BMIS2526-Data Programming with R

(Project Report)

Predicting Fake Reviews at Amazon.com using Sentiment Analysis

Submitted by: Nikhil Aggarwal

Katz Graduate School of Business

Email id: nia45@pitt.edu

Date: December 13, 2018

1 TABLE OF CONTENTS

2 INTRODUCTION

2.1 OVERVIEW

Amazon.com, Inc. is an American electronic commerce and cloud computing company based in Seattle, Washington, that was founded by Jeff Bezos on July 5, 1994. The tech giant is the largest Internet retailer in the world as measured by revenue and market capitalization, and second largest after Alibaba Group in terms of total sales. The Amazon.com website started as an online bookstore and later diversified to sell video downloads/streaming, MP3 downloads/streaming, audiobook downloads/streaming, software, video games, electronics, apparel, furniture, food, toys, and jewelry. The company also owns a publishing arm, Amazon Publishing, a film and television studio, Amazon Studios, produces consumer electronics lines including Kindle e-readers, Fire tablets, Fire TV, and Echo devices, and is the world's largest provider of cloud infrastructure services (IaaS and PaaS) through its AWS subsidiary. Amazon also sells certain low-end products under its in-house brand Amazon Basics.

2.2 IMPORTANCE OF PRODUCT REVIEW

As quoted by Michael LeBoeuf, “A satisfied customer is the best business strategy of all”, Amazon has always maintained high emphasis on customer satisfaction. Amazon customer reviews about the products are one of the main reasons to attract customers on Amazon. It basically helps them understand almost every detail of the product. Since, consumers cannot physically inspect the product while shopping online, Amazon product review is the one they can trust in order to judge a product.

A research conducted by Dimensional Research claims that 90% of consumers online believe their purchasing choices are influenced by product reviews. Many customers consider positive reviews a prerogative to purchasing items online, especially those that cost more. Therefore, it is very important for new sellers to understand that Amazon customer reviews can make or break their e-commerce careers.

2.3 SENTIMENT ANALYSIS

Today, electronic word-of-mouth (e-WOM) is one of the most important factors for digital marketing. Companies are using the digital platforms for promoting their products. Nowadays customer’s online reviews can influence the purchase decisions of a product and some of the sellers are taking advantage of this. They are posting fake or non-genuine positive reviews for their products to increase their product rating and finally increase their product sales. Here comes the application of Sentiment Analysis.

Sentiment Analysis is the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within a text. Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organizations across the world. And we will be using similar strategy to categorize reviews of Amazon Data Set into positive and negative sentiments.

3 OBJECTIVE

Customer reviews has a major impact on sales of a particular product online. A good customer review can boost up sales of a bad product and similarly, a bad customer review can reduce sale. Therefore, it is very important to eliminate bad reviews. As a data Scientist, I will be working on Amazon data to predict positive and negative reviews using Sentiment Analysis. In addition to that, I will be predicting helpfulness of a review based on its star rating, sentiment score derived and many other factors.

4 LITERATURE REVIEW

4.1 ABSTRACT

Customer reviews were supposed to be internet's one of the greatest breakthroughs. They let a customer know, before even purchasing of a product that what kind of a product it is. They are so powerful that a review can effect a sale of a product in both good and bad ways. Therefore, companies like Amazon build their success on the trust created by these reviewing systems.

Amazon has had a fake review problem for a long time. Up until late 2016, Amazon allowed sellers to give away products in return for a review. Those reviews were "honest and unbiased", at least according to the disclaimers that reviewers sometimes added.

Back then, many sellers used product giveaways to increase their positive reviews. Amazon's algorithms acted on the review data, search visibility went up, and buyers bought those items more often. Everyone went away happy, right? Well, at least the sellers did. Then Amazon prohibited all incentivized reviews, and the problem swiftly went underground. Incentives continued to be offered, but away from the official discount code system, so Amazon couldn't see the activity at all.

The fake review problem is plaguing the e-commerce industry and its criticality can be easily seen from the reference above. I am going to address this issue in my project and present a model to predict fake reviews by analyzing Amazon user reviews.

4.2 EXISTING RESEARCH

As per the previous study conducted by students of Stevens Institute of Technology, fake positive reviews are more common than fake negative reviews. They hypothesized that it is easier to self-inject positive reviews to a particular product than posting negative reviews for a competitor. This is intuitive as the culprits caught performing these practices will have to face serious repercussions. As a result, practices of posting negative reviews are performed by small Business who do not have a big brand image to lose, if get caught. For instance, in 2013, the Taiwan Fair Trade commission fined Samsung $340,000 for hiring two external companies for positing negative reviews for their competitor HTC.

A research conducted by students of University of Chicago, described this as a process where an attacker manipulates crowd opinion by using fake or descriptive reviews. Studies on Yelp found that a 1-star rating increase for restaurants can result in increase of 5-10 % of their revenue. Sites like Yelp and Amazon have been consistently engaged in a battle with fake reviews, as attackers try to adapt and bypass various defense systems. Sites like Yelp and Amazon have been consistently engaged in a battle

with fake reviews, as attackers try to adapt and bypass various defense schemes. Yelp's review filter system flags suspicious reviews and even raises an alert to the consumer if a business is suspected of engaging in large-scale opinion manipulation Recently, attacks have been known to generate highly deceptive (authentic looking) fake reviews written by paid users. Much of this comes from malicious crowdsourcing marketplaces, known as crowdsourcing systems, where a large pool of human workers provides on-demand effort for completing various malicious tasks. Some of the key sources identified for the origin of Fake Reviews:

1. The 100%-off Coupon: Lots of Facebook groups have been identified which offer buyers a juicy deal i.e. if they will buy that seller's product and will post a 5-star rating the seller will provide a coupon of worth equal to price of product and sometimes more than that. In this way they review remains Verified and for a company ecommerce company like amazon, whose business depends on these reviews, this becomes very difficult for them to track.

2. The bot Armies: Several sellers have been identified having thousands of fake accounts and are using people or scripts to write positive 5 start review.

3. The Bait and Switch: Another technique identified was that once a seller has earned a high rating for a product, he simply switches its name and description to a similar product. In this way, a similar product which was having less rating before will get higher rating and eventually sales of that product will increase.

4.3 Conclusion

It is a one of the biggest challenges for all ecommerce companies. As in Amazon, a customer can report a suspicious review 24 hours a day and 7 hours a week. The company take strict actions against both the seller and the reviewer by suspending their accounts and taking further legal actions.

But it's an arms race, a cat and a mouse game and it's not clear who is winning. Amazon and other review-based companies are fighting the same kind of trust battles that are holding every aspect of trust these days. It's no longer enough to be a good judge of value and quality while shopping, it is about your judgement of reviews that guide you while shopping.

5 Data Collection

5.1 Reference Site

Amazon Customer Reviews (or Product Reviews) is one of Amazon's iconic products. Since 1995(the period of its first review) millions of Amazon customers have contributed over a hundred million reviews to express opinions and describe their experiences regarding products on the Amazon.com website. Due to this Amazon has released their review Dataset for research in Natural Language Processing, Machine Learning etc. Specifically, this dataset was constructed to represent a sample of customer evaluations and opinions, variation in the perception of a product across geographical regions, and promotional intent or bias in reviews. The name of Website is www.s3.amazonaws.com.

5.2 Sample Data and Variables

For our Analysis, we have taken its Electronics Dataset which contains approximately 30 Lakhs rows and 15 Variables. Details of each Column is provided below.

1. Marketplace: - Location where the product is sold.
2. Customer_id:- Unique id of reviewer.
3. review_id: Unique Id of each review.
4. product_id: Unique Id of each product sold by Amazon
5. product_parent: Parent Product of item sold.
6. product_title: Title of Product sold.
7. product_category: Category of product.
8. star_rating: Rating of the review given by customer.
9. helpful_votes: no of helpful votes.
10. total_votes: Total no of votes.
11. Vine: Vine customer or not.
12. verified_purchase: Weather purchase is verified or not.
13. review_headline: Headline of review.
14. review_body: Body of review.
15. review_date: Date on which review was posted.

5.3 Data Cleaning

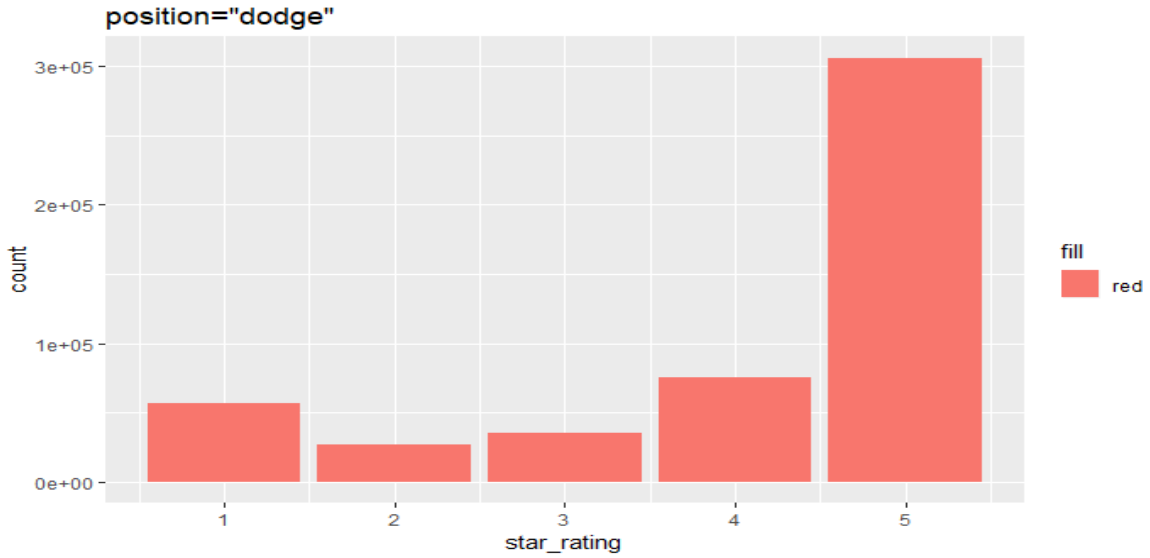
- Reducing data size: Dataset provided by Amazon consists of approximately 30 Lakhs rows but due to limitation of system specifications I am using only 5 Lakh Latest reviews. For doing that, I have first converted review_date which was earlier in character format to Date format and then arranged the reviews based on years. Finally, I have extracted top 500,000 rows for our analysis.

- Remove NA values: NA values are something which may impact our analysis in unpredictable ways. In our dataset set, I found that there are some NA values. Since, those are present in review body and the number of those rows are single digit only, therefore for easy of our analysis, I have removed those rows.

- Converting all Variables as factors or Continuous Variables: When we are reading our CSV file, everything is read as characters. For sake of analysis, I have converted all variables to their appropriate formats like star_rating to categorical Variable, Date to proper Date Format ,etc.

6 EXPLORATORY ANALYSIS

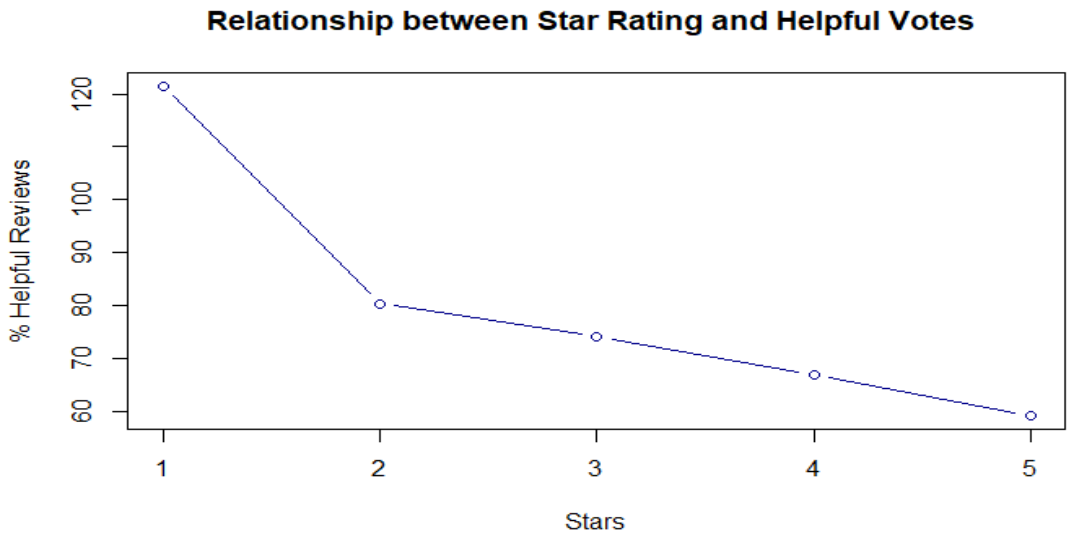
6.1 TOTAL NO OF REVIEWS PER EACH STAR RATING



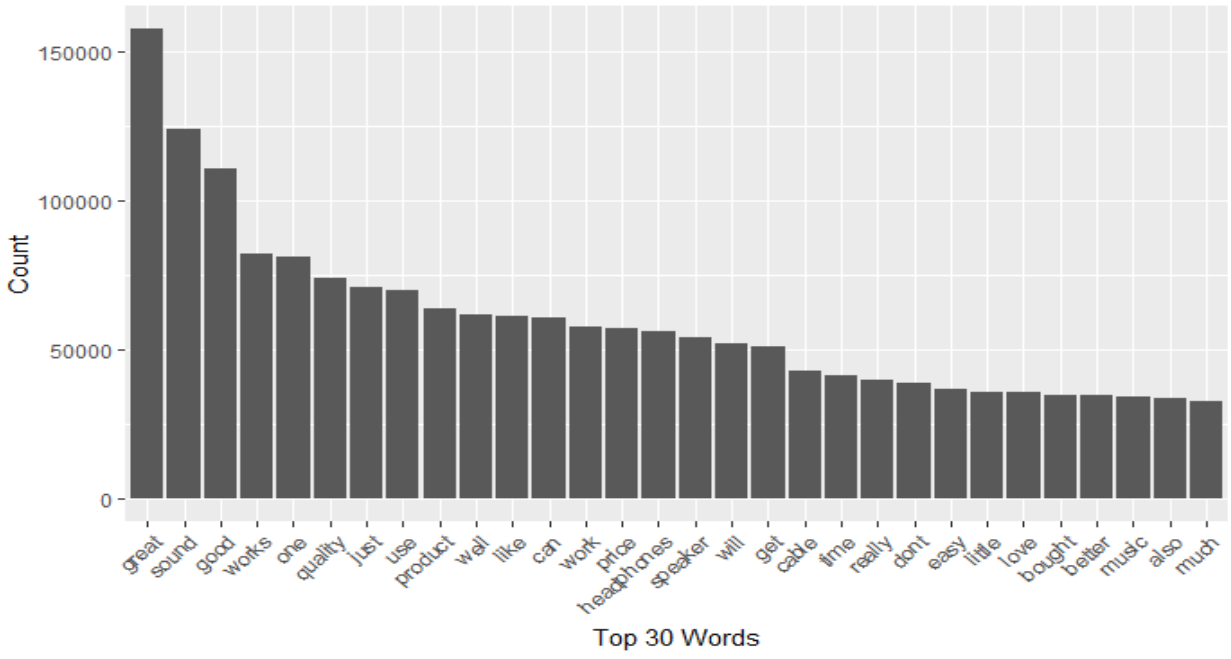
Based on the visual graphical analysis, we found that our data set has reviews with maximum 5 star ratings. Also, there is approximately equal count of 1 star and 4 star reviews. Moreover, Minimum number of reviews are there with 2 star ratings.

6.2 RELATIONSHIP BETWEEN STAR RATING AND HELPFUL VOTES

Based on the graph, we analysed that as the star rating of reviews are increasing the percentage of helpful reviews are decreasing.



6.3 TOP 30 WORDS AND THEIR COUNT



6.4 WORDCLOUD OF TOP 150 WORDS

7 DATA ANALYSIS APPROACH AND RESULTS

7.1 APPROACH

1. Text Mining: I am using Text Corpus of library package" tm" for storing reviews. Now, tm package comes with a function tm_map which is used for following transformations.
    a. Text to Lower Case: We want all text to be converted to lower text case so that our corpus should not read "good" and "Good" as different words. Converting all of them to lower case will solve this issue.
    b. Remove Punctuation Marks: Now from the cleaned data I have removed all the punctuation marks.
    c. Other operations performed on this corpus include removing numbers, white spaces and other stop words.

2. DTM and Frequent Five: I have created the Document Term Matrix of this refined corpus where each word and its occurrence is stored. For calculating the score of all these words, I am using AFFIN Lexicon. The AFINN lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment. All of this information is tabulated in the sentiments dataset, and tidy text provides a function get_sentiments() to get specific sentiment lexicons without the columns that are not used in that lexicon.

   Using this Lexicon, I have calculated the score of each review. Based on the score, I have classified reviews as positive and negative. Now, it is impossible to conduct analysis on complete dtm. Hence, I have selected frequent five words i.e. words that occur in at least five reviews and we have created our model based on that.
3. Cross Validation Approach: We have used Cross validation approach to create test and train Datasets. We have taken sample of 1,00,000 reviews and divided it equally between test and train datasets.

7.2 MODELLING: NAÏVE BAYES

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive)independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring several parameters linear in the number of variables (features/predictors) in a learning problem. Maximum likelihood training can be done by evaluating a closed form expression (mathematical expression that can be evaluated in a finite number of operations), which takes linear time.

It is based on the application of the Bayes' rule given by the following formula:

$$P(C = c|D = d) = \frac{P(D = d|C = c)P(C = c)}{P(D = d)}$$

Formula 2.4.1.1: Baye's rule

where D denotes the document and C the category (label), d and c are instances of D and C and P(D = d) = ∑ (D= d|C =c)P(C=c). We can simplify this expression by

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Formula 2.4.1.2: Baye's rule simlifed

Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.

It allows easy identification of confusion between classes e.g. one class is commonly mislabeled as the other. Most performance measures are computed from the confusion matrix. In our model, Confusion Matrix is as below.

```
##                Actual
## Predictions     Neg  None   Pos
##         Neg   14397  1060  2228
##         None   3076 10237    12
##         Pos    3433   593 14964
```

Accuracy: 79.2%

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   Neg  None   Pos
##       Neg  14397  1060  2228
##       None  3076 10237    12
##       Pos   3433   593 14964
##
## Overall Statistics
##
##                Accuracy : 0.792
##                  95% CI : (0.7884, 0.7955)
##     No Information Rate : 0.4181
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6839
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: Neg Class: None Class: Pos
## Sensitivity              0.6887      0.8610     0.8698
## Specificity              0.8870      0.9190     0.8772
## Pos Pred Value           0.8141      0.7683     0.7880
## Neg Pred Value           0.7986      0.9549     0.9278
## Prevalence               0.4181      0.2378     0.3441
## Detection Rate           0.2879      0.2047     0.2993
## Detection Prevalence     0.3537      0.2665     0.3798
## Balanced Accuracy        0.7878      0.8900     0.8735
```

7.3 MODELLING: LOGISTIC REGRESSION

Logistic regression is a method for fitting a regression curve, y = f(x), when y is a categorical variable. The typical use of this model is predicting y given a set of predictors x. The predictors can be continuous, categorical or a mix of both. The categorical variable y, in general, can assume different values. In the simplest case scenario y is binary meaning that it can assume either the value 1 or 0.

R makes it very easy to fit a logistic regression model. The function to be called is glm() and the fitting process is not so different from the one used in linear regression. In our case, I am using Logistic Regression to predict helpfulness of votes based on sentiment score derived above, star rating, weather it is vine customer or not, total votes on that review and weather this is verified purchase or not.

Confusion Matrix

Prediction	Reference No	Reference Yes
No	39393	3802
Yes	696	6109

Accuracy: 91%

```
               Accuracy : 0.91
                 95% CI : (0.9075, 0.9125)
    No Information Rate : 0.8018
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.6791
 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.9826
            Specificity : 0.6164
         Pos Pred Value : 0.9120
         Neg Pred Value : 0.8977
             Prevalence : 0.8018
         Detection Rate : 0.7879
   Detection Prevalence : 0.8639
      Balanced Accuracy : 0.7995

       'Positive' Class : No
```

8 CONCLUSION

Amazon fake review problem is a very big problem and predicting positive or negative sentiments is a small branch of this big tree. Using Naïve Bayes algorithm and our data set, I have achieved an accuracy of 80 percent. In addition to that, I have predicted helpfulness of votes based on star rating, sentiment score and various other factors. However, there are several other factors which needs to be analyzed to predict weather a review is fake or not but at this point of time, I am considering just two analyses. With continuous effort and using different models we can further increase its accuracy. Hence, we will stick to our Naïve Bayes algorithm and Logistic Regression to predict positive and negative sentiment as well as helpfulness of a review.

9 REFERENCES

(2018). In J. S. Robinson, *Text Mining with R.*

amazon-product-review-importance. (n.d.). Retrieved from https://www.amzinsight.com: https://www.amzinsight.com/amazon-product-review-importance/

Confusion Matrix in Machine Learning. (n.d.). Retrieved from www.geeksforgeeks.org: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/

Dr. Rajesh Bose, R. K. (n.d.). *Sentiment Analysis on Online Product Reviews.* Retrieved from https://www.researchgate.net: https://www.researchgate.net/profile/Sandip_Roy8/publication/326816109_Sentiment_Analysis_on_Online_Product_Reviews/links/5b649a02458515cf1d32e9fe/Sentiment-Analysis-on-Online-Product-Reviews.pdf?origin=publication_detail

How to Perform a Logistic Regression in R. (2018, June 24). Retrieved from https://datascienceplus.com: https://datascienceplus.com/perform-logistic-regression-in-r/

Lamberti, M. (2015). *Twitter Emotion Analysis.*

McCabe, C. (2018, June 15). *amazon-fake-reviews*. Retrieved from https://www.webretailer.com: https://www.webretailer.com/lean-commerce/amazon-fake-reviews/#/

Pogue, D. (2018, August 3). *The rise of fake Amazon reviews — and how to spot them*. Retrieved from https://finance.yahoo.com: https://finance.yahoo.com/news/rise-fake-amazon-reviews-spot-175430368.html

Theodoros Lappas, G. S. (n.d.). *The Impact of Fake Reviews on Online Visibility:A Vulnerability Assessment of the Hotel Industry.* Retrieved from https://pdfs.semanticscholar.org: https://pdfs.semanticscholar.org/8bc1/d7198b053849b2cebc6b9ecf19a840d94d39.pdf

wikipedia. (n.d.). *Amazon (company)*. Retrieved from https://en.wikipedia.org: https://en.wikipedia.org/wiki/Amazon_(company)

Yuanshun Yao, B. V. (2017, September 8). *Automated Crowdturfing Attacks and Defenses in Online Review System.* Retrieved from https://arxiv.org: https://arxiv.org/pdf/1708.08151.pdf