Sephora Fake Reviews Analysis

Group Name: Shake Shack
Su Liang, Lyu Lin, Ziye Meng, Tianling Wang

Introduction

A Facebook study shows that 46% of make-up shoppers discover products online because they could get more detailed information and reviews from other shoppers, which could largely influence their purchasing decisions (Facebook IQ, 2018). However, as the competition in beauty products becomes increasingly fierce, merchants are eager to boost their product ratings through reviews. In our project, we analyze the product reviews on Sephora, one of the most popular beauty products websites in the United States. It turns out that a large proportion of the reviews on Sephora are actually fake and our model result shows that this number is as high as 10%. Note that the fake reviews here refer to reviews written by professional review writers who are paid by product merchants to attract potential customers.

The purpose of this project is to identify fake reviews for products on Sephora website and address the following issues:

1. What is the difference between the rating distribution of all reviews and that of fake reviews?
2. After fake reviews are removed, how does the rating of a product change?
3. What is the relationship between the number of total reviews and fake reviews?
4. Are there any patterns between the number of fake reviews of one product and its total rating over time?

Getting Data

Since we want the latest reviews from Sephora, we decided to scrape all skincare product reviews from its website. The most technical difficulty we encountered was that Sephora website is lazy loaded, meaning that the source page of the reviews will not be fully loaded unless all reviews are visited. In order to load all reviews, one has to scroll down to the reviews section of a product web page and recursively click the 'Show more 6 reviews' button. The web-scraping technique taught in class does not apply to this situation since it only deals with static web page. Therefore, after doing some deep research, we used Selenium to create a webdriver to automate this interaction process. In our web-scraping code, we utilized ActionChains from the Selenium package to generate automatic mouse clicks and other movements so that the 'Show more 6 reviews' button can be recurrently clicked. A short video has been included to illustrate this process. Due to the large volumes of reviews, we spent one month scraping reviews, and approximately 800,000 reviews were collected.

Data Description

In total, there are 791,735 reviews from 3,400 products in our raw dataset. Because our analysis and models apply to all categories, we pick Moisturizers to do our analysis as this category is the largest category (Fig#1). We have two types of datasets: Product Information Data and Reviews Data. For Moisturizers, we have 237,300 reviews and 903 products.

Product Information Data Columns (8 in total):

▫ product category ▫ sub-category brand name ▫ product name ▫ product item id
▫ loves (the number of people who love this product) ▫ price ▫ product total rating

Reviews Data Columns (14 in total):

▫ product sub-category ▫ product item id ▫ number of total reviews for the product

▫ username ▫ user skin type ▫ user skin tone ▫ user age range ▫ review title ▫ review content
▫ review rating ▫ helpfulness (number of people who think this review is helpful)
▫ unhelpfulness ▫ recommend (whether the user would recommend this product to others)
▫ review time ▫ free product (whether the user received a sponsored free sample).

Label Fake Reviews

1. Reviews with Same Contents but from Different User ID(s)

   Because spammers are usually required to achieve some specific review volumes assigned by their employers, they may post duplicate reviews using different user ID to complete their jobs on time and reduce their workload (Chowdhary & Pandit, 2018). As a result, these reviews should be detected as fake.

2. Reviews with High Similarities to the Duplicated Review Contents

   Since professional review writers would have similar writing styles, we extract reviews that are highly similar to the abovementioned duplicated reviews. Using the duplicated reviews contents, we build a corpus and apply Latent Similarity Indexing model to calculate the similarity scores of other remaining reviews to the existing corpus. We label reviews with more than 90% similarity to the corpus as fake.

3. Multiple Reviews from One Reviewer for One Specific Product

   It is also unusual for one reviewer to post multiple reviews for a product, and the motivation behind this is questionable (Chowdhary & Pandit, 2018). Therefore, we identify these kinds of reviews as spam.

4. Reviews with Extreme Sentiment Score from Unpopular Products

   Suspicious reviews usually have extremely positive or negative sentiment due to merchants' demand of increasing their own product ratings or decreasing their competitors' product ratings (Chowdhary & Pandit, 2018). However, popular products with high review volumes might have many positive reviews from genuine users. In order to reduce the probability of mislabeling, we only look at products with review volumes below the 25th percentile or loves below the 25th percentile. We implement weighted sentiment analysis using Valence Aware Dict Sentiment Response (VADER) to calculate the compound sentiment score for each review from unpopular products. Since the compound score ranges from -1 to 1, reviews with an absolute compound score over 0.8 are deemed as fake.

5. Reviews Mentioning Brand Name

   Some reviews mention the product's brand multiple times because the reviewer participates in marketing events of the brand and receives a complimentary product for review purposes. There are also reviews comparing the product with competitors' to boost or damage the reputation of the product (Chowdhary & Pandit, 2018). We obtain the brand list of products under the same category from the product dataset and search for each brand in every review to find whether a review mentions a brand name. A mismatch could occur if the brand name appears in a common word. For example, the brand name 'tarte' would appear in the word 'started' during the search. Therefore, we add whitespace before and after the brand name to improve accuracy for searching. After searching, we find that 87.8% of reviews do not mention any brand, 11.2% mention a brand name once, and a brand is at most mentioned 8 times in a review. We decide that a review is fake if it mentions a brand name more than or equal to 2 times (Fig#2).

6. Reviews with Typos

We use the package pyspellchecker to find typos in a review. The first step is splitting a review into a list of words using regular expression (regex). The reason for using regex is that we do not want words containing numbers, like '30ml', because pyspellchecker would identify these words as misspelled. After implementing pyspellchecker, we get the list of detected misspelled words for each review. We further remove abbreviations in the typo list because pyspellchecker sometimes process abbreviations improperly: it would identity 'can't' as correct, but 'i've' as misspelled. Based on the filtered list, we calculate the ratio of typos for every review. If the typo ratio of a review exceeds the 95% quantile (i.e. more than 6.25% of words in the review are typos), we label it as fake (see Fig#3).

7. Time Interval Analysis

Since fake reviews tend to present in short periods where review number increases rapidly (Fornaciari & Poesio, 2014), we carry a weekly analysis on review numbers for each product. By visualizing time series plots for some skincare products, we find that there exist rapid increases in review volumes in certain weeks. Fig #4 in the Appendix shows a time series plot of weekly review numbers for one product.

Therefore, for each product we define a threshold as 1.96σ above average weekly review counts for that product. Weeks with review number above that threshold will be categorized as rapid growth. We then extract the reviews in those weeks and apply the following two criteria:

1) Reviews mentioning competitor brands are grouped as fake. This criterion further improves criteria 5, here we only analyze the reviews in weeks with rapidly growing review numbers, and a review is counted as fake when competitor brands are mentioned at least once, instead of twice. The function for this criterion returns a two-element tuple: the first element is the review, and the second element is a list containing competitor brands that are mentioned in the review. (e.g. ('it has completely transformed the dry areas of my skin especially on my chin area, a place where my other daily moisturizer from clarins wasn't doing the trick...', ['clarins'])).

2) Look at the 'not helpful' versus 'helpful' in each review, if more than 70% of people consider the review as unhelpful, then the review will be grouped as fake.

Finally, we highlight the weeks that could possibly have some fake reviews in the time series plot, as shown in Fig #5 in the Appendix. The result shows that weeks with rapid review growth are very likely to contain fake reviews.

Machine Learning Models

Two Methods:

1. Without TF-IDF

To better describe the content of the review text, we add several columns as features, including whether the review has a title, the complexity score of each review (length of text, average characters per word, average words per sentence, unique vocabulary percentage), and the user information integrity (whether the user gives her information about her skin type, skin tone, and age).

In total, we have 10 features as input: total_reviews, review_rating, free_product, recommendation, have_title, len_of_review, avg_chars_per_word, avg_words_per_sentence, unique_vocab_percentage, information_integrity. And the target result is the labeled fake reviews.

2. With TF-IDF

We want to further explore whether the frequency of each word will influence the prediction result. Therefore, we use TF-IDF (Term Frequency - Inverse Document Frequency) to represent the relative frequency of each word, besides the original 10 features. To reduce feature dimensionality, we remove those words that only appear once.

We test both methods using Random Forest with a small subset of the Moisturizer category, which is the Decollete and Neckcream subcategory. The accuracy of both are nearly the same at around 0.9. The feature importance graph Fig #6 of the TF-IDF method shows that the frequency of each word has limited predictive power, compared with the feature importance Fig #7 without TF-IDF where total_reviews and avg_chars_per_word have more than 0.2 importance.
Since the TF-IDF method is both time and space consuming, we decide to choose the 10 original features as the model input for the whole Moisturizer category dataset.

Seven Models:

We apply 7 classification models to predict fake reviews, including Logistic Regression, Random Forest, Bagging, Neural Network, K-Nearest Neighbors, Multinomial Naive Bayes, and Linear Support Vector Machine. The test accuracy of these 7 models is quite similar ranging from 0.89 to 0.92, among which Bagging gives the best accuracy score of 0.92. The radar graph shows the accuracy of each model Fig #8.

We further apply the Bagging model to the whole dataset and use the prediction result to relabel fake reviews for analysis.

Output Analysis

We use the graphing library plotly to visualize our results because it allows interactive charts. Issues raised in the introduction are answered in this section:

1. What is the difference between the rating distribution of all reviews and that of fake reviews?

   According to Fig #9, the portion of extreme ratings (1-star and 5-star) in fake reviews are much higher than that in all reviews. Only 63.2% of all reviews have a 5-star review rating while 81.9% of fakes reviews have a 5-star review rating. This can be addressed by the motivation of merchants: they want to elevate their product total ratings through 5-star fake reviews and damage the reputation of their competitors through 1-star fake reviews.

2. After all fake reviews in a product are removed, how does the rating of a product change?

   The answer is YES. Fig #10 shows that after we removed all fake reviews, 71% of products (512 out of 903) rating decreased, which can be explained by the heavy portion of 5-star ratings in fake reviews.

3. What is the relationship between the number of total reviews and number of fake reviews?

   Since the total review number is the most predictive feature, we further analyze how it influences the fake review result. Fig #11 in the Appendix is a scatterplot of the number of fake reviews against the number of total reviews for all moisturizer products. We can see a positive relationship between the two variables with a correlation of 0.51. This indicates that the more

reviews a product has, the higher the probability that they "recruit someone" to write batches of reviews on purpose, which are labeled fake.

4. Are there any patterns between the number of fake reviews of one product and its total rating over time?

   We have discovered **THREE** patterns:

   1) When the rating of a product is stabilized over time, the number of fake reviews of this product tends to decrease. Fig#12 and Fig#13 shows this pattern from two different products.
   2) There is a rapid growth in product rating when the number of fake reviews increases over a short period. This trend again reflects that merchants pay for full star reviews to uplift their product ratings. Fig#14 and Fig#15 demonstrates this pattern.
   3) After the rating of a product decreases, the number of fake reviews increases until the rating of this product reaches a stable and high rating. Fig#16 and Fig#17 illustrates this pattern.

Limitations

1. The LSI model does not take the order of words into consideration (it breaks the order when building a corpus). If we have more time, we can use a technique called 'Deep Siamese Text Similarity' to calculate semantic and structural similarity.
2. Some spammers might modify part of review contents they have written to create a 'new' review. And our analysis does not include those partially overlapped reviews because comparing all reviews to each other is time-consuming.
3. Criteria 5: We only consider in English here. Some reviews are in Spanish and are highly likely to be identified as misspelled by the pyspellchecker. Taking reviews of other languages into consideration will further improve the accuracy of detecting typos.
4. Criteria 7: We only examine the reviews in the weeks that are defined as rapid-growth weeks. Therefore, this criteria is biased in the sense that it ignores fakes reviews that could present in weeks with small review numbers.
5. There are three outliers in the Number of Total Reviews VS Fake Reviews figure (i.e. most of the reviews are labeled fake). This may result from imperfections of our label criteria.

Bibliography

Chowdhary, S., & Pandit, A. (2018). Fake Review Detection using Classification. International Journal of Computer Applications, 180(50), 16–21. doi: 10.5120/ijca2018917316

Fornaciari, T., & Poesio, M. (2014). Identifying fake Amazon reviews as learning from crowds. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. doi: 10.3115/v1/e14-1030

Understanding the Beauty Shopping Journey of the Connected Consumer. (2018, October 15). Retrieved from https://www.facebook.com/business/news/insights/understanding-the-beauty-journey-of-the-connected-consumer.

Appendix

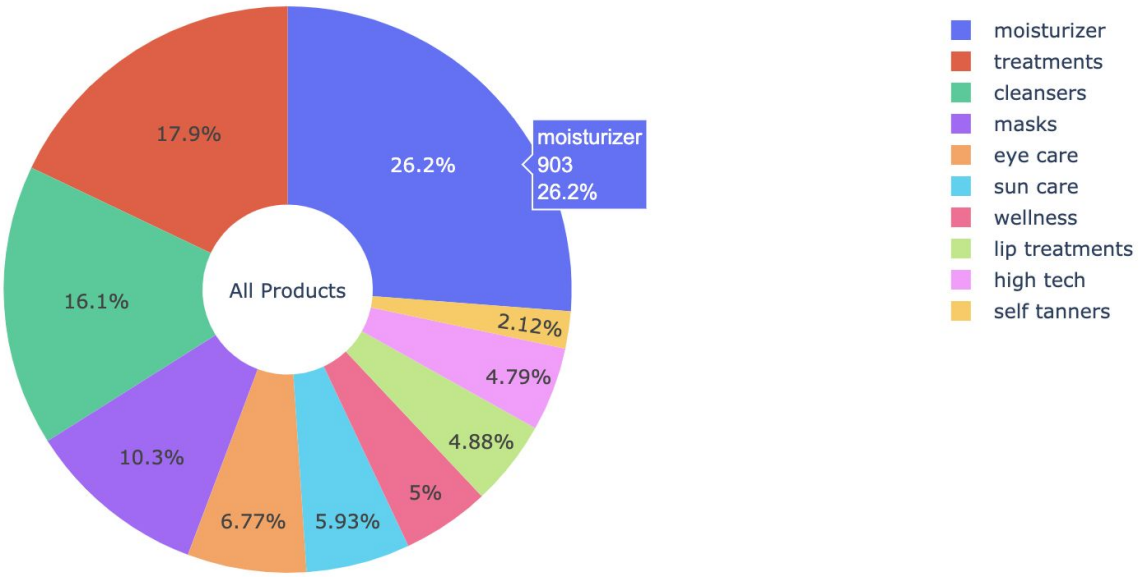

Figure 1. Category Distribution of Skincare Products ↑

user_name	title_and_review	brand	mentions	counts	review_rating
kristinareyes	belif me this is the bomb just got my @belifusa voxbox from @influenster to tryout #complimentary tonight! this aqua bomb sleeping mask is great for dry, dull, and rough uneven skin. #bedtimewithbelif #belifusa #influencer #belifaquabomb #belifaquabombsleepingmask #excited #skincare #skincarejunkie #facemask #sleepingmask #bedtimewithbelif #contest #complimentary @belifusa @influenster	belif	belif	8	5

Figure 2. A Fake Review Labeled by Brands Mentioning ↑

user_name	skin_type	skin_tone	age_range	review_title	review
nwyeung	Normal	Light	NaN		Hdhsushbsjabzghxjabaz

Figure 3. A Fake Review Labeled by Typo Detection ↑

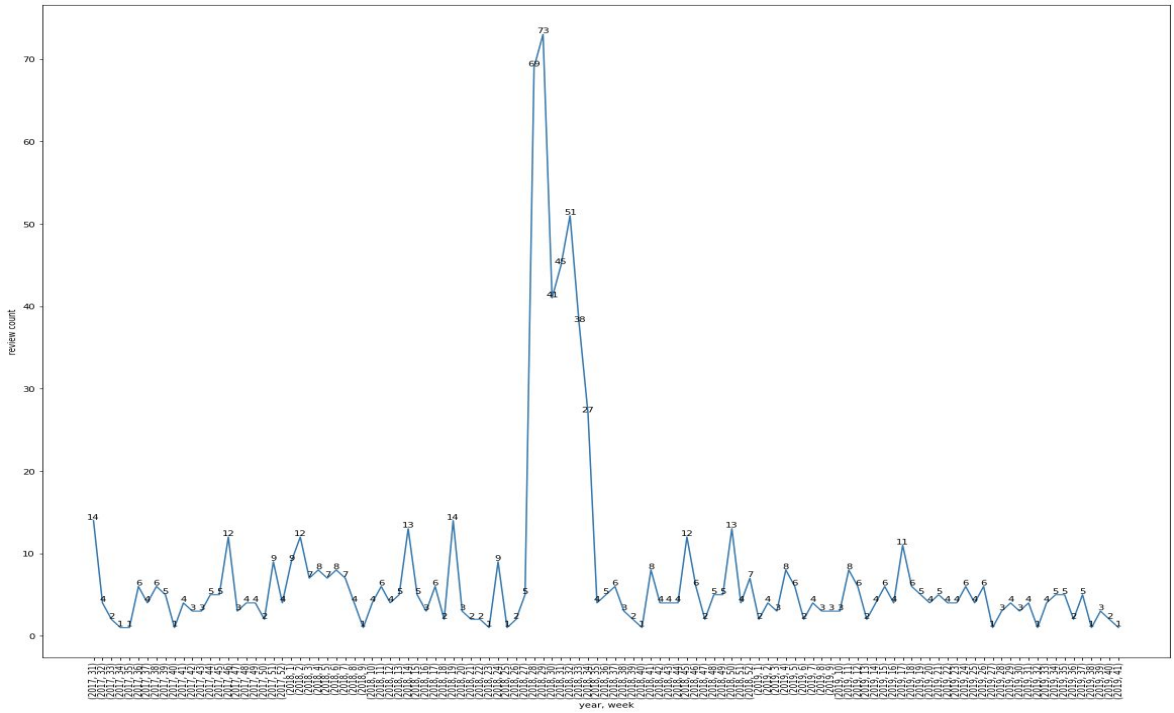

Figure 4: Time series plot of weekly review numbers for one product (KIEHL'S SINCE 1851 Ultra Facial Cream)

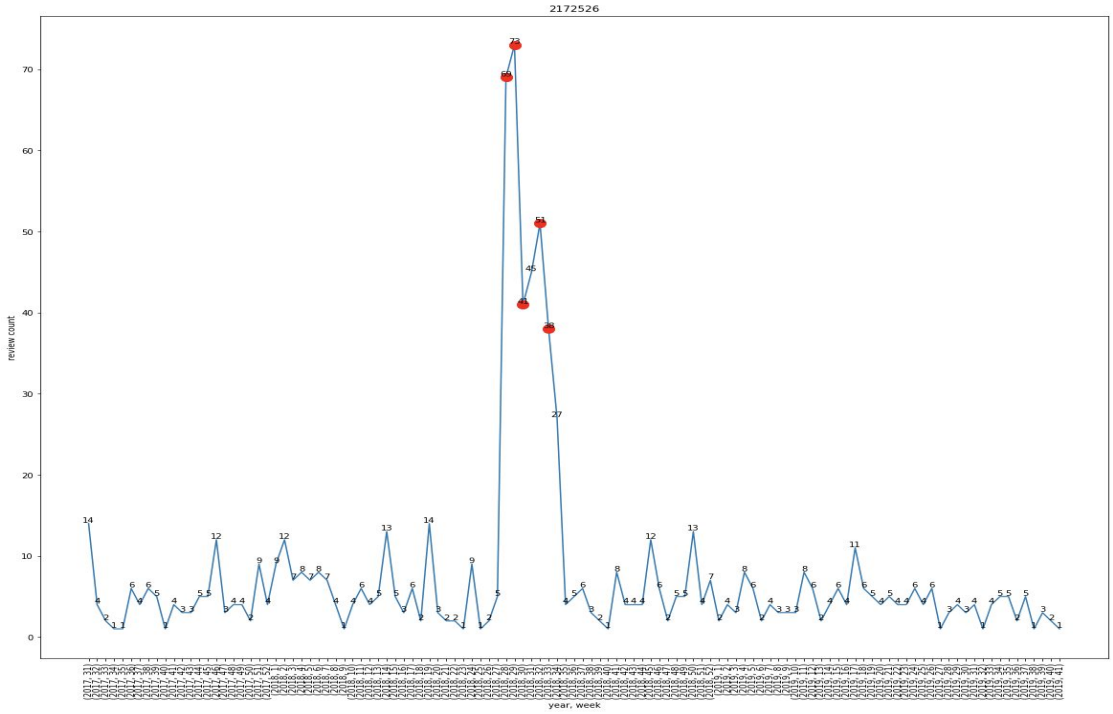

Figure 5: Time series plot of weekly review numbers with fake review labels ↑

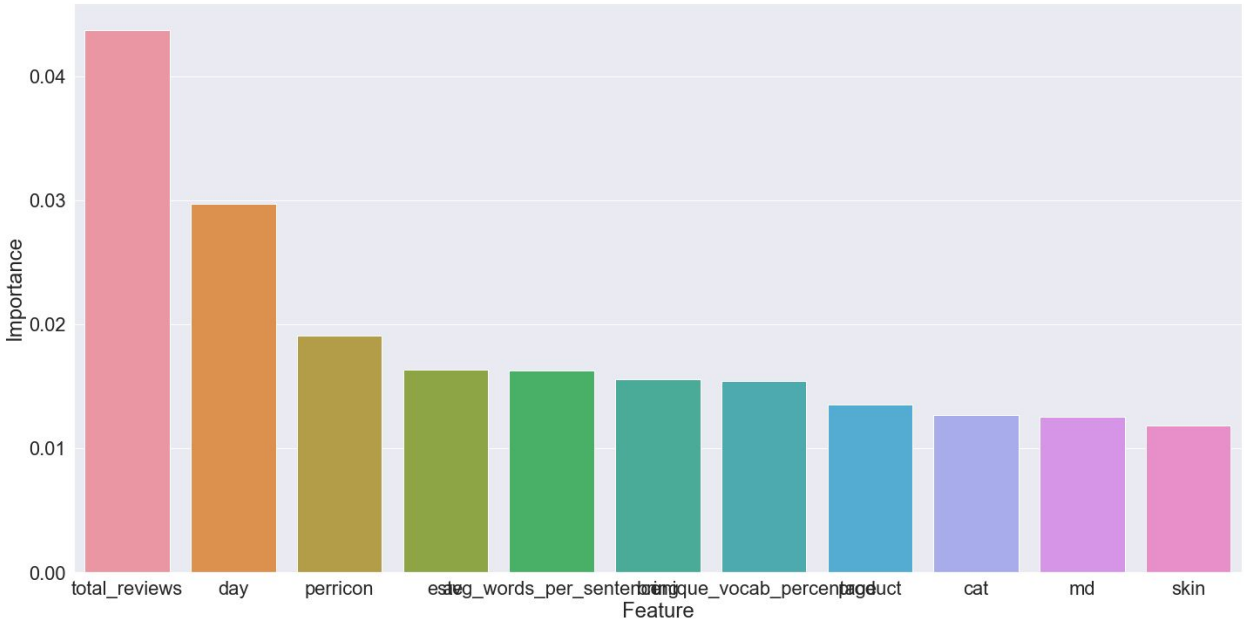

Figure 6. Importance of Features with TF-IDF ↑

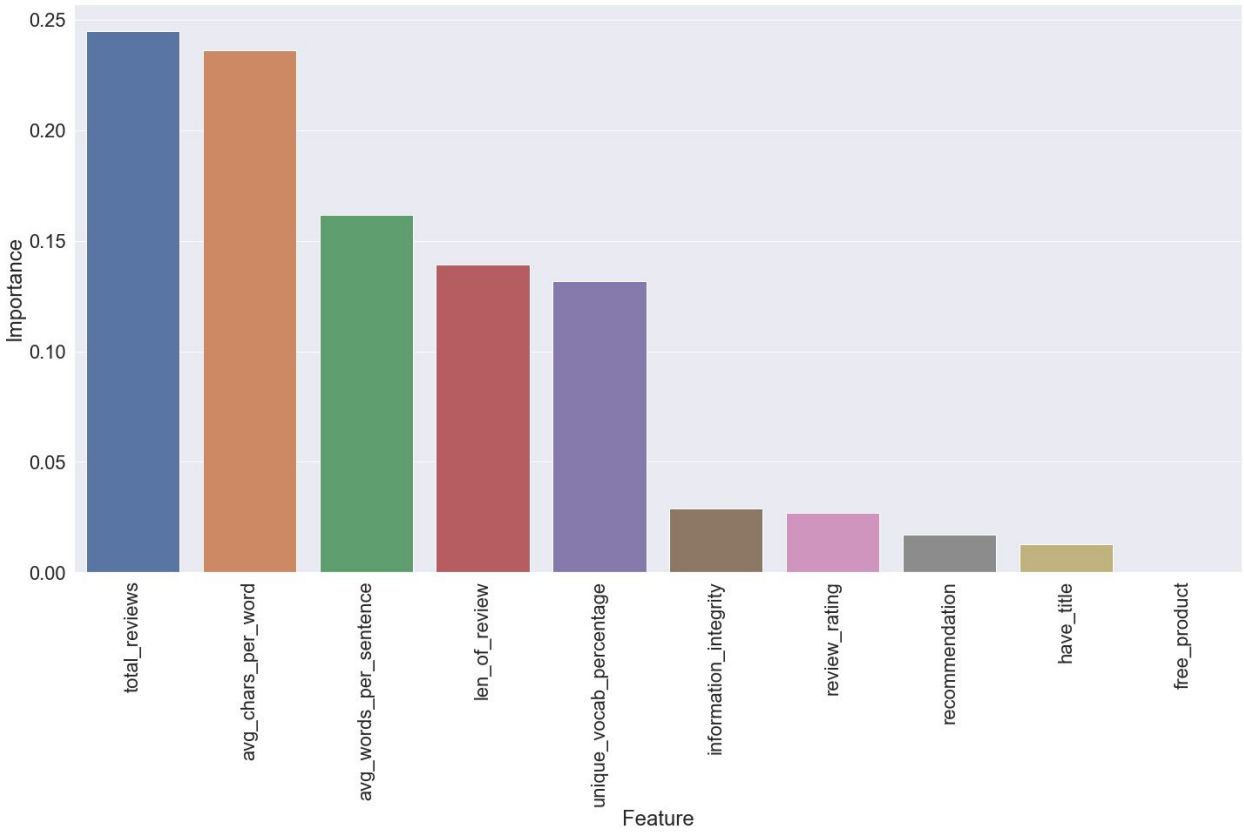

Figure 7. Importance of Features without TF-IDF ↑

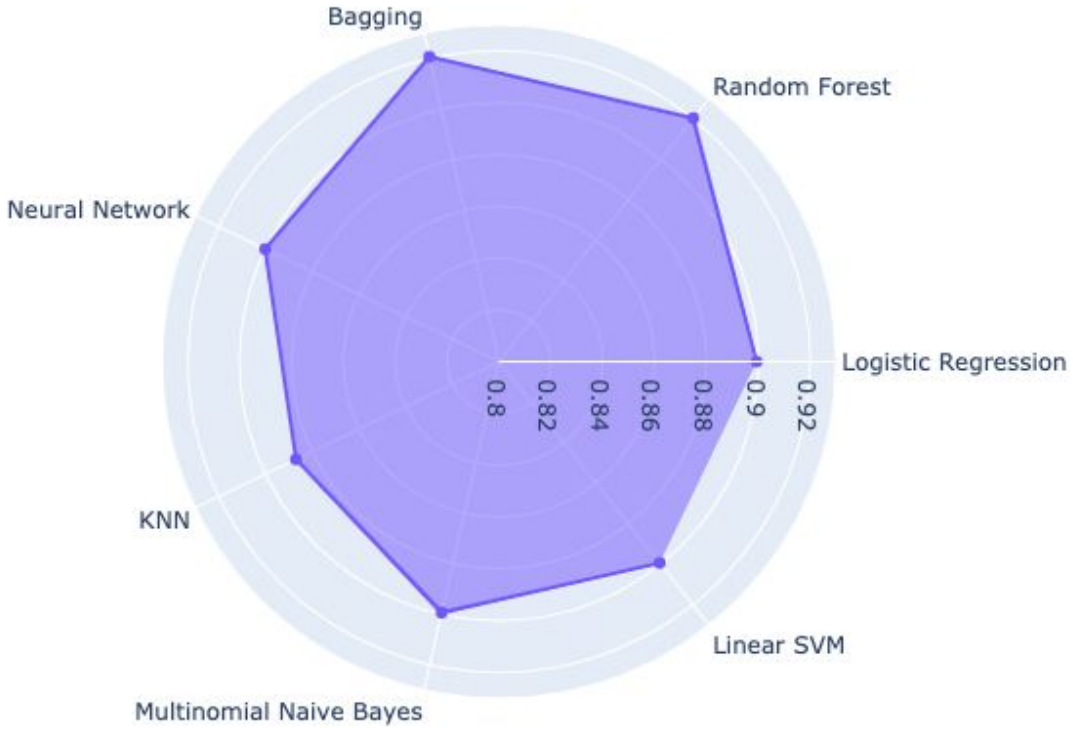

Figure 8. Model Accuracy Comparison ↑

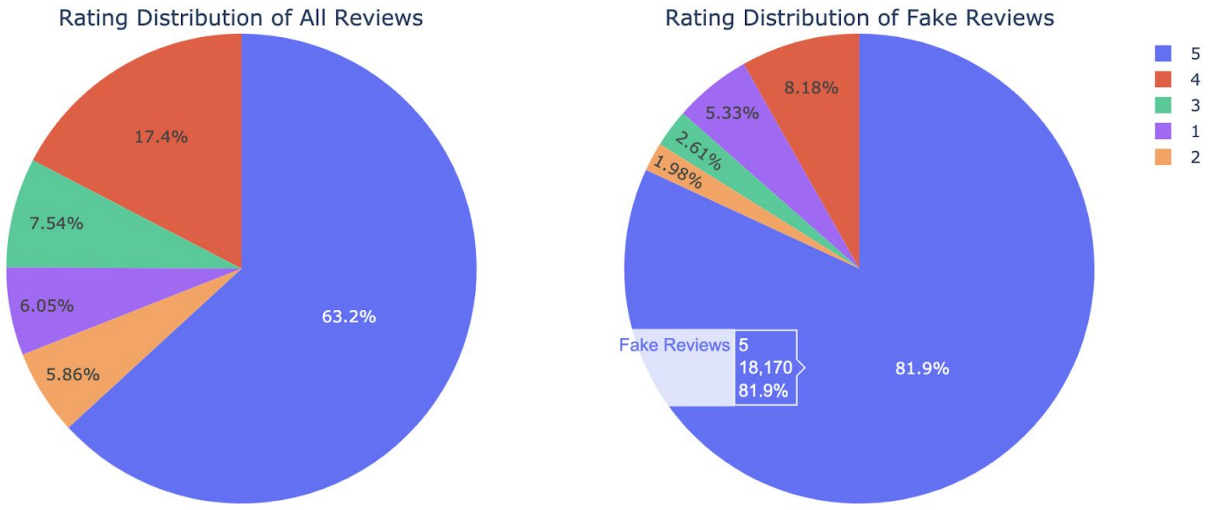

Figure 9. Rating Distribution of All Reviews VS Labeled Fake Reviews ↑

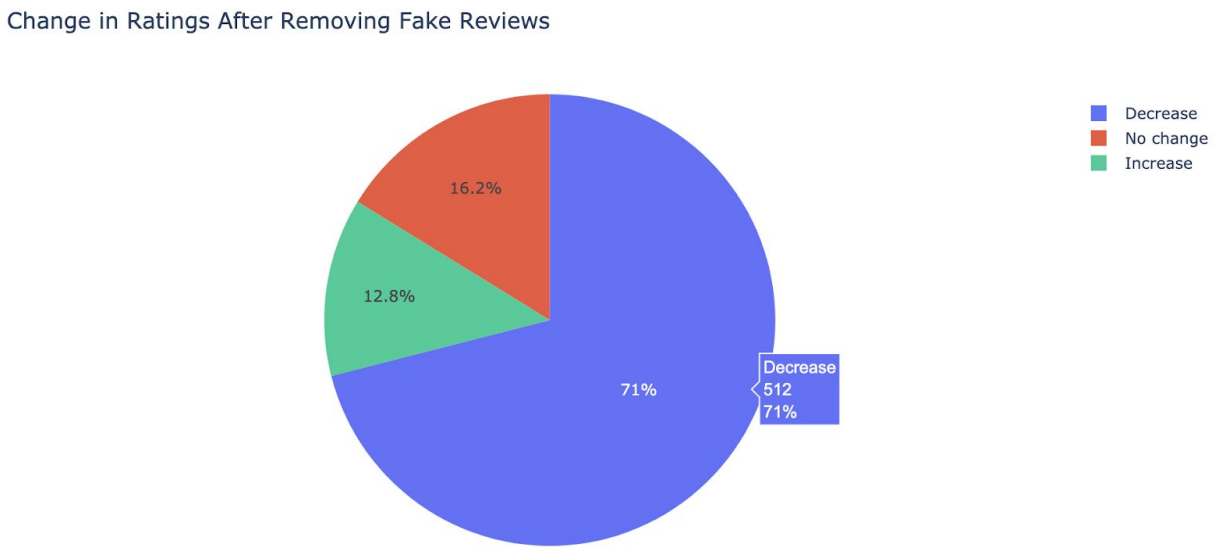

Figure 10. Change in Ratings after Removing Fake Reviews ↑

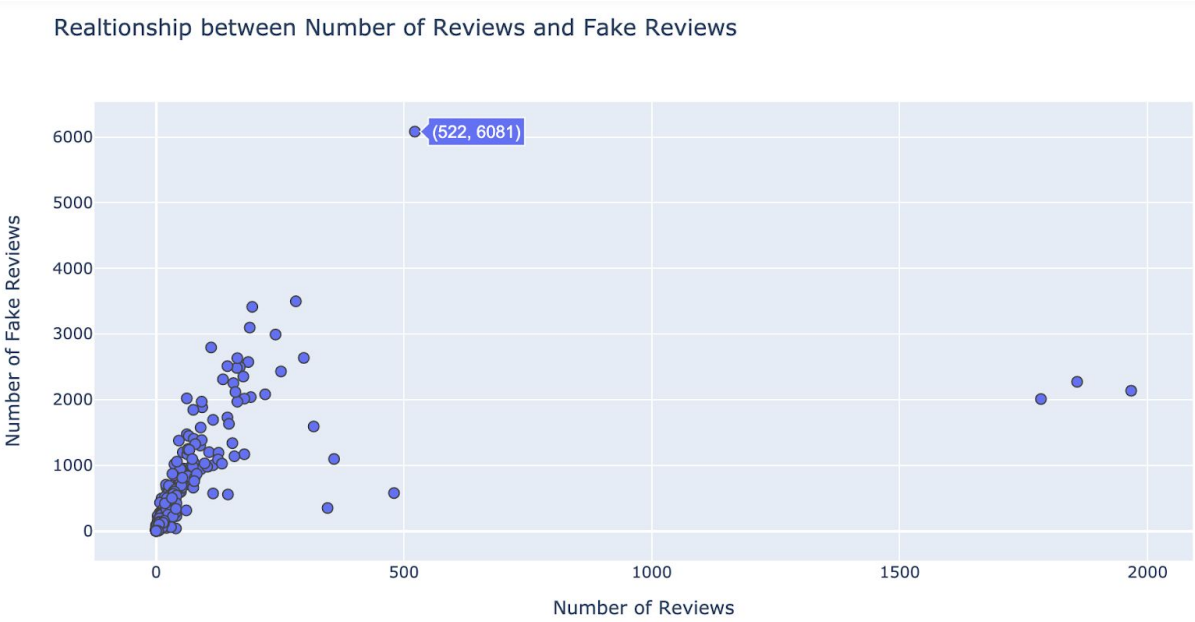

Figure 11. Scatterplot of Total Number of Reviews VS Fake Reviews ↑

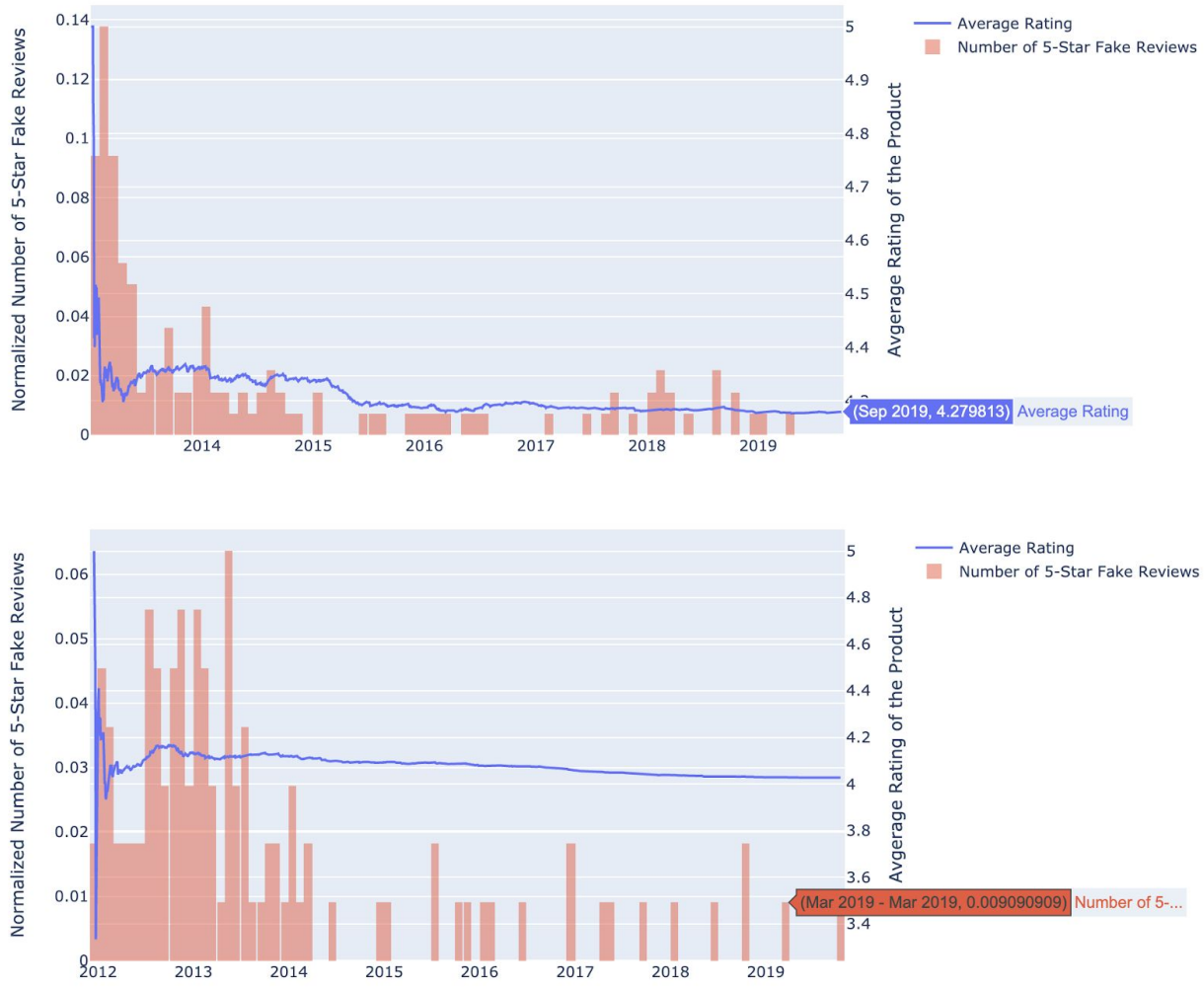

Figure 12&13. Decrease of Fake Reviews in Accordance with Stabilization of Rating ↑

Figure 14&15. Boost in Rating after Large Increase in 5-Star Fake Reviews↑

Figure 16&17. Fake Reviews to Increase Rating after Decreases in Rating ↑