Airbnb Reviews Sentiment Analysis
John Prichard, Waranya Phanphon, Jonathan Ramos

Dataset and Motivation

Airbnb started in 2007 and has since exploded into an enormous business generating billions of dollars for hosts every year. The airbnb website offers many services to both hosts and guests, not the least important of which is their system for reviews. Each time there is a booking, both the guest and the host have the opportunity to review and rate their experience. As such, it is extremely important for hosts to consistently provide an experience that is up to the standard of the guest. But what is the guest really looking for? What are the important things that a host can offer that will consistently make guests happy and allow them to maintain their business. These are the questions that we are interested in answering in this exploratory data analysis. In this study we will perform sentiment analysis on airbnb reviews and use the sentiment scores to see if we can develop a model to predict the sentiment of a review. We will explore the text within the reviews to search for features that predict the sentiment. We will also explore other features associated with listings/hosts such as bedrooms, bathrooms, airbnb status, etc.

To investigate this matter we will use datasets obtained from http://insideairbnb.com/get-the-data/. The website provides review and listings data for a wide range of cities across the globe; datasets are updated on a quarterly basis. For the study we will focus on the city of Portland Oregon. The main data sets that we will use include listings.csv, reviews.csv, and neighbourhoods.geojson. Below is the head of the reviews csv and the list of columns contained in listings.csv. The reviews csv contains 470,436 rows and 6 columns, the most important of which being the comments column, which is the text of the review.

reviews.csv head

[5]:

	listing_id	id	date	reviewer_id	reviewer_name	comments
0	12899	24767	2010-01-24	69327	Stuart	Recommended! Very good value for a spacious, a...
1	12899	29230	2010-03-13	72846	John	Our ten days visiting in Portland were enormou...
2	12899	29806	2010-03-16	84196	Lois	We had a wonderful time staying in the area of...
3	12899	32572	2010-03-31	89114	Troy	I stayed at Ali and David's place when I first...
4	12899	32862	2010-04-02	100318	Cathy	Clean, comfortable, quiet rooms; easygoing gen...

After tokens and sentiment features were extracted via natural language processing (NLP), the set of review features was merged with the set of listings to generate the final set used for machine learning, see **Table 1** for brief descriptions of each column.

Research Questions

-What are guests looking for in an airbnb listings
-What features are most associated with positive/negative reviews
-Develop a model that predicts the sentiment of a review

Literature Review

There are a few studies that we found that deal with similar research questions. This study https://medium.com/@yogi.sarumaha/airbnb-sentiment-analysis-with-python-e81e66fee6a6 performs sentiment analysis on airbnb reviews. This study https://redirect.cs.umbc.edu/courses/graduate/676/SP2021/termpapers/CMSC476676-TermPaperParulekarJugal.pdf also does a similar analysis on what appears to be a very similar data set. Both of these studies focus on vader sentiment analysis. Our study goes beyond what these have done in several ways. Firstly, our study attempts to improve the accuracy of sentiment scores by using the Roberta sentiment analysis model in addition to Vader. Secondly, our study will do a more detailed analysis of the

words within the reviews, breaking the comments into various parts of speech in an attempt to gain knowledge of words that are associated with positive/negative reviews. Thirdly, our research extracts additional features that come directly from the review comments. These include features such as the length of the review, whether the host name is mentioned in the review, and the specific adjectives and nouns within the review. Lastly, our model takes all of the features and develops models that attempt to predict the sentiment variable. All of this is done with the purpose of better understanding what causes reviews to be positive or negative.

Cleaning/preparing for Sentiment Analysis

The first thing that we needed to do was prepare the comments field for Vader sentiment analysis, word counting, and LDA topic modeling. The data was scraped from the airbnb website and the comments field contained many html tags. The first thing we did was remove these tags. The presence of such tags could potentially influence the Vader sentiment scores. For example, if the text was “This place was great
”, The vader sentiment analyzer would classify the great
 as neutral instead of positive because it doesn’t recognize the word.

The next thing that needed to be done was to remove stop words and lemmatize the text. LDA topic modeling requires these two cleaning tasks.. Lda topic modeling and Vader sentiment analysis are both based on a “bag of words” approach. Therefore, We prepared the tokens for both tasks at once. LDA is looking for general topics and not necessarily sentiment. Therefore when removing the stop words we felt it necessary to keep the word “not”, which is a stop word. Taking out this word could greatly skew the sentiment analyzer in the wrong direction. We also removed some of the punctuation, but did not remove exclamation points and emojis. These characters are used to convey meaning and the Vader sentiment analyzer is able to pick up on them. We replaced some of the contracted words with equivalent two word phrases, eg. (isn’t with is not). All of this was done in a specific order as to prepare for the topic modeling but to not hinder the sentiment analysis. In general, I think that these two tasks require separate preparation. However, with the size of the data set that we had, we wanted to prepare for both in one shot.

Feature Extraction

Once the reviews data was cleaned and the sentiment analysis was performed. We created several new columns derived from the comments field. We made columns for Tokens, Corpus_Length, Adjectives, Nouns, Nouns_with_adjectives and hostMentioned. The tokens represent all of the meaningful words that are in the review. The corpus length represents how many meaningful words in the reviews. The host mentioned column is a boolean column that states whether the host name was mentioned in the review. We then merged the reviews csv with the listings csv to see how well some of the features within it, could help us understand the sentiment scores.

We also built new columns from existing features in listings data, namely we built the sextant feature and the host_local feature. The sextant feature is a feature that places each listing into a distinct geographical region. We built six distinct regions using the neighborhood column. Lastly, we built a column called ‘host_local’, which is another boolean column that specifies if the host lives in portland or not.

Data Visualization (For all visualization please refer to the figures at the end of the text)

From pair plots generated from the set of all reviews shown in figure 10 comparing various features with our target sentiment column, we have the following conclusions:

There are no clear linear relationships between Price vs Sentiment, Accommodates vs Sentiment, Bedrooms vs Sentiment, Bathrooms vs Sentiment, Beds vs Sentiment (Figure 10 A-E). There appears to be some weak correlation with the review scores columns for Check In, Communication, Location, and

-*Value with Sentiment (**Figure 10 H-K**)*. This means that listings with higher ratings in these categories tended to also have higher positive sentiment.

*-Number of Reviews vs Sentiment (**FIgure 10 F**):*
The plots show a very slight positive correlation between number of reviews and positive sentiment scores. As the number of reviews increases, there is a subtle upward trend in positive sentiment. However, the relationship is not strong. There is no discernible pattern between the number of reviews and negative sentiment.

*-reviews_per_month vs Sentiment: (**Figure 10 G**)*
Positive sentiment dominates across all frequencies of reviews per month. There are far fewer negative sentiment points overall. For both sentiments, the points are most concentrated at lower reviews per month (under 10). Above 10-15 reviews per month, the sentiment points become much more sparse, especially for negative sentiment.

*-review_scores_location vs Sentiment: (**Figure 10 J**)*
There appears to be little correlation between a listing's review location score and the sentiment scores from the reviews. Both positive and negative sentiment is seen across the full range of location scores. The positive sentiment points are more concentrated in the upper right (high location score and high sentiment), while negative sentiment is more dispersed.

*-Corpus Length vs Sentiment: (**Figure 10 L**)*
The scatter plot for corpus length vs sentiment shows a more interesting pattern compared to the previous features.There appears to be a negative relationship between corpus length and positive sentiment scores. As the corpus length increases, the positive sentiment scores tend to decrease.
On the other hand, there seems to be a positive relationship between corpus length and negative sentiment scores. As the corpus length increases, the negative sentiment scores tend to increase slightly.

*-Sentiment vs Neighborhood Chloropleth (**Figure 11**)*
The choropleth suggests that positive sentiment is less prevalent in certain geographical areas. In particular the graph shows that the southeast side of Portland is less associated with positive reviews. After seeing this we decided to further investigate geographic location as a factor by breaking the city in sextants (6 geographic regions), (**FIgure 6,7**).

Machine Learning

In the section of the notebook preceding the machine learning, there is some additional data wrangling and EDA performed on the data set on a listing by listing basis once the necessary sets have been merged. In particular once the set of all reviews was generated, this data needed to be merged with the set of all listings. In order to do this, reviews were grouped by listing id and the sentiment scores were averaged and taken as targets. It is not necessarily the case that all listings have reviews so listings without reviews, and therefore without sentiment analysis, were dropped from the set. Other data cleaning and preparation tasks for machine learning included imputing missing price values with the median by number of guests accommodate, casting strings to floats or ints, casting ‘t’ or ‘f’ to True or False respectively, filling in missing values for Beds, Baths and Bathrooms by parsing out the Name string of the listing, as well as building dummy columns for our categorical data types, see **Table 1**. The final set of features in the merged data frame are described in **Table 1**, note that the set of engineered and NLP features from the Reviews dataset are highlighted in yellow.

Before moving on to machine learning, some further EDA was performed on the merged set. From the correlation heat map, there are four main areas of interest: Availability_30,60,90,360 are

generally well-correlated with each other, review_scores (rating, cleanliness, communication, etc) are correlated with each other, review_scores are correlated with the sentiment scoring features , and different methods of sentiment scoring are correlated with each other (that is, Roberta and Vader generally agree; positive scores are negatively correlated with negative scores and vice versa), (**Figure 1**) In particular, if we look at the correlations with posRoberta, we can see that the strongest positive correlations are with the review_scores features indicating that reviews with higher sentiment are also rated highly for things like cleanliness, accuracy, location, communication and value.

Next, pairwise scatterplots comparing Vader vs Roberta sentiment scores, each for negative, neutral and positive scores, were plotted to assess how well Vader and Roberta sentiment scores agree with each other. From visual inspection of these plots (and the correlation heat map) we can see that although there is some spread, in general Roberta and Vader scores are in agreement with each other (**Figure 2**) . It is important to note that Roberta considers a greater proportion of the sentiment scoring as positive sentiment whereas Vader leans more towards higher neutral scores. **Figure 3** shows bar plots and cumulative distributions denoting significant differences in both positive and negative sentiment scoring in local vs non-local hosts (negative sentiment, $p<0.0001$; positive sentiment, $p<0.0001$). This means that in general, reviews for local hosts are less negative and more positive than reviews for non-local hosts. In a similar way, **Figure 4** shows bar plots and cumulative distributions of negative and positive sentiment scoring in super hosts vs non-super hosts. A similar pattern arises where super hosts receive less negative sentiment and more positive sentiment than non-super hosts (negative sentiment, $p<0.0001$; positive sentiment, $p<0.0001$). The correlation heat map also suggested that corpus length may also be an important feature related to sentiment scoring so scatter plots of negRoberta and posRoberta vs corpus length along with their regression lines are plotted in **Figure 5**. Although the coefficients are quite small, p values for both negRoberta and posRoberta linear regressors are much smaller than 0.05, indicating that there is a relationship between these features. In particular the slope for negRoberta vs corpus length is slightly positive and the slope for posRoberta vs corpus length is slightly negative.

Since location is often regarded as an important feature of any property, the average posRoberta and negRoberta scores by neighborhood were plotted (**Figure 6**). In general there are some spikes in negative sentiment in some of the downtown and western neighborhoods. To make this plot easier to read, neighborhoods were grouped by sextant (either ‘Southwest’, ‘Southeast’, ‘Northeast’, ‘Northwest’, ‘North’, or ‘South’). From the one-way ANOVA, there is at least one difference in both positive and negative Roberta scores across sextants (**Figure 7**; posRoberta, $p<0.0001$; negRoberta, $p<0.0001$) with the bar plots indicating that Southwest and Northwest neighborhoods receive slightly more negative sentiment.

As for the NLP token features, the top ten tokens extracted during NLP were grouped by posRoberta quintiles. **Table 2** lists the top 20 tokens in the top quintile of posRoberta sentiment scores and their count frequency. In particular, two types of tokens were extracted: single word tokens and adjectives followed by nouns. Important qualities of listings with the highest posRoberta score include: “clean”, “comfortable”, “quiet neighborhood”, “great restaurants”, “public transportation”, “easy access”, “natural light”. This table also suggests that specific amenities such as “hot tub” can contribute to high positive sentiment.

Lastly, distributions of our potential target variables are plotted in **Figure 8**. The distribution of Roberta positive sentiment scores are skewed to the right whereas the Vader scores are centered around 0.4-0.5. As mentioned in the EDA, Vader splits more of the sentiment into Neutral. Although the Roberta scores are skewed, NLP with Roberta gave more meaningful tokens and so posRoberta will be taken as our targets for all machine learning tasks.

For machine learning, data was split into 80% train and 20% test sets. A variety of models were trained and Random Forest ended up being selected as it had the best cross-validated RMSE of 0.0457. After tuning the model we had a final cross-validated train RMSE of 0.0456 and a cross-validated test

RMSE of 0.0596. Since we are interested in investigating features of listings with high positive sentiment, it is convenient that Random Forests can output feature importances as part of the model and so **Table 3** shows the importances of the top 15 features. All the review_score metrics show up as being very important, which was expected from the correlation heat map, but it is notable that corpus length and price are also near the top of this list. Permutation feature importance is another way to assess how important features are for a given model. This approach involves shuffling a given feature many times and calculating how much the score decreased, and the intuition is that if a feature was very important, shuffling that feature (effectively destroying the relationship between that feature and the target) results in a large change in evaluation. **Figure 9** shows bar plots of the mean score decrease after 100 iterations of shuffling and model evaluation and a similar pattern arises; the review_score metrics are very important along with corpus length.

Since the logic that reviews with very high review scores are well associated with reviews of very high sentiment is a bit circular, machine learning was repeated after removing all the review scores features. Again, a Random Forest was selected for consistency even though Ridge Regression performed just slightly better. After tuning a new Random Forest, we have a final cross-validated train RMSE of 0.0743 and a final cross-validated test RMSE of 0.0693. Inspecting the feature importances of this Random Forest, we can see that Corpus_Length and price still come out as important features but notably, there are some new features at the top of the list: reviews_per_month and calculated_host_listings_count_entire_homes (**Table 4**). Following permutation feature importance of the Random Forest built on this new set without review scores features, Corpus_Length, calculated_host_listings_count_entire_homes, response_t__within an hour, host_is_superhost and host_local are among the features identified as most important (**Figure 10**). It is important to note that some of the features identified from the permutation testing agree with some of the EDA regarding the relationship between posRoberta and Corpus_Length, host_is_superhost, and host_local.

Conclusion

These analyses reveal that many qualities of both a listing and the host can contribute to positive sentiment in reviews. Hosts who are local and hosts who are super hosts have a slight edge over hosts who are non-local and non-super hosts. Additionally, host response time and the number of listings a host has seem to play important roles as well. It is likely the case that guests prefer experienced hosts who are accessible and responsive. As far as qualities of the unit that perform well, NLP and token analysis revealed that units that are clean, quiet, conveniently located nearby shops and restaurants, and have accessible public transportation are associated with high positive sentiment. Overall, qualities identified in this set of analyses would be important for any prospective Airbnb manager in Portland, OR to consider in order to maximize positive review sentiment.

FIGURES

Figure 1 - Correlation Heat Map

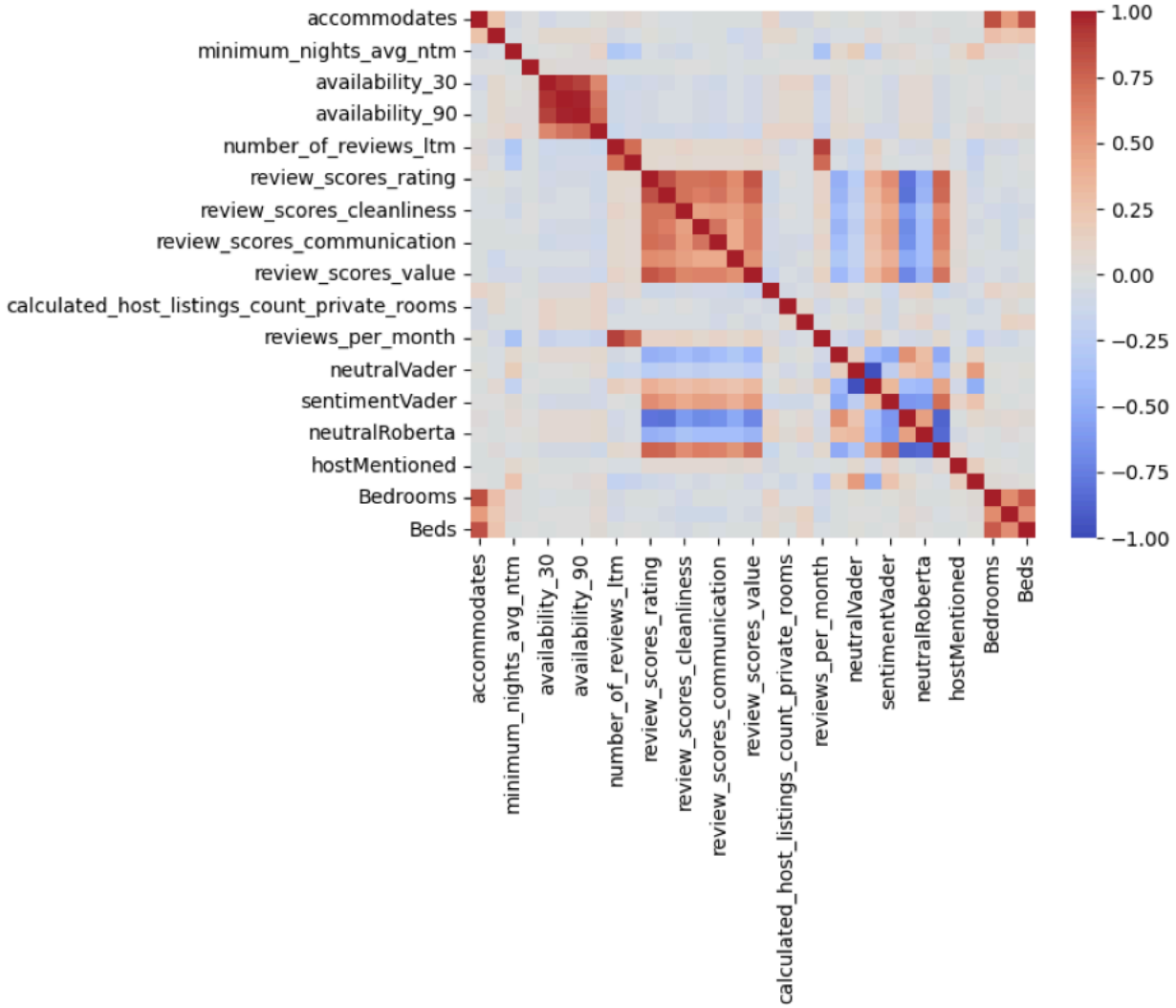


Figure 2 - Vader vs Roberta Sentiment Scoring Pairwise Plots

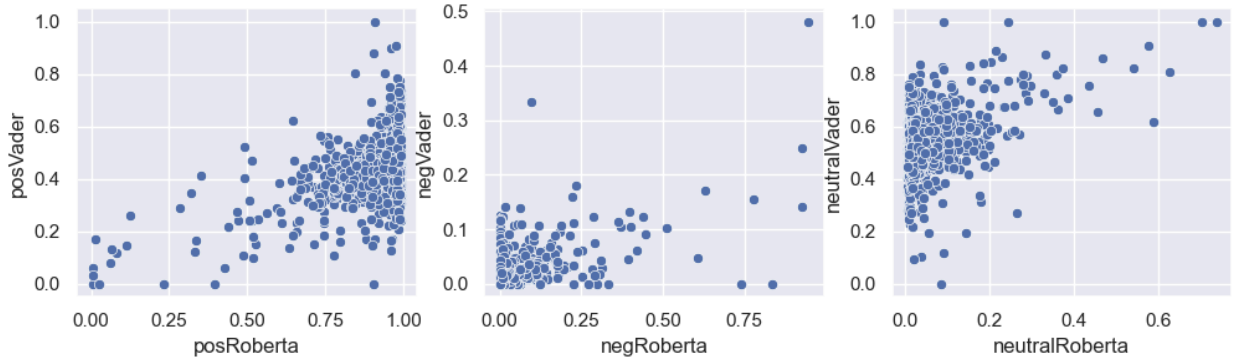


Figure 3 - Positive/Negative Roberta Scores in Local vs Non-Local Hosts

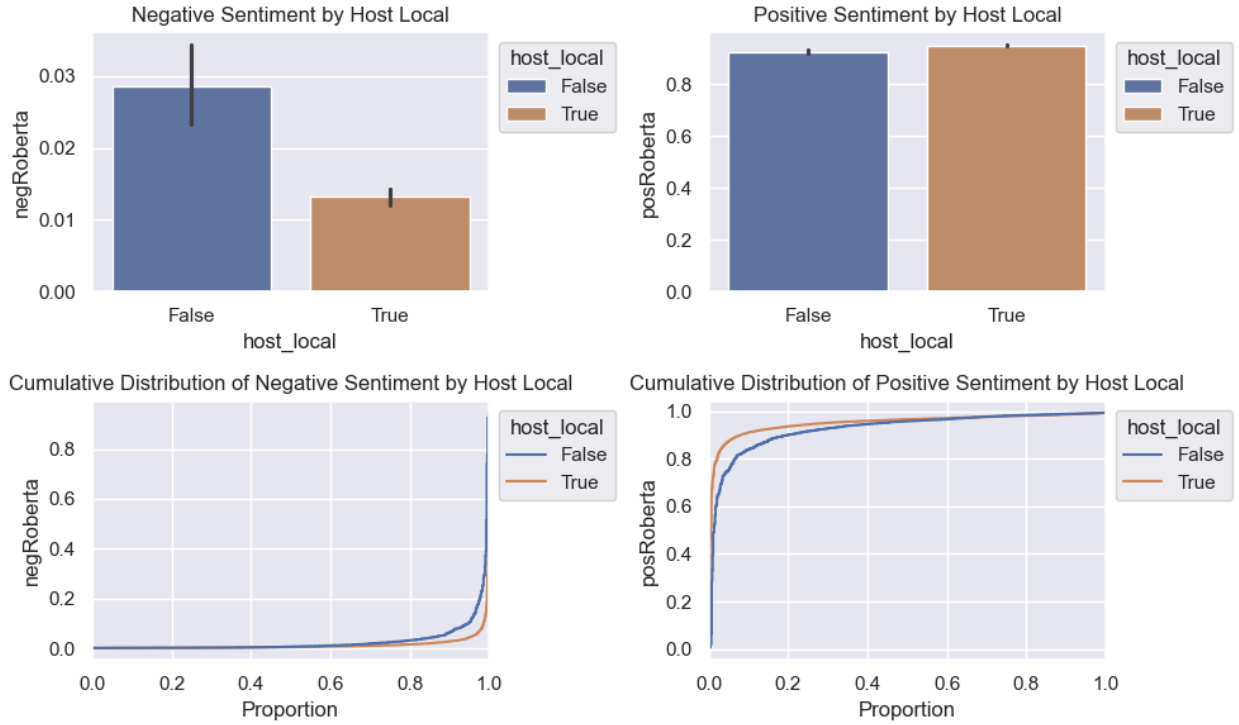


Figure 4 - Positive/Negative Roberta Scores in Super vs Non-Superhosts

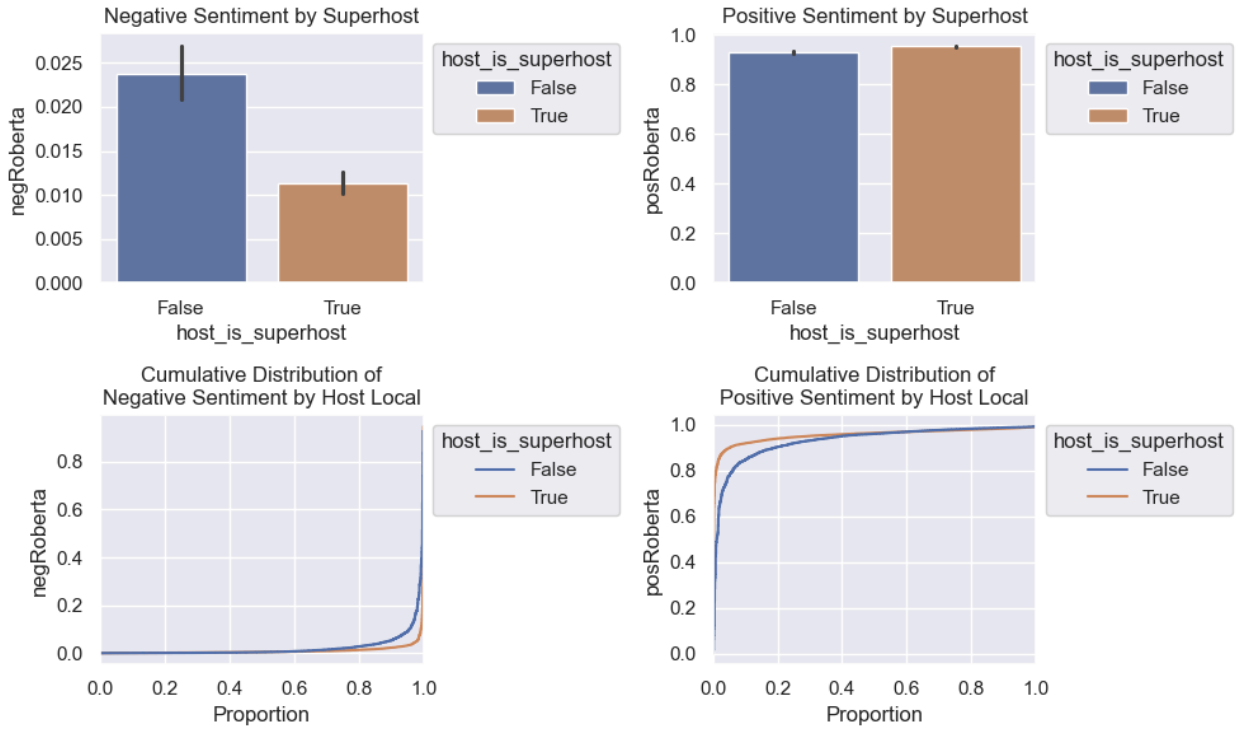


Figure 5 - Roberta Sentiment vs Corpus Length

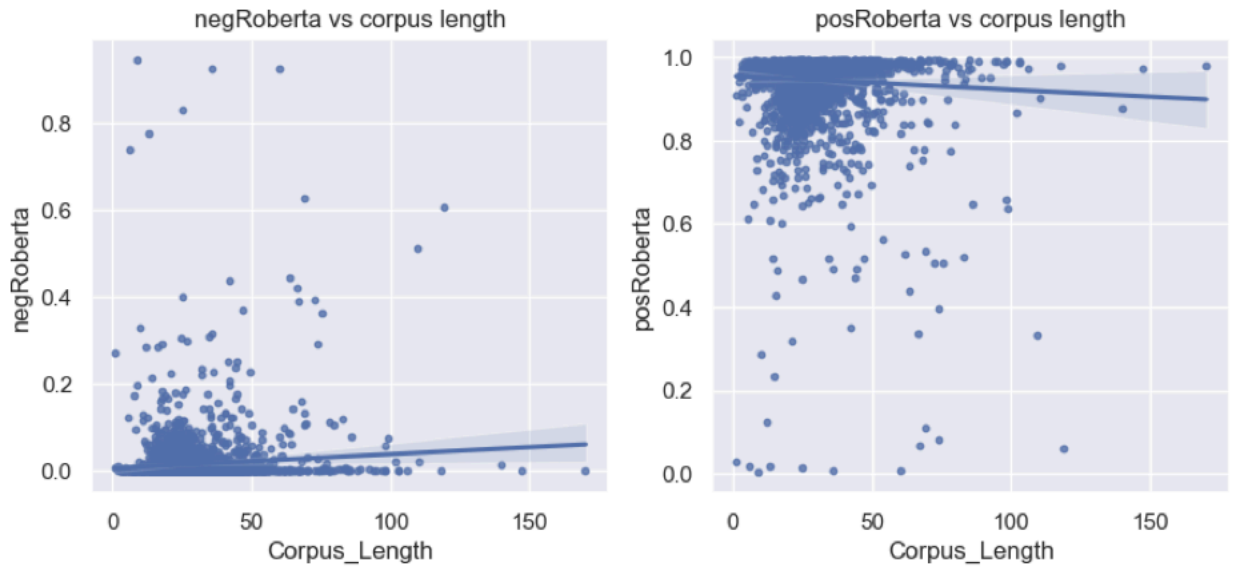


Figure 6 - Positive/Negative Roberta Sentiment by Neighborhood

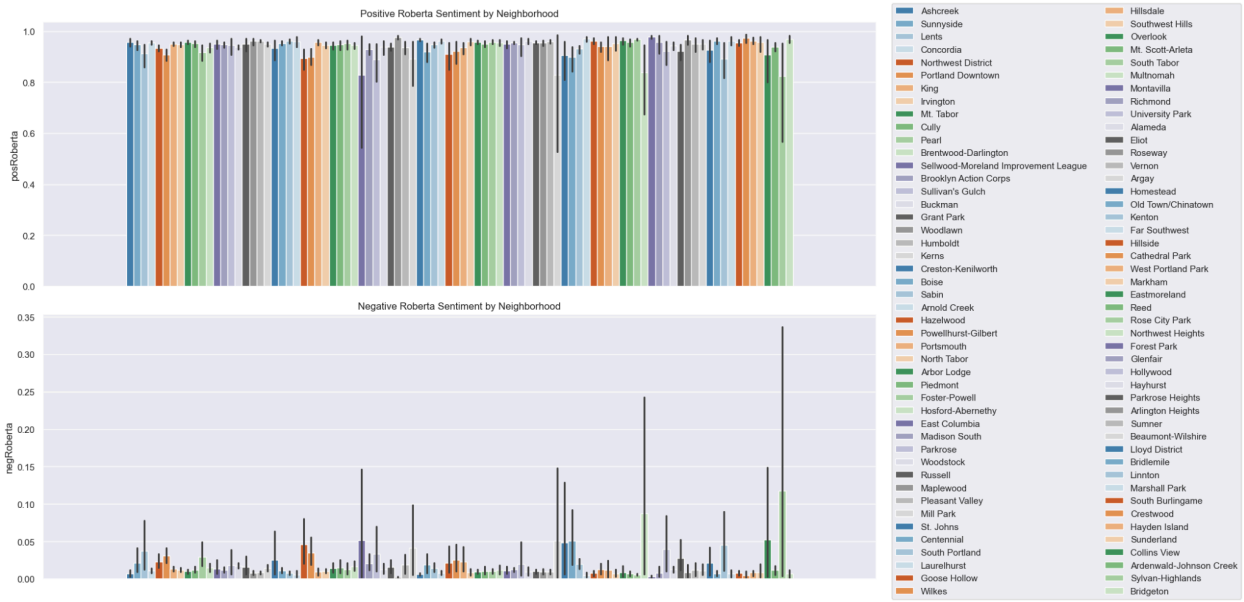


Figure 7 - Positive/Negative Roberta Sentiment by Sextant

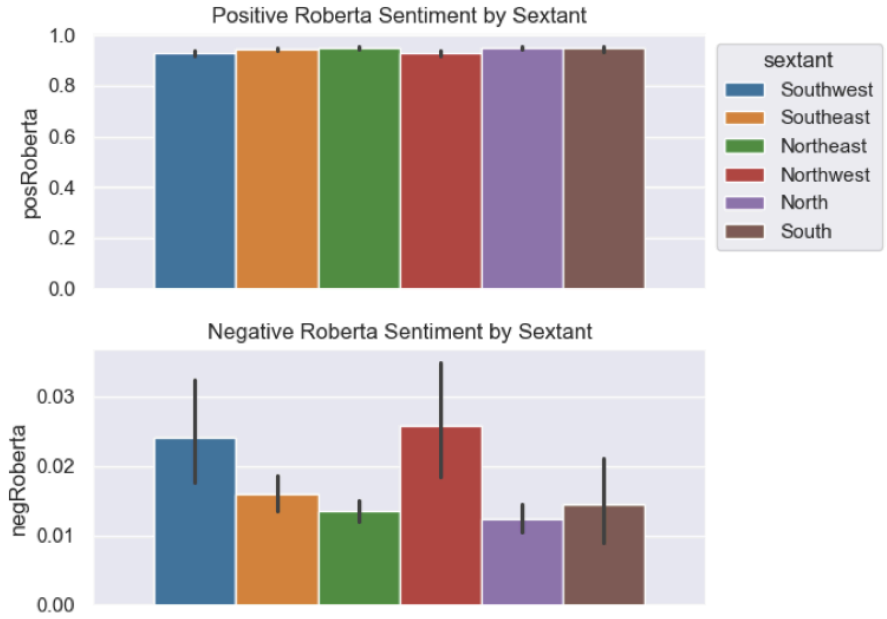


Figure - 8 Positive Sentiment Distributions

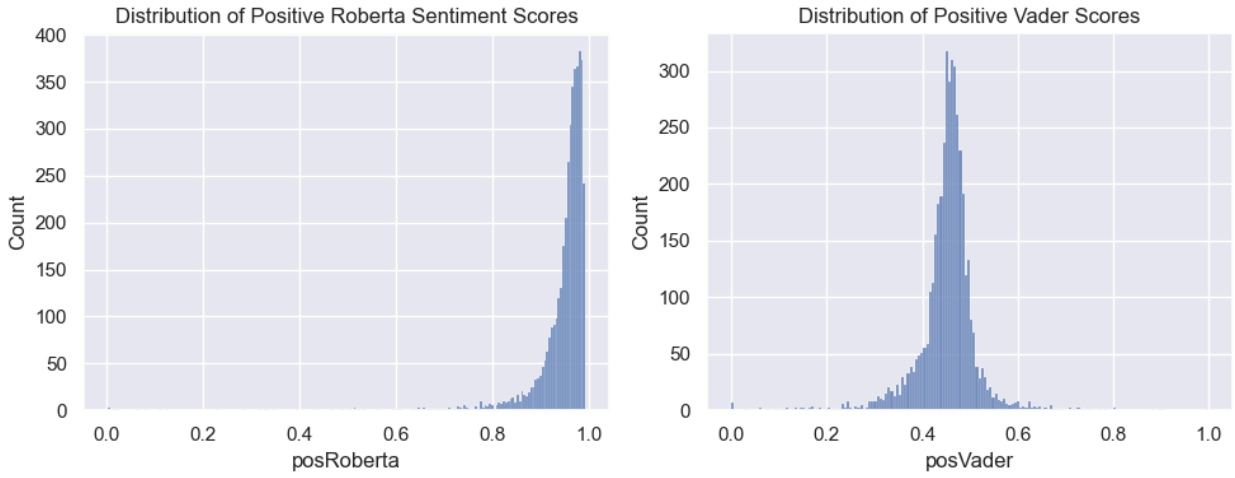


Figure 9 - Permutation Feature Importances, Full Dataset

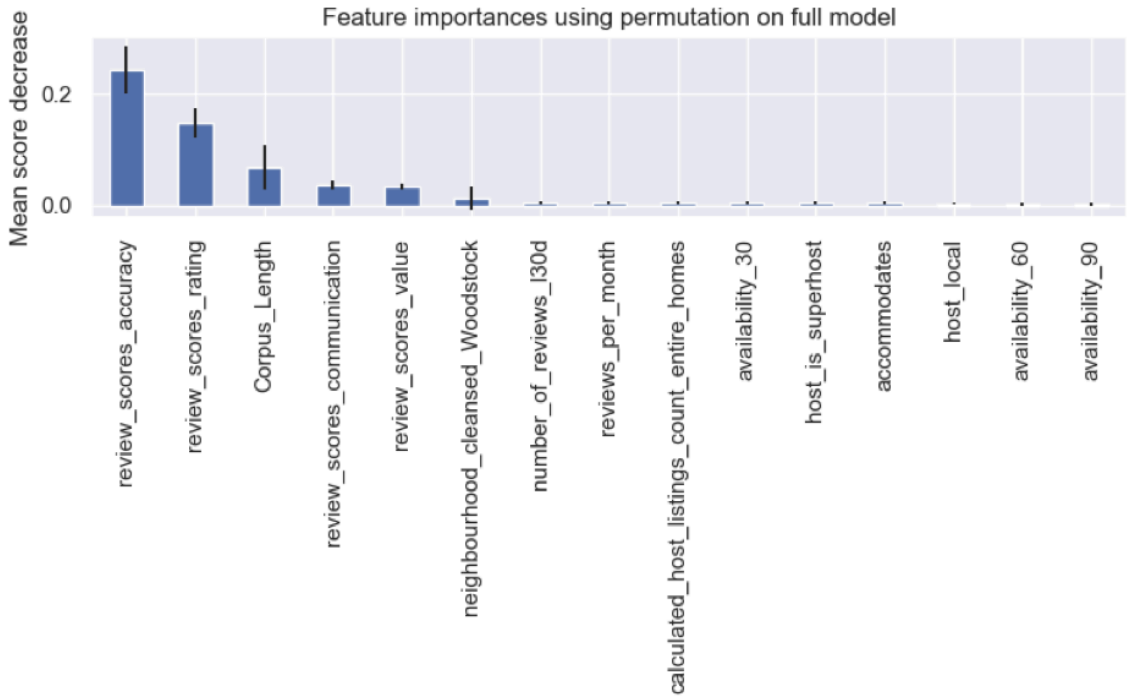


Figure 10 - Permutation Feature Importances, Subset without Review Scores

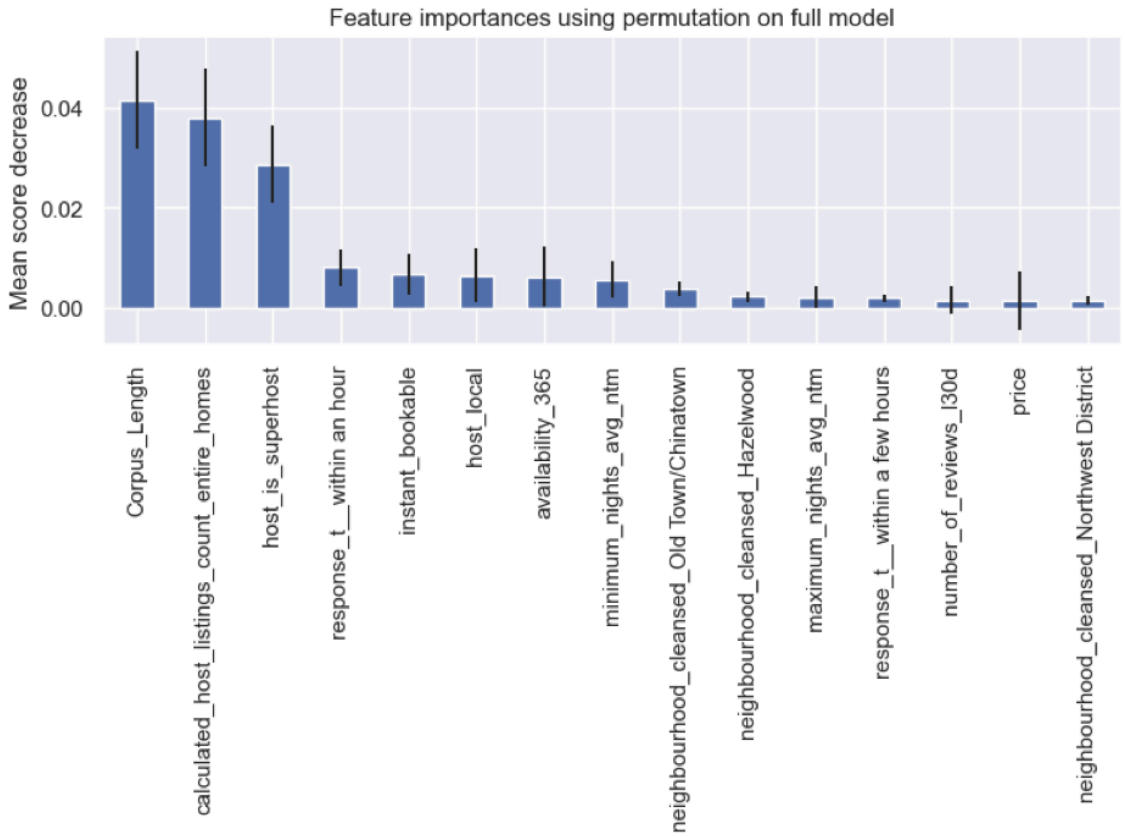


Table 1 - Feature Descriptions, merged set for Machine Learning

Feature name	Description
id	The shared listing id to merge the set of reviews with the set of listings
host_location	Self reported location of host
host_response_time	Categorical var denoting how quickly hosts respond to prospective guests: ‘within an hour’, ‘within a few hours’, ‘within a day’, ‘a few days or more’
host_is_superhost	Boolean, true if host is super host, false otherwise
host_verfications	Categorical var denoting methods by which host is verified: ‘phone’, ‘email’, ‘work email’
host_has_profile_pic	boolean, true if host has profile picture, false otherwise
host_identity_verified	boolean, true if host identity has been verified
neighbourhood_cleansed	Categorical var containing labels of neighborhood names where listing is located
room_type	Categorical var denoting what type of room listing is. All property types are categorized into the following classes: ‘Entire home/apt’, ‘Hotel room’, ‘Private room’, ‘Shared room’
accomodates	Integer denoting how many guests listing accomodates
price	Float, price of listing
minimum_nights_avg_ntm	The average number of minimum nights a listing is able to be booked within the next 12 months
maximum_nights_avg_ntm	The average number of maximum nights a listing is able to be booked within the next 12 months
has_availability	boolean, true if unit has availability
availability_30	integer, number of days listing is available within the next 30 days
availability_60	integer, number of days listing is available within the next 60 days
availability_90	integer, number of days listing is available within the next 90 days
availability_360	integer, number of days listing is available within the next 360 days
number_of_reviews_ltm	integer, the total count of reviews from the last 12 months
number_of_reviews_l30d	integer, the total count of reviews from the last 30 days
review_scores_rating	float, average overall rating
reviews_scores_accuracy	float, average score representing how accurate guests perceived listing to be compared to the actual unit.
review_scores_cleanliness	float, average score representing how clean the listing was
review_scores_checkin	float, average score representing how well check-in went

review_scores_communication	float, average score representing how well host communicates with guests
review_scores_location	float, average score representing how much guests like listing's location
review_scores_value	float, average score representing how much guests thought the unit was good value (was the listing worth the price?)
instant_bookable	boolean, true of listing is instant bookable
calculated_host_listings_count_entire_homes	integer, count of listings posted by host that are entire homes
calculated_host_listings_count_private_rooms	integer, count of listings posted by host that are private rooms
calculated_host_listings_count_shared_rooms	integer, count of listings posted by host that are shared rooms
negVader	float, negative vader sentiment score
neutralVader	float, neutral vader sentiment score
posVader	float, positive vader sentiment score
sentimentVader	float, composite vader sentiment score
negRoberta	float, negative Roberta sentiment score
neutralRoberta	float, neutral Roberta sentiment score
posRoberta	float, positive Roberta sentiment score
hostMentioned	float, proportion of reviews for a given listing that mentioned host by name
Corpus_Length	float, average corpus length of reviews for a given listing
Bedrooms	float, number of bedrooms in listing
Bathrooms	float, number of bathrooms in listing
Beds	float, number of beds in listing

Table 2 - Common Tokens, Reviews of Top Listings

Token (single word)	Count	Adjective-Noun	Count
good	754	great place	436
stay	721	great location	334
place	691	great stay	274
location	437	great host	217
clean	408	nice place	192

would	332	good location	122
host	289	next time	107
great	276	quiet neighborhood	101
comfortable	272	great restaurants	81
portland	265	great time	75
house	224	easy access	73
room	159	comfortable beds	65
home	157	perfect location	63
space	104	easy check	55
apartment	100	perfect place	52
easy	86	public transportation	51
neighborhood	84	comfortable bed	43
everything	84	next door	38
quiet	55	great space	36
u	53	great neighborhood	35

Table 3 - Random Forest Feature Importances

Feature	Importance
review_scores_accuracy	0.386356
review_scores_rating	0.269160
Corpus_Length	0.066015
review_scores_value	0.043656
review_scores_location	0.024202
review_scores_communication	0.021299
price	0.015443
reviews_per_month	0.014917
review_scores_cleanliness	0.012190
review_scores_checkin	0.011412

Table 4 - Random Forest Feature Importances, without review_scores features

Feature	Importance
Corpus_Length	0.147329
reviews_per_month	0.073344
calculated_host_listings_count_entire_homes	0.068567
price	0.060043
availability_365	0.052484
minimum_nights_avg_ntm	0.042300
number_of_reviews_ltm	0.040147
availability_90	0.038135
availability_60	0.034512
availability_30	0.033705

Figure 10 - Pair plots

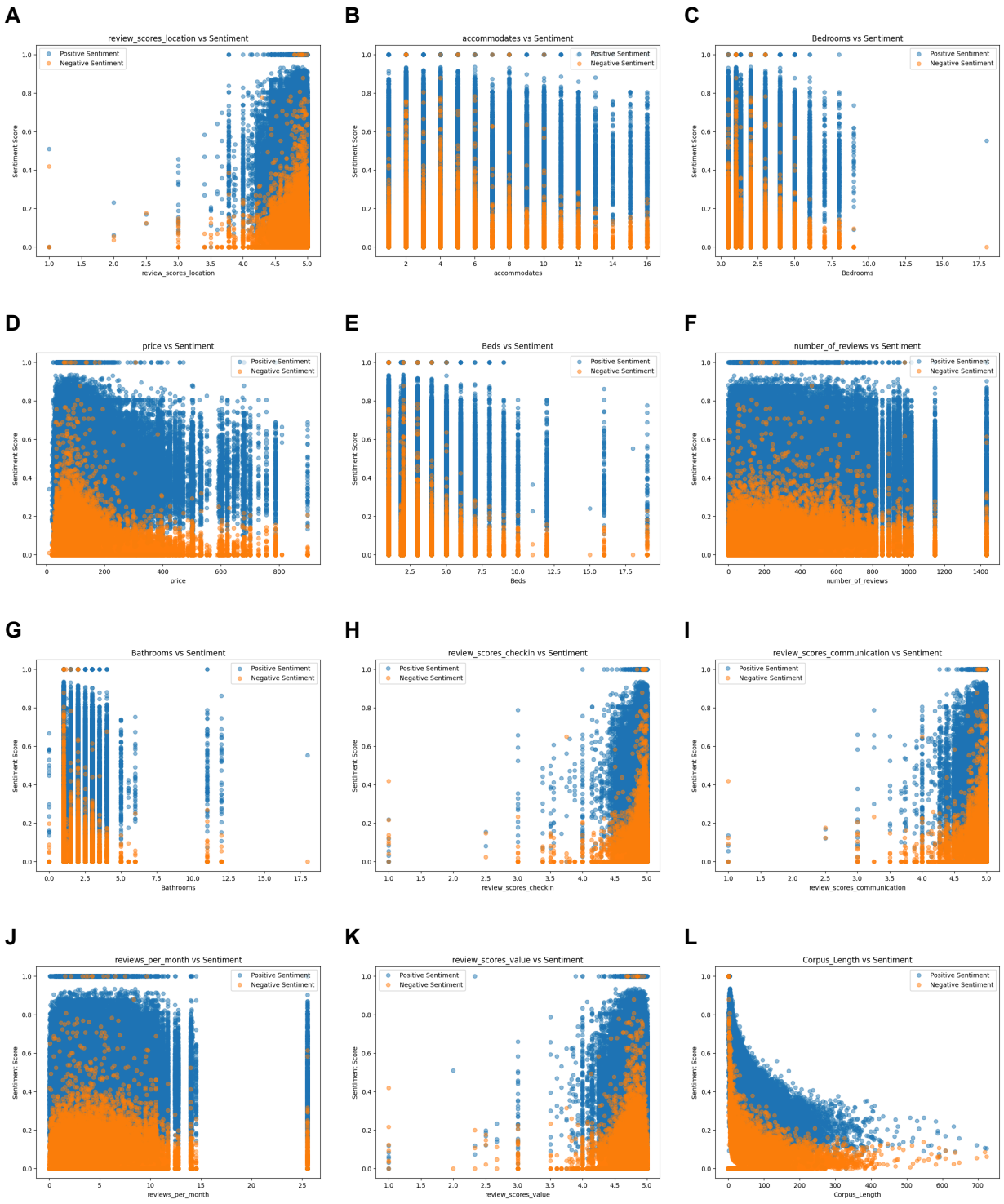
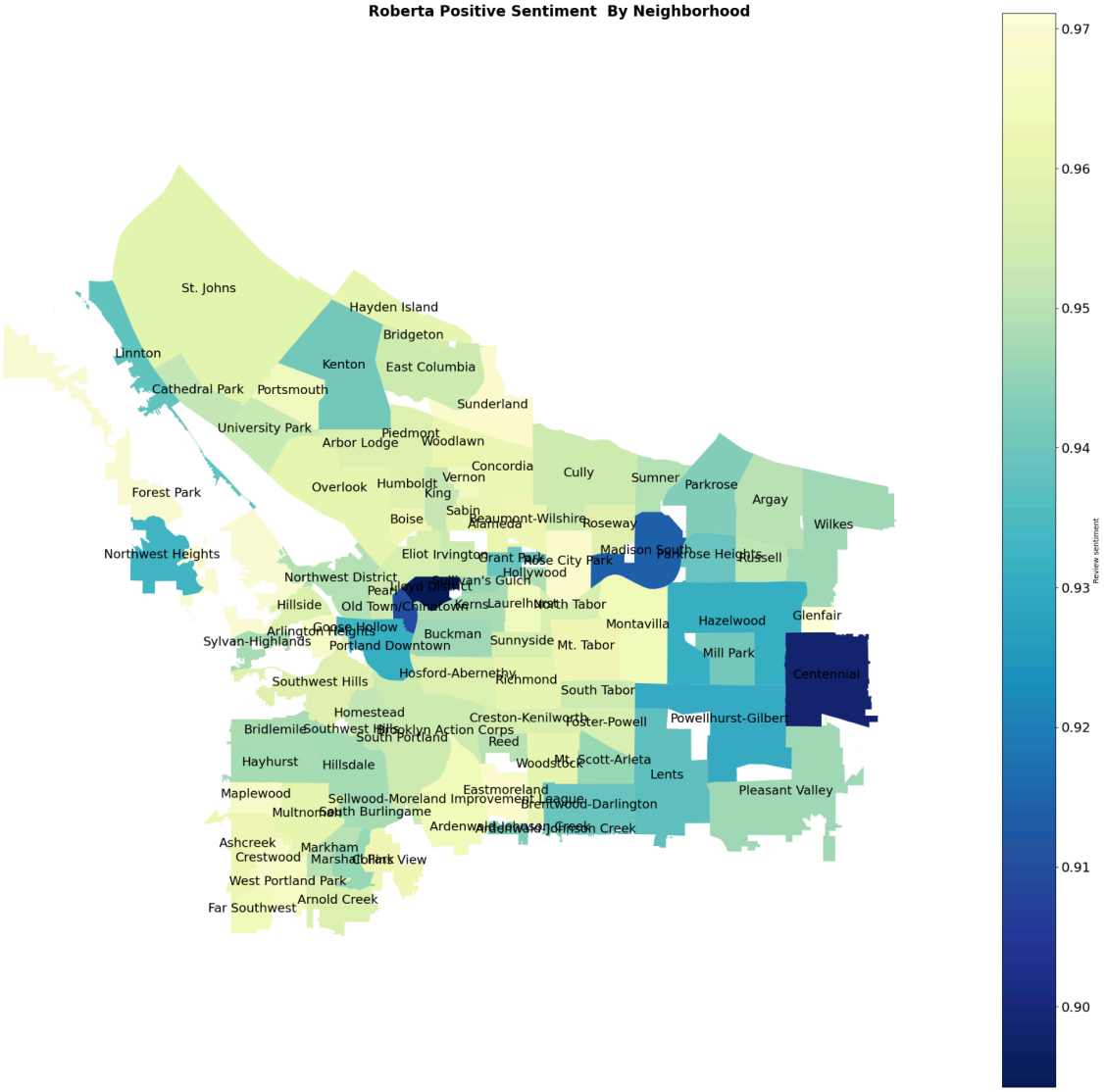


Figure 12


Roberta Positive Sentiment By Neighborhood
St. Johns
Hayden Island
Bridgeton
Linnton
Kenton
East Columbia
Cathedral Park
Portsmouth
Sunderland
University Park
Piedmont
Arbor Lodge
Woodlawn
Concordia
Cully
Sumner
Parkrose
Argay
Forest Park
Overlook
Humboldt
Vernon
King
Sabin
Boise
Beaumont-Wilshire
Alameda
Roseway
Wilkes
Northwest Heights
Eliot
Irvington
Grant Park
Rose City Park
Hollywood
Madison South
Parkrose Heights
Russell
Northwest District
Pearl District
Sullivan's Gulch
Hillside
Old Town/Chinatown
Kerns
Laurelhurst
North Tabor
Montavilla
Hazelwood
Glenfair
Arlington Heights
Goose Hollow
Sylvan-Highlands
Portland Downtown
Buckman
Sunnyside
Mt. Tabor
Mill Park
Hosford-Abernethy
Richmond
Centennial
Southwest Hills
South Tabor
Homestead
Creston-Kenilworth
Foster-Powell
Powellhurst-Gilbert
Bridlemile
Southwest Hill
South Portland
Brooklyn Action Corps
Reed
Hayhurst
Hillsdale
Woodstock
Mt. Scott-Arleta
Lents
Maplewood
Sellwood-Moreland Improvement League
Eastmoreland
Pleasant Valley
Multnomah
South Burlingame
Brentwood-Darlington
Ardenwald-Johnson Creek
Ashcreek
Crestwood
Markham
Marshall Park
Corbett
View
West Portland Park
Arnold Creek
Far Southwest
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.90
Review sentiment