

Using Review Metadata to Identify Fake Amazon Reviews

Problem Statement

The problem of fake amazon reviews has resulted in the success of fraudulent companies that sell products under the name of a real brand. When customers receive low-quality products from the fraudulent company, the reputation of the brand is negatively impacted. The purpose of this project is to investigate how false reviews can be identified by addressing the following questions:

1. Do products with high initial rates of review accumulation have less verified reviews?
 - a. If we assume that fraudulent companies will generate fake reviews from the birth of their product listing to artificially inflate their review numbers and overall rating, a high initial rate of review accumulation would be associated with a high number of false reviews. Theoretically, verified reviews are more likely to be true reviews since a transaction between the buyer and the seller is required for verification. So, a product listing with a large number of false reviews would have a high initial rate of review accumulation and a low number of verified reviews.
 - b. To determine if products with high initial rates of review accumulation have less verified reviews, we will find the proportion of verified reviews for each product listing and the rate of review accumulation over the first six months of a product's lifespan and plot the two against each other.
2. How do reviews accumulate over the lifetime of a product listing for suspicious reviews? For trustworthy reviews?
 - a. I hypothesize that for products with genuine reviews, review accumulation will follow an exponential trajectory. Customers use reviews to determine if they will buy a product. Product listings with a small number of reviews are considered a risky purchase and therefore, only a small number of customers will be willing to take the risk. Of those customers willing to take the risk, a smaller percentage will return to leave a review. Therefore, for genuine postings, reviews will be few and far between during the beginning of the product listing's lifetime. As the number of reviews increase, more customers will feel comfortable purchasing the product and the rate of review accumulation will increase exponentially.
 - b. To determine how reviews accumulate over the lifetime of a product listing, I will first determine the amount of time that a product listing has been active by sorting the reviews by the product listing identification number (asin) and finding the first and last review for the product. I will find the rate of review accumulation over the product listing's first 6 months, from 6 months to a year, from the first year to the second year, and from the

second year to the fifth year. Suspicious product listings will be defined as product listings that have a high initial rate of review accumulation and a low number of verified reviews. Trustworthy product listings will be defined as product listings with a low initial rate of review accumulation and a high number of verified reviews. Suspicious and trustworthy listings will be subsetting from the dataset and the frequency distributions of the reviews will be compared.

3. Do unverified reviews tend to have a more positive sentiment than unverified reviews?
 - a. If there is a fundamental difference between verified and unverified reviews and unverified reviews tend to have more positive sentiments, this would suggest that unverified reviews are, in fact, more likely to be fraudulently generated.
 - b. We will compare the mean overall score of verified and unverified reviews with a paired t-test to test the following hypotheses:
 - i. $H_0: \mu_1 - \mu_2 = 0$
 - ii. $H_1: \mu_1 - \mu_2 > 0$


Where μ_1 is the mean overall score of unverified reviews and μ_2 is the mean overall score for verified reviews.


Obtain


Three JSON datasets consisting of about 3.3 million amazon reviews and totalling 1.56 GB were obtained from [Jianmo Ni, Jiacheng Li, and Julian McAuley \(2019\)](#), (data is split across multiple files - 1pt). The first dataset (208.9 MB) contains 494,485 reviews in the Arts, Crafts, and Sewing category, the second dataset (747.8 MB) contains 1,711,519 reviews in the Automotive category, and the third dataset (604 MB) contains 1,128,437 reviews in the Cellphones and Accessories category. The columns of interest in the datasets were the asin number – a product listing identifier, the overall star rating (1-5), the review text, time the review was posted in Unix time, the reviewer identification number, and whether the review was verified or not (True/False) (Data contains strings with punctuation (quotes or commas): 1pt). The difficulty of the dataset was assessed to be three points in total.


Scrub


The initial data was in JSON format and the three datasets were uploaded to Google Cloud Storage and imported into Big Query. From BigQuery, the data was uploaded to Cloud Dataprep and the three datasets were automatically combined when uploaded. In Dataprep, the columns of interest were subsetting (see Obtain section), and missing values were filtered before being exported to BigQuery as a table.


 final_table


 QUERY

 SHARE

 COPY

 SNAPSHOT

 DELETE

 EXPORT

SCHEMA		DETAILS		PREVIEW	
Row	overall	verified	reviewerID	asin	reviewText
1	5	true	A3KQXBYJYV9S54	B019QEVW6Q	Work great, very bright.
2	5	true	AFWGESLQWF3TN	B019ZR6I2W	Had to ship back due to my rollbar size, was bummed. GREAT mirror!
3	5	true	A3CV8UEGCZZ384	B01A7SDWR2	IF you are changing transmission fluid on MB's or VW/Audi's then this out features. It also helps to read the instructions, the big cap on top d make that connection until I read the very good instructions. A nice qu
4	5	true	A2PXLRLYUT7I5M	B01A90SNPY	Great graphics and stands out from all the other JKUs on the road.
5	4	true	A213EXST8K5SS3	B01ACOA1P2	Doesnt come with a 1.5 inch curved brackets like the one in the photo. Light works great, very bright.
6	5	true	A2ZJRQE3HWS92H	B01AECHB3C	These are great and I stuck them on the sides of my Jeep. They have t
7	5	true	A12FFN7XD6LJ41	B01AG4NFXS	haven't used it yet but very visible air pressure measurement.
8	1	true	A2IZGDK9N370UA	B01AH007B0	These did not fit on my 2013 ZL1. They hit my Stainless Works header

Results per page: 50 1 - 50 of 3332521

<<

<

>

>>

Figure 1. Screenshot of Big Query preview screen showing sample rows. The bottom right of the screen displays the number of rows in the table: about 3.3 million.

A cluster was created in Dataproc to run the pyspark script job which answered the second question as described in the explore section.

Explore

A Pyspark script was created to answer question (2) in Dataproc. The spark.sql() function was used to select the first and last review for every product listing to estimate the product listing's lifetime:

```
# Get min and max review times for each asin
min_time_df = spark.sql("Select asin, MIN(unixReviewTime) as min_time FROM review
GROUP BY asin")
max_time_df = spark.sql("Select asin, MAX(unixReviewTime) as max_time FROM review
GROUP BY asin")
```

The product listing's lifetime in months was calculated by subtracting max_time from min_time and dividing by 259,200 seconds (i.e. 30 days). The total number of reviews during the product listing's lifetime was found using spark.sql():

```
# number of reviews in the product's lifetime
review_count_df = spark.sql("Select asin, count(distinct reviewText) as review_count
FROM review GROUP BY asin")
```

Lastly, the number of reviews for the first six months of a product listing's lifetime was found using spark.sql() by determining the number of distinct reviews where the review time occurred between the first review and up to six months after the first review:

```
six_mo_df = spark.sql("Select asin, count(distinct reviewText) as six_mo FROM review
WHERE unixReviewTime BETWEEN (min_time) AND (min_time+15770000) GROUP BY asin")
```

The same calculation was conducted to find the number of reviews that occurred during the second six months, second year, and from the third year to the fifth year. The figure 5 summarizes the output for the first twenty results. The pyspark data frame was saved to Big Query using the following code:

```
df.write.format('bigquery') \
.option('temporaryGcsBucket', 'cs512_bigquery') \
.save('final_project.edited_dataset')
```

After being saved to BigQuery, the pyspark dataframe was exported to Google Cloud Storage and saved to my local computer for use in R. The proportion of verified reviews was plotted against the rate of review accumulation over the first six months of the product's listing:

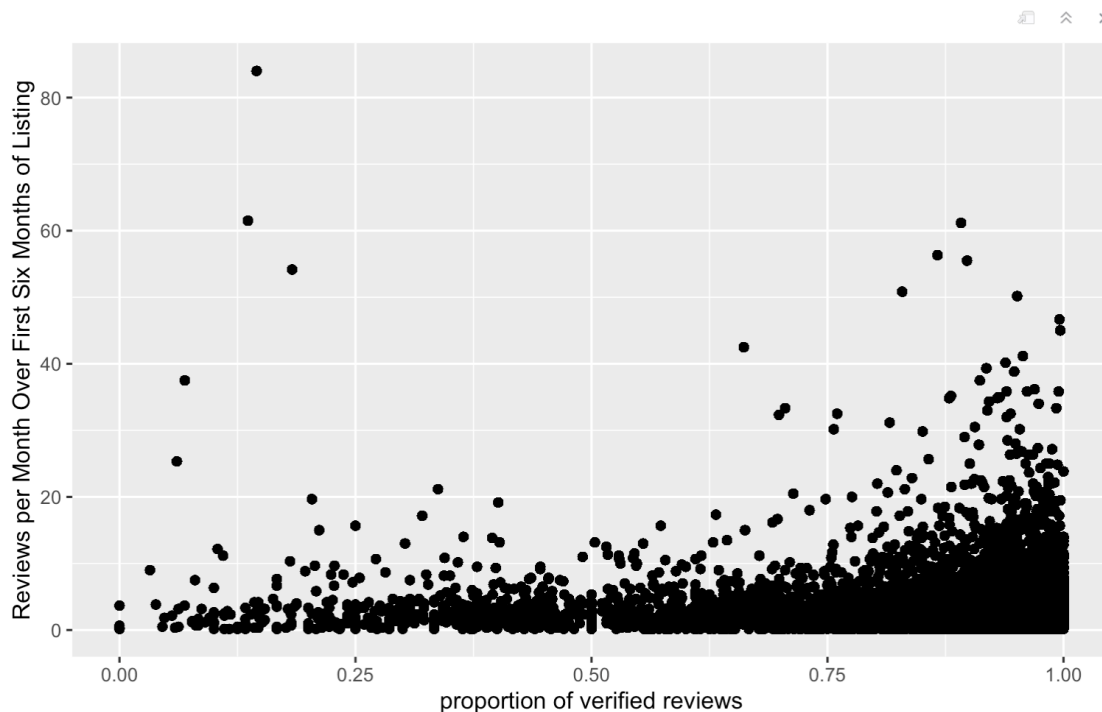


Figure 2. The rate of review accumulation over the product listing's first 6 months in reviews/month plotted against the product listing's proportion of verified reviews.

In general, it seems that product listings with more verified reviews tend to have higher initial rates of review accumulation. There are a few outliers that have high initial rates of review accumulation and a low proportion of verified reviews, these outliers are suspicious and may warrant further investigation. To compare how reviews accumulate over time for suspicious and trustworthy reviews (question 1), the outliers in the dataset were segregated from the most trustworthy product listings using the following code:

```
sus_table <- joint_table[(joint_table$sixprop > 30) & (joint_table$verprop < 0.25), ]  
trust_table <- joint_table[(joint_table$sixprop < 30) & (joint_table$verprop > 0.75), ]
```

Data points with a rate of review accumulation over 30 reviews per month and a proportion of verified reviews less than 0.25 were considered suspicious. Trustworthy data points had a rate of review accumulation that was less than 30 reviews per month and a proportion of verified reviews greater than 0.75. Rows were randomly sampled from the trustworthy dataset so that both datasets were equal in size. The frequency distribution of reviews for the suspicious and trusted product listings were plotted (Figure 3, Figure 4).

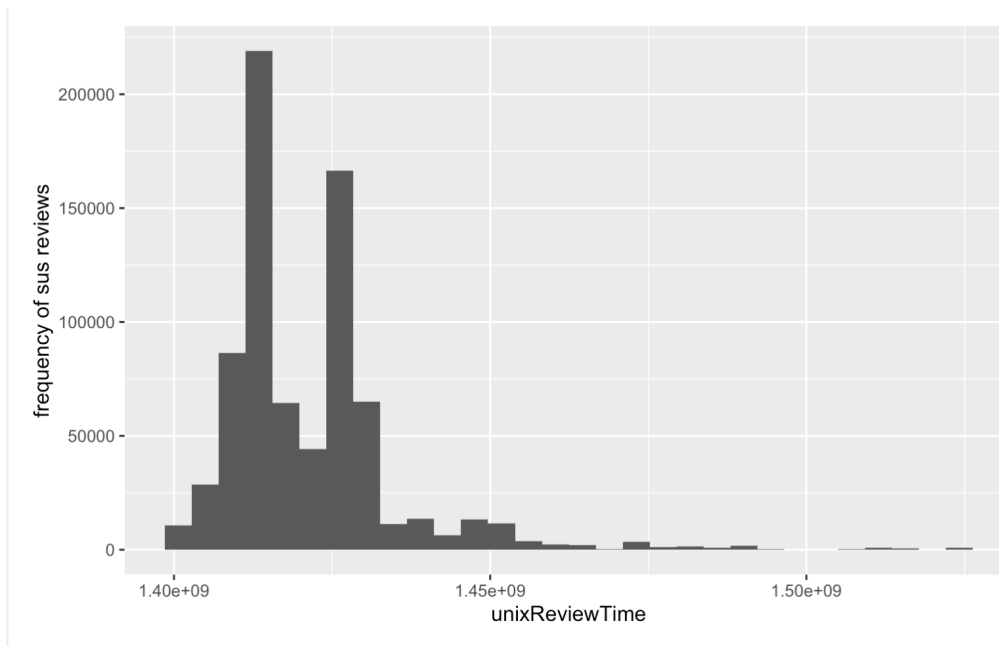


Figure 3. Frequency of suspicious reviews over time.

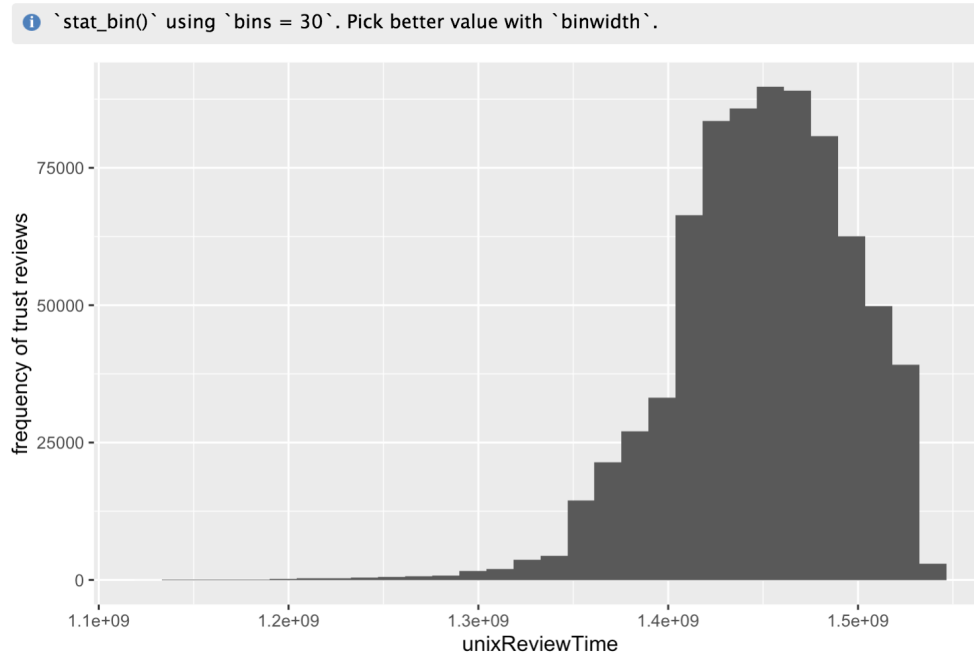


Figure 4. Frequency of trustworthy reviews over time.

It appears that the suspicious sellers had little reviews over the lifetime of their product listing but a few time frames with sudden influxes of reviews.

To answer the third question, a paired t-test was conducted to compare the mean overall score of reviews for verified and unverified reviews. First, a levene test was conducted to test for unequal variances among the variables using the following hypotheses:

$$H_0: \sigma_1 = \sigma_2$$

$$H_1: \sigma_1 \neq \sigma_2$$

Where σ_1 is the variance of the overall score for verified reviews and σ_2 is the variance of the overall score for unverified reviews.

The levene test was conducted with the `leveneTest()` function in R from the `car` package. The test resulted in a p-value of less than $2.2e-16$. At a significance level of 0.05, we reject the null hypothesis that there is no difference between the variance of the overall score for verified and unverified reviews. Since the variances were found to be unequal, we will use the Welch-Satterthwaite correction when we conduct the paired t-test.

The t-test was conducted using the `t.test()` function in R with `var.equal = FALSE` which adds the Welch-Satterthwaite correction.

```
t.test(final_table$overall, as.numeric(final_table$verified), paired = TRUE, alternative = 'greater',
var.equal = FALSE)
```

With 95% confidence, the mean overall rating for unverified reviews is, on average, between 2.469 and infinite units greater than the mean overall rating for verified reviews (p-value < 2.2e-16, paired t-test).

Output

LINE WRAP: OFF

asin	six_mo	sec_six	sec_yr	five_yr	review_count
1620921200	4	4	6	1	15
9861936831	6	5	3	1	15
B00006IFAQ	8	3	20	48	79
B0000AXNLT	1	7	14	24	46
B0000AY9W6	3	2	14	152	227
B0000AZ5ML	4	1	2	11	18
B0000DIWIM	1	1	1	15	40
B0000UZ528	2	5	11	12	30
B00026Z3E0	1	5	19	207	375
B0002MBKA0	1	1	1	14	18

Figure 5. The number of reviews for each product listing during the first 6 months, the second 6 months, the second year, and from the third year to the fifth year. The last column summarizes the total number of reviews that were received in the product listing's lifespan.

Model

Models were not used in this project because they were not required to answer the questions of interest. In future analysis, it would be interesting to create a predictive model for identifying fake amazon reviews. A regression model may be fit for this future project.

Interpret

To address the key questions of this analysis:

1. Do products with high initial rates of review accumulation have less verified reviews?

The relationship between review accumulation and verification proportion seemed to follow a positive linear trajectory: product listings with a high proportion of verified reviews tended to have higher initial rates of review accumulation (Figure 2). For data points that strayed from the linear trajectory, data points with exceptionally high initial review accumulation rates also had a low number of verified reviews (Figure 2). These outliers were deemed suspicious and they were subsetted from the data for further analysis.

2. How do reviews accumulate over the lifetime of a product listing for suspicious reviews? For trustworthy reviews?

The data that were subsetted for question (1) were investigated further by plotting the frequency of reviews against time (Figure 3, Figure 4). Trustworthy reviews tended to follow a distribution that was close to normal. Reviews accumulated exponentially at first but slowed quickly after they reached their peak (Figure 4). Suspicious reviews had a bimodal distribution where the majority of reviews were occurring over the same period of time (Figure 3). This is congruent with the assumption that fraudulent companies may be paying individuals to write reviews or writing reviews themselves which would cause most of the reviews to be written within the same period of time.

3. Do unverified reviews tend to have a more positive sentiment than verified reviews?

To understand if unverified reviews tend to be more positive, on average, than verified reviews, a paired t-test was conducted to assess the following hypotheses:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

The t-test found that unverified reviews are, on average, 2.47 units greater than verified reviews (paired t-test, p-value < 2.2e-16).

There were a few limitations to the analysis that should be considered. The lifetime of the product listing was calculated as the difference in time between the first review and the last review. The first review could have occurred a long time after the birth of the product listing. Additionally, we are only assuming that the reviews were randomly selected so we don't know if it's appropriate to extend our inference to all amazon reviews or all reviews in general. Furthermore, it's impossible to know if the unverified reviews are actually fake and there are many ways that companies can fabricate verified reviews. However, considering the limitations, it seems that a good starting point to create a predictive model for identifying fake amazon reviews is assessing their rate of review accumulation and proportion of verified reviews. Other variables may be useful in prediction like the reviewer ID and product listing ID.

Sample of Records

- Initial data (<1MB, also included in the zip file)

```
{ "overall": 4.0, "verified": true, "reviewTime": "03 29, 2016", "reviewerID":  
"AIE8N9U317ZBM", "asin": "0449819906", "style": { "Format": " Kindle Edition"},  
"reviewerName": "Zelmira, Ph.D.", "reviewText": "Contains some interesting stitches.",  
"summary": "Four Stars", "unixReviewTime": 1459209600}  
{ "overall": 5.0, "vote": "18", "verified": true, "reviewTime": "08 12, 2015",  
"reviewerID": "A3ECOW0TTLH9V6", "asin": "0449819906", "style": { "Format": " "
```



```

Paperback"}, {"reviewerName": "Dangerous when Cooking", "reviewText": "I'm a fairly
experienced knitter of the one-color or color block intarsia vein, rather than a Fair
Isle maestro, and what I loved best about this stitch guide is the multitude of
reversible stitch patterns offered and shown reverse and obverse. If you knit and love
to accumulate guides, stitch dictionaries, pattern books and design-your-own project
books, this is a great resource. I find I'm always adapting knitting patterns slightly
or significantly to swap out cables, add interesting borders so I can knit edges and
body at the same time, what have you.\n\nThis gives you enough classic stitches to
satisfy but its strength is in fresh twists on the usual or entirely new (at least to
me) options in textured, lace, cables&cross stitches, slip st., and novelties. As
others note, the stitches are arranged from simplest to most challenging in each
section-- also a great help when deciding how much sweat and tears I'm willing to
expend.\n\nThis also does not frustrate me in the ways too many other books and guides
do.\nHere, Leapman uses the symbols common in knitting magazines and most books on
knitting you've seen. Yay!\nOne of my peeves with some designers and guidebooks is the
use of their own symbology for charting various stitches. Alice Starmore leaps to mind
as a woman living in her own private Idaho filled with her own Runic symbols. I have
to translate her every chart's cable squiggles into the symbols I'm more familiar with
just to get the thing going. Gorgeous end results, but geez.\n\nThis is concise but
full of options, so I FIND something nifty quickly and easily.\nWhile I do love
browsing Barbara Walker's stitch books, sometimes I'm just looking for a simple option
and don't want to spend all day combing through umpteen volumes, each of which has its
own section for lace, k/p, cables, slip st.s, panels, etc. I wish Walker would compile
all her lace, all cables, all k/p, all color work, all panels, etc into huge sections
in just ONE encyclopedia.\n\nUntil that happens, this is my new go-to for fast,
interesting ideas to enhance my knitting.\n\nThis is not a great resource for
traditional knitting stitch patterns, such as gansey (guernsey), it's perhaps best for
'modern' classic knitting.", "summary": "My current favorite go-to guide for
inspiration", "unixReviewTime": 1439337600}
{"overall": 4.0, "vote": "3", "verified": true, "reviewTime": "04 5, 2015",
"reviewerID": "A278N8QX9TY2OS", "asin": "0449819906", "style": {"Format": "
Paperback"}, {"reviewerName": "Just us", "reviewText": "Great book but the index is
terrible. Had to write and high light my own cross ref info.", "summary": "lots of
great examples, good instructions, color pics", "unixReviewTime": 1428192000}
{"overall": 5.0, "verified": true, "reviewTime": "10 11, 2014", "reviewerID":
"A123W8HIK76XCN", "asin": "0449819906", "style": {"Format": " Kindle Edition"},

```

The Python script for running the dataproc job is attached. The output from the job is shown in Figure 5.

Sources

¹Jianmo Ni, Jiacheng Li, Julian McAuley Empirical Methods in Natural Language Processing (EMNLP),
2019