Identifying Fake Amazon Beauty Product Reviews Using Classification in Natural Language Processing

Keerthana Manivasakan, **Kevin Chen** and **Kieren Gill**
New York University
Computer Science Department
New York, NY
{km4855|klc8452|ksg7699}@nyu.edu

Abstract

When purchasing items online, customers typically rely on reviews and ratings provided by previous buyers. However, there has been an increasing number of fake reviews, many of which are hard to distinguish from actual reviews. In this paper, we compiled review data from beauty products on Amazon, which we use to explore two things. First, we delve into the different features of online reviews, identifying the best features to use when training machine learning models to distinguish genuine reviews from fake reviews. Second, we explore if a model trained on an automatically labeled dataset can predict fake reviews on a manually labeled dataset. Results suggest that the model we developed with fewer features performs nearly as well as other models with more features.

1 Introduction

Online shopping provides individuals with the convenience of having anything from groceries to clothing items delivered to their doorstep without ever having to leave the comfort of their homes. However, one downside comes with having our retail needs digitally satisfied: consumer uncertainty. The intangibility of items on a digital display makes it difficult for consumers to discern a product's authenticity, quality, and fit.

In an effort to circumvent this issue, most online retailers allow customers to provide public feedback on products and services through ratings and reviews. Ratings are usually measured on a scale of one to five stars, with a one-star rating being the worst and a five-star rating being the best. Reviews are typically written in natural language and can include supplementary images. While ratings are supposed to help quantify the value of a product, reviews also allow consumers to provide more qualitative feedback, and share their personal experiences with their purchased products or services. Hence, many customers rely on consumer reviews to fill in this gap in information.[1] Seeing a positive review gives customers a sense of insurance and trust that they will be satisfied with their purchase. On the other hand, if the buyer sees a negative review, they are likely to be discouraged or skeptical of purchasing the product.

As companies begin to realize the influence reviews can have on consumers, the importance of having a positive reputation is more desirable than ever. This growing desire has led to an increasing number of companies are generating fake reviews for their products to elevate their online presence. Here we define fake reviews as "deceptive reviews provided with an intention to mislead consumers in their purchase decision making, often by reviewers with little or no actual experience with the products or services being reviewed."[2] This is not to be confused with the incentives that many businesses provide for consumers to leave reviews on their products. While these reviews are somewhat inorganic, we still consider them authentic because the consumer has purchased the product or service.

Due to the increasing number of fake reviews, identifying a fake review is now something online shoppers have to worry about. In this paper, we use features of product reviews to develop our own classification model to identify fake reviews and compare our model to other existing models. We

[1] https://www.qualtrics.com/blog/online-review-stats/

[2] https://www.tandfonline.com/doi/full/10.1080/07421222.2016.1205907

also investigate the differences between models trained on an automatically labeled dataset and models trained on a manually labeled dataset.

The data and NLP[3] systems detailed in this paper were developed specifically for beauty products sold on the widely used ecommerce website, Amazon.com.

2 Related Works

2.1 Classification for Hijacked Reviews

Hijacked Reviews are when sellers commit fraud by "hijacking" product pages that have previously accumulated positive reviews from other products. Essentially, this is the practice of review reuse to inflate the credibility and trustworthiness of a product. A twin LSTM[4] neural network and a BERT[5] sequence pair classifier have both been tested on identifying hijacked reviews (Daryani and Caverlee 2021, 70), with the BERT classifier yielding better results. Hence, we opted to use a classifier instead of a neural network when designing our model.

2.2 Flaws in Majority Baseline Approach

In many supervised learning methods, it can be unfeasible or extremely expensive to label data objective and reliably. Hence, researchers often collect labels from multiple experts or annotators. However, this can lead to disagreements, especially when there is no gold standard for the dataset. The learning from crowds approach proposes a probabilistic approach that evaluates the different experts and also gives an estimate of the actual hidden labels (Raykar et al. 2010). This method has been used to label fake Amazon book reviews (Fornaciari and Poesio 2014), and has shown greater success than the commonly used majority voting baseline. Based on this information, we opted not to use the majority voting baseline, but instead developed our own annotation baseline for our manually labelled dataset.

2.3 Types of Spam Reviews

Spam reviews can be classified into different three different types (Jindal and Liu 2008). Type 1 reviews are untruthful opinions, and are written with the intention of misleading consumers or opinion mining systems. Type 2 reviews focus only on the brand or manufacturer of the product, and not the product itself. Type 3 reviews are non-reviews, and are typically either advertisements or irrelevant information. We used these descriptions to select Type 1 spam reviews for our manually labelled dataset.

2.4 Deceptive Features in Reviews

Information related to reviews can be broken down into three types of features: review centric features, reviewer centric features, and product centric features (Jindal and Liu 2008). Jindal and Liu identified 35 key features in Amazon reviews, splitting them into three different categories. Using these features as a reference, we designed our own model with fewer features, and compared it to Jindal and Liu's model.

2.5 Duplicate Reviews

Duplicate and near-duplicate reviews with over a 90% similarity calculated using Jaccard's distance are classified as spam (Jindal and Liu 2008). We used this information to develop our own automated labelling system.

3 Data Collection

In this paper, we used reviews from beauty products on Amazon.com. We obtained a dataset of 371,345 reviews. From this dataset, we opted to split the dataset into a manually labelled dataset and an automatically labeled dataset.

3.1 Manually Labeled Dataset

Given that we would be manually labeling the data, we randomly selected 1,000 type 1 reviews for labeling.[6] Then, we agreed to focus on three main features when labeling to establish a

[3] Throughout the paper, we use "NLP" as an abbreviation for Natural Language Processing.

[4] LSTM is an abbreviation for Long Short-Term Memory, and is a type of neural network.

[5] BERT is an abbreviation for Bidirectional Encoder Representations from Transformers, a transformer-based machine learning technique for NLP pre-training developed by Google.

[6] Annotating was done by the three authors of this paper.

baseline, and to remove noise amongst the annotators.

	Real	Fake
Context	"This moisturizer in this face mask helped my acne go away."	"This face mask is perfect you should buy it!"
Time	"Great stocking stuffer." (June 20th)	"Great stocking stuffer." (December 10th)
Tense	"My daughter really loved these and asked me to buy more."	"I'm sure she will love it."

Table 1: Manually labeled reviews with a focus on three chosen features[7]

The table shows examples of reviews that were manually labeled by annotators as real or fake. The tone indicates whether a review was persuasive or informative. We realized that deceptive reviews focused more on persuading consumers to buy the product, more often than not claiming it was "perfect" rather than informing them on why they should buy it. Thus, we can see how one educates the user on aspects of the product while the other makes outlandish claims to justify one's purchase. Next, we looked at context regarding the review. For example, we found two identical reviews regarding Christmas stockings – one in June and one in December. Contextually, it does not make sense to post a Christmas-related view in June, so we labeled the review as fake. In addition to this, the tense of a review is also crucial in determining its authenticity. Ratings that used future tense and inferred someone will enjoy or use a product meant that the writer of the review did not actually use the product yet. Therefore, it would be hard to trust a review saying someone will enjoy something when they have not even tried it yet.

The annotators read through type 1 reviews one by one and independently marked them, stopping to discuss their framework at intervals of 10, 20, 50, 100, and 200, adjusting it each time to become more accurate. After 200 reviews, the annotators labeled the remaining reviews. In total, each annotator individually labelled 1000 reviews as real or fake.

To investigate the accuracy of our annotations, we calculated our inter-annotator agreement score, where we achieved a Fleiss' kappa value of 0.623. We opted to calculate Fleiss' kappa value instead of Cohen's kappa because Cohen's kappa is calculated between a pair of annotators whereas Fleiss' kappa is calculated over a group of multiple annotators. This shows that our annotation methods were relatively clear, reproducible, and uniform.

3.2 Automatically Labeled Dataset

We opted to also include an automatically labeled dataset so that we could test our model on a larger data pool. For this dataset, we used 150,000 randomly selected reviews. From these reviews, we sorted them by product, and searched for duplicate or near-duplicate reviews within each product.[8] After obtaining 600 or so duplicates, we searched the rest of the dataset for reviews that matched our initial list of duplicates, marking them down as duplicates as well.

4 Detecting Fake Reviews

This paper uses a logistic regression model to compare which set of features produces a more accurate prediction when detecting fake reviews. Specifically, we look at three sets of features in relation to predicting a review: a list of review-based features developed by ourselves, a subset of the most optimal features from our exhaustive list, and Jindal and Liu's review-based features (a common standard for determining whether a review is fake).

4.1 Our Features

We decided to include the following features relating to reviews:

1. (F1) The number of feedback (votes) a review gets. This value is provided as feedback from consumers who may have used the review to guide their decision-making process when buying the product or by people who think the review is an accurate indicator of the product. If a review is fake, this value can also be inflated by bots upvoting the review.

[7] These reviews were obtained from our dataset of Amazon beauty reviews.

[8] Near-duplicate reviews are reviews with a 90% similarity or higher.

2. (F2) The length of a review and (F3) the length of a review title. We thought that the shorter the review, the more likely the review was generated by a bot. Particularly, we found that shorter reviews lacked any substantive details about the product, which a bot could have easily generated due to the lack of specificity.
3. (F4) The lexical diversity of the review. We defined the lexical diversity of a review as how complex a review is, dividing the number of words by the number of unique words within a review. Specifically, if a review was more complex, it was more likely to be considered a valid and valued review by customers.
4. (F5) The number of times a brand name is mentioned. We chose this feature because we believed that referring to a brand an excessive number of times implied that it was written to persuade the user to buy a product of that brand rather than to inform the user about the product itself.
5. (F6) A binary value indicating whether a review had an image. We found that reviews with an image were more likely to be real, indicating that they had tried out the product.
6. (F7) Whether or not the purchase was verified. In some cases, we found that fake reviews tended to have unverified purchases, so we found it prudent to include this information.
7. (F8) The (categorical) rating value provided for the review. Each review is provided a number from 1 to 5, depending on the rating provided by the reviewer. Because this was a categorical rating, we split using dummy variables to represent each rating from 1 to 5.

4.2 Subset of our Features

From our exhaustive list of review-based features, we decided we wanted to focus on a subset of five features that would optimize efficiency and space while not compromising model integrity.

To identify this subset, we used recursive feature elimination, which is a methodology to select and rank features by recursively considering smaller and smaller sets of features, repeating the process until we reach the desired number of features. In our case, we decided to set our desired number of features to select as 5 to parallel the number of features provided by Jindal and Liu.

We ran the recursive feature elimination process on the automatically collected data to come up with the following subset of features:

1. (F1) The lexical diversity of a review.
2. (F2) The number of times a brand name is mentioned.
3. The dummy variables of the (F3) 1, (F4) 4, and (F5) 5-star ratings. From here, we can note that most fake reviews tended to be highly rated, usually awarding a product with 5 stars.

Note, when running the recursive elimination process on the manually labeled data, the model valued whether the purchase was verified over the number of times brand name was mentioned. However, when testing model accuracy and scores, we found that the combination of features scored similarly, so we chose to only focus on the subset of features selected from the automatically labeled data.

4.3 Jindal and Liu’s Features

When looking at reviews, Jindal and Liu provided three categories of features that they used to train their model: review centric features, textual features, and rating related features. Here, we specifically focus on comparing the following review centric features:

1. (F1) The number of votes a review gets.
2. (F2) The length of review title.
3. (F3) The length of review.
4. (F4) The position of a review by date and (F5) a binary value indicating whether a review is a product’s first review. They found that reviews written earlier have more of an impact on the sale of a product which a spammer could use to their advantage.

Something to note, is that when running the recursive feature elimination model, we also included Jindal and Liu’s features. When we analyzed the results, we found that the position of when a review was posted had the least impact of all the features.

4.4 Model Development

In total, we trained two classification models per each set of features, yielding a total of six unique models. We opted to use a multiple logistic regression because it provides a probability estimate of each review being fake or real. If the probability estimate is more than or equal to 0.5, the model labeled the review as fake. More specifically, we used a weighted logistic regression to deal with the data imbalance between the low percentage of reviews labeled fake (approximately 5%) and the high percentage of reviews labeled not fake.

For each set of features, one model was trained with the manually labeled data, and one was trained with the automatically labeled data. Model A and B are based on Jindal and Liu's review centric features, Model C and D are based on features we came up with, and Model E and F are based on the reduced features of Model C and D.

Model	Features
A and B – Jindal and Liu	• number of feedbacks a review gets • length of review title • length of review • position of review by date • if the review is the first review
C and D – Our Features	• number of feedbacks a review gets • length of review title • length of review • lexical diversity • number of brand name mentions • verified reviewer • rating • if the review has an image
E and F – Subset of Features	• lexical diversity • number of brand name mentions • 1, 4, and 5-star ratings

Table 2: Summary of features for our classification models.

All the models trained on the automatically collected data were tested on both the manual and automatically labeled data, while the manual trained models were only tested on the manual datasets. For all the models, we used a 70/30 train-test split. In the cases where we used two different sets of data to test each model (i.e., automatically versus manually labeled data), we maintained the 70/30 split within their respective datasets.

As a result, we have six models, three of which are tested on both the manual dataset and the automatically labeled data. The other three models are only trained and tested on the manually labeled data. We did not test them on the automated data set because we found that the automatically labeled data yielded a higher false negative when determining if a review was fake; therefore, if we tested the automatically labeled data on our model which was trained on the manually labeled, it wouldn't offer any valuable insights regarding the best features to detect fake reviews.

Model	Training Set	Testing Set
A	Manual	Manual
B_1	Automatic	Automatic
B_2	Automatic	Manual
C	Manual	Manual
D_1	Automatic	Automatic
D_2	Automatic	Manual
E	Manual	Manual
F_1	Automatic	Automatic
F_2	Automatic	Manual

Table 3: Training and test set pairs for our models

5 Performance Evaluation

To evaluate model quality, we calculated the area under the receiver operating characteristic curve[9], which provides us with a value which is the chance that the model will be able to distinguish between the positive and negative classes, which in this case is fake and not fake reviews. We also calculated the precision (how many of the positive predictions are correct), recall (how many of the positive cases have been correctly predicted over all of the positive cases in the data), and F1-score for each model (a measure of a test's accuracy).

5.1 Jindal and Liu Features

Below are the results for each of the models trained with Jindal and Liu's features. We find that the best performing model within the set of features is the model trained on the automatically labeled

[9] Throughout the paper we use "AUC" to mean area under the curve.

data. Model B_1, where we tested using the automated data yielded an AUC of .89 (89%). The model, B_2, predicting fake reviews within the manually labeled data had an F1-score of .71, which was the highest F1-score for all the models within the feature set for predicting fake reviews. Model B_1 had the highest F1-score for predicting not fake reviews.

	Fake			Not Fake		
	P	R	F1	P	R	F1
A – Manual and Manual	.40	.79	.53	.87	.55	.67
B_1 – Automatic and Automatic	.22	.95	.36	1.0	.83	.90
B_2 – Automatic and Manual	.81	.64	.71	.87	.64	.71

Table 4: Identifying Precision, Recall, and F1 scores of each model using Jindal and Liu's features.

Model – Data Used to Train and Test	Area Under the Curve (AUC)
A – Manual and Manual	.67
B_1 – Automatic and Automatic	.89
B_2 – Automatic and Manual	.79

Table 5: AUC scores for each model using Jindal and Liu's features.

5.2 Our Features

Table 6 and Table 7 highlight the results from the models trained on our exhaustive list of features. As we can see, the best performing model within the set of our features is the model trained on the automatically labeled data. Model D_1 had the highest AUC of .89 (89%). Model D_2 had the highest F1-score predicting fake reviews within the manually labeled data of .63 while Model D_1 had the highest F1-score for predicting not fake reviews, which was .91.

	Fake			Not Fake		
	P	R	F1	P	R	F1
C – Manual and Manual	.44	.84	.58	.90	.59	.71
D_1 – Automatic and Automatic	.23	.96	.37	1.0	.83	.91
D_2 – Automatic and Manual	.70	.58	.63	.85	.90	.87

Table 6: Identifying Precision, Recall, and F1 scores of each model using our features.

Model – Data Used to Train and Test	Area Under the Curve (AUC)
C – Manual and Manual	.71
D_1 – Automatic and Automatic	.89
D_2 – Automatic and Manual	.74

Table 7: AUC scores for each model using our features.

5.3 Subset of our Features

From our comprehensive list of features, we ran several models using a subset of those features. We can see from Table 8 and Table 9 that the best performing model for detecting fake reviews is the model trained on the automatically created data set. Specifically, when testing on the manually created dataset, model F_2 can predict if a review is fake with an F1-score of .61. Model F_1 has the highest F1-score of .87 to predict if a review is not fake. Model F_1 also has an AUC with a value of .87 (87%). However, note that model F_2 has a similar F1-score to model F_1 in terms of predicting if a review is not fake, while model F_2 has a better F1-score in predicting the manually labeled data as fake or not.

	Fake			Not Fake		
	P	R	F1	P	R	F1
E – Manual and Manual	.42	.85	.56	.90	.54	.68
F_1 – Automatic and Automatic	.18	.97	.31	1.0	.78	.87
F_2 – Automatic and Manual	.60	.61	.61	.85	.84	.85

Table 8: Identifying Precision, Recall, and F1 scores of each model using a subset of our features.

Model – Data Used to Train and Test	Area Under the Curve (AUC)
E – Manual and Manual	.70
F_1 – Automatic and Automatic	.87
F_2 – Automatic and Manual	.73

Table 9: AUC scores for each model using a subset of our features.

5.4 Discussion of Results

First, we need to understand that although data labeled in an automated manner is a great baseline to determine whether a review is fake, there can still be several false negatives that appear as "not fake" in our dataset. As a result, our goal is to see if our models can accurately label our manually labeled dataset.

However, we first start by looking at which set of features can most accurately predict data labeled automatically. Both Jindal and Liu and our features yield an AUC of 89%; however, we can note that when predicting fake reviews, our comprehensive list of features has a slightly higher recall rate by .01 (.96). Note that the subset of features has an even higher recall rate (.97) than the model with the comprehensive list of features, although its model performance is below Jindal and Liu's and the model with the comprehensive list of features.

How well does this translate to when we test with the manually labeled data? Piecing together all the values, we find that Jindal and Liu's features trained using the automatically labeled data is the best model to predict the manually labeled data. However, there are several things to note.

Firstly, Jindal and Liu's model yields the lowest F1-score out of the three feature sets when it comes to determining if a review is not fake. It has an F1-score of .71 in comparison to model D_2's score of .87 and model F_2's score of .85. However, Jindal and Liu do yield a higher F1-score when determining if a review is fake, with a score of .71 in comparison to .63 and .61. This is specifically because their recall value is higher (.64 compared to .58 and .61).

Secondly, we must consider the size of the models trained using the manually data. With a dataset of 1,000 manually labeled reviews in comparison of a dataset of 150,000 automatically labeled reviews, there will be a huge disparity in terms of model accuracy, specifically model scoring.

As a result, we will only be comparing each of the manually trained models to each other. Here, we find that our features and our subset of features do better than Jindal and Liu. Specifically, our subset of features has a higher recall value to determine if a review is fake (.85 and .84), while our comprehensive list of features has a higher recall value that the subset when detecting not fake reviews (.59 and .54). Both models also yield a higher AUC (71% and 70%) than Jindal and Liu (67%).

Ultimately the main thing to note is that the difference between the accuracy of each of these models is negligible. They vary by minimal percentages. So when determining which model to use when, and which features are most beneficial, you can find no harm in selecting the subset of features we have picked out based on recursive feature elimination. In some cases, we mentioned that it produces a higher recall rate when attempting to detect fake reviews. But overall, it performs very similarly to the models developed with other features.

6 Conclusion

In this paper, we introduced a new dataset of 1,000 manually labeled fake reviews, created a dataset of 150,000 automatically labeled reviews based on similarity to other reviews, and identified which set of features were the most ideal to detect fake reviews.

We determined that the best features depend on what the task is. Specifically, we learned that the subset of features does not sacrifice model performance. This means, we will be able to produce a faster running model and require less data necessary to collect, because we can comfortably use the subset of features to guide in model prediction without sacrificing accuracy.

Of course, a topic like this is incredibly difficult to address given the lack of data on the matter. One way to address this issue could be to create a dataset of fake reviews, which could be developed by crowd sourcing the creation of fake reviews, not the annotation. We could also consider the potential of looking into the impact of textual based features and reviewer-based features. Another interesting concept to consider is determining the intent of each review (i.e., whether a review was meant to inform or persuade), as many fake reviews intend to persuade a customer to buy the product.

7 References

Barbado, Rodrigo, Oscar Araque, and Carlos A. Iglesias. "A Framework for Fake Review Detection in Online Consumer Electronics Retailers." Information Processing &amp; Management 56, no. 4 (2019): 1234–44. https://doi.org/10.1016/j.ipm.2019.03.002.

Daryani, Monika, and James Caverlee. 2021. "Identifying Hijacked Reviews." Proceedings of the 4th Workshop on E-Commerce and NLP. https://doi.org/10.18653/v1/2021.ecnlp-1.9.

Fornaciari, Tommaso, and Massimo Poesio. "Identifying Fake Amazon Reviews as Learning from Crowds." Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, https://doi:10.3115/v1/e14-1030.

Jindal, Nitin, and Bing Liu. 2008. "Opinion Spam and Analysis." Proceedings of the International Conference on Web Search and Web Data Mining - WSDM '08. https://doi.org/10.1145/1341531.1341560.

Shani, Chen, Nadav Borenstein, and Dafna Shahaf. 2021. "How Did This Get Funded?! Automatically Identifying Quirky Scientific Achievements." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). https://doi.org/10.18653/v1/2021.acl-long.2.

Zia, Haris Bin, et al. "Racist or Sexist Meme? Classifying Memes beyond Hateful." Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), 2021, https://doi:10.18653/v1/2021.woah-1.23.

Zhang, Dongsong, Lina Zhou, Juan Luo Kehoe, and Isil Yakut Kilic. 2016. "What Online Reviewer Behaviors Really Matter? Effects of Verbal and Nonverbal Behaviors on Detection of Fake Online Reviews." Journal of Management Information Systems 33 (2): 456–81. https://doi.org/10.1080/07421222.2016.1205907.