A Comprehensive Sentiment Analysis for Amazon's Appliance Product Reviews

Team Members

Moumita Sen

Debanik Chakraborty

Mahsa Nafar Sefiddashti

Sentiment analysis

Determining the general emotions in the context of a review

Models for sentiment Analysis:

- Lexicon-based models
- Machine learning models

Research Objectives:

- To find and recommend the most appropriate model for sentiment analysis of Amazon Appliance products' reviews
- Find the correlation between price and sentiment score/ price and rating in each product type
- Recognizing the product type that needs improvement and give recommendation on that

Our focus:

✓VADER(Valence Aware Dictionary and sEntiment Reasoner)

- Specifically Designed for Social Media Text
- Fast Processing
- Dependency on Lexicon
- Limited Context Understanding

✓RoBERTa(Robustly Optimized BERT Pretraining Approach)

- Contextual Understanding
- Computational Resource
- Limited text Coverage

✓Google API(Application Programming Interface)

- Multiple Language Support
- Cloud Infrastructure offers reliability, scalability
- Potential Data Privacy Concerns
- Credit Needed

Data Cleaning Approach

Importing Data

↓

Convert Json to CSV

↓

Joining the data sets

↓

Dropping the duplicate & null rows for review column

↓

Keeping the reviews from verified users

↓

Proportionate stratified to get concise dataset

↓

Final data set has 10,610 rows

This dataset consists of 90,739 reviews of various product types within the Appliances Category on Amazon gathered by UCSD

Evaluation Criteria of the Models

1. **Relationship with True Sentiment Score:**
    - Distribution Similarity
    - Correlation to True Score

Decision: The most appropriate model

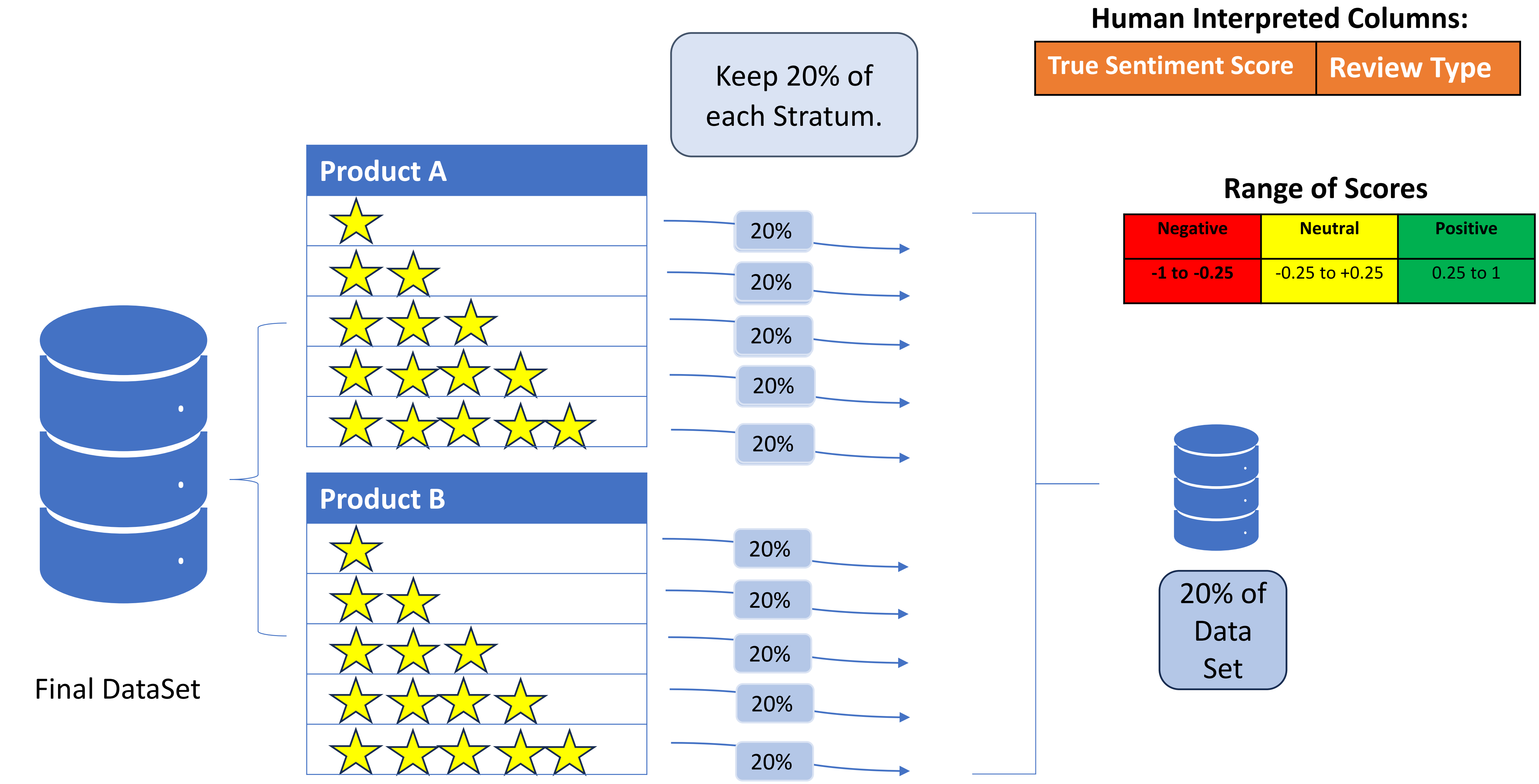
2. Relationship with the Ratings:

- Score vs Rating Bar Plot Analysis
- Correlation to Ratings

3. Precision, Recall and F1 Scores:

- True Positive, False Negative, True Negative, False Positive
- Deciding based on F1 Score

Sample Dataset for Human Interpretation and Model Comparison



Negative	Neutral	Positive
-1 to -0.25	-0.25 to +0.25	0.25 to 1

Compound Sentiment Scores

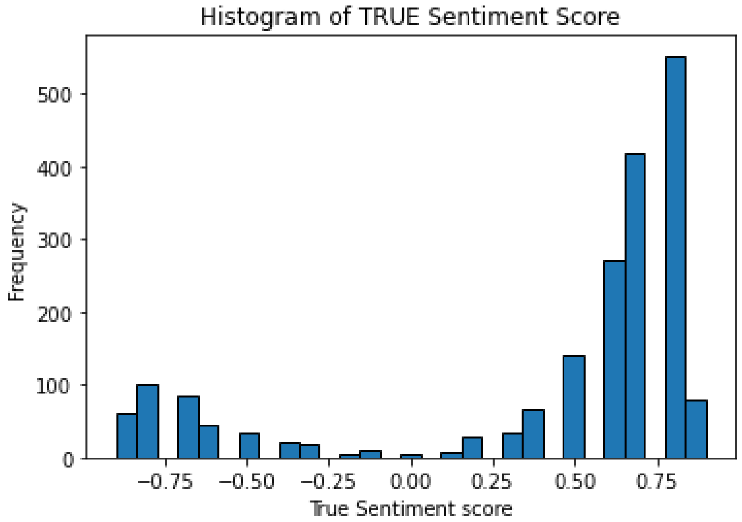
Google API

$$\tanh\left(\frac{\text{Sentiment\_Score}}{\text{Sentiment\_Magnitude}}\right)$$

RoBERTa

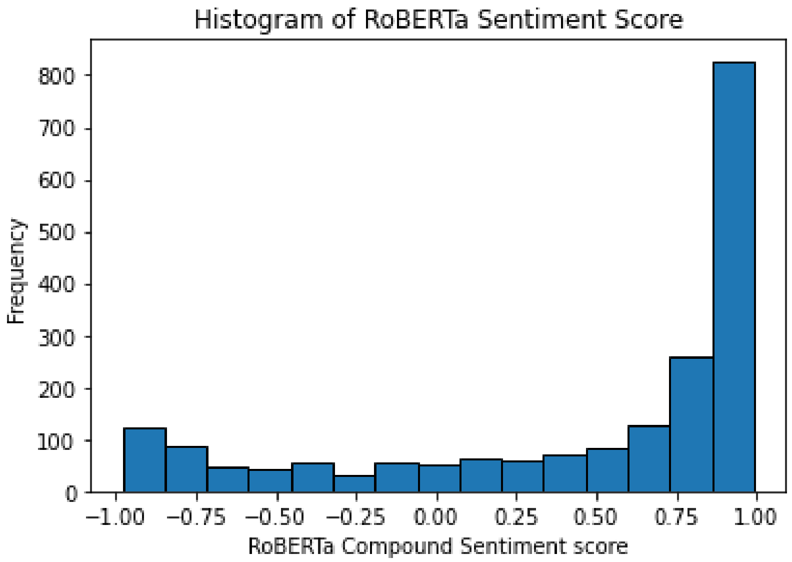
$$\text{roberta\_neg} + 2 \times \text{roberta\_neu} + 3 \times \text{roberta\_pos} - 2$$

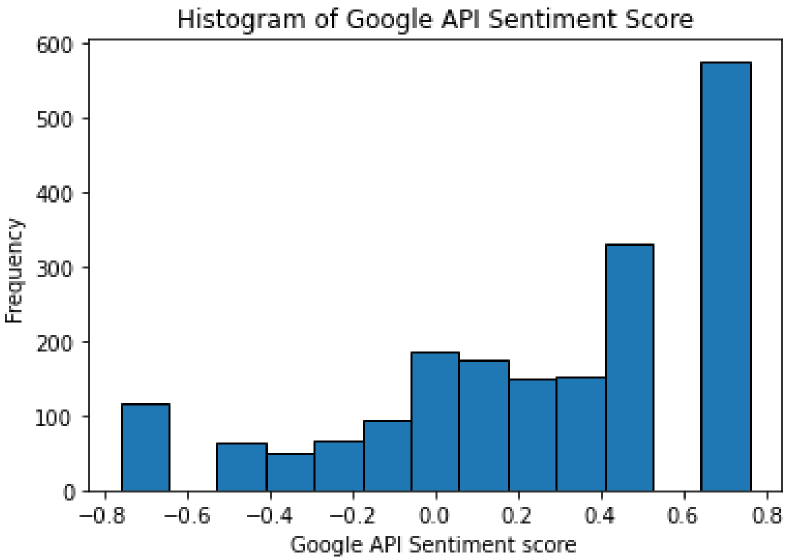
Criteria 1 :Relationship to the True Sentiment Score

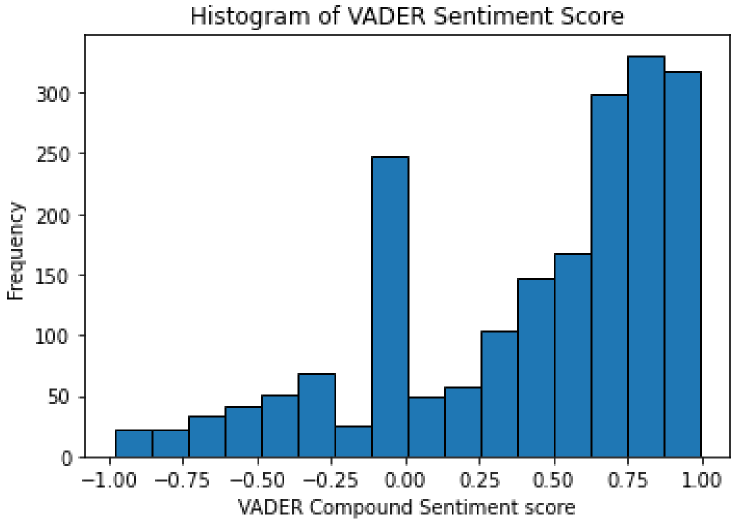


Correlation to True Sentiment Scores

Model	Correlation
VADER	0.55
RoBERTa	0.82
Google API	0.69



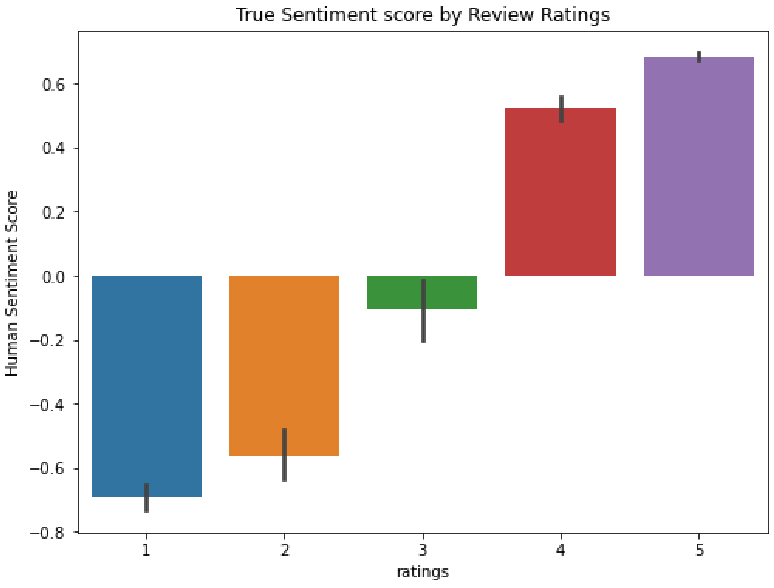




RoBERTa Senitment Score:

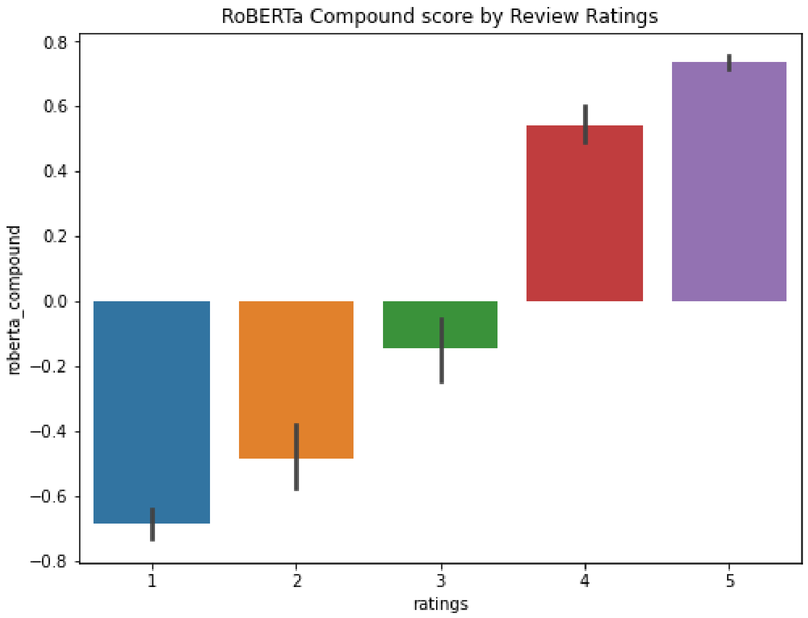
- More similar to True Sentiment Score distribution
- Maximum and Strong Positive Correlation to True Sentiment Score

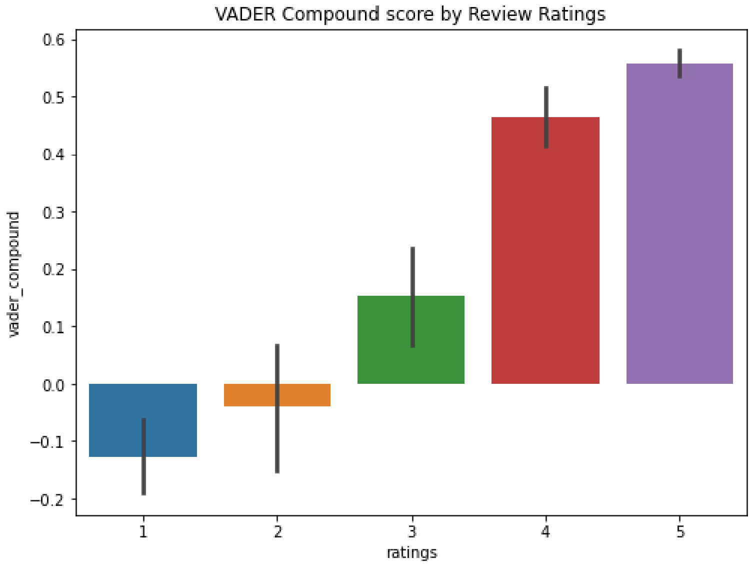
Criteria 2 : Relationship to the Ratings (Review Stars)



Correlation to Ratings

Model	Correlation
VADER	0.5
RoBERTa	★ 0.78
Google API	0.67





RoBERTa Sentiment Score:

- More similar to True Sentiment vs Rating Bar Chart
- Maximum and Strong Positive Correlation to Ratings

Criteria 3 : Precision, Recall and F-1 Score

True Sentiment Score \ Model Sentiment Score	+	−
+	**True Positive**	**False Negative**
−	False Positive	True Negative

Actually Positive

Review Type (Positive)

Positive in Model

Model Score > 0.25

Actually Negative

Review Type (Negative)

Negative in Model

Model Score < - 0.25

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Objective: Minimizing False Positive

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Objective: Minimizing False Negative

$$\text{F-1 Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Objective:
A Harmonic Mean of Precision and Recall
Finding a Balance between these two metrics.

Criteria 3 : Precision, Recall and F-1 Score

Precision, Recall and F1 Score

Model	Precision	Recall	F1-Score
VADER	0.93	0.95	0.94
RoBERTa	0.99	0.96	0.97
Google API	0.99	0.97	0.98

Key Insights:

RoBERTa and Google API:

- Exhibited Same capacity to Minimize the False Positive
- Google API performed slighty better in minimizing False Negative than RoBERTa

- VADER Model has the Least Precision, Recall and F-1 Score
- Despite having the same Precision, Google API has 1% higher Recall than RoBERTa
- Which is why F-1 Score in Google API is 1% higher than RoBERTa as well.
- Google API is slightly a better model in terms of Recall and F-1 Score than RoBERTa

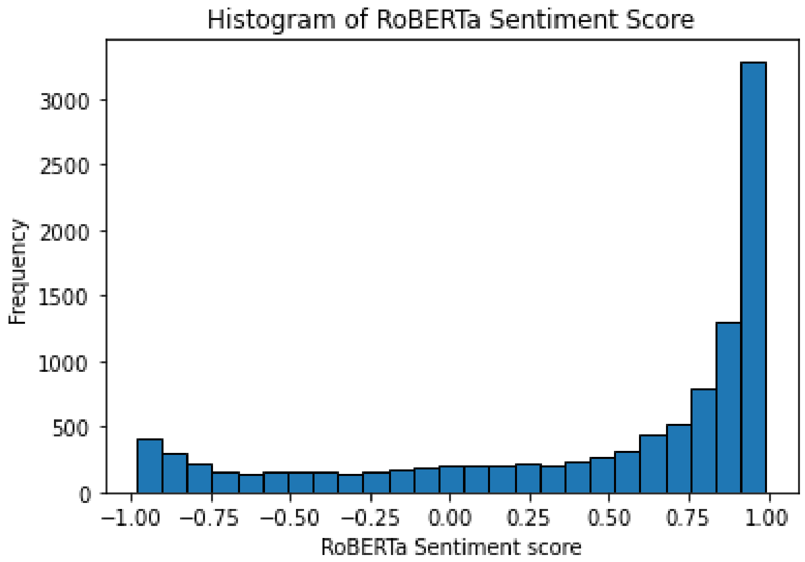
Decision on the Most Appropriate Model

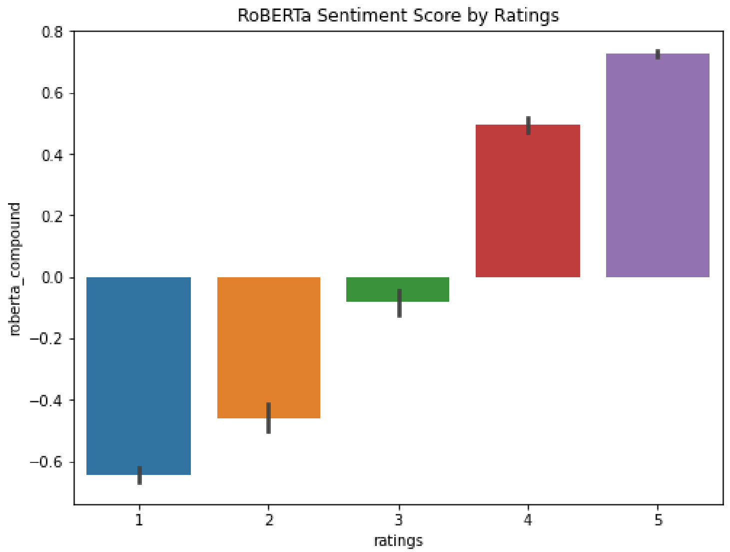
RoBERTA Model!!

Reasons:

- **RoBERTa:** Strong Positive Correlation to True Sentiment Score
- **RoBERTa:** Strong Positive Correlation to Ratings
- **Google API and RoBERTa** exhibited comparable performance in minimizing False Negative and False Positive, with marginal 1% difference considered negligible.
- **Google API:** Inconsistent Distribution, Google Credit needed

RoBERTa Sentiment Score on 10,610 rows





<u>Out of these Reviews:</u>

- 77% are Positive
- 23% are Negative

Correlation of RoBERTa Scores of our dataset to Ratings is 0.76 (which was 0.78 on Sample)

Product Type Specific Analysis
Price vs Sentiment Score & Price vs Ratings

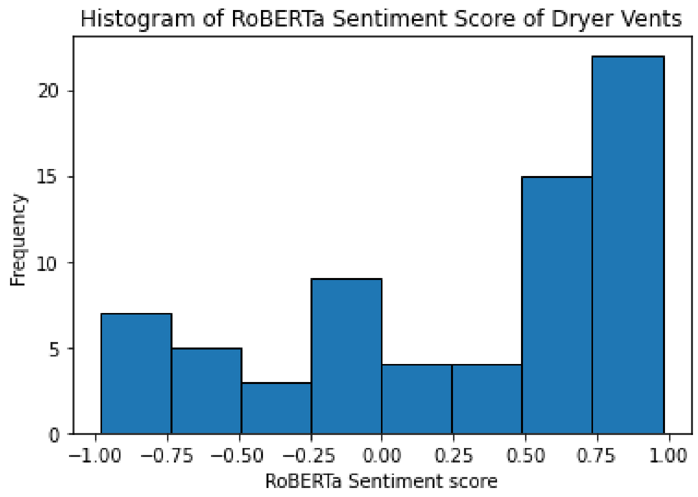
- Correlation between Prices and Ratings for each of 42 Product Types

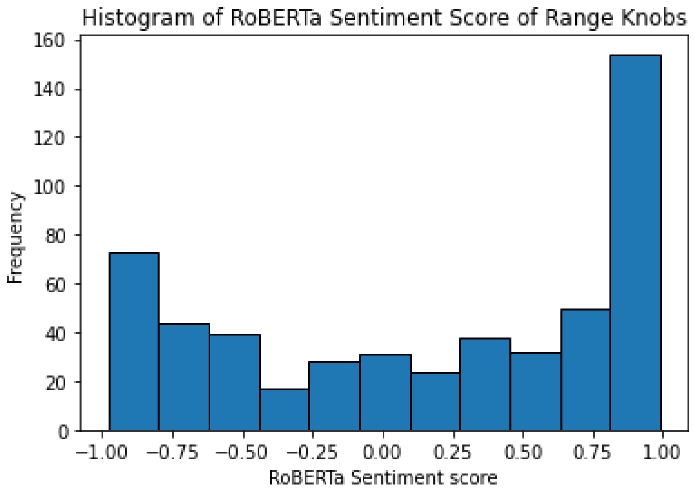
- Given the threshold of 0.3 in positive and –0.3 in negative (just above weak range)
- Only Single Wall Oven showed moderate correlation
- 0.46 in Price vs Sentiment Scores and 0.50 in Price vs Ratings.
- But Single Wall Oven only had 13 reviews, which is insignificant!

- So, we concluded that for almost all the Product types, **Price is not significantly related to Sentiment Scores or Ratings.**

Product Type Specific Analysis
High Negative Comments

- Percentage of Negative Reviews for each of the 42 product types
- Fixing 34% (above 1/3) as our threshold, we got..
- Build in Dishwasher- 39% negative reviews, but has only 23 reviews (negligible)
- Dryer Vents- 35% negative reviews out of its 69 reviews (prominent negative neutrals)
- Range Knobs- 39% negative reviews out of its 530 reviews





Range Knobs having the maximum negative sentiment from public **needs improvement!**!

Libraries
Pandas
Numpy
Matplotlib.pyplot
Seaborn
Tabulate
Language
Service_account
SentimentIntensityAnalyzer
tdqm
softmax
AutoTokenizer
AutoModelForSequenceClassification
WordCloud
ENGLISH_STOP_WORDS

Model used for RoBERTa
f"cardiffnlp/twitter-roberta-base-sentiment“

Thank You

Q&A