Using Linear Regression to Predict Sentiment Score of Amazon Review Text

Data 200

Final Project Draft

Qi Tifany Chu, Elizabeth Gilson, Giovanni Battista Alteri

Final Project Presentation Youtube Link

https://youtu.be/HNwh8hnjcWs

Abstract

This study analyzes evolving sentiment trends in Amazon reviews using the Amazon Review Dataset, aiming to understand changes in user sentiments over time and build predictive models for sentiment scores.

The investigation reveals insights into sentiment analysis development and identifies influential features impacting sentiment prediction. Utilizing a 5-core subset of 142.8 million reviews from May 1996 to July 2014, the study spans diverse product categories and employs VADER sentiment analysis.

This research employs multiple models: a linear regression model predicting overly polar reviews over time, and two linear models predicting polarity scores, the second using Lasso regression with regularization. Model 1 displays remarkable accuracy, significantly reducing standard deviations for overly positive and negative reviews, while Models 2 and 3 also demonstrate acceptable accuracy, enhanced by incorporating one-hot encoded categories and regularization.

This research emphasizes the dynamic nature of sentiment trends in Amazon reviews and underscores the importance of sentiment analysis in understanding consumer feedback nuances. Future studies may explore alternative sentiment analysis methods and advanced modeling techniques for enhanced predictive accuracy in comprehending online consumer sentiments.

Research Questions

How has the sentiment of Amazon review text changed over time? Which model will best predict the sentiment score of review based on select features?

Introduction

Understanding consumer behavior through online reviews has garnered significant attention in contemporary research, particularly in the domain of online shopping platforms such as Amazon. This study delves into the analysis of consumer reviews trends with the Amazon Review Dataset. In doing so, it aims to find how the sentiment of user reviews has changed over time. Additionally, we aim to create a predictive model to forecast sentiment scores.

This investigation is important because it reveals key information about how the sentiment analysis of user reviews has developed over time. It also uncovers which features are most important for predicting the sentiment of a review and which have little to no impact. All of this data can be used by Amazon to ensure that reviews remain relevant and useful to users. For example, if the majority of reviews are negative, then Amazon might try to incentivize leaving positive reviews or try to gain more reviews in general so the feedback is more balanced.

There have been many explorations into the sentiment analysis of Amazon reviews, and other review datasets. *Predicting Amazon customer reviews with deep confidence using deep learning and conformal prediction* (Norinder, 2022) uses machine learning to predict the sentiments across different types of products offered by Amazon. In comparison to our project this paper uses a deep neural network to make predictions and attempts to categorize reviews into negative (1-2 overall score) or neutral-positive (3-5 overall score). They found that deep

learning in combination with conformal prediction yielded accurate predictions across different product categories. Similar work has been done on predicting Yelp reviews using supervised learning algorithms (Xu, 2014).

Those papers focus on classifying reviews into two potential categories: positive or negative, however other studies have been done taking into account "neutral" reviews. In *Using Machine Learning to Predict the Sentiment of Online Reviews: A New Framework for Comparative Analysis*, three different datasets are examined and deep learning is used to classify reviews as positive, negative, or neutral (Budhi, 2021). Since our model predicts the sentiment score it provides a continuous quantitative result compared to the two or three categories commonly used in the current literature.

Additionally, multiple papers have used VADER as a metric to measure sentiment, such as *Using VADER sentiment and SVM for predicting customer response sentiment* (Borg, 2020). As it has been used in recent publications, we feel as if it is a good choice for a sentiment analysis tool.

Description of Data

Our project focuses on analyzing consumer review trends in data from Amazon. This dataset was originally made up of 142.8 million Amazon reviews from May 1996 to July 2014. This data was collected by Professor Julian McAuley at UCSD. To make data analysis more manageable we are using the 5-core of the dataset which means each user and each item has five reviews each. The available information in this dataset include: ReviewerID, ID of the product,

name of the reviewer, the helpfulness rating, the text of the review, the overall rating, the summary of the review, and the time of the review (McAuley, 2015).

Since we are using the 5-core dataset, the reviews will likely be more balanced. However, taking a 5-core does introduce bias as users or products with less than five reviews are automatically excluded (selection bias).

The dataset is extremely large, so to focus our analysis we chose a subset of datasets that could show a balanced view of the variety of products that Amazon sells. This is another point at which selection bias could have an impact on the outcomes of our data. The product types that we decided to focus on are: Tools, Sports, Lawn, Kindle, Movies, Video Games, and Toys.

We decided to use a VADER score sentiment analysis of the review text as our main metric for how positive/negative the reviews are. VADER or Valence Aware Dictionary for sEntiment Reasoning is a rule-based model which maps lexical features to an emotional score to determine the sentiment of a given piece of text (Hutto, 2014).

Here is a table to briefly summarize the features of the dataset we examine.

Feature	**Description**	**Feature Type**	**Storage type**
overall	Rating of the product, discrete values from 1 to 5	Quantitative-discrete	Double
vote	Number of helpful votes on the review	Quantitative-discrete	Int
verified	Whether of not the user	Qualitative -	Boolean

	giving the review has a verified purchase or not	dichotomous	
asin	ID of the product	Qualitative - nominal	Int
reviewerName	Username of the reviewer	Qualitative - nominal	String
reviewText	Main text portion of the review	Qualitative - nominal	String
summary	Summary text of the review	Qualitative - nominal	String

Figure 1. Features of Amazon Dataset

Here is a table showing the features we have added to the dataset. Not all of them have been used in our final model, however they may be referenced in this paper.

Feature	**Description**	**Feature Type**	**Storage type**
polarity_review	The VADER sentiment score of the reviewText	Quantitative - continuous	double
length_of_review	The number of characters in the reviewText	Quantitative- discrete	double
length_of_username	The number of characters in the reviewerName	Quantitative- discrete	double
count	The number of reviews for each asin	Quantitative - discrete	double
percentile	The percentile rank of values in the 'polarity_reviews' column	Quantitative- continuous	double

Figure 2. Added Features of the Amazon Dataset

Methodology

EDA Work

When we examined the overall scores, we found that they were highly skewed to the left. The vast majority of reviews had an overall score of five-stars despite the fact that the sentiment analysis of the review text showed a variety of positive and negative sentiments (Figure 3G). Because of this, and the fact that the discrete values of the overall score were limiting, we decided to create models to predict the sentiment of the review text.

For further EDA focusing on the polarity scores of the reviews, we decided to look at the polarity score trends over time by year (Figure 3A) and by month (Figure 3B). Over the years, there seemed to be an increase in the range of polarity scores with more outliers present in the later years. Whereas, there seems to not be any noticeable differences in the range of scores when viewed by months. This is also related to an increase in popularity of Amazon where the number of reviews have also increased over time (Figure 3F). We also viewed the spread of polarity scores by different variables, such as, ranges of verified status (Figure 3C), word count range (Figure 3D) and Vote count range (Figure 3E). There seemed to be a trend where as the number of words in the review increased so did the reviews (Figure 3D). The vote column did not seem to have a different spread of polarity data between ranges (Figure 3E). The verified users also seemed to have a smaller range and median of their polarity scores compared to the unverified users (Figure 3C).

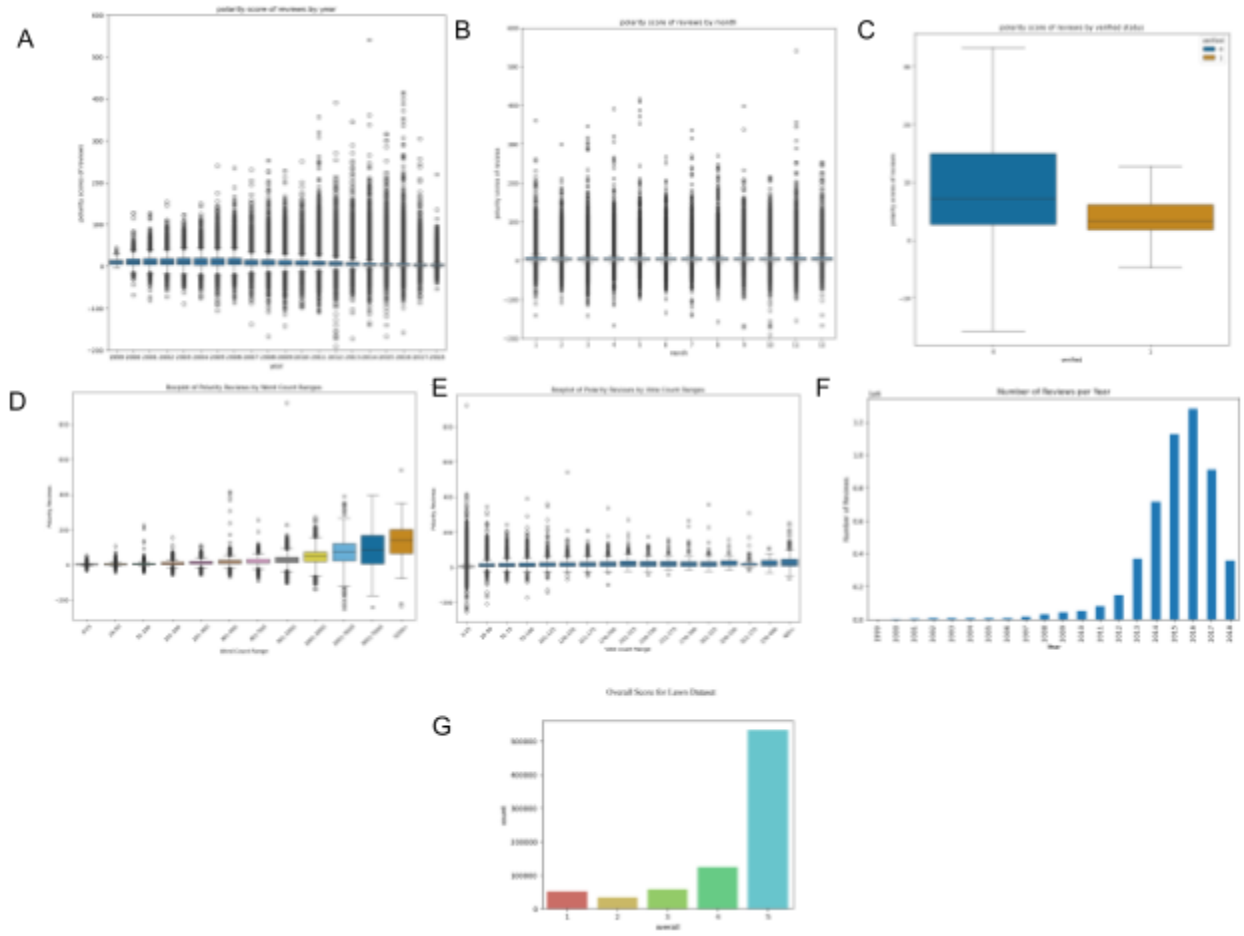

Figure 3. Selected EDA Visualizations

A) Boxplot - Polarity Score of Reviews by Year
B) Boxplot - Polarity Score of Reviews by Month
C) Boxplot - Polarity Score of Reviews by Verified Status
D) Boxplot - Polarity Score of Reviews by Word Count Range
E) Boxplot - Polarity Score of Reviews by Vote Count Range
F) Barplot - Number of Reviews by Year
G) Barplot - Counts of Overall Score

Description of Models

For the first model, we used a simple linear regression model to predict the proportions of overly polar (positive or negative) reviews given the year. We defined "overly polarizing" as having a polarity score in the top or bottom 25 percentile of all polarity scores over all years. When looking at the trends of the overly polarizing reviews, we noticed that the proportion of

overly negative reviews stayed fairly consistent over the years at around 20% of total reviews per year, whereas the proportion of overly positive reviews decreased over the years from ~60% in the early 2000s to 20% by 2014. Interestingly, before 2014, the proportion of overly positive reviews were much greater than that of negative reviews but for all data sets, 2014 was the crossover point where negative reviews were more frequent. We saw this trend across all 7 categorical datasets and the trend persists when the 7 data sets are merged together.

Using the combined data set, we created two models, one for the positive and another for negative. We used two methods (manuel and sklearn.LinearRegression) to calculate the model equation of: **proportion = theta_0 + theta_1 * year.** We are hopeful that this model will illuminate how the sentiments of reviews have changed overtime and predict how consumer sentiments will change in the future.

For the second model, we used a Multiple Linear Model to predict the polarity scores of reviews. We decided to group the data by product (or "asin" number) and find the means of numerical columns. This will be more useful for the Amazon sellers as it gives an overall view of the polarity of their item. Next, we added two columns "count" which represents the number of reviews per item and "length of review" which represents the number of words in the review. Even though the "verified" column is categorical, by one hot encoding it (0=False, 1=True), the mean gives us a proportion of verified reviews for that item. We believe this is a good representation of the verified variable for each item. To choose which numerical parameters would be best to use for this model, we began by constructing a heat map which shows the pairwise correlation between all the numerical variables. Focusing on the "polarity_reviews" column, the "overall", "vote", "length of review", "year", "verified" parameters seem to have at

least a weak correlation. Playing around with the different combinations of features, we decided on using “year”, “length of review” and “verified” as our parameters for the linear model.

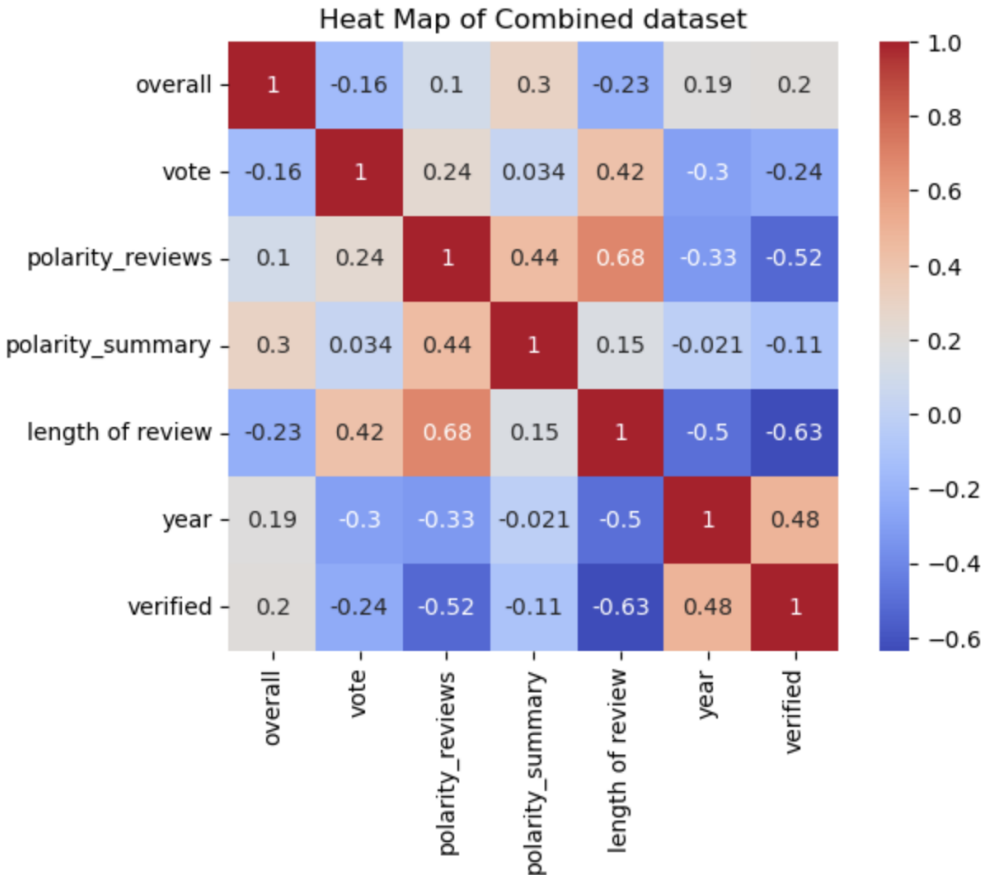

Figure 4. Heat Map of Features for Model 2

For the third model, we used a similar method to the second model but with all seven of our chosen categories. We one hot encoded these categories, dropping one to avoid multicollinearity. Then we replace the nan values in votes with 0, as a nan value means there were no votes given to that particular review. Then we drop any na values from the dataframe. All numerical features are normalized, this includes polarity summary, length of review, and year. We utilize scikit-learn to use train/test splitting for the model, with a test size of 33%. Then we create a linear regression model and fit it using the training data.

Further expanding on this model, we also attempted to use regularization using the Lasso method from scikit-learn. First the features are standardized using StandardScaler to ensure mean is zero and standard deviation is one for each feature. This scaling is essential for regularization techniques like Lasso, which penalize coefficients. Then for the actual Lasso Regression, we iterate through a range of alpha values and for each value we fit a Lasso model on the scaled training data, collect the coefficients and calculate the MSE on the test data. When we plot the MSE against the alpha values, we can observe that when the alpha value reaches approximately five, the penalty is so great that the lasso regression is fitting a constant model to the data, and that is why the MSE doesn't change after this value (Figure 5).

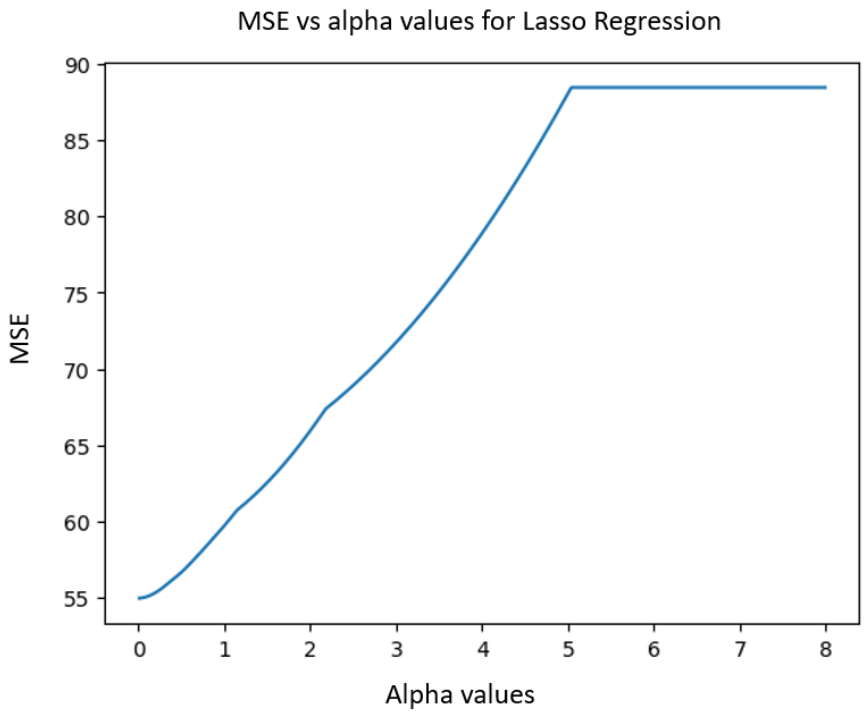

Figure 5. Mean Squared Error for different alpha values

To adjust this hyperparameter, we decided to use cross validation. We use scikit-learn's LassoCV model to perform a 5-fold cross-validation for alpha selection. This determines the optimal alpha value, which was found to be 0.01. Then we use this value to predict the test data.

Summary of Results & Discussion

Our results from Model 1 show that the model is extremely successful in predicting the change of overly positive reviews over time and somewhat successful at predicting the change of overly negative reviews over time (Figures 6 and 8). The model for overly negative reviews was able to reduce the standard deviation by 14.6% (Figure 6). The model for overly positive reviews was able to reduce the standard deviation by 64.9% (Figure 8). Additionally, when we plot the residuals, there are no significant trends and they are approximately symmetric (Figures 7 and 9). Overall, our model 1 was accurate and there is clearly a correlation between average sentiment score of review text and year.

```
~~~ Overly Negative Reviews ~~~
(Lower) Model: -4.53 + 0.00*Year
Correlation Coefficient: 0.5209554836277837
model SD = 0.021832813006335078
actual SD = 0.025577818839695718
percent decrese of SD: 14.64161528718213%
```

Figure 6. Assessment statistics of overly negative reviews

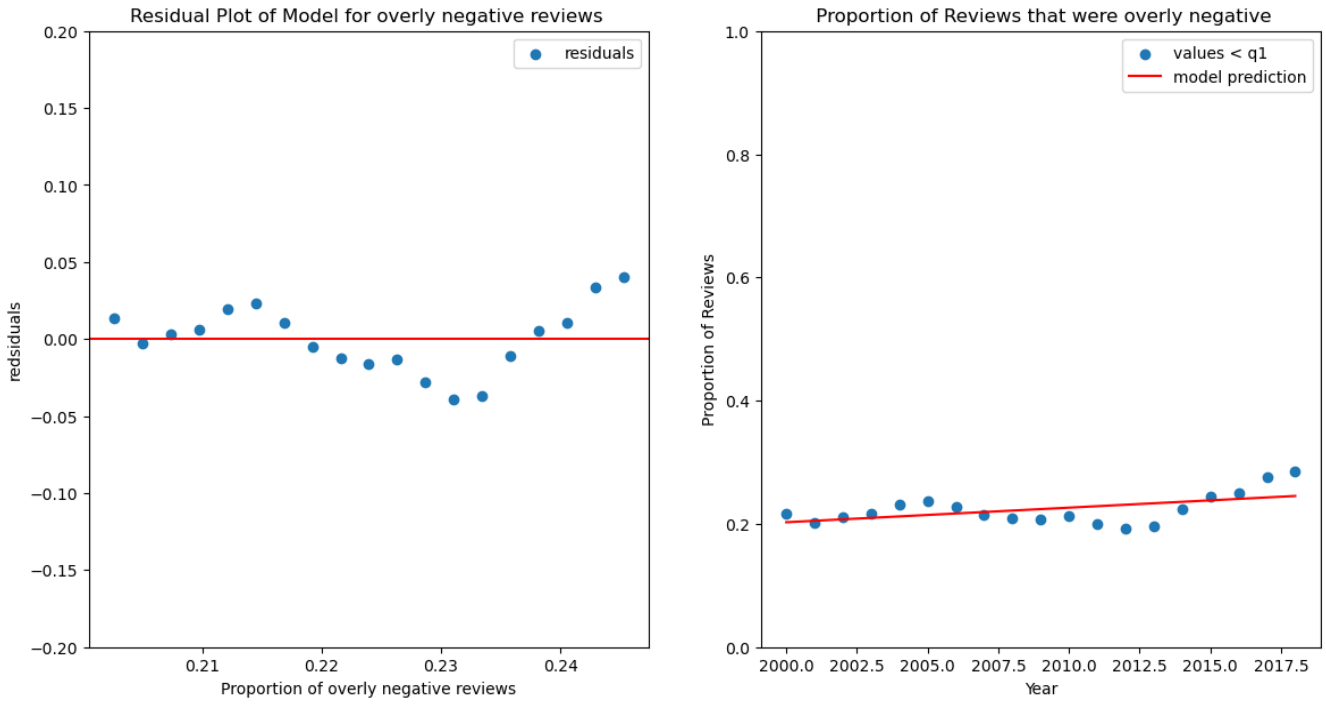

Figure 7. Graphs showing residuals of Model 1 (left) and how the proportion of overly negative reviews changes over time (right)

```
~~~ Overly Positive Reviews ~~~
(Upper) Model: 51.18 + -0.025*Year
Correlation Coefficient: -0.9363500784835717
model SD = 0.0533018430622533
actual SD = 0.15182781823953914
percent decrese of SD: 64.8932299230179%
```

Figure 8. Assessment statistics of overly negative reviews

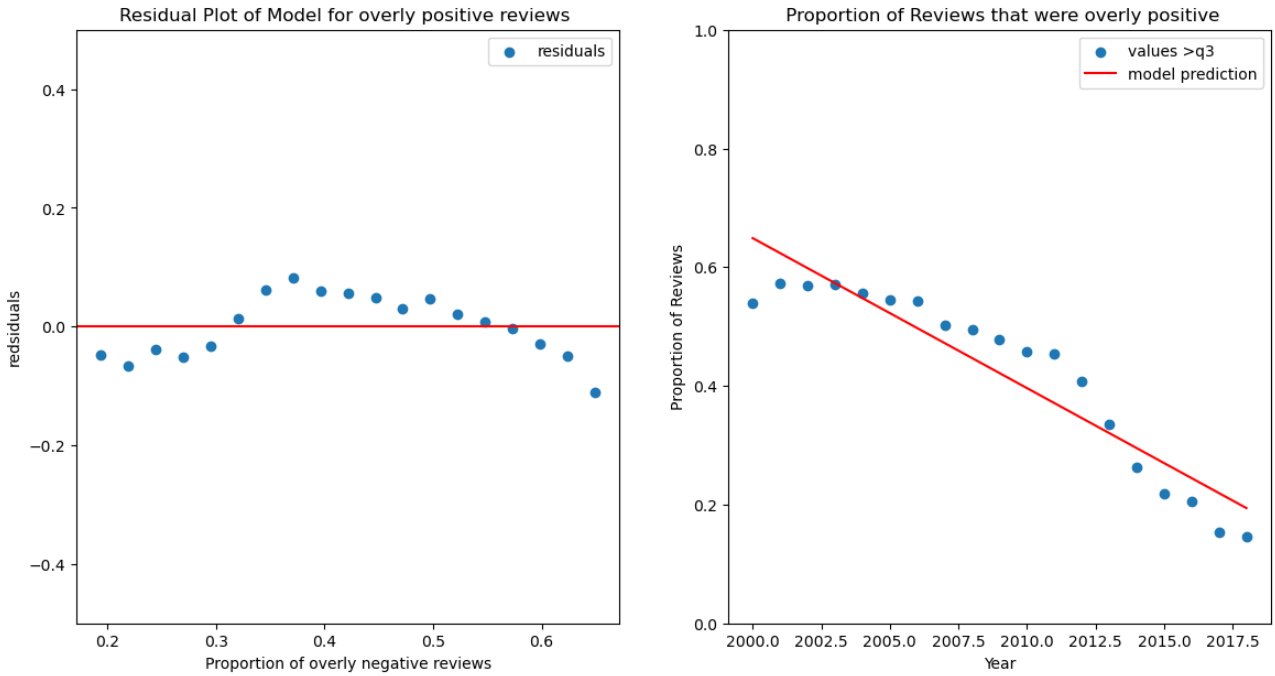

Figure 9. Graphs showing residuals of Model 1 for overly positive reviews (left) and how the proportion of overly positive reviews changes over time (right)

Our results for Model 2 and Model 3 also show some level of accuracy. Both have residual plots with no distinct patterns and symmetry (Figures 10 and 12). Model 2 had an average RMSE of 2.5 and Model 3 had an RMSE of 7.30. This indicates that including the one hot encoded categories or the regularization has helped improve the accuracy of the model. To evaluate the fit of model 2, we used the average polarity score of the full data set as a "baseline" value for RMSE and standard deviation (SD) comparisons (Figure 11A). Compared to the RMSE of the average polarity score to the actual polarity values, there was a 28% reduction in model predicted RMSE. Whereas in comparing the standard deviation of the errors to the baseline, there was also a ~28% reduction in the standard deviation (Figure 11A). We plotted the actual vs

predicted values to see how accurate our model 2 predictions were (Figure 11B).

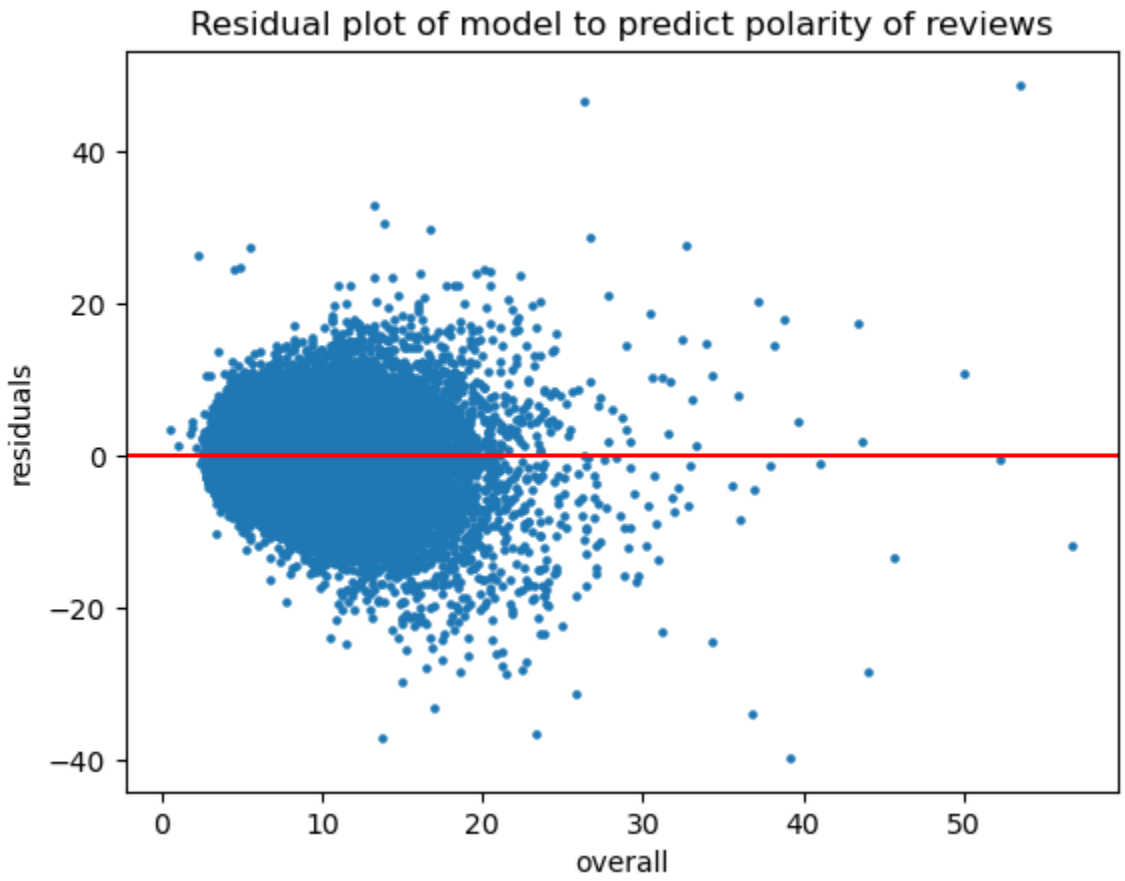

Figure 10. Residual plot for Model 2 on the full dataset

(A)

```
Intercept: -182.5
    Slope: 0.1
     SD(errors): 2.57
SD(polarity_reviews): 3.56
size of error decreased by 27.859941870137806 %
                  mse (model): 6.582876742369346,                    rmse (model): 2.5657117418699524
mse (avg polarity of dataset): 12.649385384396934,   rmse (avg polarity of dataset): 3.5565974448054893
rmse reduced by 27.860496396148328%
```

(B)

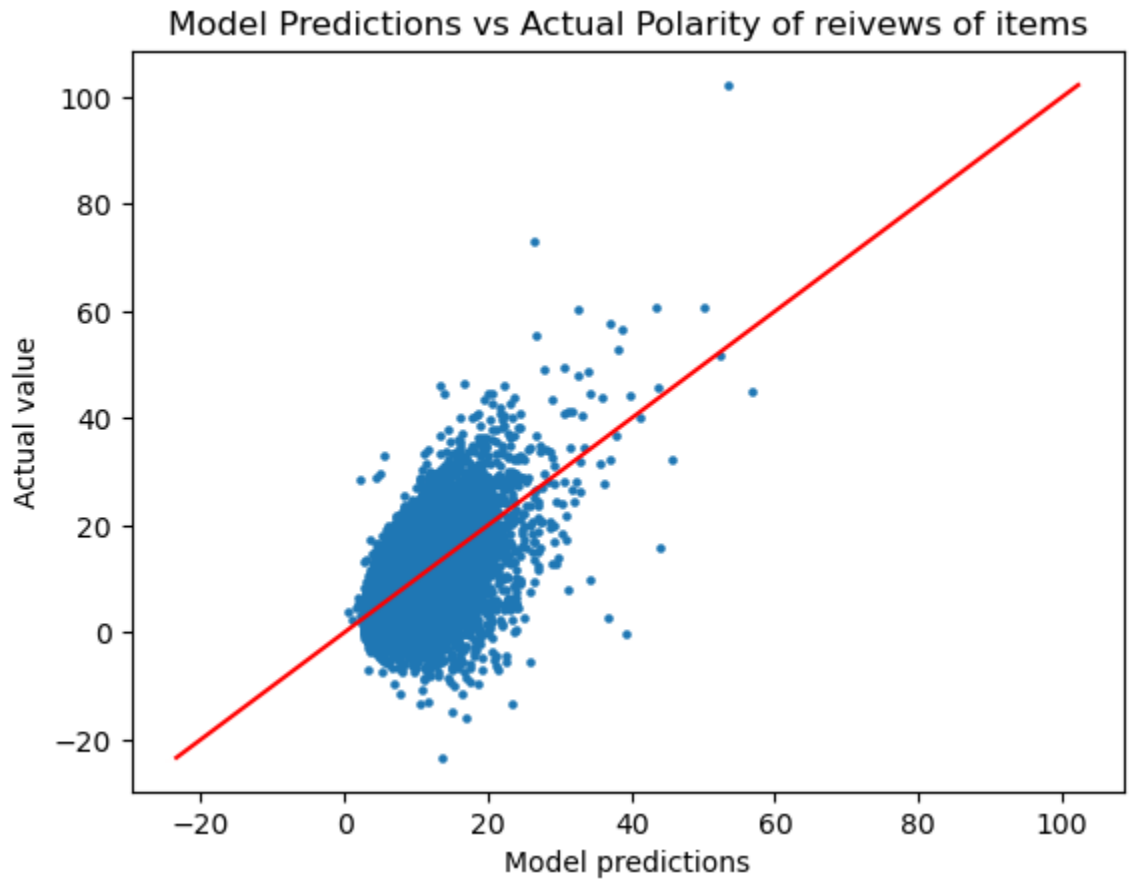

Figure 11. Results of Model 2: Multiple Linear Model on combined data. (A) SD and RMSE calculations of model and baseline. (B) Plot of actual vs predicted values of polarity scores.

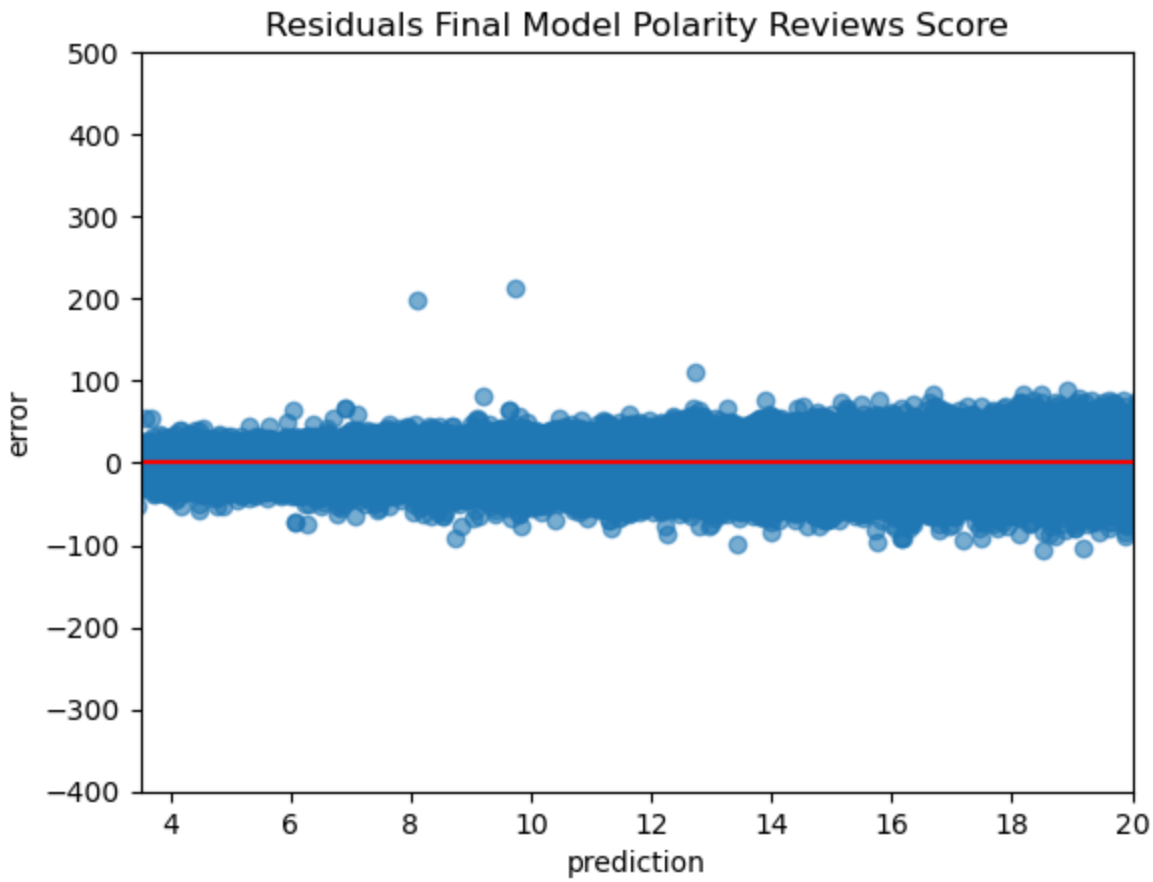

Figure 12. Residual plot for Model 3 on combined data

Conclusion

In our research analyzing Amazon Review Dataset, we uncovered evolving sentiment trends in user reviews, revealing critical insights for understanding consumer feedback. Our study highlighted the changing landscape of sentiment analysis and identified pivotal features influencing sentiment prediction.

We recognized a notable skew towards five-star ratings in the overall review scores and devised models to predict sentiment scores, rather than relying solely on discrete ratings. Our

models, including linear regression and Lasso regression with regularization, exhibited success in tracking changes in overly positive and negative reviews over time.

Model 1, particularly, demonstrated remarkable accuracy, achieving significant reductions in standard deviation for overly positive and negative reviews, correlating sentiment scores with temporal shifts.

Model 2 showed acceptable accuracy, and we found that incorporating one-hot encoded categories or regularization in Model 3 doesn't further improve predictive performance, evidenced by a greater error values, but a balanced residual plot.

Ultimately, our study sheds light on the dynamic nature of sentiment trends within Amazon reviews and emphasizes the importance of leveraging sentiment analysis for a nuanced comprehension of consumer feedback. Future avenues could explore alternative sentiment analysis methods, to confirm we get the same results. Additionally, as shown in the introduction section, the current research generally uses machine learning and neural network-based approaches for enhanced predictive accuracy in understanding online consumer sentiments, so we could try to use more sophisticated methods to improve performance.

References

Bonta, Venkateswarlu, Nandhini Kumaresh, and N. Janardhan. "A comprehensive study on lexicon based approaches for sentiment analysis." Asian Journal of Computer Science and Technology 8.S2 (2019): 1-6.

Borg, Anton, and Martin Boldt. "Using VADER sentiment and SVM for predicting customer response sentiment." Expert Systems with Applications 162 (2020): 113746.

Budhi, G.S., Chiong, R., Pranata, I. et al. Using Machine Learning to Predict the Sentiment of Online Reviews: A New Framework for Comparative Analysis. Arch Computate Methods Eng 28, 2543–2566 (2021). https://doi.org/10.1007/s11831-020-09464-8

He, R., & McAuley, J. (2016). Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. Proceedings of the 25th International Conference on World Wide Web, 507–517. https://doi.org/10.1145/2872427.2883037

Hutto, C., & Gilbert, E, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". Proceedings of the International AAAI Conference on Web and Social Media, 8(1), (2014), 216-225. https://doi.org/10.1609/icwsm.v8i1.14550

Hutto, Clayton, and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." Proceedings of the international AAAI conference on web and social media. Vol. 8. No. 1. 2014.

McAuley, J., Targett, C., Shi, Q., & van den Hengel, A. (2015). Image-based Recommendations on Styles and Substitutes.

Ni, J., Li, J., & McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). https://doi.org/10.18653/v1/d19-1018

Shrestha, N., & Nasoz, F. (2019). Deep Learning Sentiment Analysis of Amazon.Com Reviews and Ratings. International Journal on Soft Computing, Artificial Intelligence and Applications, 8(1), 01–15.

T. U. Haque, N. N. Saber and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," 2018 IEEE International Conference on Innovative Research and Development (ICIRD), Bangkok, Thailand, 2018, pp. 1-6, doi: 10.1109/ICIRD.2018.8376299.

Tan, W., Wang, X., & Xu, X. (2018). Sentiment Analysis for Amazon Reviews. Stanford University.

Ulf Norinder & Petra Norinder (2022) Predicting Amazon customer reviews with deep confidence using deep learning and conformal prediction, Journal of Management Analytics, 9:1, 1-16, DOI: 10.1080/23270012.2022.2031324

Xu, Y., Wu, X., & Wang, Q. (2014). Sentiment Analysis of Yelp‘s Ratings Based on Text Reviews. Stanford University. [Original source: https://studycrumb.com/alphabetizer]