CSCI 5302 - Final Project

Patrick Connelly Aneesh Khole Uttara Ketkar

2024-04-28

Table of contents

1 Introduction

1.1 Abstract

We explore the research performed in Guha Majumder, Dutta Gupta, and Paul (2022), and seek to further it via their recommendations for predicting the perceived usefulness of online customer reviews to potential customers. Our work focuses on the expansion and generalization of their multiple linear regression model. To check the model's general applicability, we collect additional products and reviews, and do so for the same products from multiple e-commerce websites to examine whether such models are applicable to any platform, or if the models may be platform-specific. Furthermore, we explore use of additional features and coefficients, and use of other prediction and classification models to assess the degree to which a customer review is useful to future customers.

1.2 Concept and Motivation

Customers, when searching for products with specific features and aspects, need sufficient information to make a decision as to whether to procure a specific product. According to research by Guha Majumder, Dutta Gupta, and Paul (2022), if a customer can gather and understand product quality before the purchase, it is considered a search good, while experience goods are those which must be purchased or experienced to evaluate them. When a product is more in the directon of experience vs. search-based, other customers' experiences can shed light on its features and return on investment than information directly from the vendor can. Having reviews from reliable sources with sufficiently detailed information can enable greater confidence in a purchase, improved customer satisfaction, and smooth the process of ecommerce for customers.

We seek to expound upon the research of (Guha Majumder, Dutta Gupta, and Paul 2022) to explore additional recommended research areas to improve upon and increase the general applicability of the model.

1.3 Our Research Plan

Guha Majumder, Dutta Gupta, and Paul (2022) provided the following summary model for what aspects and features they took into consideration in predicting the perceived usefulness of a customer review in Figure 1.1.

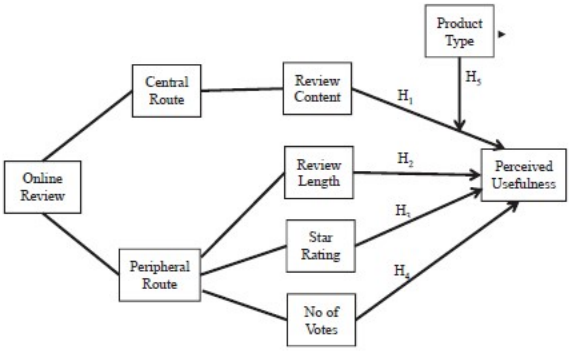

Figure 1.1: Model Overview

Furthermore, the authors provided the following areas for recommended additional research at the conclusion of their paper:

1. Expand the number of products beyond 3 items (one search, one experience, one mixed) to better generalize the model.
2. Explore customer or reviewer metadata for classifying reviewer types to enhance model performance.

We seek to examine the above two above items, and to explore the possibility of assessing a scale for products to determine the extent to which they are a search or experience-based product. We further seek to inspect additional potential modifiers to the underlying model for statistical and operational applicability; we've sought out work from other research teams to identify potential methods we can leverage to pursue these ends.

- Determining the polarity of a customer review by employing a classifier such as Naive Bayes.
- Using Kansei engineering approaches to convert unstructured product-related texts into feature–affective opinions.
- Attempting to assess the reliability of a customer's review based on star-rating and a 'sentiment score' of their textual feedback.

Exploring methods employed within each of combinations of these research efforts, we will pursue potential improvements on the models outlined in Guha Majumder, Dutta Gupta, and Paul (2022). We will examine additional products and product types between multiple e-commerce websites (BestBuy, Target, Amazon). A summary of our explorations are depicted in Figure 1.2.

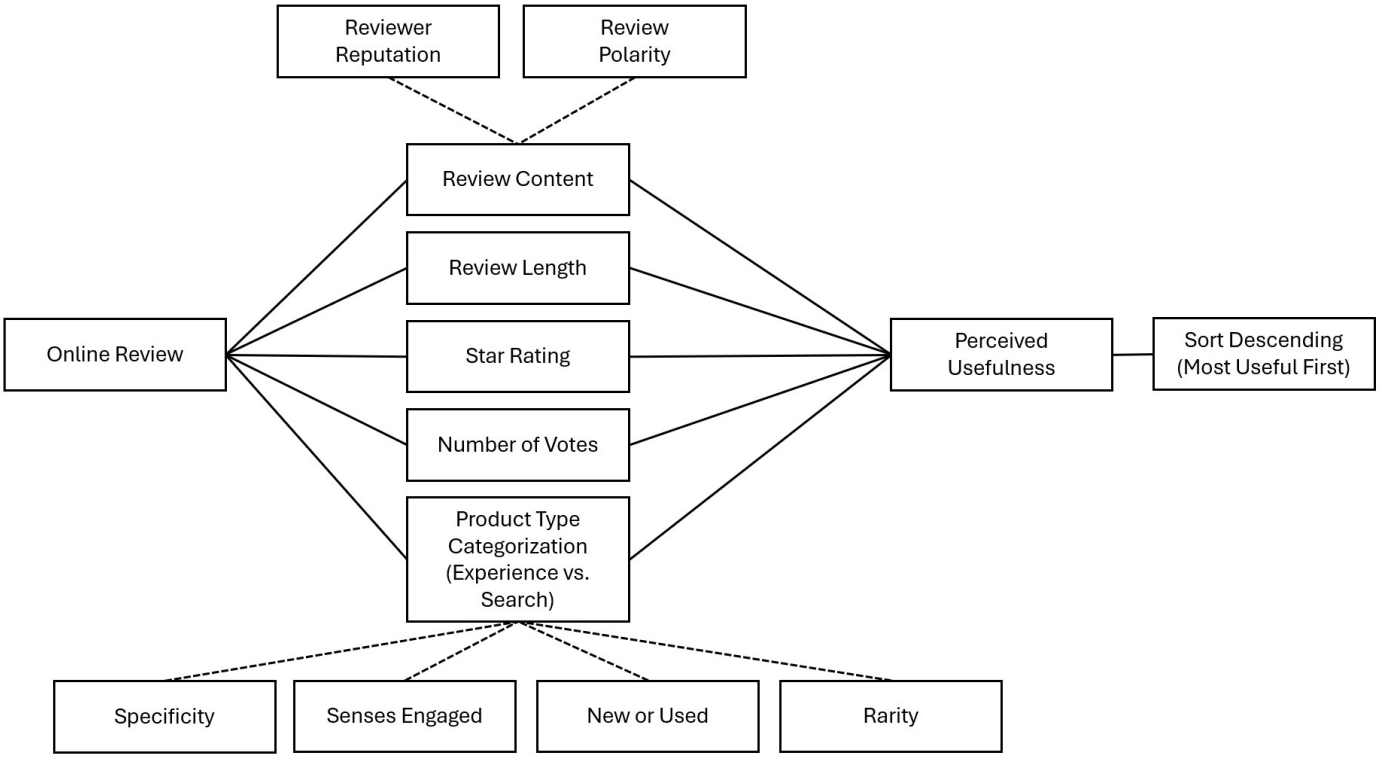

Figure 1.2: Model Modification Goals

This is not final, but what we plan to explore. If any metrics or measurements are found to not be significant in analysis and prediction of usefulness of a review, we will seek to explain the relationships (or lack thereof) and modify the final model accordingly. By incorporating these additional measures, we may be able to improve upon and generalize the original model to multiple product types across multiple e-commerce vendors.

1.4 Why It Matters

Feedback from customers can be beneficial to both vendors and consumers, but it is not always ordered by the most informative or beneficial feedback first. Certain features of products, reviews, and reviewers (such as reviewer reliability, review quality, product quality, specificity and detail of the product, amongst others) can impact the usefulness of the feedback on a customer-by-customer basis. Level of detail, star-rating, and number of votes that support the review as being useful to a customer can all help determine its usefulness to other customers. Leveraging metrics and data associated with a product, a review, and a reviewer together may allow for online vendors to improve consumer e-commerce experience, support identificaion of issues with product quality and sales, and enable vendors to adjust practices in product marketing, inventory, and manufacture.

Examining additional product types could support a generalization of the authors' methodology to other products. Furthermore, the exploration of a sliding scale for search vs. experience-based products can further support generalization and business goals. Producing a reliable scale and methods for classifying a products' degree of being experienced-based can inform vendors on:

- How to best sort product reviews.
- Examine what are the most helpful reviews to know the performance of the product alongside customer experience and sentiment.
- Adjust the product, its marketing, or future production based upon market efficacy.
- Understand the emotions a customer wants to express through a review is crucial as it will affect the "recommendation score" of that particular product or a different one from a similar category.
  - To contribute in determining this recommendation score, we can use a probabilistic machine learning algorithm like Naive Bayes to determine the polarity (positive, negative, or neutral) of customer reviews.
  - Typically used for amending product design, Kansei Engineering can be used to incorporate human emotional responses into evaluation of a customer review.
- Determine which customer is trustworthy, meaning who has actually purchased the product versus a customer who gave a false review. Based on the 'customer reputation score', our aim is to classify customers into groups to judge reviewer reliability. This has two main aspects:
  - Star-rating score which is a discrete scale that tells the inclination of a customer.
  - Text review 'sentiment score' using NLP that explains customer opinions based on words.

1.5 Literature Survey

- Guha Majumder, Dutta Gupta, and Paul (2022)
  - Examined multiple-linear regression modeling to calculate the usefulness of an online review based upon type of product (search vs. experience), review sentiment, review star rating, review length, and number of votes for the review as being "useful". Suggested exploration using larger number of products as well as customer/reviewer metadata.
- Hu, Gong, and Guo (2010)
  - The proposed system employs a two-step process for opinion mining: identifying opinion sentences using a SentiWordNet-based algorithm and extracting product features from all reviews in the database. This feature extraction function focuses on identifying commonly expressed positive or negative opinions before extracting explicit and implicit product features.

- Rajeev and Rekha (2015)
  - This paper presents techniques like Opinion mining, feature extraction and Naives Bayes classification for review polarity determination. The authors suggest performing both Objective and Subjective analysis of features by considering qualitative and quantitative features of the data respectively.
- Wang et al. (2018)
  - Authors have proposed a solution by implementing Kansei engineering and text mining simultaneously which will help customers in decision making process. It helps to categorize reviews into multiple sections and perform text mining by NLP techniques like Sentence segmentation, Tokenization, and POS tagging.

1.6 Research Questions

- Can the model from Guha Majumder, Dutta Gupta, and Paul (2022) be generalized with:
  - larger volume of products and product types from which to mine data?
  - a sliding scalar multiplier representing the degree to which a product is a "search" (0) or "experience" (1) product?
  - Adding modifiers to review content based upon:
    - Customer / Reviewer reliability and reputation?
    - Review Polarity?
- Can the polarity of reviews be judged accurately by using a Naive Bayes classification model? Hu, Gong, and Guo (2010)
  - What is the impact of different feature extraction methods (e.g., bag-of-words, TF-IDF) on the performance of Naive Bayes classification model? Wang et al. (2018)
- Can products be classified on their degree of being search or experience based by examining product variables such as:
  - Degree of specificity in the product description? (e.g. level of detail, length, numeric values, descriptive values may suggest the product is more search than it is experience-based)
  - Whether the product is offered in brand-new condition only, or offered as new, used, or refurbished? (e.g. refurbished products may be more search products than they are experience products)
  - Which of the 5 senses the product engages? (e.g. engagement of more senses, or engagement of solely specific senses like hearing and vision may suggest more experience-based than search based; examine relationship between search and experience vs. senses engaged)
  - Item rarity (limited production or unique items vs. bulk-produced items)? (e.g. limited production products may be more experience-based than search-based)
- Can newer natrual language processing libraries provide a better fit for Review Content metrics examined by Guha Majumder, Dutta Gupta, and Paul (2022)?
- How does sentiment in customer reviews correlate with customer satisfaction metrics or sales figures for a particular product?
- Can we categorize customer reviews based on customer experience and sentiment?

- Do specific product star ratings tend to incite more reviews, and if so, how does this impact the overall reputation measurement?
- Are specific quality descriptors in text-based reviews (e.g., 'enthusiastic', 'disappointed') strongly associated with certain rating levels, and how does this association affect product reputation?

1.7 Goals / Definition of Success

- Replicate similar results to Guha Majumder, Dutta Gupta, and Paul (2022) with similar product types
- Expound upon Guha Majumder, Dutta Gupta, and Paul (2022) with additional products, including:
  a. Original products from (paper): Digital Music, Video Game, and Grocery Item
  b. Additional products (Amazon and Target): Furniture Items, Clothing Items, Home Appliances, Books, Cosmetics, Cleaning supplies
  c. Additional Proucts (Amazon, Target, BestBuy): Electronics
  d. Verify goodness of fit of original model
- Determing best metrics and/or modifiers for Review Content and Customer Reliability
- Achieving similar or better fit than original paper's modeling; extrapolate to other product types.
- Determining strength of correlation metrics (support, confidence, lift) between Naive Bayes' classifier for review polarity Hu, Gong, and Guo (2010)
  – Integrate with the model and test if Naive Bayes shows strong correlation metrics.
  – Compare and contrast the model with and without incorporation.
- Successful computation of reputation scores for reviewers
  – Check applicability across all sites used for determination of validity within the model.
  – If valid and applicable, execute model against testing data set to ensure it holds.

1.8 Project Schedule / Timeline

Below in Table 1.1, we lay out the major tasks, deliverables, and their respective due dates for this effort.

Table 1.1: Major Project Tasks

Table 1.1

Task	Due Date
Milestone 1 Submission	Feb 26 2024
Product Identification and Selection	Feb 28 2024
Vendor Identification and Selection	Feb 28 2024
Data Collection	Mar 8 2024
Data Cleaning/Pre-Processing	Mar 17 2024
Milestone 2 Submission	Mar 20 2024
Review Classification (Naive Bayes, Kansei)	Mar 27 2024
Product Classification	Mar 27 2024
Reputation Classification	Mar 27 2024

Task	Due Date
Exploratory Data Analysis	Mar 31 2024
Milestone 3 Submission	Unknown
Model Selection	April 7 2024
Model Testing:	April 11 2024
Complete Final Paper / Milestone 4	April 17 2024

2 Data Collection and Exploration

2.1 Data Collection Overview

The original efforts by Guha Majumder, Dutta Gupta, and Paul (2022) selected three products, all listed on Amazon for sale. In our efforts, we leveraged python Selenium, urllib, and Beautiful Soup to scrape data from 20 different products across multiple websites (Amazon, BestBuy, and Target). Where possible, we sought to collect the exact same 20 products from each site and customer feedback associated with each.

As part of collection, to the greatest extent we were able, we cleaned information *during* the scraping process. Doing this enabled us to have minimal cleaning efforts after collection. Post collection, remaining items such as handling and removing special characters, unicode characters, addressing customer reviews written in foreign languages, and addressing misspellings remained necessary.

In terms of simplicity for scraping our data, we manually identified a list of products from each of the aforementioned sites. Our team divided responsibilities to produce scraping code customized for each of the three websites.

2.2 Data Collection Details

In collecting our data, in order to adhere to the model implemented by Guha Majumder, Dutta Gupta, and Paul (2022), we required the following data points:

Table 2.1

Variable	Data Type
Product Title	string
Product Category*	string
Product Details/Specs	string
Product Cost	float

For the product category variable - we may add our own manual categorization. Guha Majumder, Dutta Gupta, and Paul (2022) manually set the value for this variable. Part of the intent of our research is to seek out means and methods to replace this variable with a continuous scale (ranging from 0 for a "search" good, to a 1 for an "experience" good).

As an initial proxy for this variable and to operationalize it, we leverage a measure of subjectivity for the product - namely how subjective (e.g. how many adverb, adjective, and other word modifiers) are present within the details and specifications of a product. A product that more aligns to a "search" product, we hypothesize, will have fewer modifying words and be oriented toward the facts of the object.

For example, a desk has specific dimensions for length, width, and height, an associated weight, and material from which the desk is made, and possibly some warranty information - all of which are likely to be contained within the product description and specifications. We would characterize such a good as a "search" good (or a 0 on our scale). Leveraging existing language processing tools should allow us to calculate a value for subjectivity in the product's description and specifications.

Initially, we'll explore product subjectivity in the combination of the specification and the description, though it may be necessary to explore product subjectivity solely within one of these fields or the other to pursue our modeling.

Table 2.2: Review Data Required

Table 2.2

Variable	Data Type
Verified Purchase	boolean
Star Rating	float
Review Content	string
Useful Votes	integer

In Table 2.2, we outline the specific datapoints we sought out for reviews across each website. Guha Majumder, Dutta Gupta, and Paul (2022) leveraged star rating, review content (specifically the review length), and the number of votes for the review being useful as key measures in their research. To further their work, we plan on exploring the impacts of verified product purchasers and the impact of verification on how useful a review may be to potential customers.

Table 2.3: Additional Calculated Columns, Post Data Collection

Table 2.3

Variable	Data Type
Product Subjectivity	float
Review Length (Words)	integer
Review Subjectivity	float
Review Polarity	float

Post collection, we added the calculations listed in Table 2.3 to our review data and product data (less reputation score). Each of these calculations will allow us to better understand our underlying data and explore possibilities of where and how each may fit into models for review usefulness.

We have also established a master listing of all products for which we collected data and have associated arbitrary identifiers with the products. In instances where we've successfully pulled data for *identical* products from multiple websites, it can allow us to explore the impact on product and review metrics and investigate the listing site as a treatment variable.

For instance - exploring the impact of review subjectivity, polarity, length, and usefulness, based upon which site the product was listed.

2.3 Data Collection Procedures

We wrote code to allow us to gather information from each website. The general process for each e-commerce platform is similar. To alleviate any unnecessary burden for any of these websites, we manually identified URLs to the specific products we sought out to gather, and wrote our code to iterate through those URLs and pull the necessary data and features we sought. This manual identification also allowed us to ensure, in most cases, that we were getting the *exact* same product during data capture. This hybrid approach enabled higher certainty in getting the same product while also accelerating collection, structuring, and cleaning of product review information.

- Gathering from Amazon (All Products)

  - Product & Review data was scraped from Amazon's website using Python and Selenium. A Selenium WebDriver was utilized to automate web browser interactions. After navigating to product categories like electronics, home appliances, furniture, books, and grocery, Selenium's functions were employed to locate review elements. These elements were then parsed and collected, storing the data in a structured format i.e. a CSV file. Pagination handling was implemented to scrape reviews from multiple pages.
    * Challenge: Amazon's product "All Reviews" webpage HTML structure had 10 reviews per page with a "Next Page" navigation button that was clickable only up-to 10 review pages. This restricted our scope of the number of reviews being scraped per product to a maximum of 100.
    * Solution: Instead of scraping based on the "All Reviews" webpage, we decided to scrape reviews based on "star-rating" thereby, increasing our scope from a total of 100 reviews per product to having a maximum of 100 reviews per star rating i.e. 5*100 = 500 reviews per product.

- Gathering from BestBuy (Electronic Products, Furniture Item(s)? - no grocery or clothing)
  - Just like Target and Amazon, even BestBuy has dynamic content on its web page. We employed Python with Selenium to automate the exploration of product pages, unveiling hidden content, and harvesting essential data. Employing Selenium's functionalities, we initiated the traversal process, enabling the program to automatically expand pertinent sections to uncover additional information. By targeting elements such as product details and reviews, we orchestrated the seamless extraction of critical fields from each product's page. This automated approach allowed us to efficiently parse through an extensive array of reviews, ensuring a comprehensive analysis of user feedback for the products under scrutiny. We systematically stored the extracted data in our records tables for further analysis and reference.

- Gathering from Target (All products)
  - Target has dynamic content on their webpages. We used Python Selenium to navigate to product pages and automate the selection of items needed to expand sections to reveal additional data. We also automated the process of expanding out all reviews so as to iterate through and parse the content of every review for each product in question.

2.4 Data Exploration

After collection and cleaning, we plan to explore our data via visualization, seeking to answer key research questions.

- Is the price of a product higher, given it's offered on Amazon, BestBuy, or Target?
- Is a product's star rating affected by which e-commerce platform is selling it?
- Is there a substantial difference in number of product reviews on one e-commerce platform vs. another?
- Is one e-commerce platform more likely to have input and feedback on reviews (i.e. higher proportion of "this review is helpful" votes to total number of reviews)?
- What is the difference in the level of detail provided in product descriptions (e.g. for the same product) across each e-commerce platform?
- Do certain product categories perform better on specific platforms?
- Are users more likely to leave reviews on one platform over another?
- Do customers show different purchasing behaviors based on promotional strategies employed by platforms?

Structuring our data properly during the collection process will enable us to explore and answer these questions.

2.5 Data Exploration and Visualization

For our data exploration, we plan to examine solely the reviews for which we have data from all of our websites. Due to the nature of the vendors, not all offer the same products online. We've included some unique products from each site (and may even gather more), but will exclude them from initial analysis.

The common items between all 3 websites include the following:

Table 2.4

product_title
Samsung Galaxy S22 Ultra 5G Unlocked (128GB) Smartphone - Burgundy
HP DeskJet 2755e Wireless All-In-One Color Printer, Scanner, Copier with Instant Ink and HP+ (26K67)
JBL Charge 5 Portable Bluetooth Waterproof Speaker - Target Certified Refurbished
TurboTax 2023 Deluxe Federal and State Tax Software
Hamilton Beach 4 slice Toaster 24782
LG 65" Class 4K UHD 2160p Smart OLED TV - OLED65C3
GE JES1460DSBB 1.4 Cu. Ft. Black Counter Top Microwave
Doritos Nacho Cheese Flavored Tortilla Chips - 14.5oz
Crest Cavity & Tartar Protection Toothpaste, Baking Soda & Peroxide - 5.7oz/3pk
OXO POP 3pc Plastic Food Storage Container Set Clear
Hogwarts Legacy - Xbox Series X
Star Wars Jedi: Survivor - PlayStation 5

The reason for only examining common products is to check for comparability and similarity of the products associated variables (e.g. product subjectivity, review subjectivity, review polarity, star rating, and so forth) between the websites. If they are similar or comparable, it may mean that we could use single models to make predictions on the usefulness of customer feedback. If they are substantially dissimilar, it may mean that modifiers are needed based upon the e-commerce platform in which the product is listed.

We'll start by looking at distributions of some of these key variables, and check some of the common trends between them, potentially moving on to hypothesis testing of these variables to check for statistically significant differences.

2.5.1 Univariate Plots and Distributions

First, we want to examine the review content across all websites in a single, simple visual - a Wordcloud. Seeing common words and phrases can prime us for what we might expect to see in more detailed statistical plots.

Examining the Wordcloud, some larger words stick out ("easy", "good", "love", "need" and "great"). There don't seem to be very many negative singular words here as it pertains to these reviews. This may suggest that the content of reviews, generally, gravitates toward positivity in reviews. We will proceed to examine this with appropriate statistical plots.

Examining the histogram plots for star-rating by website, we can see that, generally, reviews tend to provide more positive than negative feedback for the selected products, supporting what we see coming out of Figure 2.1

Across all three websites, there appears to be consistency with adherence to, and issues with, the normal distribution for subjectivity. These charts suggest sufficient normal distribution of review subjectivity (degree of inclusion of word modifiers such as adverbs and adjectives).

Therre seems to be slight skewness in the tails of these Q-Q distributions. Filtering off some of the outliers may grant us reasonable relevance and assurance to perform hypothesis testing and evaluation of these variables across sites (e.g. ANOVA, F-Testing, etc).

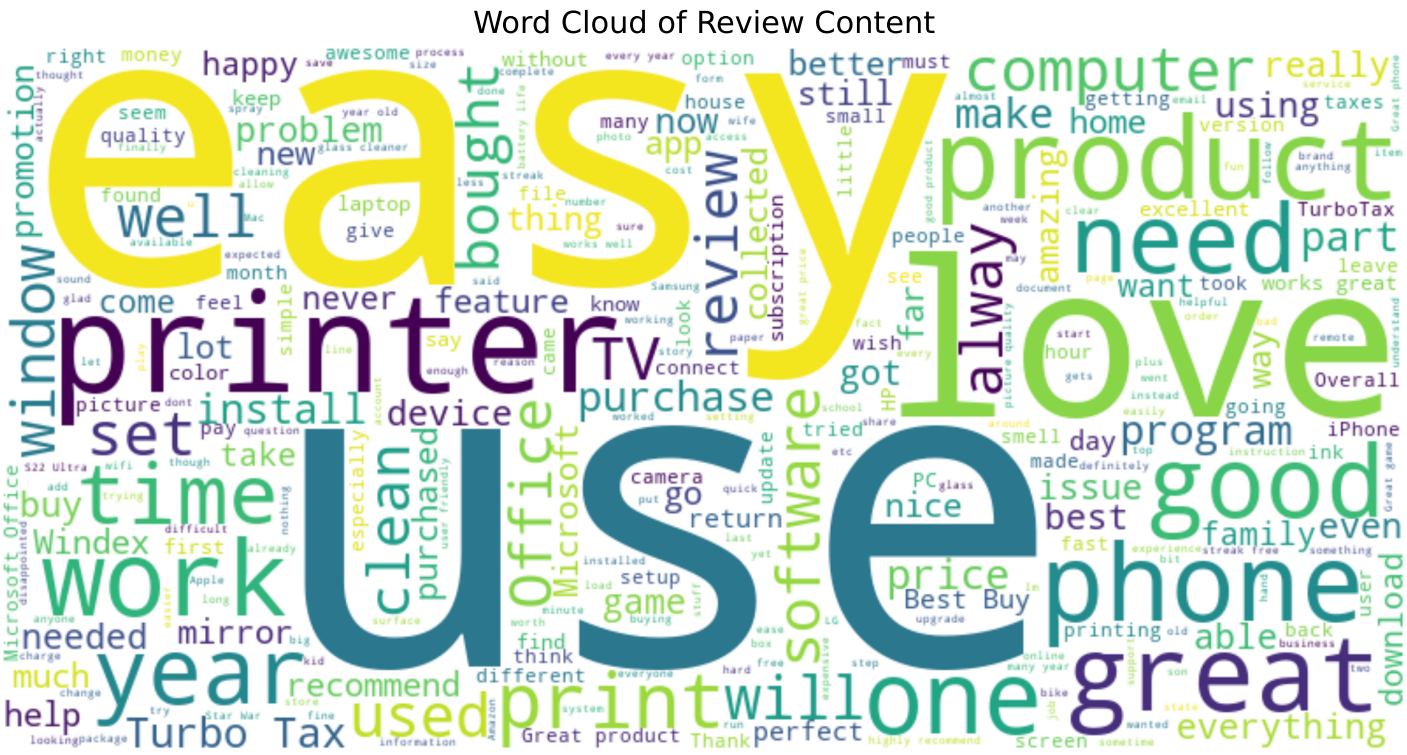

Figure 2.1: Wordcloud of review content

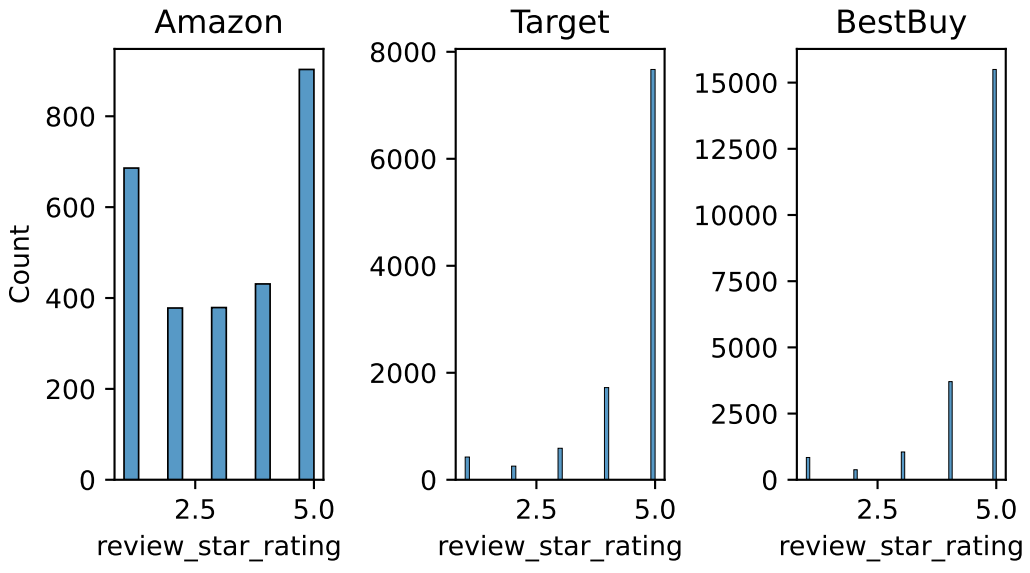

Figure 2.2: Histogram Plot (star-rating, by-site)

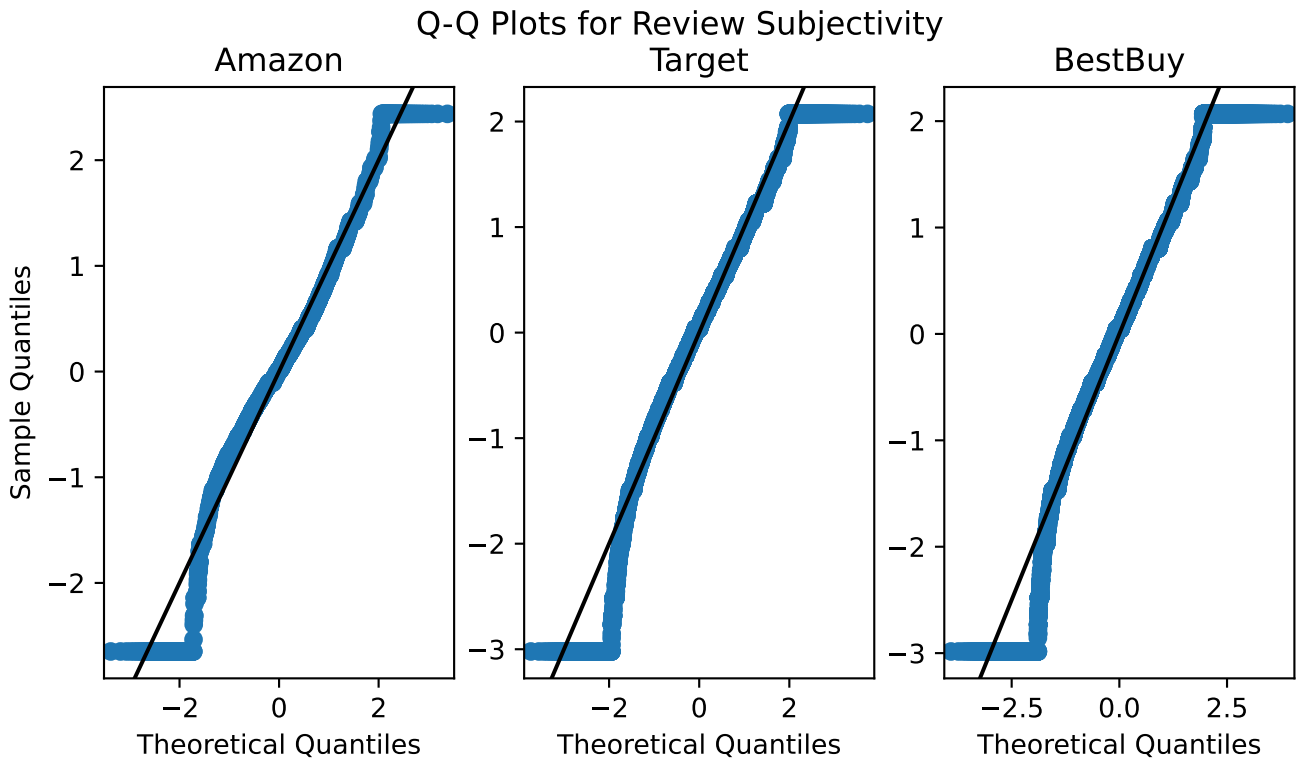

Figure 2.3: Q-Q plots (star-rating, by-site)

We'll try plotting the same Q-Q plot with outliers removed. To remove outliers, initially, we leveraged the inter-quartile range of each variable and excluded any records for which the variable was more than $1.5 \cdot \text{IQR}$ away from the 1st and 3rd quartiles.

Before we proceed to re-examining the Q-Q plot with outliers removed, we'll examine boxplots for these variables to examine the prevalence of outliers.

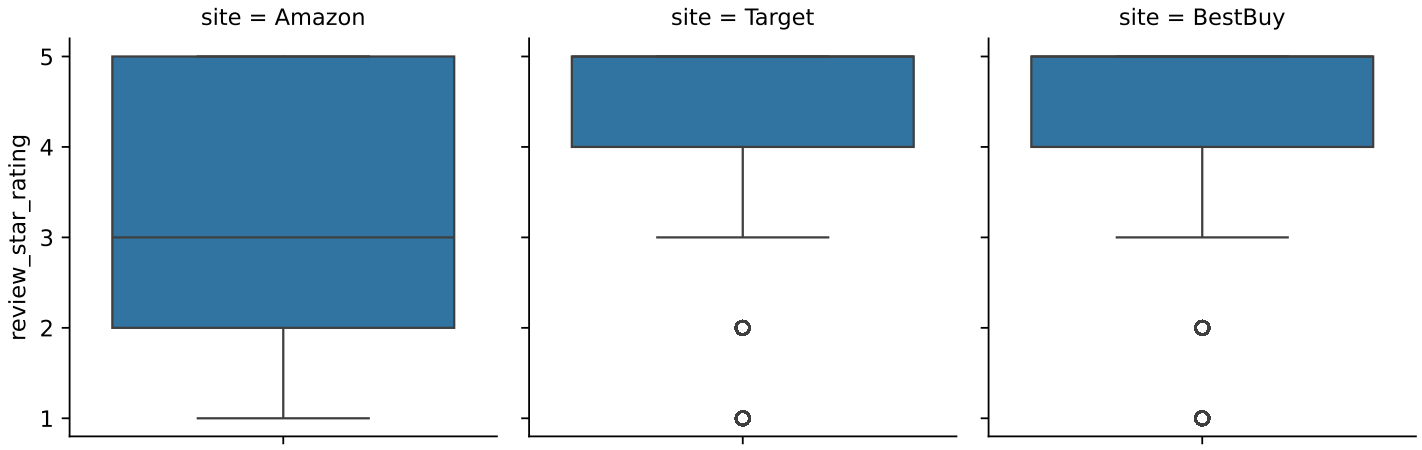

Figure 2.4: Boxplots - Review Star Rating

Boxplots for star ratings on both Target and Amazon are generally higher with outliers on the lower-end of the 1 to 5 scale. Amazon, however, seems to have a wider spread of information

Boxplots for reviewe polarity suggest common threads between BestBuy and Target in terms of the number summary (min, max, quartiles, and outliers at the lower end). More notably, the polarity (or how positive or negative the content of the reviews are) generally tends toward positive. Amazon, on the otherhand, seems to

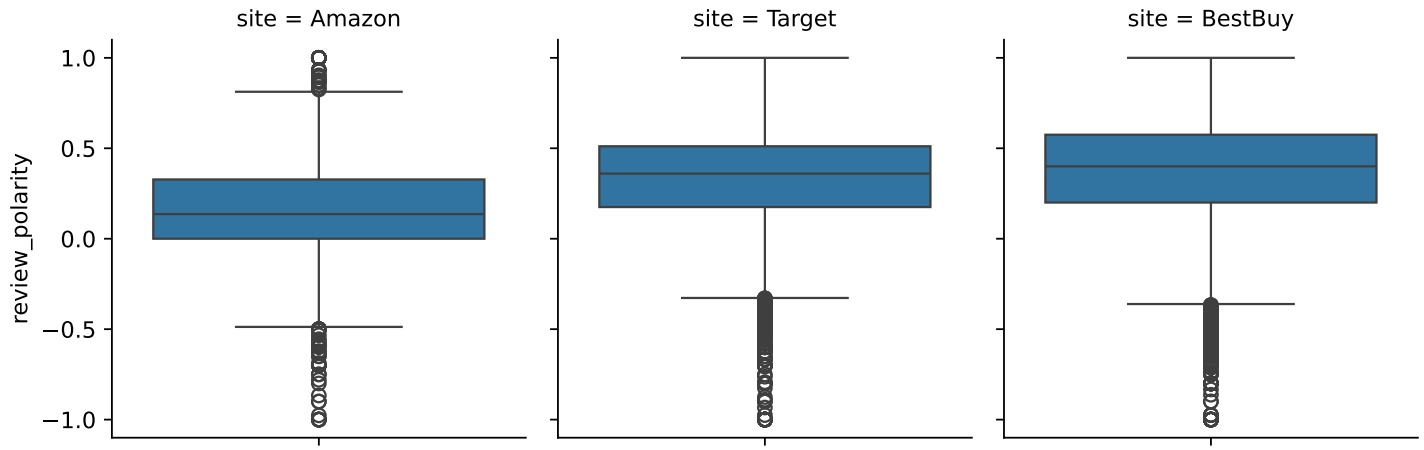

Figure 2.5: Boxplots - Review Polarity

show a lower center of mass and a narrower spread, with outliers to both extremes for positive and negative polarity.

Next, we'll examine subjectivity in the same fashion.

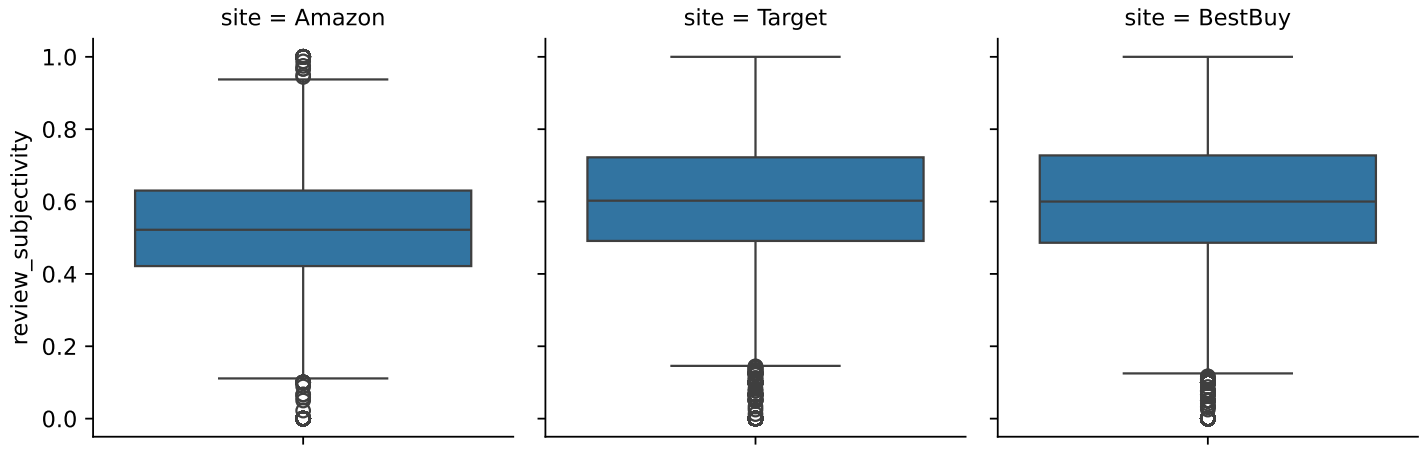

Figure 2.6: Boxplots - Review Subjectivity

Subjectivity, generally, seems to follow the same trends as review polarity. This suggests that these reviews could come from similar or the same population in terms of polarity and subjectivity. Further statistical analysis would be needed to make a definitive determination here.

Now that we've examined the centers and spread for these variables and understand where some of their outliers may exist, we'll examine filtering those outliers from their Q-Q plots.

First - Subjectivity.

It seems that our adjustment for outliers sufficiently made corrections for normality across the sites to better adhere to the normal distribution on the lower tail. We may need to make further adjustments on the upper tail to further refine data selection for our training dataset. Amongst the over 34K reviews in the common dataset, approximately 25.4K reviews remain after removing these outliers using this method.

After identifying additional means to filter the data, these methods should suffice in support of using review subjectivity as a feature within various models.

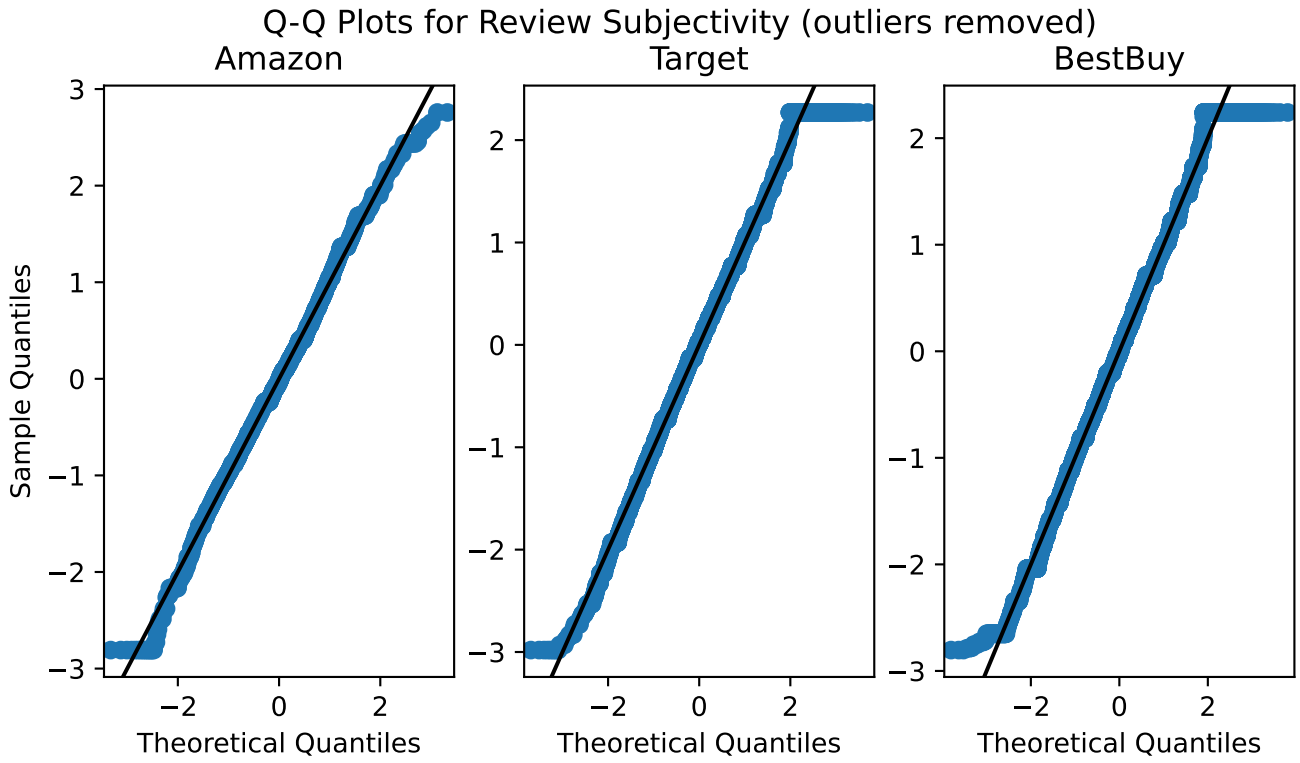

Figure 2.7: Q-Q plots (star-rating, by-site, outliers removed)

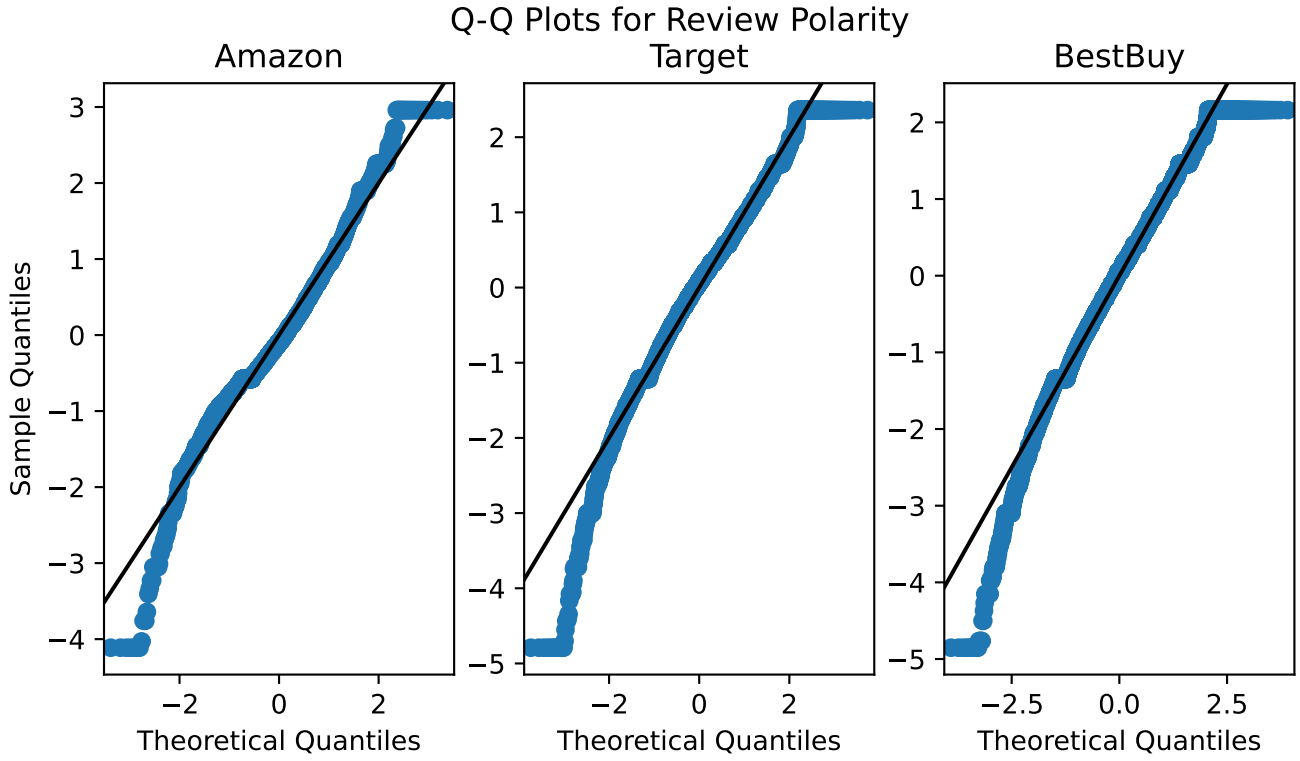

Figure 2.8: Q-Q plots (star-rating, by-site)

Similar to review subjectivity, review polarity has good adherence to the normal distribution (particularly on the quantile interval of [-2,2]). There are similar issues in the tails of these distributions as exist for review subjectivity.

As such, reduction in outliers may enable us to perform hypothesis testing during our model design and implementation. We'll examine the same methods of outlier removal as we did for review subjectivity.

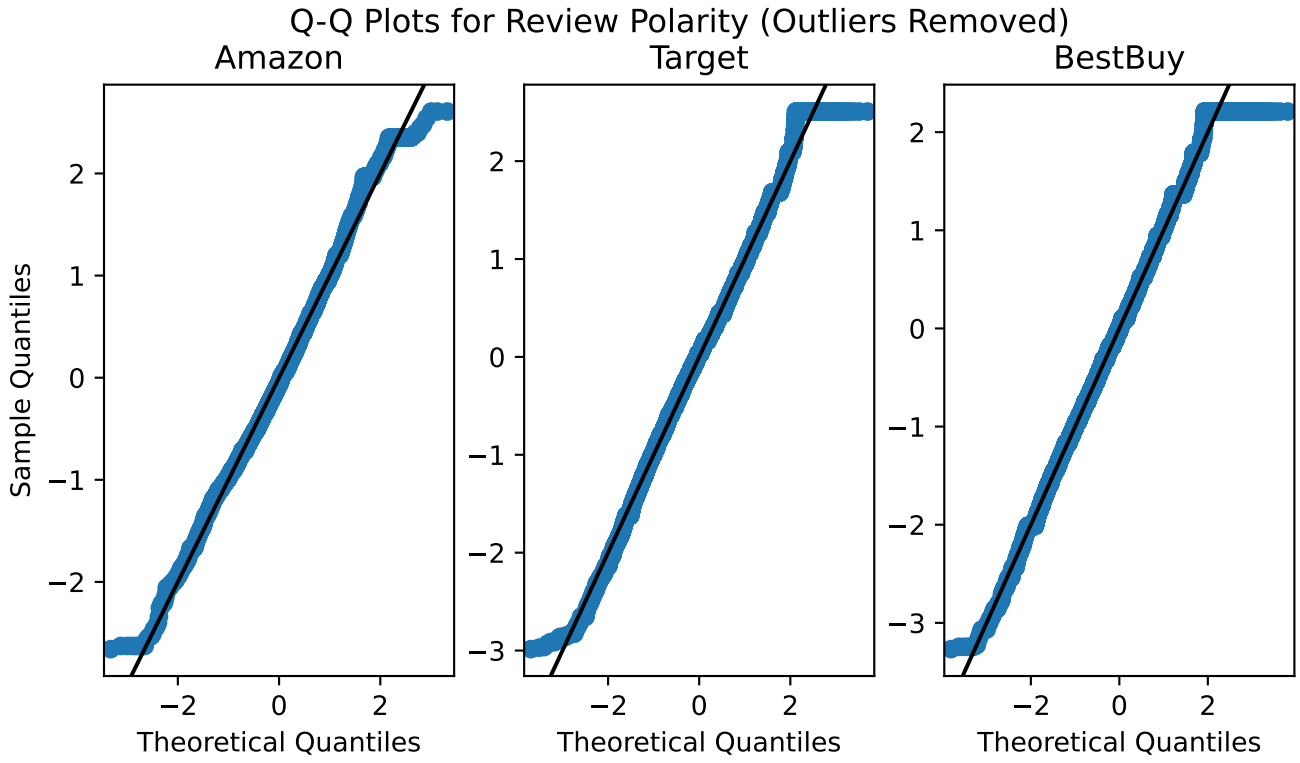

Figure 2.9: Q-Q plots (polarity, by-site, oultiers removed)

This method of removal seems to mirror that of review subjectivity, and as such, additional fitration of the dataset will be necessary to enable this feature's use within various models.

Another key distribution we must understand is that of our targeted response variable - how useful a review is, as voted by other customers. We'll plot the response as a pure density plot to explore it's shape.

The black lines represent random samples from the exponential distribution (with $E[X] = 1.9 \cdot \bar{V}$ with $\bar{V}$ being the mean for helpful votes within the distribution), and the green lines represent the distribution of helpful votes. It seems that, roughly, the distribution of helpful votes does follow the exponential distribution in the case of Amazon and Target.

Examining the plot of Figure 2.10, the distribution of helpful votes appears to be exponentially distributed on a per-website basis, with many reviews having an expected total count of helpful votes centered fairly low.

Knowing the distribution of our selected response variable will assist us in the modeling process. The nature of the response variable's distribution may require us to perform transformations on features and responses (e.g. if we pursue a multiple linear regression model).

2.5.2 Bi/Multivariate Plots

In Figure 2.11 and Figure 2.12, we observe the comparison of multiple features like review polarity, review subjectivity and verified purchases in the form of a Kernel Densoty Plot. The 2 different visuals depict the difference between the whole dataset and after the filter of `verified_purchase` = 1 is applied. This difference may lead to

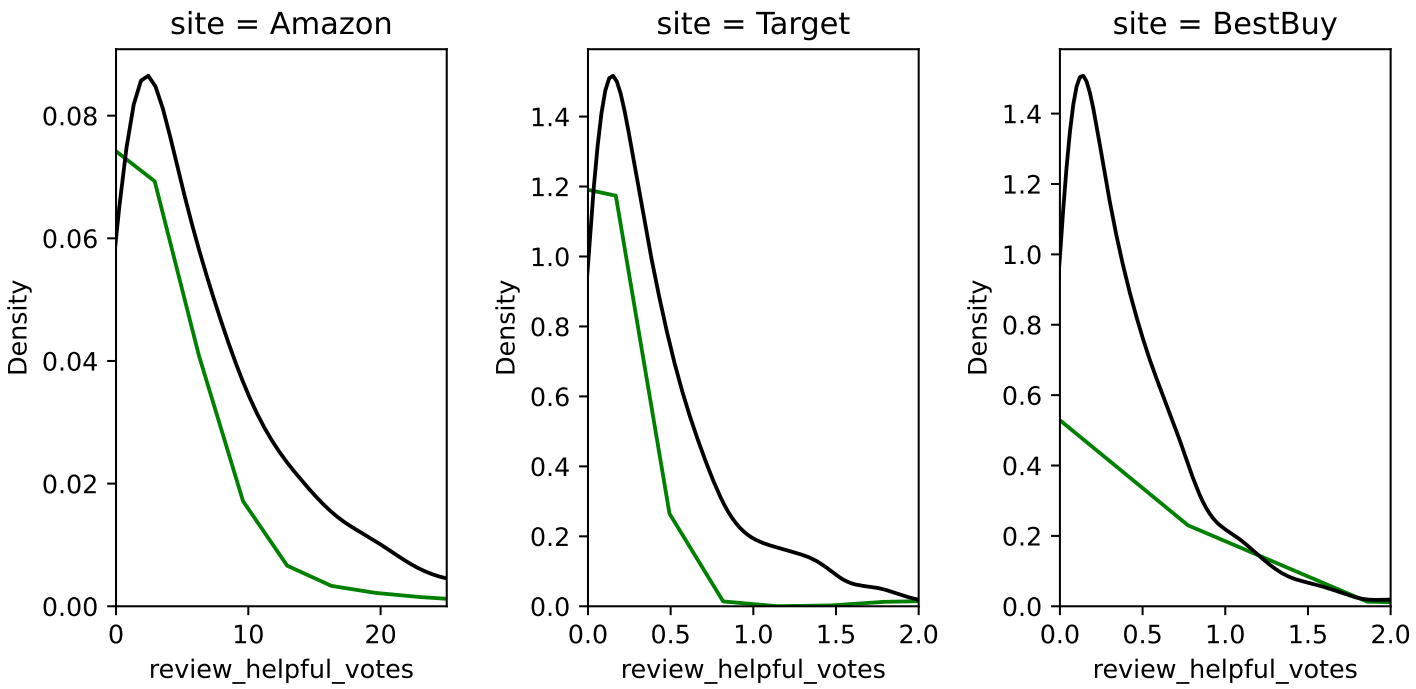

Figure 2.10: Distribution of Review Helpful Votes

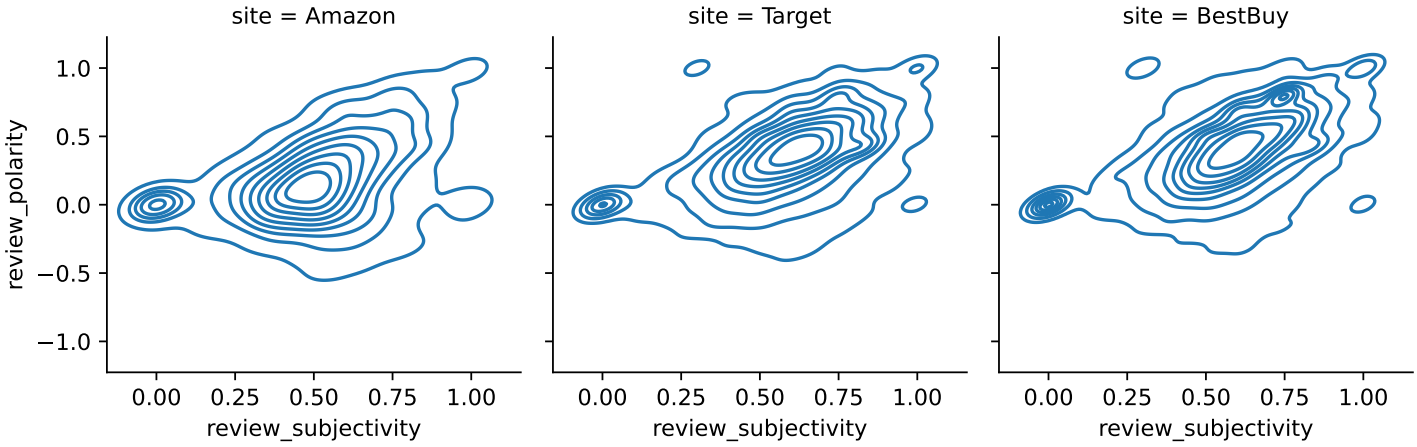

Figure 2.11: Bivariate Plot for Sensitivity and Polarity, by Site (all purchases)

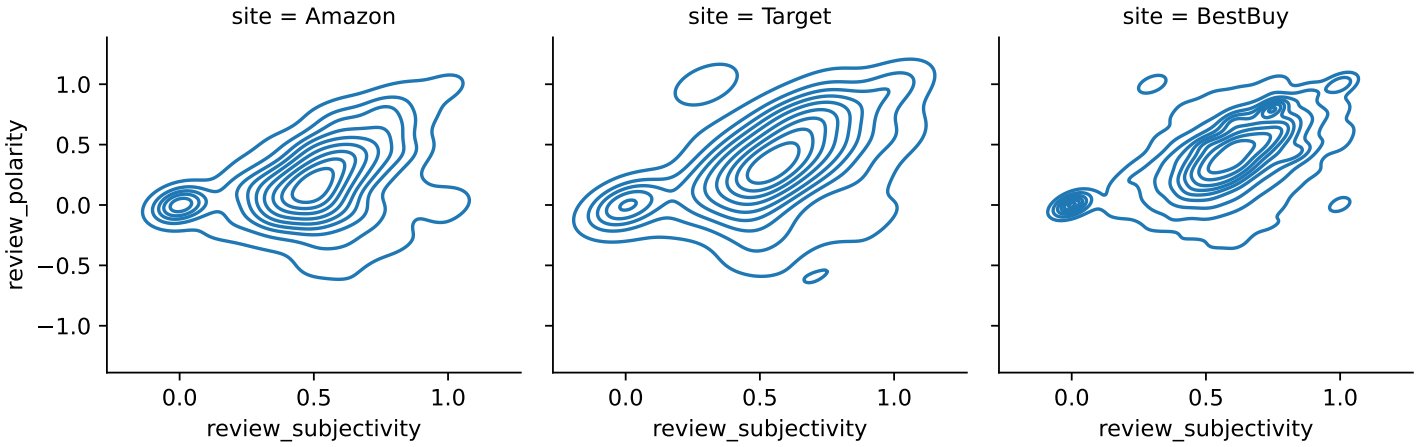

Figure 2.12: Bivariate Plot for Sensitivity and Polarity, by Site (verified purchases)

variations in the distribution and relationship between subjectivity and polarity of reviews across different sites, particularly if there are differences in the characteristics of verified and non-verified purchases.

We can see 2 major clusters at (0,0) which are mostly outliers where a review is very short in length. The second cluster around the area where subjectivity is about 0.5 suggests that as the review increases in subjectivity, i.e. the higher an opinionated a review is, the polarity also increases.

The isolated data points or "islands" outside of the main clusters suggest outliers or unique instances within the dataset. These isolated points may represent reviews that deviate significantly from the overall patterns observed in the data. They could indicate rare or extreme cases that warrant further investigation. For instance, these outliers might correspond to highly subjective or polarized reviews that are not typical of the majority of reviews.

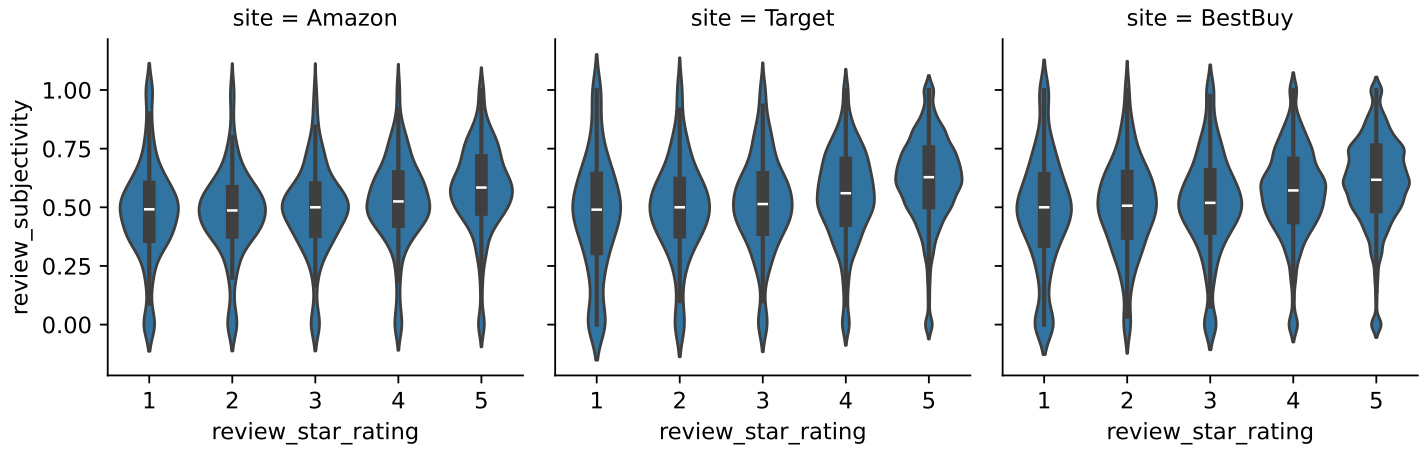

Figure 2.13: Violin Plots of Review Star-Rating vs. Subjectivity, by Site

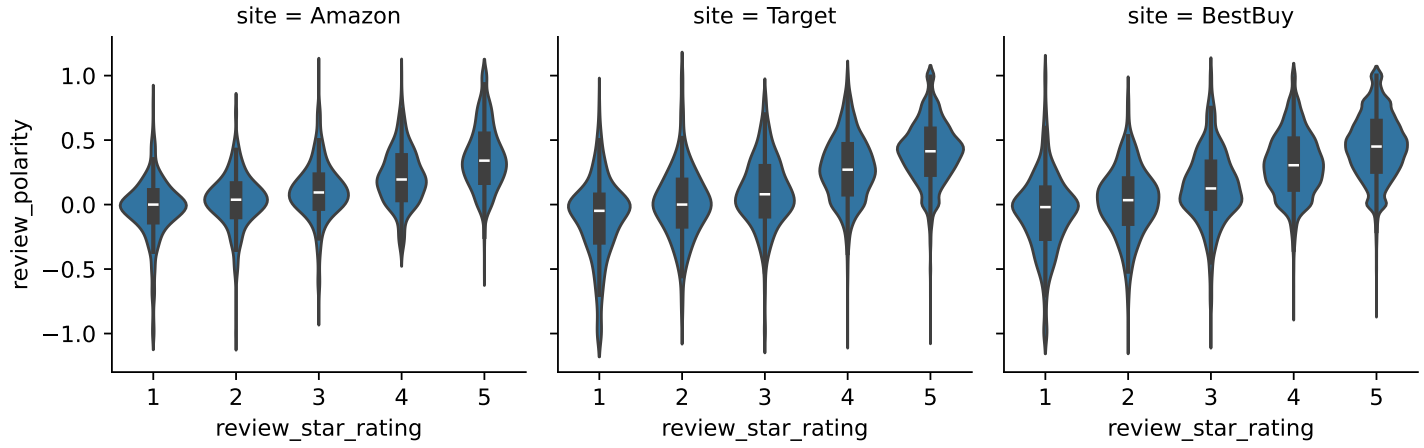

Figure 2.14: Violin Plots for Star Rating vs. Polarity, by Site

Generally, in Figure 2.13 and Figure 2.14, we see a trend for the median polarity and subjectivity of each review to increase as the star rating increases. We also see that, generally, the data suggest that we have a minimum of neutral polarity that tends towards positive as star rating increases.

Since both median subjectivity and polarity seem to increase with respect to star rating, such a correlation could be useful to us in multiple linear regression, and is generally useful to us for consideration when pursuing model development.

2.5.3 Hypothesis Testing for Key Feature and Response Variables

Some key features we plan to explore in our modeling include review subjectivity and review polarity. Knowing whether or not there is a significant difference for these features between the websites on which they're hosted will inform us during model selection, design, and implementation. As such, we'll perform ANOVA and Tukey Honest Significant Difference Tests on these variables between each site.

2.5.3.1 ANOVA Testing

To perform our ANOVA testing, we'll evaluate each dataset's review polarity and subjectivity as the mean measure, and the website on which the review was posted as the treatment variable. Prior to performing our one-way ANOVA, we'll filter the datasets down to eliminate outliers, such that the data may represent the outcomes depicted in Figure 2.7 and Figure 2.9. An assumption of ANOVA testing is that the source data (and its respective groups) adhere to the normal distribution.

We are leveraging Welch ANOVA and operating under an assumption that the variances between the groups are not equal, as visually evidenced in Figure 2.5 and Figure 2.6.

Hypotheses:

- Test 1:
  - $H_0 : \mu_{\text{Subj,Amazon}} = \mu_{\text{Subj,BestBuy}} = \mu_{\text{Subj,Target}}$
  - H_1 : at least one mean for review subjectivity is different.
- Test 2:
  - $H_0 : \mu_{\text{Polr,Amazon}} = \mu_{\text{Polr,BestBuy}} = \mu_{\text{Polr,Target}}$
  - H_1 : at least one mean for review polarity is different.
- For both tests, $\alpha = 0.003$

Table 2.5: Welch ANOVA Results (Polarity)

Table 2.5

Source	ddof1	ddof2	F	p-unc	np2
site	2	6112.904850	1362.569459	0.000000	0.106472

Table 2.6: Welch ANOVA Results (Subjectivity)

Table 2.6

Source	ddof1	ddof2	F	p-unc	np2
site	2	6353.711848	368.490784	0.000000	0.025819

The output of both Welch ANOVA tests suggests that the means for review subjectivity and review polarity, given the website on which it was posted, have a statistically significant difference. We'll seek to visualize these differences using a plot of the Tukey Honest Significance Test.

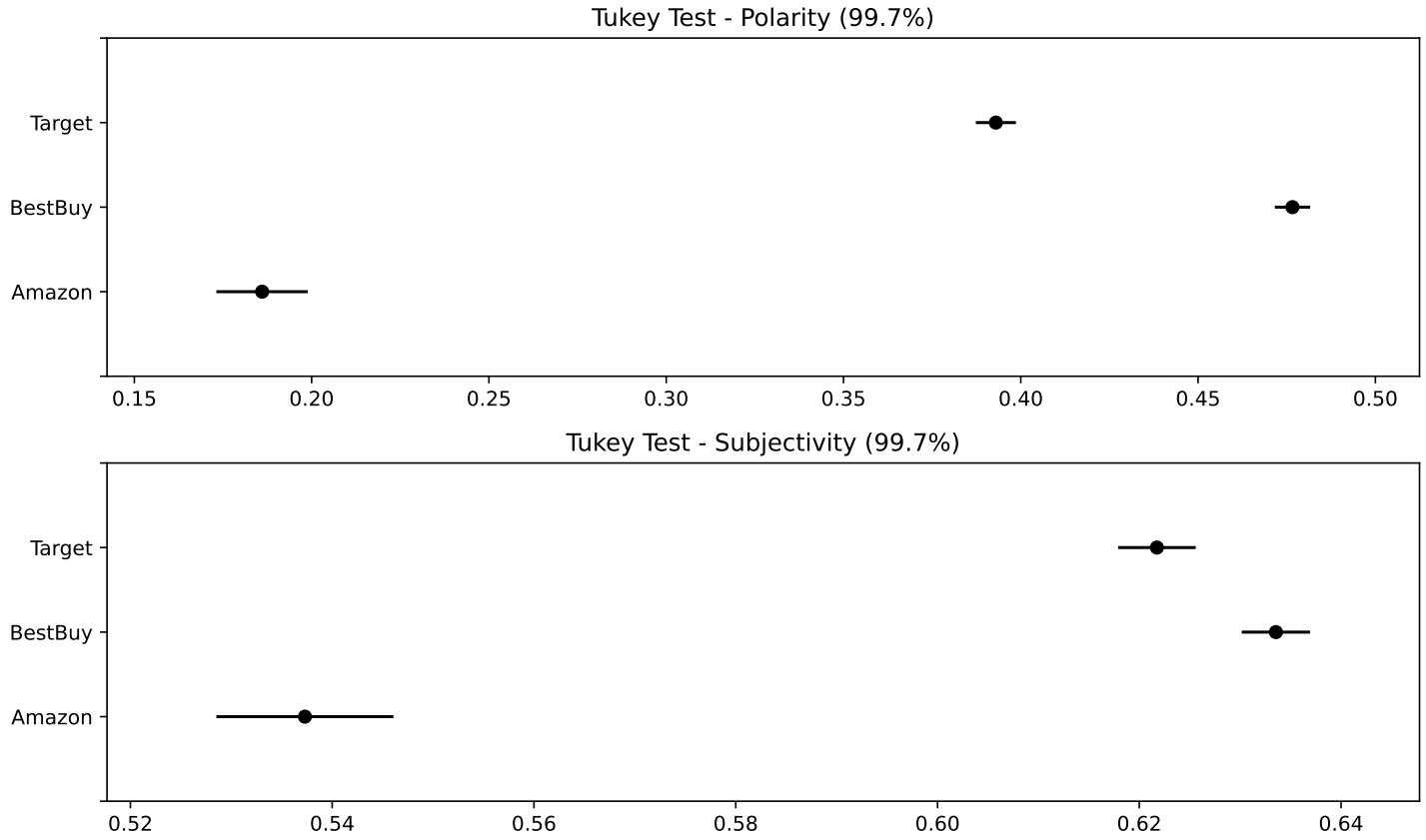

Figure 2.15: Tukey Tests for Star Rating, Polarity, and Subjectivity, by-Site

2.5.3.2 Tukey Tests with E-Commerce Platform as Treatment

The Tukey honest significance tests, depicted in Figure 2.15 suggest some interesting patterns between the three websites. Namely, target and best buy seem to have higher polarity, and subjectivity than the same variables for Amazon! Additionally, for each variable and each website, it seems there is no overlap in the variables at the 99.7% confidence level.

These statistically significant differences between the reviews, treated by website, indicate to us that we should proceed with caution in our modeling phase. Namely, it may be necessary to include an explicit variable or feature accounting for the source website in our modeling process as a predictor for the response variable.

2.6 Data Before / After

Much of our data cleaning occured during the collection process. Our team took specific steps to pursue cleaning during collection to simplify the process of bringing all information together:

- Using regular expressions to extract key values from text blocks
- Leveraging XPATH, class names, and element IDs to identify HTML fields in which our desired data points resided

Post-scraping, we had to pursue some additional cleanup

- Removal of unicode characters from review content where possible through coding and scripting.
- Conversion of numbers, stored as strings, to integers (i.e. star ratings, cost/dollar amounts)

- Handling of missing values (i.e. no ratings, no star ratings, no cost listed)

A particular challenge we came across during the data cleaning process was the handling foreign language reviews, highly repetitive reviews, and misspelled reviews. To better support our calculated measures for subjectivity and polarity, we leveraged the langdetect library to attempt to classify the languages of each of our 45,000+ reviews collected.

Table 2.7: Examples of reviews written in foreign languages

	site	reviewer_name	review_content
11	Amazon	Moldea muy bien, me gustó mucho! Es cómodo de ...	En perfectas condiciones, 100% el estado de la
13	Amazon	Diego Sanchez	Todo estuvo muy bien
22	Amazon	Daniel831	Llevo un da usndolo y aparecer funciona bien
29	Amazon	Carlos Tocto	Excelente producto y llego bien embalado
46	Amazon	Rocio castrellon	Me encanto llego en muy buen estado\r\nLa
20063	BestBuy	Iris	I love Apple, amazing calidad camera and per
20087	BestBuy	ErickL	Amazing phone:.:::::::::::::..
20458	BestBuy	senti	great.....
20569	BestBuy	Jaimerecios25	Very Good Printer yesssssssssssssssssssssssss.
20709	BestBuy	SilviaC	Excellent product. Excellent price. Will recon
5482	Target	Do it	Great upgrade for my teens
7284	Target	daisy78228	good for now.....
7303	Target	Put Jesus first	I don't want to say anything, thank you. I do
9575	Target	yuenkai	use it all the time, kind of simple to me.aaaa.
11164	Target	A	I just got it it is awesome

In some cases, the language classification by langdetect was a false negative (i.e. classified as a language other than english, when it was indeed English). In our data exploration, we found that many of these false positives were outliers in other categories (whether for review length, review subjectivity, review polarity, or star rating). As such, we find it prudent to exclude these reviews from our dataset when pursuing model development.

In total, langdetect classified fewer than 440 reviews (accounting for less than 1% of our collected reviews) as being non-English, or being repeated words or gibberish. Excluding these reviews should have minimal impact on the pursuit of model development.

TextBlob also offers us the ability to attempt to correct the spelling of reviews. Due to the amount of time it would take us to pursue spelling corre

Here are some additional examples of gibberish or non-contributional text that impact calculations for review subjectivity and polarity. While some of these could potentially provide value with deeper analysis, we find that these will not contribute significantly to our research.

	site	reviewer_name	review_content
962	Amazon	Kiran Kumar	NaN
980	Amazon	Ervey Gomez	's s s s s ! s s s !
1907	Amazon	Kathya De Alvarenga	NaN
2691	Amazon	Cristopher Leyva	10-Oct
2892	Amazon	Amazon Customer	10-Oct
2997	Amazon	Joe Zuppardo	NaN
3567	Amazon	HAMZAH ALGHAMDI	NaN
21293	BestBuy	Andy	:) :) :) :) ...
23207	BestBuy	Andy	:) :) :) :) ...

We are retaining the totality of the data we've collected, and will filter the data based upon our findings here so as to keep the most relevant and supportive data in building our models.

2.7 Insights from Collection and EDA

As in Guha Majumder, Dutta Gupta, and Paul (2022), we find that online reviews, this time across multiple websites, tend toward positivity. The values for star rating tend toward the 3 to 5 out of 5 star range, the polarity tends toward positive as star rating increases, and subjectivity (and one might argue, expressiveness) increases with star rating as well. These correlations can prove useful to us in our research.

We've also witnessed, tested, and verified that there is a statistically significant difference for review subjectivity and polarity, given the comment was hosted on a specific website. In light of these significant test results, we believe it may be necessary to account for the specific website from whence a review originates in training, testing, and validation data. The stark differences, without adjustment, could negatively impact the performance of any models if we fail to account for these differences therein.

The nature of the distribution of useful votes for a review poses a potential challenge to our research. The exponential distribution of useful votes could prove difficult to predict, as more and more useful votes become exceedingly rare for a given customer comment. As such, we may have an easier time with *categorizing* a review as being useful or not useful, in lieu of *predicting* a numeric value to represent how useful a comment is (e.g. predict the number of votes in favor of a comment as being useful to other customers).

Furthermore, exclusion of outliers could also pose challenges to our research. When excluding outliers, since the distribution of star rating tends toward the more positive reviews and results (reference Figure 2.4), we could inadvertently build models that perform the same way and are less able or unable to effectively categorize the usefulness of a lower star-rated review comment. With the nature of these outliers and the fact that having a high number of useful votes in and of itself is an outlier, we may need to examine building models upon the normalized data (e.g. the filters applied in Figure 2.7 and Figure 2.9) as well as the negation of that normalization, focusing on the outliers, so as to ensure that all the cases for our models are covered.

3 Models Implemented

In our modeling of the collected data, we seek to investigate several models for the generalization of the work performed by Guha Majumder, Dutta Gupta, and Paul (2022).

We will examine, compare, and contrast the use of the following models:

- Multiple Linear Regression Prediction
- Logistic Regression Classification
- K-Nearest Neighbors Classification
- Support Vector Machine Classification

3.1 Included Variables

Table 3.1: Selected Variables for Model Training

Table 3.1

Variable	Pre-Transformation Type	Post-Transformation Type	Purpose	Reason for Inclusion
review_star_rating	int	PCA scaled float	feature	literature survey
review_subjectivity	float	PCA scaled float	feature	literature survey
review_polarity	float	PCA scaled float	feature	literature survey
review_length	int	PCA scaled float	feature	literature survey
prod_subjectivity	float	PCA scaled float	feature	intuition
total_star_rating	float	PCA scaled float	feature	literature survey
site	dummy float	PCA scaled float	feature	intuition from EDA
review_helpful_votes	int	PCA scaled float	response	literature survey

3.2 Data Adjustments

As noted in our exploratory data analysis, each individual site has statistically significant differences in key variables we're considering in our modeling. To mitigate the potential for under or overfitting, and misrepresentation due to variable scale we perform the following transformations to our data:

1. Variable outlier adjustment. We noted in our EDA that each of the e-commerce platforms had high volumes of outliers with respect to the inter-quartile range. We applied a transformation to our data to map any outlier variable value on a per-website basis from its value to $\mu + 3 \cdot sd(\text{variable})$ for high-end outliers, and $\mu - 3 \cdot sd(\text{variable})$ for low-end outliers. In the event that either of these values exceeded the minimum or maximum value of the dataset, we mapped the value to the minimum or maximum accordingly.
2. Standard scaling of variables. After adjusting outliers, we re-mapped all of our feature variables to be on the scale of the standard normal distribution $N \sim (0, 1)$

3. Response variable transformation to binary value. We denoted a single useful vote as meaning that the review was useful to customers, and mapped the value to True/1, and False/0 otherwise.

Here is a sample (first 10 observations) of our data prior to the transformation:

Table 3.2: Data (pre-transformation)

Table 3.2

review_star_rating	review_subjectivity	review_polarity	review_length	prod_subjectivity	total_star_rating
5	0.638000	0.461000	19.000000	0.350850	4.800000
5	0.633333	0.211111	60.000000	0.617105	4.700000
5	0.916667	0.787778	13.000000	0.480135	4.100000
5	0.608333	0.266667	15.000000	0.658157	4.300000
5	0.400000	0.180000	28.000000	0.658157	4.300000
4	0.650000	0.433333	21.000000	0.617105	4.700000
5	0.683333	0.300000	29.000000	0.350850	4.800000
5	0.300000	0.750000	13.000000	0.460182	4.500000
5	0.635417	0.516250	20.000000	0.350850	4.800000
5	0.745887	0.470522	28.000000	0.435398	4.700000

And here is a sample of our data after the applied transformations:

- Snippet of first 10 observations in our training dataset, post-transformation:

Table 3.3: Data (post-transformation)

Table 3.3

review_star_rating	review_subjectivity	review_polarity	review_length	prod_subjectivity	total_star_rating
0.527771	0.264387	0.336418	-0.347482	-1.353087	1.088999
0.527771	0.240213	-0.563168	0.998209	1.428196	0.716223
0.527771	1.707921	1.512800	-0.544412	-0.002591	-1.520431
0.527771	0.110709	-0.363171	-0.478769	1.857024	-0.774879
0.527771	-0.968488	-0.675166	-0.052086	1.857024	-0.774879
-0.481349	0.326548	0.236820	-0.281839	1.428196	0.716223
0.527771	0.499220	-0.243173	-0.019265	-1.353087	1.088999
0.527771	-1.486502	1.376802	-0.544412	-0.211017	-0.029328
0.527771	0.251005	0.535315	-0.314660	-1.353087	1.088999
0.527771	0.823259	0.370697	-0.052086	-0.469912	0.716223

Classes in the response variable (the number of helpful votes a review received) set were mapped as follows for all non-linear models:

- 0: if the review had no helful votes
- 1: if the review had one or more than 1 helpful votes

Our reasoning for this transformation is that, across the totality of our data, a comment receiving more than one vote as being useful is quite rare in our dataset, as uncovered during our exploratory data analysis. As such, even a single vote for being useful should put the comment in the running for being considered useful.

3.2.1 Dimensionality Reduction

For all models outside of the multiple linear regression, we performed a principal component analysis on the scaled data.

Table 3.4: Cumulative Variance of Principal Components (selected features)

Table 3.4

Principal Component	Cumulative Variance	Explained Variance
PC1	0.350576	0.350576
PC2	0.548099	0.197523
PC3	0.673668	0.125569
PC4	0.788246	0.114578
PC5	0.878823	0.090576
PC6	0.945954	0.067131
PC7	1.000000	0.054046

To reduce dimensionality for Logistic Regression, Support Vector Machine, and K-Nearest Neighbor models, we elected to reduce from 7 principal commponents to 6. From the above table, we see that these 6 components explain approximately 95% of the variation within the training data. We projected our training and testing data from their 7-dimensional feature space to a reduced 6-dimensional principal component vector space and simplify computations. The only exception here is for our multiple linear regression, which used all 7 dimensions in its perscribed vector space.

3.2.2 Training Data

To train our dataset, we leveraged the data, post-transformation to train each of our models, including the adjustments of outlier datapoints to being within 3 standard deviations of the mean of each variable. We selected an 80% sample of this data and leveraged the same dataset to train each model.

In the interest of generalization, we sought to take our training dataset solely from products that were common across all three e-commerce platforms. By working with common data from each site, the data and the models may be able to formulate a more generalized construct of how certain variables behave within the different contexts of each platforms, and assist in better prediction and classification of comments as being useful or not.

3.2.3 Testing Data

For testing, we evaluated each model against transformed data, omitting the transformation of any outliers to being within 3 standard deviations of the mean. We performed this action to enable a fair comparison of each model against one another when working with real-world data.

3.2.4 Desired Outcomes and Objectives

Predictive modeling for the usefulness of a user comment on a product, in and of itself, cannot be conducted 100% objectively. We are interested in the exploration of misclassifications - particularly of false positives.

Our data exploration revealed that having even a single vote for a comment as being useful was exceedingly rare, with a median and mean number of votes hovering at or about 0 regardless of the website on which the comment was posted.

Additionally, from a technological perspective, the ability of a web user interface programmer or designer to simply filter or arrange comments by the number of votes they received is trivial.

With the above considerations in mind, we are interested in a model that provides reasonable accuracy while also having an apprpriate amount of recall and a reasonable F1 score. Having a model that perfectly maps predictions to actual outcomes is not useful to this research. Having a number of classified false positives for exploration and subjective evaluation is what interests us.

When exploring test results, we seek a percent of total positive predictions (true and false positives) matching that of the testing dataset. With this in mind, we won't have the best accuracy, F1, precision, or recall. Having values for precision and recall hovering around 0.5 should support us in classifying reviews that have no votes as being useful as potentially being useful.

3.3 Examination of the Original Multiple Linear Regression

3.3.1 Linear Model Construction

We leveraged a similar formulation to that which was used within the research of Guha Majumder, Dutta Gupta, and Paul (2022).

Our version of the linear model is designed as follows: $\hat{y} = \beta_0 + \beta_1 X_{rsr} + \beta_2 X_{rs} + \beta_3 X_{rp} + \beta_4 X_{rl} + \beta_5 X_{ps} + \beta_6 X_{tsr} + \beta_7 X_s$

Where each of the following variables have been standard-scaled to a range between 0 and 1 for the input data:

- X_{rsr} corresponds to the individual review's star rating
- X_{rs} corresponds to the review's subjectivity score
- X_{rp} corresponds to the review's polarity score
- X_{rl} corresponds to the review's length (in words)
- X_{ps} corresponds to the product description subjectivity score
- X_{tsr} is the overall star rating for the product.
- X_s is the site on which the comment was found (converted to a dummy variable for each website).

We leveraged X_{ps} as a proxy for the previous model's binary attribute for whether or not a good was search-based or experienced based. Guha Majumder, Dutta Gupta, and Paul (2022) leveraged a set of binary variables to classify a good as being search, experience, or mixed products. Our intuition was that, given the subjectivity of a product's description and/or specifications, a higher subjectivity score would correspond to an experience-based good, a lower subjectivity would correspond to a search-based good, and everything in-between would be a mixed product. This construction allows for any product to have a continuous potential range, and for most products to be mixed products (some tending more toward experience or search).

3.3.2 Inspection of Linear Model Assumptions

Generally, our examination of the multiple linear regression performed by Guha Majumder, Dutta Gupta, and Paul (2022) failed to meet the assumptions of linear regression (normality of residuals, linear pattern in fitted vs. observed values, and constant variance of residuals).

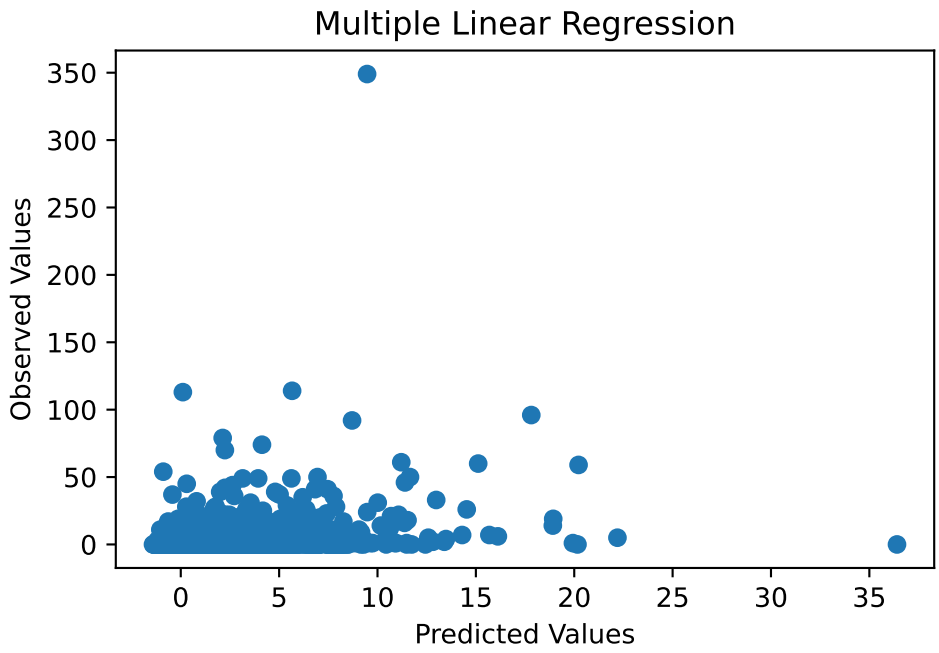

Figure 3.1: Linear Relationship Between Predictors and Response

3.3.2.1 Linearity of the Model

Our model failed to achieve any clear form of linearity between fitted and observed values.

The fitted vs. observed values for this plot are not indicative of a linear pattern between the feature and response variables. It provides a high mean square prediction error of 28.4055, an R^2 of 0.0884 and adjusted R^2 value of 0.0883. The lack of even a moderate correlation here suggests one of the following:

- The wider spread of data from multiple websites and wider range of products reduced the correlation found by Guha Majumder, Dutta Gupta, and Paul (2022)
- The linear model is not generalizable.
- The model is no better than randomly guessing the number of votes a comment could or should have associated with it.

What about if we filter down the testing data solely to reflect the positive cases - where the review comment has at least 1 vote for being useful?

The linearity issue remains even in this case.

3.3.2.2 Homoscedasticity on Normalized Data

This model has substantial challenges with heteroscedasticity. Let's examine a plot of fitted vs. residuals in the model:

As expected for a model that does not have a clear linear pattern, the residuals for this linear model are heteroscedastic. We should expect to see a constant variance in a plot of predicted vs. residual values, with no correlation between the errors and the predictions. Here, we witness this issue directly, lending to the idea that the model is a poor fit for the data, and the nature of the model (e.g. additional data, additional features, or other feature/response transformations) would need to change substantially to produce effective predictive results.

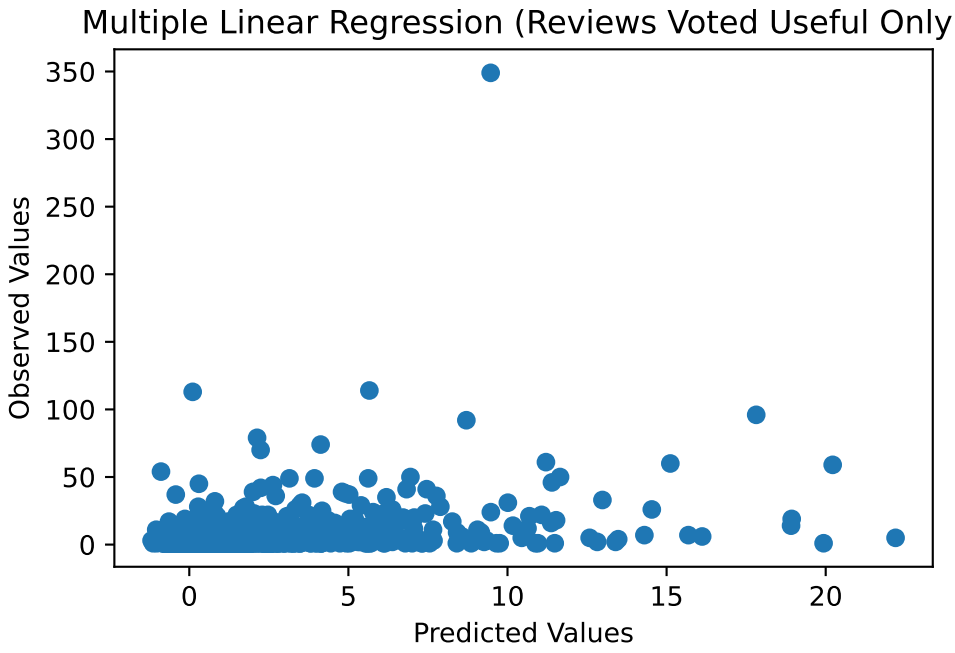

Figure 3.2: Relationship Between Predictors and Response (Reviews with at least 1 useful vote)

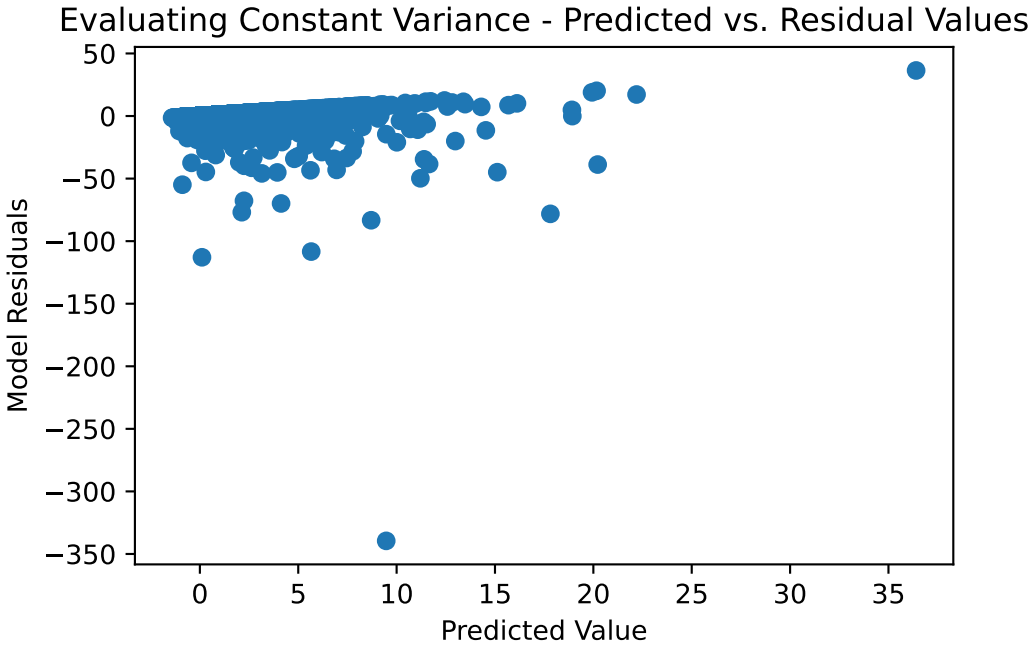

Figure 3.3: Heteroscedasticity in Model

3.3.2.3 Normality in Residuals

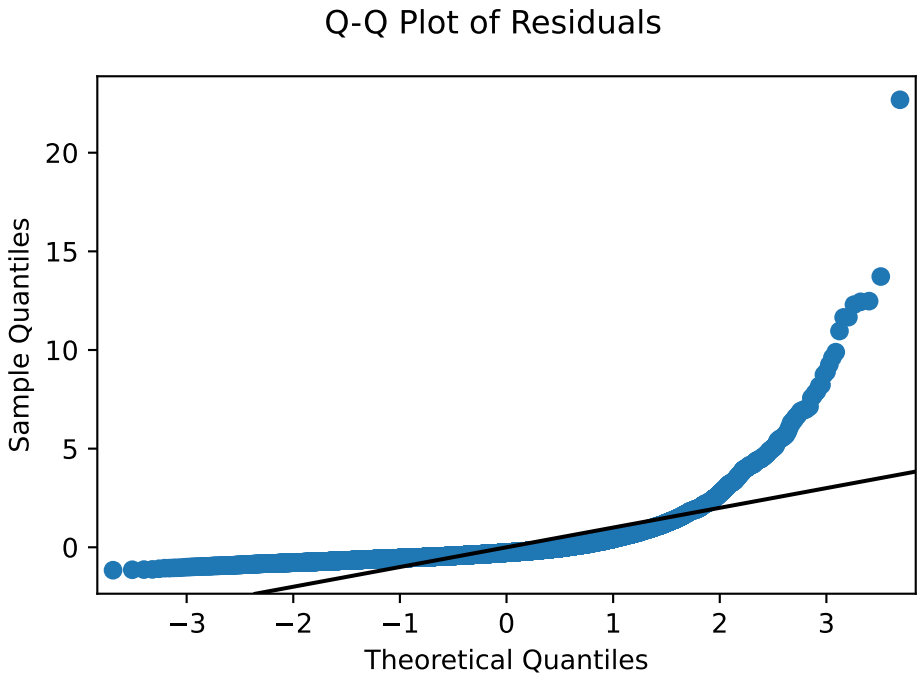

Figure 3.4: Model Residuals are Abnormal

From the above plot, it is clear that this model fails to adhere to the multiple linear regression requirement for residual normality.

3.3.2.4 Conclusion on MLR model

Expanding the Multiple Linear Regression beyond the scope of the study performed by Guha Majumder, Dutta Gupta, and Paul (2022) seems to fail all of the assumptions of a linear model.

Irrespective of this finding, we will output the results of this model sorted in descending order of the score provided by the model, to compare and contrast with the findings of other models. A simple examination of the MLR with respect to its linearity may not be an effective means to evaluate this model's performance.

Examining instead how the model predicts usefulness in comparison to other methods, alongside a subjective evaluation of the comments themselves may offer greater insight to its performance.

3.4 Logistic Regression Classification

Given the aforementioned challenges with the linear model, our next choice for examination was logistic regression. The MLR called for use of only numeric or continuous variables. Logisitic regression enables us to examine the inclusion of additional categorical variables as part of the regression consideration.

3.4.1 Hyperparameter Tuning

Logistic regression is, by our assessment, among the best performing models in our research effort. We inspected 3 logistic regression models - one tuned using the class weight hyperparameter to address the class imbalance present in the dataset. By assigning a higher weight to the minority class (1, or at least 1 vote for being useful) and a

lower weight to the majority class (0, or no votes for the comment as being useful), the model was able to better capture the patterns associated with the minority class, leading to metrics in our target range.

3.4.2 Oversampling techniques

Additionally, 2 more logistic regression models are trained using two oversampling techniques, namely ADASYN and SMOTE. ADASYN, which generates synthetic samples for the minority class based on their difficulty in learning regions, and SMOTE, which creates synthetic samples by interpolating between existing minority class samples, were used to address the class imbalance problem. These techniques help to provide the model with more balanced training data, allowing it to learn the characteristics of both classes more effectively.

It is a well know fact that oversampling techniques can be employed in the event of a severe class imbalance if hyperparameter tuning does not improve the model performance. However, despite the heavy class imbalance, the tuned model and the SMOTE model achieve great result with approximately 80% accuracy indicating that it is proficient at making correct predictions. However, the models do differ in their scoring for precision and recall.

The tuned model delivers a higher precision and a lower recall, bouth around our target range of 0.4 to 0.55 for both, leaning us in the direction of examining the tuned model's outputs for predicted useful comments. The "incorrect" prediction of elements from the minority class, which is what we're looking for. We don't want an over-prediction for false positives, but a reasonable degree of comments that could potentially be useful from a customer standpoint.

3.4.3 Logistic Regression Test Results

Table 3.5: Test Scores for Logistic Regression

Table 3.5

Model	Accuracy	F1	Precision	Recall
Logistic Regression (TUNED)	0.827358	0.474599	0.433375	0.524491
Logistic Regression (ADASYN)	0.783330	0.473885	0.370796	0.656368
Logistic Regression (SMOTE)	0.800134	0.484988	0.393074	0.633007

Comments, sorted by probabability of being useful in descending order, are located here (link).

3.5 K-Nearest Neighbors Classification

K-Nearest Neighbors is used to learn and identify the target class instances. It makes predictions by calculating distance (usually, Euclidean distance) between a given instance and all other instances in the dataset in feature space.

3.5.1 Hyperparameter Tuning

The value of k is the most critial hyperparameter in the KNN Calssification algorithm. It determines the performance of the model. Usually a small k value leads to overly complex understanding of the data that might result into overfitting, however, a higher k values can lead to underfitting.

We have looked at multiple values of k for our given data and compared a set of model metrics - accuracy, F1 score, precision, and recall to find the most optimal model to be the model with `k = 3`.

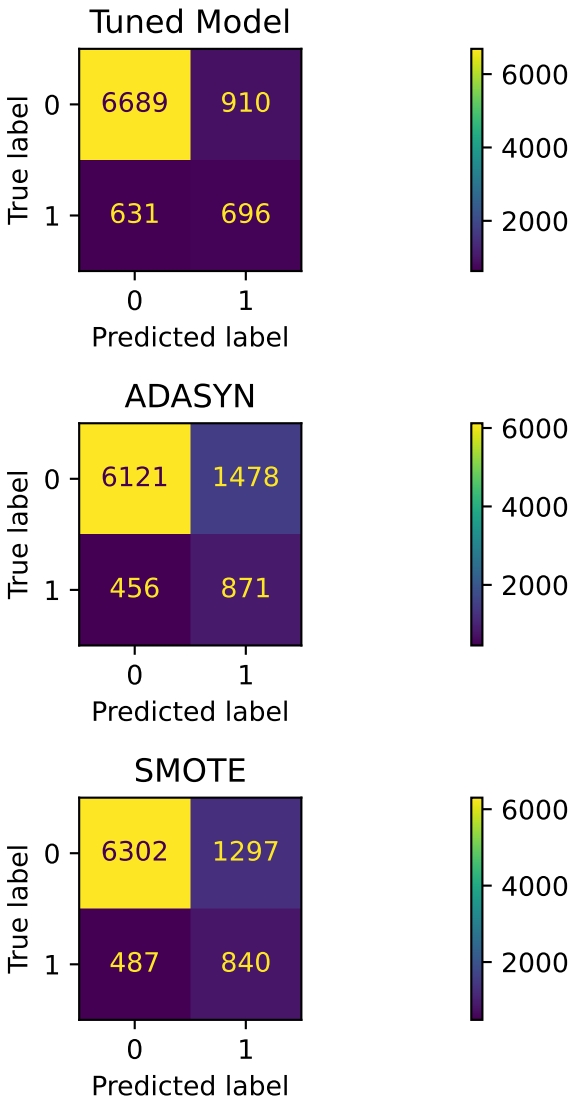

Figure 3.5: Logistic Regression Confusion Matrices

As we did for our logisitc regression models, we leveraged the SMOTE oversampling technique to support each KNN model in better predicting true positive cases. Having additional co-located true-positive neighbors will impact the distance metric of test data points and the number of closest neighbors in a particular class.

3.5.2 KNN Test Results

Looking the results shown below we can say that:

1. KNN at each neighbor level performs poorer in comparison to logistic regression. The accuracy and precision do not reach sufficient levels. Furthermore, the total percent prediction of positive cases (ranging from 21.5-25.6%) far exceed the percent of true positive cases in the testing dataset (approximately 17%). These over-optimistic prediction levels (in combination with a high number of false negatives) suggest we may not be meeting the mark with this model, as once again, votes for a comment as being useful is relatively rare in this dataset.
2. For KNN with n=3 neighbors, we are the closest to the actual percent of positives in the source dataset. This model, however, has poor performance for precision, recall, and F1.
3. For KNN with n=9 neighbors, we improve the F1 score, but the precision and recall remain substantially low.
4. KNN, for any number of neighbors and given our selected features, may be an insufficient model for our use case. All of the KNN models far exceed our target % of total positive predictions. The source testing data has around 17% of the reviews as being "useful", whereas each of these models' positive predictions exceed that rate by at least 4%.

Table 3.6: Test Scores for K-Nearest Neighbors

Table 3.6

Model	Accuracy	F1	Precision	Recall
KNN (k=3)	0.772687	0.370462	0.314873	0.449887
KNN (k=5)	0.770222	0.395520	0.324782	0.505652
KNN (k=7)	0.765516	0.405229	0.325274	0.537302
KNN (k=9)	0.765068	0.411451	0.327818	0.552374

Simply examining the above confusion matrices, especially the upper-right hand corner false posistives, we see that KNN is likely over-optimistic about the usefulness of customer feedback. For the purpose of comparison to other models, we will examine the results of KNN `n=3`, as it held the closest value to the overall percent of true positives (17% true positives, 21.4% predicted).

3.6 Support Vector Machine Classification

Support Vector Machines work on classification problems by finding an optimal hyperplane that best classifies the target classes in the given feature space. Because of its flexibility of moving the hyperplane and adapting to the intricacies of the data, SVM could be a useful and powerful algorithm for our use case.

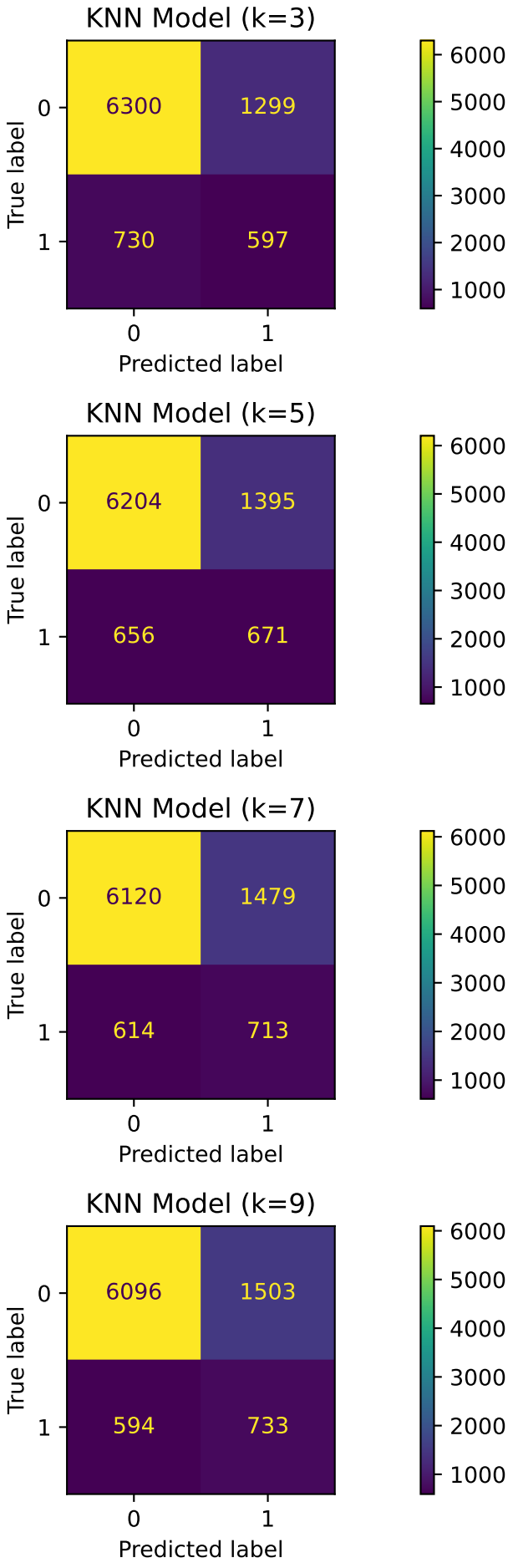

Figure 3.6: Confusion Matrices for KNN Models

3.6.1 Hyperparameter Tuning

The complexity of a SVM model is determined by the choice of kernel function that supports the capturing of nuances within the data. Below are our implmentation results of the performance metrics of Support Vector Machine (SVM) models trained with different kernel functions: polynomial (SVM-Poly), radial basis function (SVM-RBF), and sigmoid (SVM-Sigmoid).

One note on our SVM-Poly implementation - it is equivalent to a standard SVM linear kernel, as we have implemented it with `degree=1`. We found during our evaluation, for reasons unknown to us, that polynomial implementation with degree 1 ran faster than that of the standard linear kernel. This was simply crafted this way to reduce execution time.

To tune each version of this model, we leveraged a reduced set of principal components, and adjusted class weighting, as having useful votes for a product is a relative rarity across each of the e-commerce platforms. To boost our recall, we elected to assign weights of 0.2 to class 0 (not useful) and 0.8 to class 1 (useful).

Furthermore, for the tuned model, we adjusted the probability threshold to 0.1 vs. the default 0.5 and class weighting set to better adjust the model for mis-classification of the minority class (as we did for the logistic regression). The combination of this shift with the class weighting allowed us to achieve a total positive prediction rate in close proximity to the actual true positives.

3.6.2 SVM Test Results

Table 3.7: Test Scores for Support Vector Machine

Table 3.7

Model	Accuracy	F1	Precision	Recall
SVM-Poly	0.827134	0.465905	0.430858	0.507159
SVM-RBF	0.849989	0.418079	0.493840	0.362472
SVM-Sigmoid	0.768205	0.367085	0.308960	0.452148

Looking the SVM testing results, we can say that:

1. SVM with a polynomial kernel (set to degree 1, or linear) predicts a total percent of true positives close to the underlying source data (approximately 17%).
2. SVM with radial basis function kernel appears to under-predict positive cases and fails to meet our F1 threshold.
3. SVM with sigmoid kernel over-predicts false positives and fails to meet our F1 threshold.

SVM Poly (linear, degree 1) appears to achieve reasonable accuracy while hitting an appropriate level for F1 and Recall for our use case. The presence of false positives gives us something to subjectively examine for its usefulness as a potential customer.

3.7 Model Comparison

We examine the following table to compare and contrast our implemented models on our collected data.

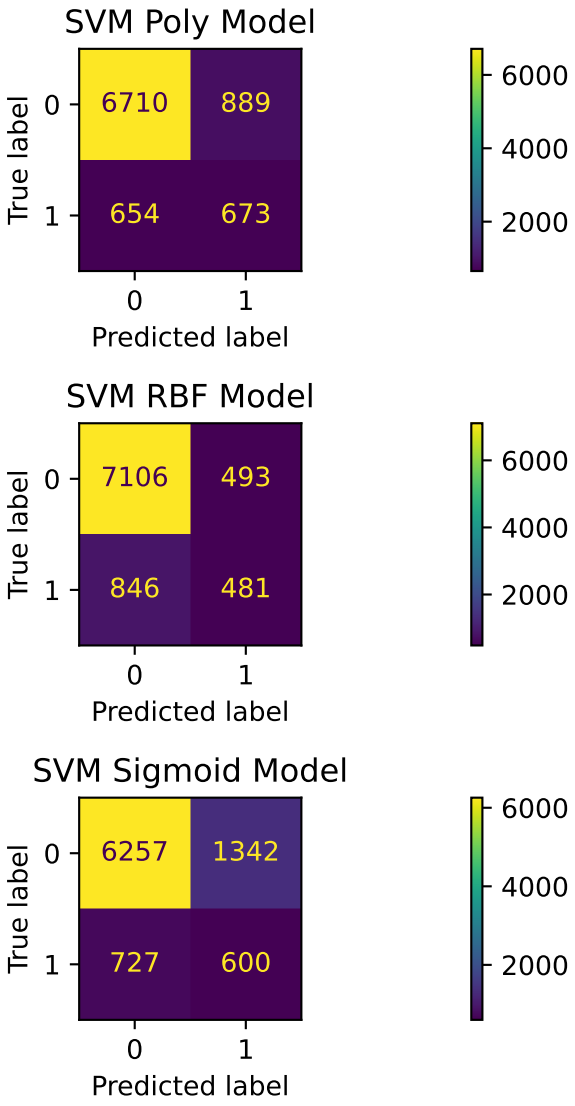

Figure 3.7: Confusion Matrices for Support Vector Machines

Table 3.8: Summary Metrics (all evaluated models)

Table 3.8

Model	Accuracy	F1	Precision	Recall
Logistic Regression (TUNED)	0.827358	0.474599	0.433375	0.524491
Logistic Regression (ADASYN)	0.783330	0.473885	0.370796	0.656368
Logistic Regression (SMOTE)	0.800134	0.484988	0.393074	0.633007
KNN (k=3)	0.772687	0.370462	0.314873	0.449887
KNN (k=5)	0.770222	0.395520	0.324782	0.505652
KNN (k=7)	0.765516	0.405229	0.325274	0.537302
KNN (k=9)	0.765068	0.411451	0.327818	0.552374
SVM-Poly	0.827134	0.465905	0.430858	0.507159
SVM-RBF	0.849989	0.418079	0.493840	0.362472
SVM-Sigmoid	0.768205	0.367085	0.308960	0.452148

The most interesting results, we find, come from the tuned Logistic Regression model and SVM-Poly models. These seem to hit a sweet spot when it comes to F1 and recall scores. Exceeding certain thresholds (at or around 0.5), seems to have too high a percentage of false positives.

50% of our oversampled training data held votes as being useful comments, and our regular training data contained samples with approximately 7% being voted as useful.

For our testing data, approximately 17% of the records held votes for being useful.

The Tuned Logistic Regression predicted a total number of positive (true and false positives) of 1606, or 17.99% of the available samples. This result is very close in proximity to the actual total for true positives within the data.

The SVM Poly model also produced a prediction of approximately 17.5% of the testing data being useful (comparable to the actual value of 17%).

None of the other model formulations or permutations achieved a percentage of total positive prediction rate as in close proximity to the actual underlying data.

The false positives for SVM-Poly and Tuned Logistic Regression are located here:

- Logistic Regression Prediction Results
- Support Vector Machine Prediction Results
- K-Nearest Neighbor Prediction Results
- Multiple Linear Regression Prediction Results
- For each of the models, the files are filtered to solely contain false positive reviews (those for which the model predicted the review is useful, but had no votes in favor of it in the source data).

Amongst the ranking of outputs from MLR, Logistic Regression, and SVM, we see some common threads.

- Each recommended the same top 3 review comments for the following products:
  - Mario Kart 8 from Amazon.
  - Dyson Ball Vacuum from Target.
  - HP Deskjet 2755e Printer from Amazon.

Table 3.9: Comparison of Metrics, top 30 Reviews for 3 models.

	model	metric	review_subjectivity	review_polarity	review_star_rating	review_length
0	Multiple Linear Regression	mean	0.515744	0.175872	4.100000	328.766667
1	Multiple Linear Regression	std	0.084931	0.097946	1.268994	175.919786
2	Logistic Regression (Tuned)	mean	0.521513	0.108089	2.266667	258.933333
3	Logistic Regression (Tuned)	std	0.089206	0.135156	1.484014	211.463200
4	SVM-Poly (Tuned)	mean	0.515099	0.100752	2.066667	244.266667
5	SVM-Poly (Tuned)	std	0.083521	0.130734	1.412587	212.249192

- Examining some metrics from the top 30 recommended "useful" reviews -
  - Logistic Regression and SVM had similar means and standard deviations for review star rating, review subjectivity, and review polarity. Each seemed to prefer slightly positive reviews (as follows from our EDA), and an even split on the subjectivity of the text (not overly precise, not overly vague). Similarly, they made more false positive predictions where the products subjective description of itself was near center, just like the review's subjectivity.
  - Each model seemed to hold a preference for longer reviews, and for product descriptions that we assess may be considered "mixed" products (part search, part experience). Each model also held a bias for higher total star ratings for the product, with the MLR holding the greatest bias.
  - The MLR holds a bias toward higher star ratings (4.1) in its predictions. Similarly, it had a slightly greater bias toward positivity in the review. However, it held comparable results for subjectivity in comparison to SVM and Logisitic regression.
  - Generally the near "neutrality" with regards to polarity, subjectivity, and star rating for logistic regression and SVM could be useful to prospective customers. Allowing users to filter results based off of these predicted classifications could bring balance to other reviews presented when sorted in descending order by number of useful votes. The neutrality aspect can give the customer insight into unknown positive or negative reviews and shed light on useful information as a potential customer.
  - The presence of common reviews within the top 30 suggests that each of these models could hold a degree of validity. The interpretation and assessment, however, is subjective. Performing a further study with participants to evaluate the prediction of useful comments would support determining the validity of these models and their effectiveness.

4 Conclusion

Evaluating the performance our models in comparison to the source data, we find that our selected features are likely weak in the prediction of a feedback comment being useful. There are substantial issues with the strength of our selected methods and metrics for both the training and testing data.

We attempted multiple iterations of examining our data, examining models and how to adjust them, and how to optimize performance. Ultimately, our work did not deliver a system that could definitively be used to predict usefulness of a review.

Despite that, we were able to explore and answer several of our planned research questions. This study can prove to be a basis of further examinations into how to better classify sentiment and assess the usefulness of reviews based on customer experience.

4.1 Review of Research Questions

At the start of this effort, we set out to answer the following questions:

- Can the model from Guha Majumder, Dutta Gupta, and Paul (2022) be generalized with
  - **larger volume of products and product types from which to mine data?**
    - We find that a larger mix of products across multiple platforms does support a wider-reaching and more generalizable model. Their model, with some adaptations, was able to identify several subjectively useful comments. However, the methods we used may not be fully appropriate or meet sufficient performance thresholds to establish confidence in our selected models.
    - We ascertained that the Linear Regression model used in Guha Majumder, Dutta Gupta, and Paul (2022) could not capture the underlying patterns in our case from the larger volume of products. It is found that the results from the tuned Logistic Regression and SVM-Poly models which have credible scores up to a certain threshold. These models were able to capture insights from the large data which was not seen from the logistic regression model
  - **a sliding scalar multiplier representing the degree to which a product is a "search" (0) or "experience" (1) product?**

    Leveraging a proxy, the subjectivity of a products description, as a means of measurement for a search or experience product may serve as a useful strategy in future research. This approach provides a useful way to comprehend and classify products across the consumer goods spectrum, facilitating more perceptive analysis and strategic decision-making across a range of industries.
  - **Adding modifiers to review content based upon:**
    - Customer / Reviewer reliability and reputation?

    This question could not be answered from our analysis as for all the websites - Amazon, Target and BestBuy, only a verified buyer was allowed to review a product. Due to this factor, we were forced to accept that every reviewer was reliable. As going around this was to go through each and every review looking for jargon content which was not feasible.
    - Review Polarity?

Positive, negative and neutral reviews all offer insightful information about the varying subjective experiences that people have with a product. Positive reviews are classified by positive attitudes and recommendations, and they are usually linked to products that provide remarkable experiences or successfully meet particular needs. On the other hand, unfavorable reviews could indicate shortcomings, contradictions, or mismatches between what customers expect from a product and how well it performs.

- Can the polarity of reviews be judged accurately by using a Naive Bayes classification model? Hu, Gong, and Guo (2010)
  - **What is the impact of different feature extraction methods (e.g., bag-of-words, TF-IDF) on the performance of Naive Bayes classification model? Wang et al. (2018)**

    While formulating the problem statement, we listed using Naive Bayes classification for possible solutions but did not explore them in our analysis due to personnel and time resource constraints.

- Can products be classified on their degree of being search or experience based by examining product variables such as:
  - **Degree of specificity in the product description? (e.g. level of detail, length, numeric values, descriptive values may suggest the product is more search than it is experience-based)**
    - Products with extremely detailed descriptions typically offer precise and comprehensive details about their characteristics, features, and functionalities. Precise measurements, technical details, and explicit information about the attributes of the product are frequently included in these descriptions. The focus on specificity makes it easier for customers to look for and assess products according to their own requirements, tastes, and standards. Consequently, goods that have extremely detailed descriptions are usually linked to the "search" category, as consumers can readily find and assess them based on objective criteria.
    - So indeed, depending on factors like the level of specificity in the product description, products can be categorized as "search" or "experience" in varying degrees.
  - **Whether the product is offered in brand-new condition only, or offered as new, used, or refurbished? (e.g. refurbished products may be more search products than they are experience products)**

    While we lacked comprehensive data on this aspect, generally, products offered in brand-new condition may incline more towards the search-based category, whereas items available in used or refurbished conditions could incline towards being more experience-based due to the potential history of priduct use associated with them.
  - **Which of the 5 senses the product engages? (e.g. engagement of more senses, or engagement of solely specific senses like hearing and vision may suggest more experience-based than search based; examine relationship between search and experience vs. senses engaged)**

    While the engagement of the senses is undoubtedly a crucial aspect of the consumer experience, it was not within the scope of our current investigation. However, examining the relationship between the engagement of the senses and the classification of products as "search" or "experience" goods could yield valuable insights for future classification efforts.
  - **Item rarity (limited production or unique items vs. bulk-produced items)? (e.g. limited production products may be more experience-based than search-based)**

    Most of the items in our research were common household products i.e. bulk-produced items. We did not give sufficient research time to dig into limited production or unique items. Solely based on human psychology, it might be possible that limited production or unique items typically exhibit

more experiential qualities, as they offer distinctive and potentially rare experiences compared to bulk-produced items.

- **Can newer natrual language processing libraries provide a better fit for Review Content metrics examined by Guha Majumder, Dutta Gupta, and Paul (2022)?**

  We found that there was no clear linear model from the data which could be an outcome of we how we processed sentiment as a numeric variable as opposed to a category. Based on this assumption, newer natural language processing libraries might provide an improved fit for analyzing review content metrics like sentiment. For instance, using these libraries could involve more nuanced sentiment analysis techniques that avoid treating sentiment solely as a numeric variable, potentially providing clearer insights into customer feedback.

- **How does sentiment in customer reviews correlate with customer satisfaction metrics or sales figures for a particular product?**
  Sentiment in customer reviews, characterized by higher star ratings, tends to correlate positively with customer satisfaction metrics. Higher star ratings often reflect stronger polarity and subjectivity in sentiment, indicating greater overall satisfaction of the customer and potentially influencing purchasing decisions for future buyers.

- **Can we categorize customer reviews based on customer experience and sentiment?**

  Our study results show that it is possible to categorize customer reviews based on customer experience and sentiment. One way of doing this is using machine learning algorithms like classifiers that inculcate subjectivity and polarity as model features. The results of the classifiers implemented in this project might not be robust but can act as a base for methodological adaptations to better classify sentiment and assess the usefulness of reviews based on customer experience.

- **Do specific product star ratings tend to incite more reviews, and if so, how does this impact the overall reputation measurement?**

  There was no distinct connection between star ratings and number of reviews for a given item. Similarly to our examination of review subjectivity and polarity, there seemed to be central points where our data coalesced, namely a lot of products in the 4-5 star range with relatively few reviews (only several hundred) and around the same star rating, an island of products that held in the tens of thousands of reviews. This area almost represents change in one variable irrespective of another, this suggested a degree of independence, and as such we did not pursue further research on the matter.

- **Are specific quality descriptors in text-based reviews (e.g., 'enthusiastic', 'disappointed') strongly associated with certain rating levels, and how does this association affect product reputation?**

  Our findings suggest a positive trend between star ratings and positive polarity in reviews. In review content, specific quality descriptors like 'enthusiastic' or 'disappointed' seem to be associated with certain rating levels. However, further analysis is needed to assess how these associations impact overall product reputation, including how sentiment and specific descriptors contribute to consumer perceptions and purchasing decisions.

4.2 Interesting Findings, In Spite of Model Performance

Despite what we perceive as weakness in the models, the degree of detail and specificity within some of the predicted "useful" comments is enlightening. Let's have a look at a small selection of comments, near the top of the prediction list for our implemented models -

A Dyson vacuum is an expensive product. Hearing of someone's challenges, particularly an extreme case of having a vacuum *melt* while the customer was using it, may make one think twice about the investment. While appearing

I purchased this item from a big box store and not Amazon. I'm leaving a review here because I hope others will see it. My old Animal Complete died this morning after 13 years of use. I decided to replace it with this one. What a major mistake.

As others have mentioned, it's hard to get the vacuum head to stay attached. It fell off twice after I heard it click into place. The handle and tools are considerably more flimsy than they were on my old Dyson and the vacuum head is smaller. The dirt cup seems smaller, too.

After using it for about 5 minutes, I smelled burning plastic. When I took a step forward to take a look at it, I stepped on a chunk of molten blue plastic on my carpet. The red brush strip had come off the roller (ripped off somehow, because blue strips from the roller were present as well) and wrapped around the axle of the vacuum head rendering it completely frozen and unable to rotate.

The only positive thing that I can say about it is that for the few minutes it did work without spewing molten plastic, it worked really well for both dirt and pet hair even on places I thought were already clean.

Figure 4.1: Comment on Dyson Vacuum Cleaner - predicted useful by Logistic Regression

negative in sentiment, the commenter appears to provide a degree of objectivity to their review - and does talk about both the good and the bad of the product. These are all things we believe may be useful to a customer.

Wow. I really dont know where to begin. The holes dont line up, the construction is remarkably shoddy, the parts are uneven and dont fit together properly. The feet dont sit flush against the bottom of the couch, theres a 4mm gap on one half of the center feet, so its only a matter of time until it fails.
Its also super uncomfortable. The smaller cushions are less dense foam than the large cushion, and slightly different thicknesses. The back is also much shorter than it looks, which means sitting in it is awkward.
You also cant return it without paying $75, and its impossible to get back into the box anyway since all the cushions are vacuum sealed and they triple in size once opened.
Do yourself a favor and look elsewhere. This couch will make you very sad.
UPDATE: its been nine months, and it just gets worse. The center of the couch is collapsing, the cushion foam is deteriorating, and when my girlfriend and I sit on the couch together, we shock each other with static electricity whenever one of us moves.
My roommates are like well, I guess thats what we get for $200. Except it was $420.
I wish I had never even considered buying furniture online.

Figure 4.2: Comment on a Sofa - predicted useful by SVM

An approximately $400 sofa is also quite the investment, and knowing about issues with returns, issues with the product and its craftsmanship is likely a useful datapoint for a prospective customer.

This customer looks like they had a good experience - so good in fact that they went online to a commerce platform that they didn't even use to purchase the item, just so that they could post about their experience. Knowing the settings used, how quickly the oven managed to cook their food, and the ease of use likely all give utility to a prospective customer.

None of these reviews had any votes for the comment as being useful to other buyers. That being said - it's all subjective in terms of determining utility.

There are several other comments within our results that fall along these lines. Further exploration and refinement of our research could produce better modeling and reliable results for customers.

I'm leaving my review on Amazon because, doesn't everyone look at Amazon for reviews before buying. I bought mine at the Big Box S club store (It was less and had the cover and pizza peel inc). Also note the price of these just dropped Everywhere (03-10-24)! Could there be a new model on the horizon?
This oven made fabulous pizza out of the Box (Heating it on Max for 30 min before starting Pizza setting). We got the take and bake pizza from California Pizza Kitchen for $8, Used the Neapolitan setting (with and without smoke depending on diners preference) and pulled it at 3 min or less getting a little edge char but the perfect crunchy base crust, melted but not burned cheese, and crispy edged pepperoni. We're not new to pizza ovens having had the Ooni gas oven for a few years but their volt(?) was more than I wanted to pay. This electric oven was as good as the gas version without the terror of an inferno that one has with early uses of the gas oven as well as no constant turning and babying the pizza. In fact you can't baby the pizza.. the door is closed and there's NO window!
We usually use our own dough and make our pizzas from scratch so we'll see how those are, but I'm expecting it will be great, and so much easier to do than before. Of course I'm also anxious to try the smoking feature etc and will update if there's anything to say about those.

Figure 4.3: Comment on a Pizza Oven - predicted useful by MLR

References

Guha Majumder, Madhumita, Sangita Dutta Gupta, and Justin Paul. 2022. "Perceived Usefulness of Online Customer Reviews: A Review Mining Approach Using Machine Learning & Exploratory Data Analysis." *Journal of Business Research* 150 (November): 147–64. https://doi.org/10.1016/j.jbusres.2022.06.012.

Hu, Weishu, Zhiguo Gong, and Jingzhi Guo. 2010. "Mining Product Features from Online Reviews." *2010 IEEE 7th International Conference on E-Business Engineering*, November. https://doi.org/10.1109/icebe.2010.51.

Rajeev, P Venkata, and V Smrithi Rekha. 2015. "Recommending Products to Customers Using Opinion Mining of Online Product Reviews and Features." *2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]*, March. https://doi.org/10.1109/iccpct.2015.7159433.

Wang, W. M., Z. Li, Z. G. Tian, J. W. Wang, and M. N. Cheng. 2018. "Extracting and Summarizing Affective Features and Responses from Online Product Descriptions and Reviews: A Kansei Text Mining Approach." *Engineering Applications of Artificial Intelligence* 73 (August): 149–62. https://doi.org/10.1016/j.engappai.2018.05.005.