawled the text, star, and the review source of all its textual reviews according to reviews’ Xpath. The python library we use is undetected_chromedriver, which can bypass the Cloudflare that is used to limit crawling of the website. At the same time, the script sends a notification when captcha occurred. We were able to manually bypass the captcha because its frequency of occurrence is based on time rather than how much data we had visited. An IP pool was prepared to avoid ban of our IPs by the website. Finally, we collected 36347 reviews of 289 project management software products on website including both incentivized and organic ones. We put all the reviews of each product into a csv file, which is named as product_name.csv. However, analysis on products containing only small number of reviews
might not be general. Therefore, we mainly used 3 individual products where each contains more than 1000 reviews to conduct the analysis and comparison as well as cross-validation of the analysis results. We also conduct the analysis of all reviews from all products to validate the reviews on individual products and multiple products. We added all the data we have (i.e. 36347 reviews of the 289 software products) into one csv file and we will name it as ALL_DATA. Finally, we got 4 csv files as our dataset, and we analyze them separately. The initial dataset attributes contain review text (or simply called review), star, review source into ‘Incentive’. Meanwhile, we rename ‘review source’ as ‘label’. Therefore, we only one type of Organic reviews, so we transferred ‘review source’ contains several types of Incentivized reviews but can largely reflect the level of satisfaction of the reviewer and other noise. We also found the column ‘review source’ of each type of incentivized reviews into ‘Incentive’. Meanwhile, we rename ‘review source’ as ‘label’. Therefore, ‘label’ contains two types value: Incentive and Organic. The average length of the review is: 60 words.

B. DEFINITION OF REVIEW QUALITY AND SENTIMENT
Some of the review quality attributes are included in the review, such as the overall review rate, but other review quality attributes, such as ARI, have to be calculated. More complex review attributes, like sentiment and subjectivity, have to be analyzed using machine learning methods.

1) DEFINITION OF REVIEW QUALITY
The definition of review quality was mainly based on the number of words of the review, the number of sentences of the review, and the number of letters of the review, and we denote them as word_num, sentence_num, letter_num, respectively. These three attributes are easily calculated after standardizing the dataset. To better define the review quality, we synthesized the characteristics of these three attributes by using The Automated Readability Index (i.e. ARI) [30], as defined in (1).

\[
ARI = 4.71 \left( \frac{\text{letter}_\text{num}}{\text{word}_\text{num}} \right) + 0.5 \left( \frac{\text{word}_\text{num}}{\text{sentence}_\text{num}} \right) - 21.43
\]

(1)

The ARI outputs a number which approximates the grade level needed to comprehend the text. For instance, if the output number of a review is 10.62, it indicates that a person below 11th grade will have no difficulty to understand the review. Therefore, it can be used to characterize whether a comment is concise and easy to understand and thus evaluating the publicity of the comment. Finally, we define review quality as ARI to summarize the attribution of word_num, sentence_num, and letter_num.

2) DEFINITION OF SENTIMENT
The definition of sentiment is divided into two aspects: one is the star rating given by a reviewer, the other is the review’s polarity and subjectivity, which can be calculated based on the text content. The star rating is the level of sentiment, since it can largely reflect the level of satisfaction of the reviewer with the product. TextBlob is an open source text processing library written in Python that can be used to perform many natural language processing tasks, including sentiment analysis through the TextBlob.Sentiment functionality. This function returns a tuple of the form (polarity, subjectivity), where polarity defines negativity or positivity in a text, and subjectivity implies that the expression of a text is vague or positive. The value of the polarity is a float type that vary from −1 to 1, and refers to ‘very negative’ to ‘very positive’ and the value of the subjectivity is also a float type that vary from 0 to 1, and refers to ‘objectivity’ to ‘subjectivity’. Here we use star, sentiment.polarity and sentiment.subjectivity to represent the sentiment of the review.

C. EXPLORATION OF DIFFERENCES BETWEEN INCENTIVE AND ORGANIC REVIEWS IN THE PRODUCTS
In this section, we first investigate the difference between incentivized and organic reviews in terms of the mean of star, mean of word_num, mean of polarity, mean of subjectivity and mean of ARI (i.e. review_quality) to assert this difference in a statistical way using A/B testing.

Second, in order to analyze the differences in text content between the incentivized and organic reviews, we will use the LDA method to generate a specific model, both for incentive and organic, thus comparing the similarities in text content between two types of reviews with topic modelling.

1) CONDUCT A/B TESTING TO INVESTIGATE THE DIFFERENCE BETWEEN THE TWO TYPES OF REVIEWS
We are going to perform a non-parametric permutation test (i.e. A/B testing) with the goal of asserting whether or not there is any statistically significant difference of the mean of

| Variable         | Data Type |
|------------------|-----------|
| review text      | string    |
| star             | integer   |
| review source    | string    |

| Variable         | Data Type |
|------------------|-----------|
| ARI(review_quality) | double   |
| label            | string    |
| polarity         | double    |
| review           | string    |
| star             | integer   |
| subjectivity     | double    |
| word_num         | integer   |
each attribute (i.e. star, word_num, polarity, subjectivity, ARI) between the two types of reviews. An A/B test was applied since it is not feasible to perform all possible permutations and calculate the exact $P_{\text{value}}$.

The hypothesis is defined based on each attribute: \{star, word_num, polarity, subjectivity, ARI\}.

**Null hypothesis** $H_0$: The mean values of each attribute between incentive and organic are equal because all of the data belongs to the same population.

**Alternate hypothesis** $H_1$: The mean values of each attribute between incentive and organic are different.

The details of the A/B test are summarized as follows:

For reviews from one product, we labelled incentive and organic reviews into group$_1$ and group$_2$, respectively. For one attribute in both groups, we take out the specific attribute column from both groups. Then we take the average of the both columns and denote them as $\mu_{\text{inc}}$ and $\mu_{\text{org}}$. Finally, we got the real difference of this attribute between incentive and organic reviews and we denote the real difference as $\Delta \mu$. After getting the $\Delta \mu$, the process of the resampling can be elaborated as described below.

Let $M$ refer to the number of organic reviews and $N$ refer to the number of incentive reviews. Further, we take $M$ and $N$ data as two groups. Finally, we get mean of both groups and the difference of them as $\Delta \mu_{\text{true}}$. At last, we will assert if $|\Delta \mu_{\text{true}}|$ is greater or equal to $|\Delta \mu|$ or not. The resample process is repeated a specific number of times (in our study, 100000 time) and the $\Delta \mu_{\text{true}}$ generated each time is recorded.

Notice that with the A/B testing we can find the $P_{\text{value}}$ as in (2), where $N$ is the number of replicate samples that have been simulated, and $r$ is the number of these replicates that produce a test statistic ($|\Delta \mu_{\text{true}}|$) greater than or equal to that calculated for the real data ($|\Delta \mu|$). The addition of the value 1 in the numerator and denominator comes from imposing that the estimated $P_{\text{value}}$ is unbiased.

\[
P_{\text{value}} = \frac{r + 1}{N + 1} \tag{2}
\]

Under the null hypothesis, each resample has a probability of getting a test statistic at least as extreme as the one for the original sample that is equal to the permutation test $P_{\text{value}}$.

Consequently, out of $N$ resampling simulations, the number of resamples with a test statistic at least as extreme as the one in the original sample has a binomial distribution with parameters $N$ and $P_{\text{value}}$. As a result, we can form a confidence interval for the $P_{\text{value}}$ using standard methods for a binomial proportion confidence interval. Particularly, we use Wilson score interval as in (3), (4).

\[
\text{Interval}_{\text{lower}} = \frac{p + \frac{z^2}{4n} + \frac{z}{2n} \sqrt{p(1-p) + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \tag{3}
\]

\[
\text{Interval}_{\text{upper}} = \frac{p + \frac{z^2}{4n} + \frac{z}{2n} \sqrt{p(1-p) + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \tag{4}
\]

We assume that there will have different values between two types reviews in the null hypothesis. We assume that all the data belong to the same population and the fact of resampling them does not generate anything more than a random variation of the same data, but not a new distribution.

We will conduct A/B test individually on every attribute in \{star, word_num, polarity, subjectivity, ARI\} to assert whether the mean value of the attribute between incentive and organic reviews are different in section 3.

2) USE LDA MODEL TO GET THE DIFFERENCE IN TEXT BETWEEN TWO TYPES OF REVIEW

In addition to statistically analyzing the difference between the means of each attribute of the two types reviews, we look directly into the text content itself to compare the difference between incentive and organic reviews, using Latent Dirichlet allocation (LDA) [32] model.

**Latent Dirichlet allocation** is a generative probabilistic model of a corpus. The basic idea is that a document is represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA assumes the following generative process for each document $w$ in a corpus $D$:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the $N$ words $w_n$:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word $w_n$ from $p(w_n | z_n, \beta)$, a multinomial probability conditioned on the topic.

The training process of the LDA model can be summarized in 4 steps:

**Step1**: For every word $w_n$ in corpus, randomly assign a topic id $z$ for it.

**Step2**: Traverse every word in the corpus again, using Gibbs Sampling formula to sample it and update its topic in the corpus.

**Step3**: Conduct Step 2 again until Gibbs Sampling converges.

**Step4**: The statistical corpus topic-word co-occurrence frequency matrix is the LDA model.

In our study, we use LDA model to compare the similarity between incentive and organic reviews. Commonly, similarity scores vary from 0 to 1, and if two texts are very similar, the scores will tend to approach 1. Just as shown in Figure 1, our method can be divided into 4 stages: Data Pre-processing, Model Training, LDA space vector generation and Similarity score calculation.

In **Data Pre-processing**, we will process the ‘reviews’ of the dataset again and this includes: stemming the verbs, removing the stop-words, and transforming reviews into the sparse word frequency vector. Let corpus$_{\text{inc}}$, corpus$_{\text{org}}$ refer to incentive reviews and organic reviews after the processing above.

In **Model Training**, we input corpus$_{\text{inc}}$ and corpus$_{\text{org}}$ to train two LDA models: LDA$_{\text{inc}}$ and LDA$_{\text{org}}$, respectively, with number of the topics so that the model achieves the highest coherence scores.
In **LDA space vector generation**, two LDA spaces are generated by LDA$_\text{inc}$ and LDA$_\text{org}$ respectively. Next we input corpus$_\text{inc}$, corpus$_\text{org}$ to the LDA spaces, in order to transform corpus into LDA vector. After we get LDA vector, we are able to create index to get similarity scores.

In **Similarity score calculation**, the flowchart in Figure 1 can be summarized as: create index → conduct queries on all incentive or organic reviews with the opposite label index → form a Similarity score list → get the final similarity scores.

Let’s take the generation of Similarity$_\text{inc}$ (i.e. The mean similarity value of all incentive reviews in index$_\text{org}$) as an example to illustrate the idea. We first denote $M, N$ as the number of organic and incentive reviews respectively, and here are the four steps.

**Step1**: We create index$_\text{org}$ for Query by gensim.similarities.MatrixSimilarity, where the input of gensim.similarities.MatrixSimilarity is the LDA vector of corpus$_\text{org}$.

**Step2**: We traverse the corpus$_\text{inc}$. For every review in the corpus$_\text{inc}$, specifically, we take review$_\text{inc}_i$, and conduct a Query using index$_\text{org}$, and the Query will return a similarity score list between review$_\text{inc}_i$ and all the organic reviews. The operation of taking the mean value of the list is Mean$_0$, and the mean value is Similarity$_\text{inc}_i$ as defined in (5).

**Step3**: We do the Iteration (i.e. Do the same operation as in Step 2 until the last review in corpus$_\text{inc}$). Thus, we can form the Similarity score list as in Figure 1.

**Step4**: We take the mean value of Similarity score list as Similarity$_\text{inc}_i$ (We refer to this operation as Mean$_1$) as defined in (6).

$$
\text{Similarity}_\text{inc}_i = \frac{\sum \text{index}_\text{org}(\text{review}_\text{inc}_i)}{N} i = 0, 1, 2 \ldots N - 1 \quad (5)
$$

$$
\text{Similarity}_\text{inc} = \frac{\sum_{j=0}^{N-1} \text{Similarity}_\text{inc}_j}{N} \quad (6)
$$

To ensure the completeness, we will get Similarity$_\text{org}$ (i.e. The mean similarity value of all organic reviews in index$_\text{inc}$) with the similar process above. (7,8)

$$
\text{Similarity}_\text{org}_j = \frac{\sum \text{index}_\text{inc}(\text{review}_\text{org}_j)}{M} j = 0, 1, 2 \ldots M - 1 \quad (7)
$$

$$
\text{Similarity}_\text{org} = \frac{\sum_{j=0}^{M-1} \text{Similarity}_\text{org}_j}{M} \quad (8)
$$

Finally, do a weight summation on Similarity$_\text{org}$ and Similarity$_\text{inc}$ according to the quantity of the two types.
reviews (9) and we denote the summation as $Score$, which refers to the similarity score between two types of reviews.

$$
Score = \text{Similarity}^{inc} \times \frac{N}{M+N} + \text{Similarity}^{org} \times \frac{M}{M+N} 
$$

(9)

IV. RESULTS

In this section, we describe the data analysis conducted and the comparison study results. First, we present the distribution of our dataset. Then, we discuss the $A/B$ testing on the chosen attributes: star, words_num, polarity, and ARI index and use LDA model to get the difference in text between two types of reviews, thus leading the final results on the difference between incentive and organic reviews.

A. DATASET

As mentioned in Section III.A, after crawling the data on G2, we chose three products ‘Clickup’, ‘Asana’, ‘Mavenlink’, and ‘ALL_DATA’ (i.e., the all 35431 reviews of 289 software products) to analyze. Next, we pre-process the dataset; the data transformation from TABLE 1 to TABLE 2 is described earlier. TABLE 3 shows basic information and distribution of the dataset.

B. A/B TEST

In this part, we conducted A/B test on star, word_num, polarity, subjectivity, and ARI index on all three products and ALL_DATA, thus 24 tests are conducted. For each test, we resample 100000 times to make sure the $P_{value}$ we gain is reliable. We set $\alpha = 0.05$ for all tests above. If the $P_{value}$ is below 0.05, we are able to reject the null hypothesis (i.e., supports the alternative hypothesis) that: the difference of the mean of the feature under test of the two type of reviews are statistically significant.
Specifically, observations regarding the difference of the two types of reviews can be drawn from Figure 3 and TABLE 6:

1. $P_{value}$ of features star and ARI of each of the three products and ALL_DATA is below 0.05, then we can conclude that the difference of the mean of star and ARI of the types of reviews are statistically significant.

2. $P_{value}$ of feature Word_num of product Asana, Mavenlink, or ALL_DATA is below 0.05, but $P_{value}$ of product Clickup is 0.3851, which is much higher than 0.05. We cannot make a 100% agreed conclusion, although majority of the products including ALL-DATA support the alternative hypothesis. However, exception do happen, such as the reviews collected on product Clickup.

3. $P_{value}$ of features Polarity and Subjectivity of product Asana, Mavenlink, or ALL_DATA is above 0.05, but $P_{value}$ of product Clickup for Polarity and Subjectivity is 0.0001 and 0.0084 respectively, which is much lower than 0.05. We cannot make a 100% agreed conclusion, although majority of the products including ALL_DATA support the NULL hypothesis.

4. From above observations, it is safe to conclude that the difference of overall rating/stars between incentivized reviews and organic reviews is significantly different. The conclusion is consistent to the results conducted on different products by other researchers. Probably the reviewers who took incentives for writing the reviews might give more favorable rating. The difference of the review text in syntax between incentivized reviews and organic reviews is also significantly different although exception exist. One of the possible explanations is the reviewers who took incentives for writing reviews may write the review in a more careful way. However, the sentiment calculated from review text between incentivized reviews and organic reviews is not significantly different although exception exist. It has to be mentioned that the conclusion could be different when the sentiment analysis is calculated differently.

5. The analysis results from ALL_DATA is always consistent to the analysis results produced by the majority of the three products. In order to achieve low variant results, it is necessary to conduct the analysis in a larger dataset that shall contain more products with more reviews, and the number of incentive reviews and organic reviews should be balanced.

C. LDA MODEL

In order to generate LDA model with highest coherence, we first find the best k for all the LDA model we use later. The best quantity of topics for the three products and the total reviews are listed in TABLE 4.

When the best quantity of the topic is found, the final step is ready to be conducted. We generate the LDA model with the corresponding topic’s quantity. Next, we compare the similarity between two types reviews for the 4 datasets separately with the algorithm mentioned in section Methodology. The final results of the similarity scores are shown in TABLE 5. Based on similarity score of “ALL_DATA”, one can see the topics discussed by incentive reviews and organic results are fairly different, and the results from other products do not against the observation. One of the explanations is that incentive reviews are more structured and cover mode broader topics, and organic reviews are more casual and focused on couple of topics. The observation was confirmed by manual inspection of many reviews.

V. CONCLUSION

Online product reviews are important for customers to make purchase decision. Yet, bad reviews could be misleading and even cause problems. That being the case, some vendors and manufacturers provide incentives for writing product reviews and hope they are useful and objective. Some research work showed the incentive reviews could be more biased compared to the organic reviews, and other scholars showed different results. One of the reasons explaining the observation differences could be due to the quality of the reviewers and reviews. Hence, it is necessary to control the quality of the reviewers and reviews to establish a more objective comparison of the incentivized and organic reviews. In this
This project was initiated and led by two high school students Andrew Yang and Allen Peng. They proposed the project idea, collected and cleaned the data, designed the experiment method and conducted the experiments, and drafted the research report including the literature review. Undergraduate student Hongyu Zhang offered the technical support for programming and LDA analysis, and he also drafted the first version of this article. Ph.D. student Lavinia Pieptea was responsible for advising the students in research. Jian Yang provided onsite research support and advising for Andrew and Allen, and Junhua Ding led the team for the project development. The success of the project showed the high school students can do excellent research as soon as they are well advised. The authors hope the success of the project could inspire more high school students to participate in research.

REFERENCES

[1] L. H. Al-Abbadi, D. M. K. Bader, A. Mohammad, A. Z. Al-Quran, F. M. F. Alidhaiani, S. I. S. Al-Hawyari, and F. F. Alathmaneh, “The effect of online consumer reviews on purchasing intention through product mental image,” Int. J. Data Netw. Sci., vol. 6, no. 4, pp. 1519–1530, 2022, doi: 10.5267/j.idnss.2022.5.001.

[2] A. Almansour, R. Alotaibi, and H. Alharbi, “Text-rating review discrepancy (TRRD): An integrative review and implications for research,” Future Bus. J., vol. 8, no. 1, Dec. 2022, doi: 10.1186/s43093-022-00114-y.

[3] S. Bae and T. Lee, “Product type and consumers’ perception of online consumer reviews,” Electron. Markets, vol. 21, no. 4, pp. 255–266, 2011, doi: 10.1007/s10572-011-0072-0.

[4] H. Bauck, J. Ahn, and Y. Choi, “Helpfulness of online consumer reviews: Readers’ objectives and review cues,” Int. J. Electron. Commerce, vol. 17, no. 2, pp. 99–126, Dec. 2012, doi: 10.25374/ijec-4415170204.

[5] Q. Cao, W. Duan, and Q. Gan, “Exploring determinants of voting for the ‘helpfulness’ of online user reviews: A text mining approach,” Decis. Support Syst., vol. 50, no. 2, pp. 511–521, Jan. 2011, doi: 10.1016/j.dss.2010.11.009.

[6] C. M. K. Cheung, M. K. O. Lee, and N. Rabjohn, “The impact of electronic word-of-mouth: The adoption of online opinions in online customer communities,” Internet Res., vol. 18, no. 3, pp. 229–247, Jun. 2008, doi: 10.1108/10662240810883290.

[7] A. Costa, J. Guerreiro, S. Moro, and R. Henriques, “Unfolding the characteristics of incentivized online reviews,” J. Retailing Consum. Services, vol. 47, pp. 272–281, Mar. 2019, doi: 10.1016/j.jretconsserv.2018.12.006.

[8] G. Cui, Y. Chang, L. Peng, and W. Zheng, “The importance of being earnest: Mandatory vs. voluntary disclosure of incentives for online product reviews,” J. Bus. Res., vol. 141, pp. 633–645, Mar. 2022, doi: 10.1016/j.jbusres.2021.11.068.

S. Dixon. (2019). Topic: Online Reviews. Statista. [Online]. Available: https://www.statista.com/topics/4581/online-reviews/

[10] H. Fang, J. Zhang, Y. Bao, and Q. Zha, “Towards effective online review systems in the Chinese context: A cross-cultural empirical study,” Electron. Commerce Res. Appl., vol. 12, no. 3, pp. 208–220, 2013, doi: 10.1016/j.elerap.2013.03.001.

[11] R. Filieri, “What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in E-WOM,” J. Bus. Res., vol. 68, no. 6, pp. 1261–1270, Jun. 2015, doi: 10.1016/j.jbusres.2014.11.006.

[12] R. Filieri and F. McLeay, “E-WOM and accommodation,” J. Travel Res., vol. 53, no. 3, pp. 44–57, Jan. 2014, doi: 10.1177/0047287513481274.

[13] C. Forman, A. Ghose, and B. Wiesenfeld, “Exploring the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets,” Inf. Syst. Res., vol. 19, no. 3, pp. 291–313, 2008, doi: 10.1287/isre.1080.0193.

[14] M. Ghasemaghaei, S. P. Eslamian, K. Deal, and K. Hassanein, “Reviews’ length and sentiment as correlates of online reviews’ ratings,” Internet Res., vol. 28, no. 3, pp. 544–563, Jun. 2018, doi: 10.1108/lnr-12-2016-0394.

[15] A. Ghose and P. G. Ipeirotis, “Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics,” IEEE Trans. Knowl. Data Eng., vol. 23, no. 10, pp. 1498–1512, Oct. 2011, doi: 10.1109/TKDE.2010.188.

[16] P. B. Goes, M. Lin, and C. M. A. Yeung, “‘Popularity’ effect in user-generated content: Evidence from online product reviews,” Inf. Syst. Res., vol. 25, no. 2, pp. 222–238, 2014, doi: 10.1287/isre.2013.0512.

[17] K. Kaur and T. Singh, “Impact of online consumer reviews on Amazon books sales: Empirical evidence from India,” J. Theor. Appl. Electron. Commerce Res., vol. 16, no. 7, pp. 2793–2807, Oct. 2021, doi: 10.3390/ta16070153.

[18] W. Khem-Am-Nuai, K. Kannan, and H. Ghasemkhani, “Extrinsic versus intrinsic rewards for contributing reviews in an online platform,” Inf. Syst. Res., vol. 29, no. 4, pp. 871–892, 2018, doi: 10.1287/isre.2017.0750.

[19] S. J. Kim, E. Maslowska, and A. Tamaddoni, “The paradox of (dis)trust in sponsorship disclosure: The characteristics and effects of sponsored online consumer reviews,” Decis. Support Syst., vol. 116, pp. 114–124, Jan. 2019, doi: 10.1016/j.dss.2018.10.014.

[20] S. Liang, M. Schuckert, and R. Law, “How to improve the stated helpfulness of hotel reviews? A multilevel approach,” Int. J. Contemp. Hospitality Manage., vol. 31, no. 2, pp. 953–977, Feb. 2019, doi: 10.1080/07421222.2018-0134.

[21] Z. Liu and S. Park, “What makes a useful online review? Implication for travel product websites,” Tourism Manage., vol. 47, pp. 140–151, Apr. 2015, doi: 10.1016/j.tourman.2014.09.020.

[22] R. R. Nelson, P. A. Todd, and B. H. Wixom, “Antecedents of information and system quality: An empirical examination within the context of data warehousing,” J. Manage. Inf. Syst., vol. 21, no. 4, pp. 199–235, Apr. 2005, doi: 10.2753/JMIS2104009.

[23] D.-H. Park, J. Lee, and I. Han, “The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement,” Int. J. Electron. Commerce, vol. 11, no. 4, pp. 125–148, 2007, doi: 10.1605/IJEC.2008.2104.

[24] J. Pitman, (Jan. 26, 2022). Local Consumer Review Survey 2022: Customer Reviews and Behavior. BrightLocal. [Online]. Available: https://www.brightlocal.com/research/local-consumer-review-survey/

[25] N. Purnawirawan, M. Eisend, P. De Pelsmacker, and N. Dens, “A meta-analytic investigation of the role of valence in online reviews,” J. Interact. Marketing, vol. 31, pp. 17–27, Aug. 2015, doi: 10.1016/j.intmar.2015.05.001.
[26] P. Racherla, M. Mandivivalla, and D. J. Connolly, “Factors affecting consumers’ trust in online product reviews,” J. Consum. Behaviour, vol. 11, no. 2, pp. 94–104, Mar. 2012, doi: 10.1002/cbh.385.

[27] P. Resnick and R. Zeckhauser, “Trust among strangers in internet transactions: Empirical analysis of eBay's reputation system,” Adv. Appl. Microecon., vol. 11, pp. 127–157, May 2002, doi: 10.1016/S0278-0094(02)11030-3.

[28] M. Reza Jalilvand and N. Samiei, “The effect of electronic word of mouth on brand image and purchase intention,” Marketing Intell. Planning, vol. 30, no. 4, pp. 460–476, Jun. 2012, doi: 10.1108/02634501211231946.

[29] J. Muntermann, B. Rajagopalan, and M. Siering, “Explaining and predicting online product recommendations on consumers’ online choices,” J. Retailing, vol. 80, no. 2, pp. 159–169, Jan. 2004, doi: 10.1016/j.jretai.2004.04.001.

[30] R. J. Senter and E. A. Smith, “Automated readability index,” AMRL-TR Defense Tech. Inf. Center, Univ. Cincinnati, Cincinnati, OH, USA, May 1967, pp. 1–14.

[31] S. J. Stanton, J. Kim, J. C. Thor, and X. Deng, “Incentivized methods to generate electronic word-of-mouth: Implications for the resort industry,” Int. J. Hospitality Manage., vol. 78, pp. 142–149, Apr. 2019, doi: 10.1016/j.ijhm.2018.11.004.

[32] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, Jan. 2003.

[33] H. Janssen, “Monte-Carlo based uncertainty analysis: Sampling efficiency and sampling convergence,” Rel. Eng. Syst. Safety, vol. 109, pp. 123–132, Jan. 2013.

[34] M. R. Shihab and A. P. Putri, “Negative online reviews of popular products: Understanding the effects of review proportion and quality on consumers’ attitude and intention to buy,” Electron. Commerce Res., vol. 19, no. 1, pp. 159–187, Mar. 2019, doi: 10.1007/s10660-018-9294-y.

[35] M. Siering, J. Muntermann, and B. Rajagopalan, “Explaining and predicting online review helpfulness: The role of content and reviewer-related signals,” Decis. Support Syst., vol. 108, pp. 1–12, Apr. 2018, doi: 10.1016/j.dss.2018.01.004.

[36] T. H. Tseng, S.-H. Chang, Y.-M. Wang, Y.-S. Wang, and S.-J. Lin, “An empirical investigation of the longitudinal effect of online consumer reviews on hotel accommodation performance,” Sustainability, vol. 13, no. 1, p. 193, Dec. 2020, doi: 10.3390/su13010193.

[37] R. Y. Wang and D. M. Strong, “Beyond accuracy: What data quality means to data consumers,” J. Manage. Inf. Syst., vol. 12, no. 4, pp. 5–33, 1996, doi: 10.1080/07421222.1996.11518099.

[38] K. Z. Zhang, S. J. Zhao, C. M. K. Cheung, and M. K. O. Lee, “Examining the influence of online reviews on consumers’ decision-making: A heuristic-systematic model,” Decis. Support Syst., vol. 67, pp. 78–89, Nov. 2014, doi: 10.1016/j.dss.2014.08.005.

[39] Z. Zhang, Q. Ye, R. Law, and Y. Li, “The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews,” Int. J. Hospitality Manage., vol. 29, no. 4, pp. 694–700, Dec. 2010, doi: 10.1016/j.ijhm.2010.02.002.

[40] L. Zheng, “The classification of online consumer reviews: A systematic literature review and integrative framework,” J. Bus. Res., vol. 135, pp. 226–251, Oct. 2021, doi: 10.1016/j.jbusres.2021.06.038.

HONGYU ZHANG is currently pursuing the undergraduate degree with the University of Electronic Science and Technology of China.

LAVINIA F. PIEPTEA received the B.S. degree in mathematics from the University of Bucharest, Romania, in 2019, and the M.S. degree in mathematics from the University of North Texas (UNT), Denton, TX, USA, in 2022, where she is currently pursuing the Ph.D. degree in data science with the Department of Information Science. Her research interests include biomedical computation, information security, and privacy.

JIAN YANG received the B.Sc. degree from Sichuan University, the M.S. degree from Florida International University, and the Ph.D. degree from Arizona State University. He joined as an Adjunct Professor with the Department of Information Science, University of North Texas, in 2018. His research interests include data visualization, data warehouse and business intelligence, biomedical natural language processing, and biomedical big data.

JUNHUA DING received the B.Sc. degree in computer science from the China University of Geosciences, the M.S. degree in computer science from Nanjing University, and the Ph.D. degree in computer science from Florida International University. He is the Reinburg Endowed Professor of data science with the University of North Texas. His research interests include software engineering, machine learning, and data analytics.

ALLEN PENG is with the Hamilton High School, Chandler, Arizona. He has experience working in the biomedical and bioengineering with the University of California Irvine, under Professor Kulinsky, as well as the data and computer science fields with Professor Junhua Ding and Jian Yang with the University of Northern Texas.

ANDREW YANG joined at the School Robotics Team, in 2019, where he has been involved in many competitions. He is with the Hamilton High School. His interest includes computer programming began in high school.

** * *