Flavorful Insights

A Customer Sentiment Analysis on Bakery & Confectionery Products from Amazon Fine Foods Reviews

Econ 425: Machine Learning

Master of Quantitative Economics,

University of California, Los Angeles

Oishika Kar

03.21.2025

Acknowledgements

I would like to sincerely thank several individuals and resources who have been instrumental in this project's success.

First and foremost, I extend my heartfelt gratitude to my Professor, Mr. Cheng Zhou, and Teaching Assistant, Andre Voronin. Their unwavering support, guidance, and valuable insights have significantly helped me shape this project and ensured its timely completion and delivery.

I want to acknowledge that this project builds upon the methods presented in the "Sentiment Analysis Python YouTube Tutorial" on Kaggle by Mulla (2021). The tutorial was essential in developing sentiment analysis models for this project (Mulla, 2021). However, this project has made notable modifications and enhancements to improve the efficiency of the methods, expand the analysis, and address the problem statement.

I also recognize the assistance of tools such as ChatGPT and Grammarly, which facilitated the refinement of my writing approach by playing a crucial role in ensuring my written work's clarity, accuracy, and professionalism, mainly through grammar corrections.

Finally, I thank the broader sentiment analysis and machine learning communities. Their research and contributions have formed the foundation of this project, and their collective efforts have greatly influenced my understanding of this field.

Table of Contents

1. Executive Summary

This project sheds light on the significance of machine learning in solving real-world economic problems. The project applies and compares the performance of sentiment analysis models like **VADER**, **Roberta**, and the **Hugging Face Pipeline (Transformers)** to identify which approach is most effective in capturing sentiment nuances. Comprehending the pros and cons of these models is crucial for businesses that want to understand customer preferences better. This project addresses an economic question relevant to large-scale e-commerce businesses and analyzes and provides practical insights that could benefit small business owners. It underscores the importance of sentiment analysis as a valuable decision-making tool in the food industry, particularly baking and confectionary products.

2. Problem Statement

Customer satisfaction is a critical factor for purchasing decisions in the e-commerce industry. Sentiment analysis gives businesses actionable insights by interpreting customer opinions in product reviews. E-commerce platforms like Amazon have transformed how customers interact with products and provide feedback. Online reviews reflect individual purchasing choices essential to a brand's overall reputation. Customers' opinions regarding taste, freshness, and presentation in the bakery business significantly impact purchasing decisions.

Traditional sales data generally does not capture qualitative insights that customer reviews express.

Considering this, this project aims to answer the following:

"How does sentiment analysis of customer reviews influence market strategies, product development, and pricing decisions for both large e-commerce platforms like Amazon and small businesses in the food industry?"

Let's break this down into more specific questions:

1. What are the consumer sentiments towards bakery & confectionery products on Amazon, and how do these sentiments affect market demand and pricing strategies?
2. To what extent do different sentiment analysis models impact the accuracy and cost-effectiveness of identifying consumer preferences in the bakery & confectionery products market?
3. How can insights from sentiment analysis inform strategic decisions in the food industry, such as product development, pricing, and competitive positioning?

3. Background

The problem statement is particularly important to me as my mother runs a small home-based baking business. Customer sentiment has always been a key part of her business decisions to refine product offerings and set competitive prices.

However, like many small businesses, due to small scale operations she lacks extensive data to run a comprehensive sentiment analysis on. While customer feedback can be collected, the limited volume of reviews makes it challenging to apply machine learning techniques effectively.

To overcome this limitation, the Amazon Fine Foods Reviews dataset was a good substitute as it offers a comprehensive collection of customer reviews from Amazon, a large-scale e-commerce platform. **A sentiment analysis of bakery product reviews from this dataset can help *mimic what a small business might face* if they had access to more extensive customer feedback.** This project offers important takeaways on how sentiment analysis helps to make informed decisions, optimize product quality, and enhance customer satisfaction.

Sentiment analysis helps businesses to:

- Identify major pain points of customers.
- Track shifts in customer preferences over time.
- Plan real-time strategies to boost the business.

This makes it an essential tool, in the competitive online market, for bakery businesses, both large and small, to efficiently meet consumer demand.

4. Data

Given my interest in applying sentiment analysis to support small businesses like my mother's bakery, the **Amazon Fine Foods Reviews** dataset (link in References) was an appropriate choice. While I couldn't directly use data from her business due to its limited size, this dataset provides a collection of 500,000+ customer reviews on food products, including numerous bakery items from 10/1999 to 10/2012.

This dataset is suitable for sentiment analysis as it contains both authentic text reviews and numerical ratings of products for a robust evaluation of model accuracy. Genuine customer feedback from diverse regions and demographics makes it representative of real-world consumer sentiment.

4.1 Features present in the data set:

- **Id** - Unique identifier for each review.
- **ProductId** - Unique identifier for the product being reviewed.
- **UserId** - Unique identifier for the user who wrote the review.
- **ProfileName** - Display name of the user who submitted the review.
- **HelpfulnessNumerator** - How many users found the review helpful.
- **HelpfulnessDenominator** - Total users who voted on the helpfulness of the review.
- **Score** - Rating given by the user, typically on a scale of 1 to 5.
- **Time** - Timestamp of when the review was submitted.
- **Summary** - Short summary of the review.
- **Text** - Elaborate review provided by the user.

5. Methodology

The focus of the models used in this project lies in using different NLP techniques to extract sentiment from customer feedback and classify them. The accuracy with which the classification is done demonstrates their performance and efficiency. The three sentiment analysis models used are as follows:

- VADER **(Valence Aware Dictionary and sEntiment Reasoner)**
- Roberta **(Robustly optimized BERT approach)**
- Hugging Face Pipeline (Transformers)

Rationale for Model Selection and Key Differences

Model	Type	Sentiment Output	Rationale for Use	Strengths	Key Differences
VADER	Rule-based, lexicon-based sentiment analysis	Sentiment scores: Negative Neutral Positive Compound	Best for short, informal text (e.g., customer feedback)	• Simple • Fast • Designed for short texts • Doesn't require training data	• Focuses on basic sentiment categorization • Looks for intensity of words and predefined list for which sentiment is attributed (lexicons)

Model	Type	Sentiment Output	Rationale for Use	Strengths	Key Differences
Roberta	Transformer-based pre-trained model	Sentiment scores: Negative Neutral Positive	Effective for analyzing longer and more complex text (e.g., customer reviews, product feedback)	• Provides state-of-the-art results, • Captures complex relationships in sentences. • Fine-tuned for sentiment analysis tasks.	• Pre-trained on diverse datasets • Captures intricate nuances • Gives more accurate sentiment classifications
Hugging Face Pipeline	High-level API using pre-trained transformer models (e.g., Roberta, BERT)	Sentiment labels: POSITIVE NEGATIVE sentiment score	General-purpose model, can be used for different text types (short or long)	• Easy to use • Goes well with Hugging Face's repository of models. • Very fast implementation.	• User-friendly API for sentiment analysis • Less customization • Conceals underlying technical configurations

Table 1

6. Analysis and Insights

6.1 Data Preprocessing

The data preprocessing began with loading the dataset with 568,454 rows and 10 columns including customer reviews. Initial data exploration steps were assessing the dataset size, examining column names, and renaming the 'Score' column to 'Rating' for more clarity, handling the missing values like identifying minor gaps in the 'Profile Name' and 'Summary' columns, with 26 and 27 missing entries, respectively. These were imputed with 'Unknown' for 'Profile Name' column and 'No Summary' for 'Summary' column.

Next, step was checking duplicate entries using the duplicated() function to confirm that no duplicate records were present. To refine the data set further for analysis of bakery-related products, a filtering step was applied using a keyword-based approach. A list of bakery-related keywords such as 'cake,' 'bread,' 'cookie,' and 'muffin' were defined. After matching these keywords within the 'Summary' column using a case-insensitive search, a subset containing only bakery-related product reviews was obtained. Finally, the cleaned and filtered data frame was displayed, ensuring it contained relevant and accurate information for further sentiment analysis.

Here's a slice of the final data frame after preprocessing:

```
      Id   ProductId          UserId     ProfileName  HelpfulnessNumerator  \
116  117  B0026Y3YBK  A3P60QLFDDCHOY  Giordano "GB"                      2
117  118  B0026Y3YBK  A38BUM0OXH38VK    singlewinder                     0
183  184  B001KUUNP6  A262Z0S6PT9U16    Lee Thombley                     3
375  376  B0087HW5E2  A139RTDNMU3WY5    blanket lady                     2
394  395  B001ELL6O8  A13T2G4T8LR8XA  First Time Mom                     2

     HelpfulnessDenominator  Rating        Time  \
116                       2       5  1304899200
117                       0       5  1347667200
183                       3       5  1292716800
375                       2       5  1339977600
394                       2       5  1189296000

                                         Summary  \
116                                Great cookies
117                        Best everyday cookie!
183  Perfect for gluten-free chocolate chip cookies
375           Greatest Oil since slice bread !!!!!!!
394                            Delisious Pancakes

                                                Text
116  I'm Italian and I lived in Italy for years. I ...
117  In the 1980s I spent several summers in Italy....
183  We made chocolate chip cookies with BRM Garban...
375  I have used this oil for several years and it ...
394  Before I discover this mix on Amazon I always ...
```

Table 2

6.2 Exploratory Data Analysis (EDA)

Initial EDA consisted of:

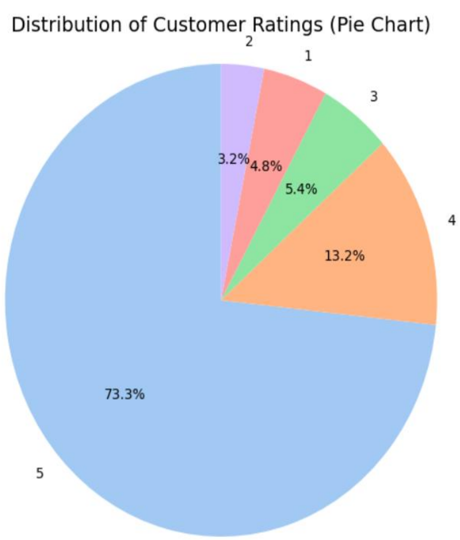
- Statistical summary table,
- Pie Chart
- Count Plot

Statistical summary

	Id	HelpfulnessNumerator	HelpfulnessDenominator	Rating	Time
count	11048.000000	11048.000000	11048.000000	11048.000000	1.104800e+04
mean	276918.953295	1.421705	1.747556	4.470221	1.281748e+09
std	165211.115639	3.706097	4.155602	1.062806	5.240040e+07
min	117.000000	0.000000	0.000000	1.000000	9.828000e+08
25%	143946.000000	0.000000	0.000000	4.000000	1.244160e+09
50%	263833.000000	0.000000	0.000000	5.000000	1.294445e+09
75%	420714.750000	2.000000	2.000000	5.000000	1.325635e+09
max	567573.000000	96.000000	96.000000	5.000000	1.351210e+09

Table 3

The mean rating of 4.47 indicates a strong inclination towards positive reviews. The standard deviation of 1.06 indicates a moderate variation in ratings, suggesting that while most products receive high ratings, there are still some products with low ratings (near the minimum of 1).



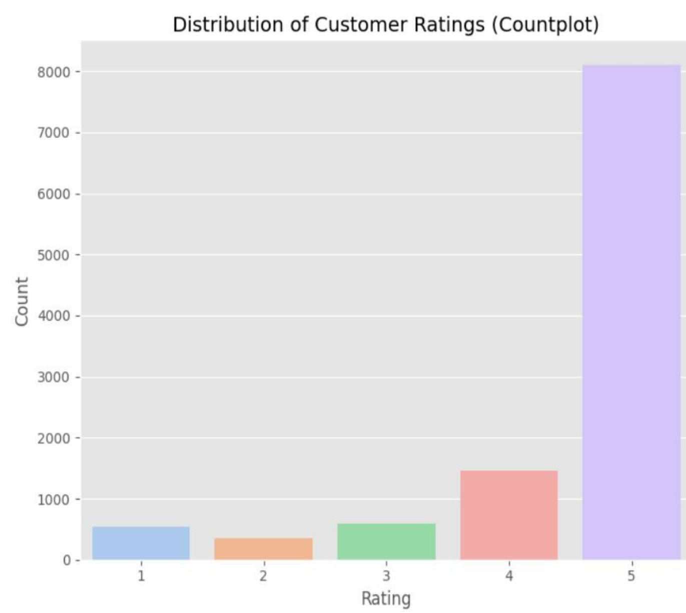


Fig 1

The ***pie chart*** revealed that about 73.3% received 5-star ratings, signifying high levels of satisfaction. Lower ratings, particularly 1-star and 2-star reviews, were minimal, reflecting a smaller proportion of dissatisfied customers.

The ***count plot*** shows the stark contrast between the number of 5-star ratings and those rated lower. The prevalence of positive reviews suggests a strong product market fit and generally favorable customer experience.

6.3 Vectorization and Keyword Analysis

Vectorization was used to identify bakery keywords in the text data to interpret sentiment scores. A TfidfVectorizer was used to convert the text data to numbers for easy analysis and interpretation. This step helps to identify the bakery keywords in the 'Summary' column. When these keywords are matched with text in the 'Summary' column, it can be determined if specific terms correlate with positive or negative sentiments and which products were frequently mentioned in reviews, enhancing model interpretability.

6.4 VADER (Valence Aware Dictionary and sEntiment Reasoner) Model

The first model implemented is VADER (Valence Aware Dictionary and sEntiment Reasoner) effective for performing sentiment analysis, which provides insight into the emotional tone of textual data. The ***SentimentIntensityAnalyzer*** from NLTK was implemented to analyze the sentiment of text in the Summary column of the dataset. This model produced four sentiment scores for each text: ***negative (vader_neg), neutral (vader_neu), positive (vader_pos), and compound (vader_compound).*** The compound score shows the overall sentiment polarity.

The function ***apply_vader_sentiment*** was used to produce these scores for each review in the dataset, which were subsequently visualized and analyzed.

```
      Id   ProductId          UserId    ProfileName  HelpfulnessNumerator  \
116  117  B0026Y3YBK  A3P60QLFDDCHOY  Giordano "GB"                     2
117  118  B0026Y3YBK  A38BUM0OXH38VK    singlewinder                     0
183  184  B001KUUNP6  A262Z0S6PT9U16    Lee Thombley                     3

     HelpfulnessDenominator  Rating        Time  \
116                       2       5  1304899200
117                       0       5  1347667200
183                       3       5  1292716800

                                          Summary  \
116                                 Great cookies
117                          Best everyday cookie!
183  Perfect for gluten-free chocolate chip cookies

                                                  Text  vader_neg  vader_neu  \
116  I'm Italian and I lived in Italy for years. I ...        0.0      0.196
117  In the 1980s I spent several summers in Italy....        0.0      0.308
183  We made chocolate chip cookies with BRM Garban...        0.0      0.575

     vader_pos  vader_compound
116      0.804          0.6249
117      0.692          0.6696
183      0.425          0.5719
```

Table 4

6.5 Plot of compound rating for the Vader Model

The compound score is a metric that combines positive, neutral, and negative sentiment scores into one value. It is a more intuitive way to analyze sentiments. Compound scores range from -1 (most negative) to +1 (most positive).

VADER Compound Score:

VADER uses a predefined lexicon of words with sentiment values and applies heuristics to calculate sentiment.

Formula for VADER compound score:

$$\textit{compound score} = \sum (\textit{positive score} - \textit{negative score}) \, / \, \textit{total words in text}$$

The compound sentiment scores were plotted against the ratings using a bar plot.

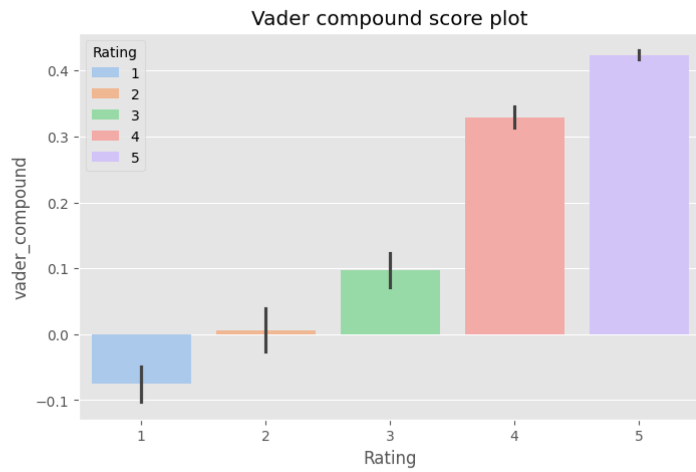


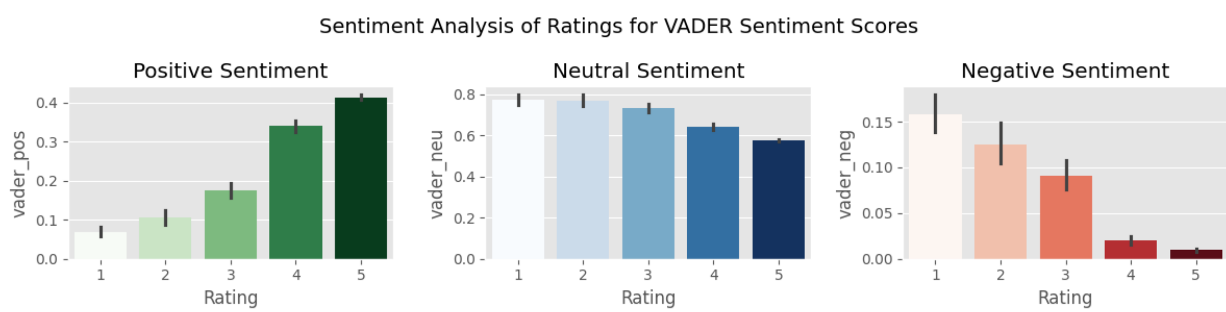
Fig 2

- Higher ratings (4 and 5 stars) show higher compound scores, implying more positive sentiments.
- Lower ratings (1, 2, and 3 stars) show lower compound scores, implying more negative or neutral sentiments.
- The error bars show variability in sentiment estimation
- Scores gradually increase with rating and sentiment becomes more positive for higher ratings and more negative for lower ratings.

6.6 Plot of Sentiment Analysis of Ratings for VADER Sentiment Scores

The plots show the relationship between ratings and the following sentiment scores from the VADER model:

- negative (vader_neg)
- neutral (vader_neu)
- positive (vader_pos)



Fig

- **Positive Sentiment:** Higher ratings (4 and 5) show higher positive sentiment scores indicating customer satisfaction. Darker shades of green indicate more positive sentiment and lighter shades of green indicate lesser positive sentiment. As customer ratings increase, the average positive sentiment score also increases. This indicates that higher-rated reviews tend to have stronger positive sentiment. The gradual shift from light green to dark green further emphasizes this trend. The error bars are relatively small, suggesting that the average sentiment scores are consistent within each rating category.

- **Neutral Sentiment:** Mid-range ratings (3) are neither very positive nor very negative. So, they indicate the highest neutral sentiment. Darker shades of blue indicate more neutral sentiment and lighter shades of blue indicate less neutral sentiment.

- **Negative Sentiment:** Lower ratings (1 and 2) show higher negative sentiment scores, indicating customer dissatisfaction. Darker shades of red indicate more negative sentiment and lighter shades of red indicate lesser negative sentiment. As customer ratings increase, the average negative sentiment score slightly decreases. This indicates that higher-rated reviews tend to have slightly less negative sentiment. The shift from light pink to dark red is subtle. The error bars are again relatively small.

How is this information useful to address the problem statement?

For Amazon, from the compound score one can understand the trends in consumer sentiments to optimize demand forecasting for popular bakery goods and strategic pricing. Amazon can also use the sentiment breakdown to optimize supply with trends in consumer preferences and improve their net promoter score for bakery product market. The plots also confirm prevailing information of sentiment towards bakery products that as customer ratings increase, the average negative sentiment score slightly decreases and the average positive sentiment score also increases. So, this can help Amazon position itself better in the bakery products marketplace to promote those bakery products which have a higher positive sentiment score and address concerns with those which have a higher negative sentiment score.

For my mother's small business, if she decides to expand her customer base and sell on Amazon or in general identify and introduce those bakery items which are most popular among Amazon customers from the sentiment breakdown, she can enhance her market positioning. These products can be easily identified as we now know the corresponding compound sentiment scores of items in the data set. The higher the compound score, the more popular is the product and higher is the customer satisfaction. This will help with increasing sales of products that align with consumer preferences.

6.7 Word Cloud for Vader Model

The word clouds generated based on the **VADER compound scores** provide valuable insights into customer sentiment. The **large size of a word shows that** it appears **more frequently** in the dataset. ***Color intensity*** shows the **strength of sentiment**. For example, ***a deeper red or green color could imply a more strongly expressed negative or positive sentiment.***

Positive Sentiment Word Cloud: Prominent words like "best", "great", "gluten free", "excellent", "yummy" show positive reception towards taste and dietary preferences.

Neutral Sentiment Word Cloud: Prominent words show lack strong emotion showing neutral sentiment. This cloud mainly shows a description of the variety of products.

Negative Sentiment Word Cloud: Prominent words like "broken cookies", "bad packaging", "hard", "terrible", "worst" show negative reception and dissatisfaction with product quality, particularly regarding taste.

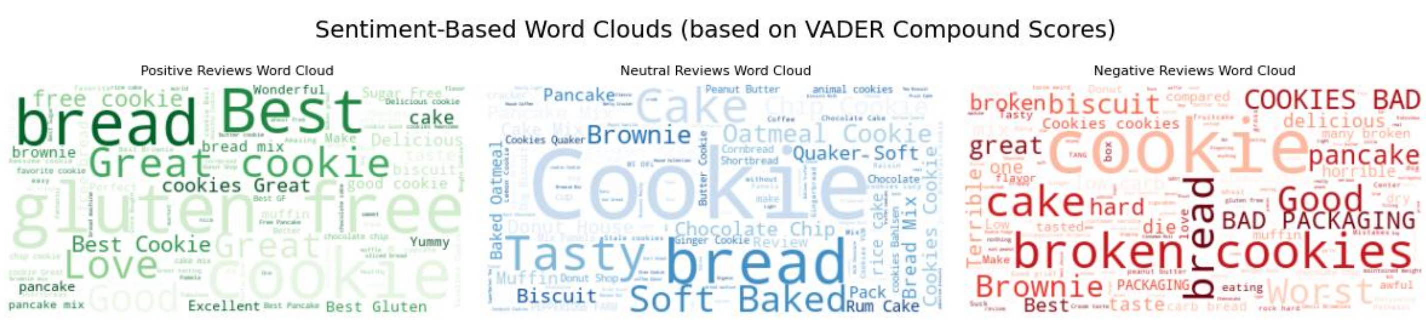


Fig 3

How is this information useful to address the problem statement?

Favorable terms like *"best," "great," "gluten-free," "excellent,"* and *"yummy"* show a strong approval of taste and dietary options. In contrast, negative words like *"broken cookies," "bad packaging," "hard,"* and *"terrible"* show dissatisfaction with packaging and product quality. Neutral terms like *"variety," "packaging,"* and *"price"* are areas where consumers seek more clarity or improved value.

For a small home-based baking business like my mother's, this information can be used to address customer complaints, promote products that have positive reception, and innovate new products based on neutral reviews. If she expands on platforms like Amazon Marketplace, she can use these trends to know what resonates with consumers. At the same time these reviews would help her identify product quality or packaging issues. She can also take into consideration popular diet choices like gluten-free or premium-quality baked goods to align her products with market demand and enhance her brand presence.

Similarly, Amazon could use this information to change product visibility and influence purchasing behavior. Amazon can promote items with positive sentiment and demoting poorly rated products, thereby enhancing overall customer experience.

6.8 ROBERTA (Robustly optimized BERT approach) Model

The second model implemented is **Roberta** (Robustly optimized BERT approach), an advanced transformer-based model for sentiment analysis. It is highly effective for understanding the sentiment of textual data. Hugging Face's **transformers** library was used to load a pre-trained **Roberta** model fine-tuned for sentiment analysis on Twitter data. The model produced three sentiment scores for each text: **negative (Roberta_neg)**, **neutral (Roberta_neu)**, and **positive (Roberta_pos).** The ***compound score*** was manually calculated.

The function ***apply_Roberta_sentiment*** was used to produce these scores for each review in the dataset, which were subsequently visualized and analyzed.

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator
116	117	B0026Y3YBK	A3P60QLFDDCHOY	Giordano "GB"	2
117	118	B0026Y3YBK	A38BUM0OXH38VK	singlewinder	0
183	184	B001KUUNP6	A262Z0S6PT9U16	Lee Thombley	3
375	376	B0087HW5E2	A139RTDNMU3WY5	blanket lady	2
394	395	B001ELL6O8	A13T2G4T8LR8XA	First Time Mom	2

	HelpfulnessDenominator	Rating	Time
116	2	5	1304899200
117	0	5	1347667200
183	3	5	1292716800
375	2	5	1339977600
394	2	5	1189296000

	Summary
116	Great cookies
117	Best everyday cookie!
183	Perfect for gluten-free chocolate chip cookies
375	Greatest Oil since slice bread !!!!!!!
394	Delisious Pancakes

	Text	vader_neg	vader_neu
116	I'm Italian and I lived in Italy for years. I ...	0.0	0.196
117	In the 1980s I spent several summers in Italy....	0.0	0.308
183	We made chocolate chip cookies with BRM Garban...	0.0	0.575
375	I have used this oil for several years and it ...	0.0	0.482
394	Before I discover this mix on Amazon I always ...	0.0	1.000

	vader_pos	vader_compound	roberta_neg	roberta_neu	roberta_pos
116	0.804	0.6249	0.004493	0.070814	0.924693
117	0.692	0.6696	0.001362	0.034527	0.964110
183	0.425	0.5719	0.001589	0.059519	0.938892
375	0.518	0.7482	0.013173	0.099402	0.887424
394	0.000	0.0000	0.004090	0.254064	0.741847

	roberta_compound
116	0.920200
117	0.962748
183	0.937303
375	0.874251
394	0.737757

Table 5

6.9 Plot of compound rating for the RobertaModel

RobertaCompound Score:

Formula for Robertacompound score:

$$compound\ score = \sum (positive\ score - negative\ score) \,/\, positive\ score + negative\ score + neutral\ score$$

The compound sentiment scores were plotted against the ratings using a bar plot.

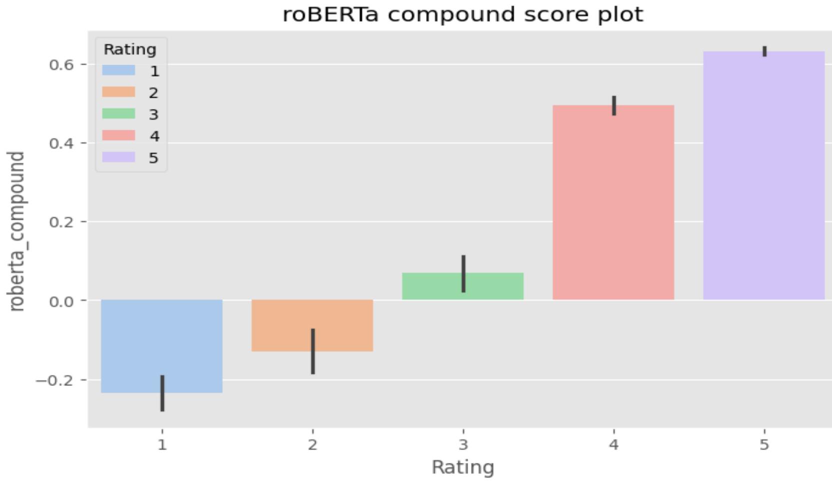


Fig 4

- Higher ratings (4 and 5 stars) show higher compound scores, implying more positive sentiments.
- Lower ratings (1, 2, and 3 stars) show lower compound scores, implying more negative or neutral sentiments.
- The error bars show variability in sentiment estimation
- Scores gradually increase with rating and sentiment becomes more positive for higher ratings and more negative for lower ratings.

6.10 Plot of Sentiment Analysis of Ratings for RobertaSentiment Scores

The plot shows the relationship between ratings and the following sentiment scores from the Robertamodel:

- negative (roberta_neg)
- neutral (Roberta_neu)
- positive (Roberta_pos)

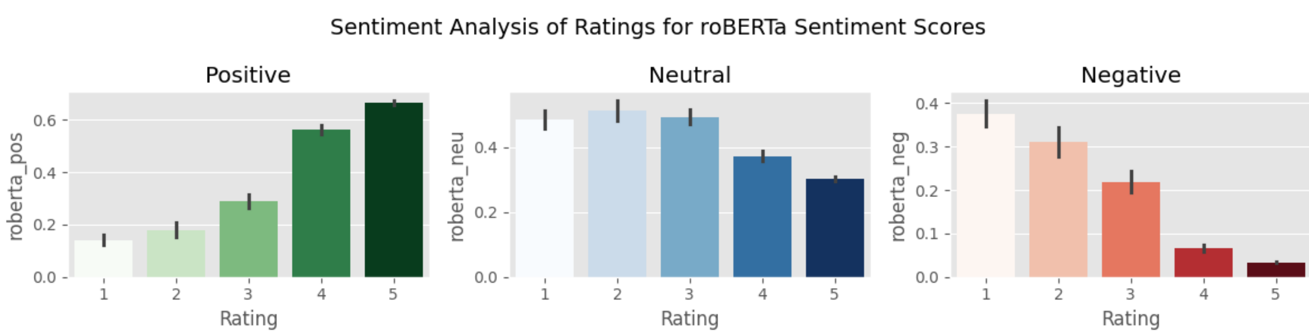


Fig:

- **Positive Sentiment:** Higher ratings (4 and 5) show higher positive sentiment scores indicating customer satisfaction. Darker shades of green indicate more positive sentiment and lighter shades of green indicate lesser positive sentiment. As customer ratings increase, the average positive sentiment score also increases. This indicates that higher-rated reviews tend to have stronger positive sentiment. The gradual shift from light green to dark green further emphasizes this trend. The error bars are relatively small, suggesting that the average sentiment scores are consistent within each rating category.

- **Neutral Sentiment:** Mid-range ratings (3) are neither very positive nor very negative. So, they indicate the highest neutral sentiment. Darker shades of blue indicate more neutral sentiment and lighter shades of blue indicate less neutral sentiment.

- **Negative Sentiment:** Lower ratings (1 and 2) show higher negative sentiment scores, indicating customer dissatisfaction. Darker shades of red indicate more negative sentiment and lighter shades of red indicate lesser negative sentiment. As customer ratings increase, the average negative sentiment score slightly decreases. This indicates that higher-rated reviews tend to have slightly less negative sentiment. The shift from light pink to dark red is subtle. The error bars are again relatively small.

6.11 Word Cloud for RobertaModel

The word clouds generated based on the **Robertacompound scores** provide valuable insights into customer sentiment.

Positive Sentiment Word Cloud: Prominent words like “best”, “great”, “gluten free”, “excellent”, “yummy” show positive reception towards taste and dietary preferences.

Neutral Sentiment Word Cloud: Prominent words show lack strong emotion showing neutral sentiment. This cloud mainly shows a description of the variety of products.

Negative Sentiment Word Cloud: Prominent words like “broken”, “bad packaging”, “bad”, “terrible”, “tasted stale” show negative reception and dissatisfaction with product quality, particularly regarding taste.

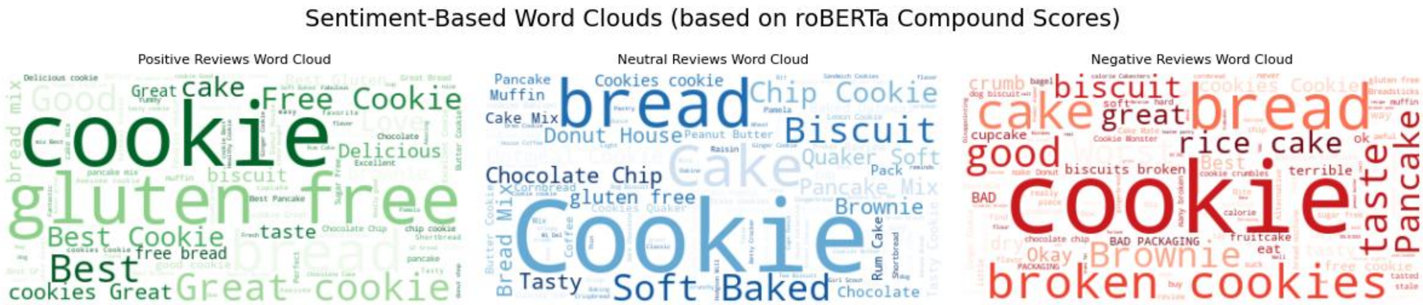


Fig 5

Like the VADER model, this analysis indicates room for improvement in product's quality, diversifying product range, better packaging to address prevalent customers’ needs.

6.12 Pair plot for comparison of VADER and Robertamodels

A pair plot for comparing the VADER and Robertamodels was created using the pair plot method in Seaborn to get a comparison between the two models for sentiment detection and their correlation with customer ratings.

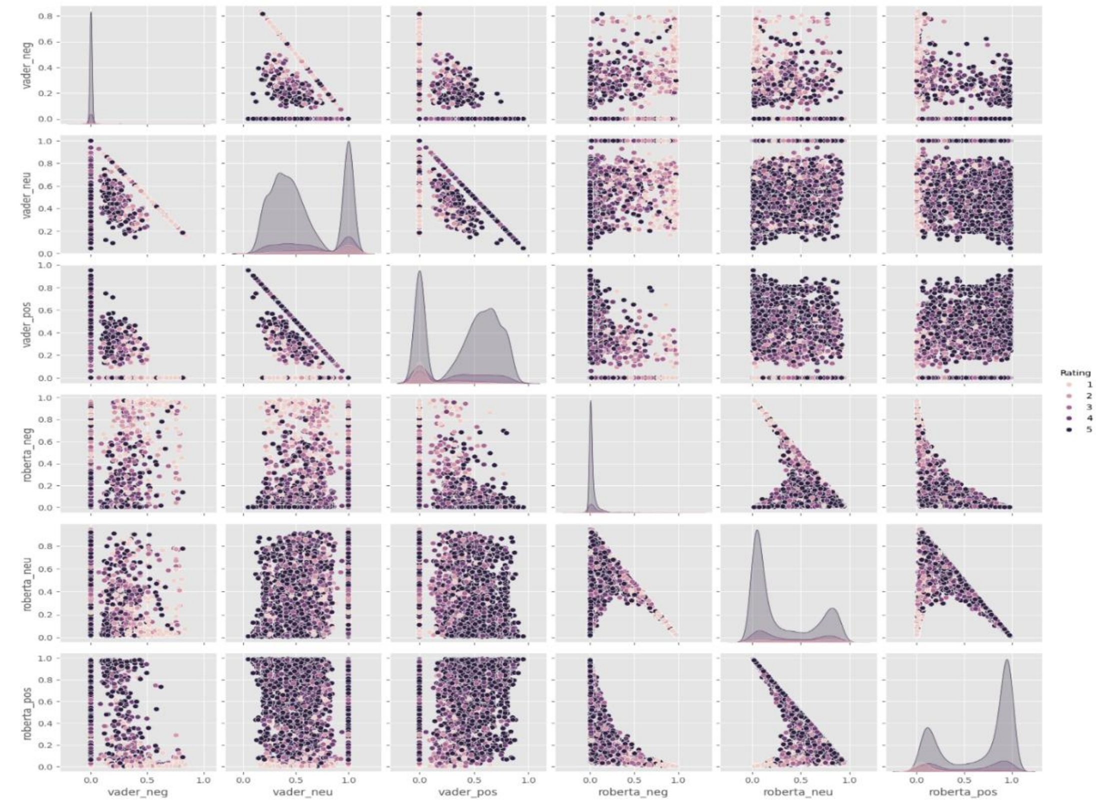


Fig 6

6.13 VADER vs. RobertaAgreement: *Addressing the Problem Statements*

- A positive **correlation** exists between the sentiment scores (positive and negative) of both the models implying that they are on the same page for sentiment classification.
- **The neutral sentiment** scores also show correlation, but they have a clustered distribution appears more clustered implying that that both the models classify reviews as neutral many a times.

What This Means for Consumer Sentiment and Market Demand?

The models show a positive correlation between the sentiment scores implying that they both are on the same page as far as prevailing customer sentiment of bakery products is concerned. Both the models can pick on consumer sentiment which is crucial for understanding how reviews impact market demand and pricing strategies. But the clustered neutral sentiment points challenge the identification of more subtle or balanced consumer opinions. This can influence pricing decisions for less polarizing products.

For example, if a bakery product has positive reviews majorly then it is fair that it may have a higher price whereas a product with mostly negative reviews could indicate that it need rebranding and other adjustments depending on what consumers want.

Sentiment and Rating Relationship

- **Positive ratings** (4 and 5 stars) correspond with **higher positive sentiment** scores -> consistent with expectations.
- **Lower ratings** (1 and 2 stars) show a correlation with **higher negative sentiment ->** dissatisfaction.
- The **neutral sentiment** scores show weak or no significant relationship with the ratings -> assigning neutral sentiments regardless of rating.

What is the key takeaway for sentiment analysis models and consumer preferences?

Positive correlation between higher ratings and positive sentiment, and lower ratings with negative sentiment, matches the prevailing consumer behavior for bakery products. **The findings are important because they show that sentiment analysis models are accurate in identifying consumer satisfaction based on ratings, essential for identifying consumer preferences.** However, neutral sentiments have a weak relationship with the ratings. So, the models might miss subtle preference nuances, affecting the cost-effectiveness of sentiment analysis of bakery products. **Roberta, which is better at distinguishing sentiments across ratings, is more accurate in identifying nuanced preferences.**

Differences

- Robertaclearly separates positive and negative sentiments between different rating groups.
- VADER has an inclination toward assigning **more neutral scores**, regardless of the rating, making it **less sensitive to sentiment**.
- Robertahas **better negative sentiment detection** compared to VADER.

What are the implications for market accuracy and cost-effectiveness?

Roberta's **improved accuracy in** differentiating sentiments based on ratings **is important for identifying shifts in consumer preferences that directly impact market demand, pricing, and product strategies.** VADER classifies more reviews as neutral **so it may be less cost-effective** in capturing subtle consumer sentiments, missing out on key insights for market positioning or differentiated product development.

For example, Roberta's differentiates between highly positive and slightly positive reviews so one can have more tailored pricing strategies, while VADER's fails to notice subtle distinctions in consumer sentiment.

Score Distribution:

- The positive sentiment scores for both models are skewed. More reviews have scores toward higher end (0.8-1.0).
- The negative sentiment scores are clustered around the lower end (0.0-0.2). This shows extreme negativity in the reviews.
- Neutral scores are concentrated at higher values. The models tend to classify a lot of reviews as neutral (especially VADER).

What insights can be gained for strategic decisions in the food industry?

The skewed distribution of sentiment scores, with more reviews being classified as positive or negative, suggests that consumer sentiments toward bakery products are generally polarized. So strategic decisions could be made related to product development. Products with strong positive or negative sentiments require more attention. The concentration of neutral sentiments in VADER may indicate that consumers find certain products less distinct in their preferences impacting their market positioning and pricing strategies. Having an idea of how neutral and extreme sentiments are distributed can help brands better align their products with customer expectations. This also helps to ensure that pricing strategies are competitive and responsive to consumer needs.

Which model is better?

- Robertacaptures sentiment nuances more effectively making it a **more sensitive and accurate tool for sentiment analysis**, primarily in recognizing negative sentiment, and differentiates well across rating categories.
- VADER often assigns neutral scores and shows less differentiation between sentiment categories, so it might not be as finely tuned to detect sentiment variations.

6.14 Hugging Face Pipeline (Transformers) Model

The third model implemented is the **Hugging Face Pipeline** which uses pre-trained transformer models such as Robertaor BERT for sentiment analysis. The ***pipeline function from Hugging Face*** was implemented for sentiment. This model uses the pre-trained distilbert-base-uncased-finetuned-sst-2-english model, fine-tuned for sentiment classification. This model produced two outputs: ***sentiment label (POSITIVE or NEGATIVE)*** and ***sentiment score***.

The function ***apply_transformers_sentiment*** was used to produce these labels and scores for each review in the dataset, which were subsequently visualized and analyzed.

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator \
116	117	B0026Y3YBK	A3P60QLFDDCHOY	Giordano "GB"	2
117	118	B0026Y3YBK	A38BUM0OXH38VK	singlewinder	0
183	184	B001KUUNP6	A262Z0S6PT9U16	Lee Thombley	3
375	376	B0087HW5E2	A139RTDNMU3WY5	blanket lady	2
394	395	B001ELL6O8	A13T2G4T8LR8XA	First Time Mom	2

	HelpfulnessDenominator	Rating	Time \
116	2	5	1304899200
117	0	5	1347667200
183	3	5	1292716800
375	2	5	1339977600
394	2	5	1189296000

	Summary \
116	Great cookies
117	Best everyday cookie!
183	Perfect for gluten-free chocolate chip cookies
375	Greatest Oil since slice bread !!!!!!!
394	Delisious Pancakes

	Text	vader_neg	vader_neu \
116	I'm Italian and I lived in Italy for years. I ...	0.0	0.196
117	In the 1980s I spent several summers in Italy....	0.0	0.308
183	We made chocolate chip cookies with BRM Garban...	0.0	0.575
375	I have used this oil for several years and it ...	0.0	0.482
394	Before I discover this mix on Amazon I always ...	0.0	1.000

	vader_pos	vader_compound	roberta_neg	roberta_neu	roberta_pos \
116	0.804	0.6249	0.004493	0.070814	0.924693
117	0.692	0.6696	0.001362	0.034527	0.964110
183	0.425	0.5719	0.001589	0.059519	0.938892
375	0.518	0.7482	0.013173	0.099402	0.887424
394	0.000	0.0000	0.004090	0.254064	0.741847

	roberta_compound	sentiment	sentiment_score
116	0.920200	POSITIVE	0.999866
117	0.962748	POSITIVE	0.999811
183	0.937303	POSITIVE	0.994876
375	0.874251	POSITIVE	0.999745
394	0.737757	NEGATIVE	0.994825

Table 6

6.15 Plot of sentiment distribution across ratings for the Huggingface Pipeline (Transformers) Model

The plots show the distribution of sentiment labels (POSITIVE and NEGATIVE) across different product ratings from the Huggingface Pipeline model:

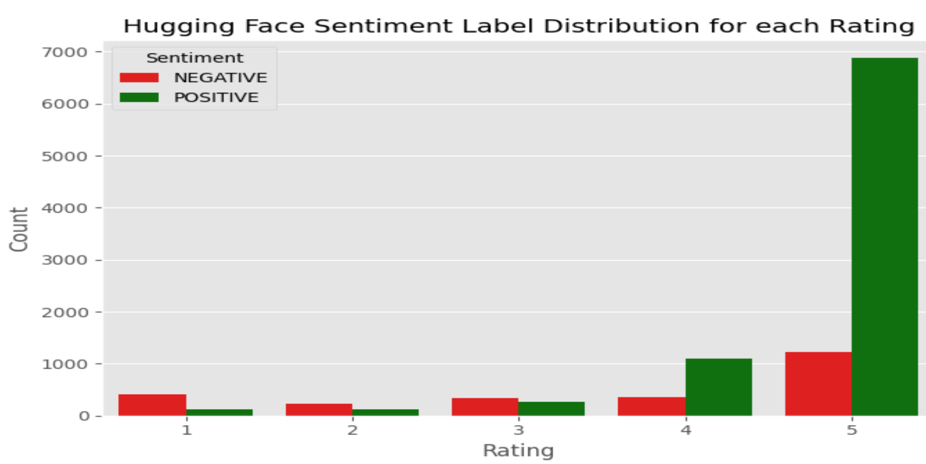


Fig 7

- With the increase in product ratings, the proportion of positive sentiment goes up.
- 5-star ratings are associated with positive sentiment. This shows strong customer satisfaction.
- Lower ratings (1-3 stars) have a more balanced sentiment distribution.

6.16 How is this information useful to address the problem statement?

This pattern suggests that the model can help Amazon identify high-performing products, which could be promoted or a small business like my mother's. Amazon

Amazon can optimize its marketplace by improving its product recommendations and ensuring that high-performing products are shown. This will directly affect customer satisfaction and drive both sales and long-term success.

The plot also highlights areas for improvement. By retrieving information about products with negative sentiment, Amazon and small businesses can adjust their product base to enhance customer satisfaction. Products with negative sentiment reviews can be refined to boost customer satisfaction.

6.17 Plot of Sentiment Analysis of Ratings for Huggingface Pipeline (Transformers) Sentiment

The plot shows the relationship between ratings and the following sentiment scores from the Huggingface Pipeline model

- negative ([sentiment_df['sentiment'] == 'POSITIVE')
- positive ([sentiment_df['sentiment'] == 'NEGATIVE')

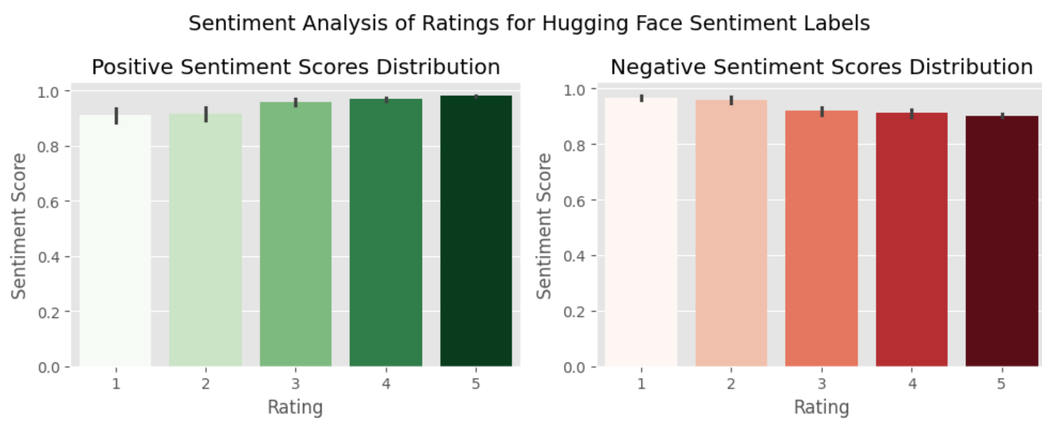


Fig 8

- **Positive Sentiment:** Higher ratings (4 and 5) show higher positive sentiment scores indicating customer satisfaction. Darker shades of green indicate more positive sentiment and lighter shades of green indicate lesser positive sentiment. As customer ratings increase, the average positive sentiment score also increases. This indicates that higher-rated reviews tend to have stronger positive sentiment. The gradual shift from light green to dark green further emphasizes this trend. The error bars are relatively small, suggesting that the average sentiment scores are consistent within each rating category.

- **Negative Sentiment:** Lower ratings (1 and 2) show higher negative sentiment scores, indicating customer dissatisfaction. Darker shades of red indicate more negative sentiment and lighter shades of red indicate lesser negative sentiment. As customer ratings increase, the average negative sentiment score slightly decreases. This indicates that higher-rated reviews tend to have slightly less negative sentiment. The shift from light pink to dark red is subtle. The error bars are again relatively small.

6.18 Word Cloud for Hugging Face Pipeline Model

The word clouds generated based on the **Huggingface sentiment scores** provide valuable insights into customer sentiment.

Positive Sentiment Word Cloud: Prominent words like "best", "love", "gluten free", "sugar free", "great cookie", "perfect" show positive reception towards taste and dietary preferences.

Negative Sentiment Word Cloud: In this case a lot of words which have positive sentiment are labelled as negative. This is a limitation of this model which was earlier stated in Section 5 (Methodology). This model has less control on fine-tuning and it may not handle some niche or domain-specific language well.

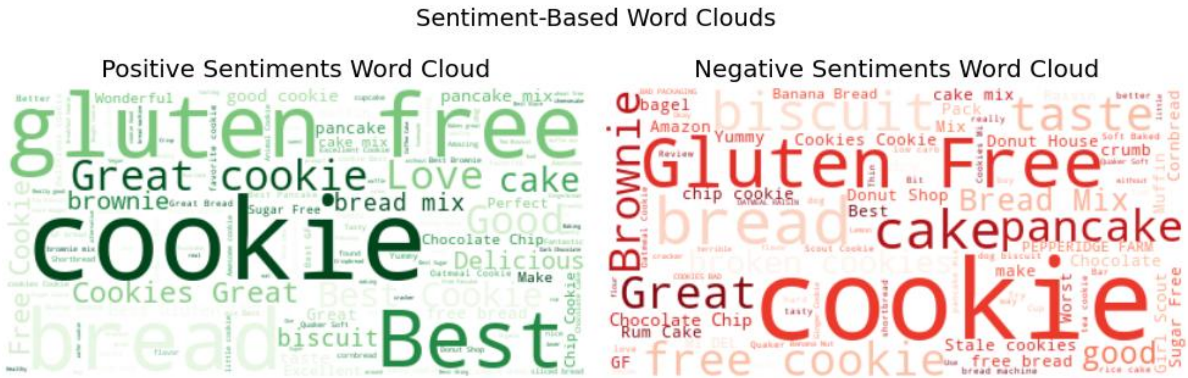


Fig:

However, Hugging Face Pipeline Model ***labels the sentiment of most words as positive*** as seen below in the summary statistics of sentiment score column and histogram plot of sentiment score distribution.

```
Positive Sentiments Count: 8467
Negative Sentiments Count: 2581

count    11048.000000
mean         0.964097
std          0.089119
min          0.500198
25%          0.984166
50%          0.998952
75%          0.999817
max          0.999892
Name: sentiment_score, dtype: float64
```

Table 7

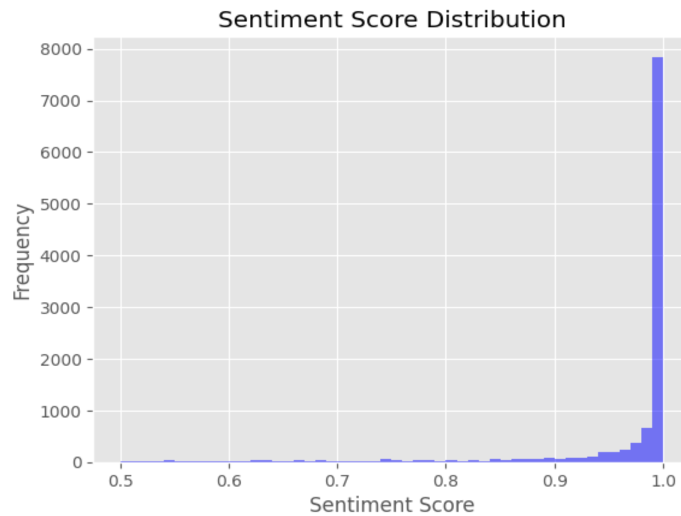


Fig 9

Does this affect business decisions?

Amazon can use the Hugging Face Pipeline Model to identify products with high customer satisfaction with an increased visibility for well-received items. But this may result in an overrepresentation of positive sentiment, overlapping areas for improvement in products with mixed reviews.

My mother, for her small business, can use this model to identify products that are generally perceived positively. This shall help to make data driven marketing and scaling decisions. However, the model's overemphasis on positive sentiment may overlook customer concerns, so addressing negative feedback and refining products effectively can be challenging.

6.19 Model Performance Evaluation and Comparison

6.19.1 Manually Assigning Actual Sentiment Scores

To evaluate the model, sentiment scores (true sentiment) were manually assigned to the ratings given and based on that later evaluate if the models have predicted the sentiment correctly. If the rating is more than 3 then it is a positive sentiment (1) else negative (0). In simpler words, we check to see if the model can classify a particular rating as positive/neutral/negative as appropriately as a human.

```
      Id   ProductId          UserId      ProfileName  HelpfulnessNumerator  \
116  117  B0026Y3YBK  A3P60QLFDDCHOY  Giordano "GB"                      2
117  118  B0026Y3YBK  A38BUM0OXH38VK    singlewinder                      0
183  184  B001KUUNP6  A262Z0S6PT9U16    Lee Thombley                      3

     HelpfulnessDenominator  Rating        Time  \
116                       2       5  1304899200
117                       0       5  1347667200
183                       3       5  1292716800

                                          Summary  \
116                                 Great cookies
117                          Best everyday cookie!
183  Perfect for gluten-free chocolate chip cookies

                                                 Text  ...  roberta_neu  \
116  I'm Italian and I lived in Italy for years. I ...  ...     0.070814
117  In the 1980s I spent several summers in Italy....  ...     0.034527
183  We made chocolate chip cookies with BRM Garban...  ...     0.059519

     roberta_pos  roberta_compound  sentiment  sentiment_score  \
116     0.924693          0.920200   POSITIVE         0.999866
117     0.964110          0.962748   POSITIVE         0.999811
183     0.938892          0.937303   POSITIVE         0.994876

     True_Sentiment  VADER_Prediction  RoBERTa_Prediction  \
116               1                 1                   1
117               1                 1                   1
183               1                 1                   1

    HuggingFace_Prediction  TS
116                      1   1
117                      1   1
183                      1   1

[3 rows x 25 columns]
```

Table 8

6.19.2 Performance Evaluation

The following columns are created using ***apply()*** and a ***lambda function.***

- **VADER_Prediction**: 1 for positive where vader_compound is greater than 0, 0 for negative where vader_compound less than or equal to 0.
- **Roberta_Prediction**: 1 for positive where Roberta_compound greater than 0, 0 for negative where Roberta_compound less than or equal to 0.
- **HuggingFace_Prediction**: 1 for positive where sentiment_score greater than 0.5, 0 for negative where sentiment_score < less than or equal to 0.5.

```
      Id   ProductId          UserId      ProfileName  HelpfulnessNumerator  \
116  117  B0026Y3YBK  A3P60QLFDDCHOY   Giordano "GB"                      2
117  118  B0026Y3YBK  A38BUM0OXH38VK     singlewinder                      0
183  184  B001KUUNP6  A262Z0S6PT9U16     Lee Thombley                      3
375  376  B0087HW5E2  A139RTDNMU3WY5     blanket lady                      2
394  395  B001ELL6O8  A13T2G4T8LR8XA  First Time Mom                      2

     HelpfulnessDenominator  Rating        Time  \
116                       2       5  1304899200
117                       0       5  1347667200
183                       3       5  1292716800
375                       2       5  1339977600
394                       2       5  1189296000

                                          Summary  \
116                                 Great cookies
117                          Best everyday cookie!
183  Perfect for gluten-free chocolate chip cookies
375           Greatest Oil since slice bread !!!!!!!
394                             Delisious Pancakes

                                                 Text  ...  roberta_neg  \
116  I'm Italian and I lived in Italy for years. I ...  ...     0.004493
117  In the 1980s I spent several summers in Italy....  ...     0.001362
183  We made chocolate chip cookies with BRM Garban...  ...     0.001589
375  I have used this oil for several years and it ...  ...     0.013173
394  Before I discover this mix on Amazon I always ...  ...     0.004090

     roberta_neu  roberta_pos  roberta_compound  sentiment  sentiment_score  \
116     0.070814     0.924693          0.920200   POSITIVE         0.999866
117     0.034527     0.964110          0.962748   POSITIVE         0.999811
183     0.059519     0.938892          0.937303   POSITIVE         0.994876
375     0.099402     0.887424          0.874251   POSITIVE         0.999745
394     0.254064     0.741847          0.737757   NEGATIVE         0.994825

     VADER_Prediction  RoBERTa_Prediction HuggingFace_Prediction  TS
116                 1                   1                      1   1
117                 1                   1                      1   1
183                 1                   1                      1   1
375                 1                   1                      1   1
394                 0                   1                      1   1

[5 rows x 24 columns]
```

Table 9

These predictions are necessary for evaluation metrics like accuracy, precision, recall, and F1 score to gauge the model performance. The continuous or probabilistic outputs (vader_compound, Roberta_compound, sentiment_score) are converted into binary predictions so that they can be compared with the true sentiment of ratings (assigned earlier).

6.20 Evaluation Metrics

Evaluation metrics measure how well each model classifies sentiment predictions compared to the manually assigned true sentiments:

- Accuracy
- Precision
- Recall
- F1 score

Model	Accuracy	Precision	Recall	F1 Score
VADER	69.09%	94.15%	68.55%	79.33%
RoBERTa	87.16%	92.45%	92.73%	92.59%
HuggingFace Pipeline	86.56%	86.56%	100.00%	92.80%

Table 10

6.21 Interpretation of results

6.21.1 Metrics for VADER

From the results, VADER has an accuracy of 69.09%, high precision (94.15%) but a much lower recall (68.55%). It correctly identifies positive sentiment in many cases. However, it misclassifies a significant number of positive instances which in turn lead to a low F1 score (79.33%).

6.21.2 Metrics for Roberta

From the results, Roberta has an accuracy of 87.16% , balanced precision (92.45%), balanced recall (92.73%) and a high F1 score of 92.59%. Roberta correctly identifies both positive and negative sentiments, making it a reliable model.

6.21.3 Metrics for HuggingFace Pipeline (Transformers)

From the results, HuggingFace Pipeline has an accuracy of 86.56%, lower precision (86.56%), perfect recall (100.00%). and an F1 score of 92.80%. It identifies all true positive sentiments in the dataset but often misclassifies negative sentiments as positive, resulting in false positives.

6.22 Confusion Matrix

The confusion matrix for each model shows the following:

- **True Positive (TP)**: Correctly predicted the positive class
- **False Negative (FN)**: Incorrectly predicted the negative class
- **False Positive (FP)**: Incorrectly predicted the positive class
- **True Negative (TN)**: Correctly predicted the negative class

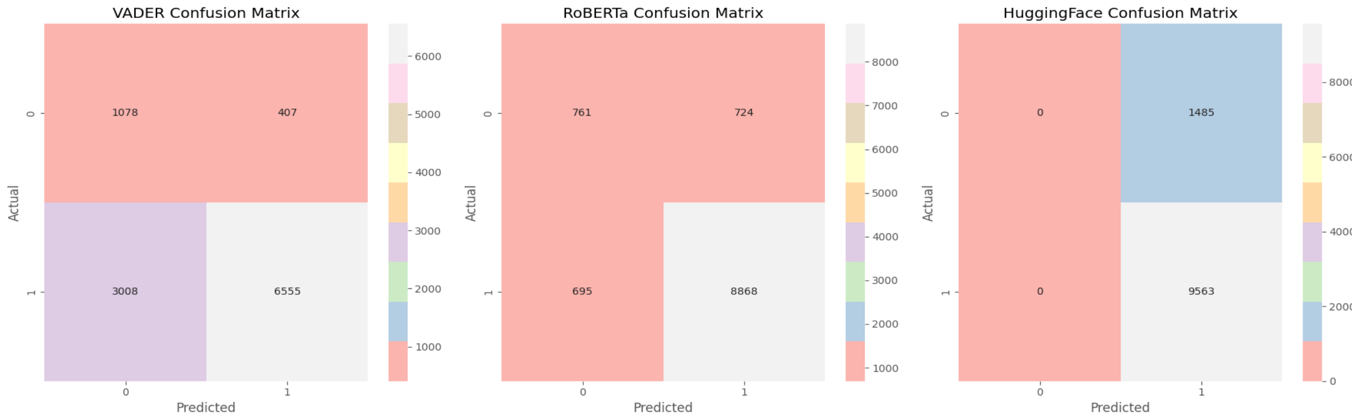


Fig 10

6.22.1 Which model is better?

- Again, Roberta is the most balanced and accurate model.
- HuggingFace Pipeline has perfect recall but the lowest precision.
- VADER is precise but less effective as it has lower recall.

Transformer-based models (Roberta and HuggingFace Pipeline) outperform the lexicon-based VADER model in sentiment analysis accuracy and F1 score.

For this data set, if it is more important to find every positive sentiment, even if some negative sentiments are incorrectly labeled, then the Huggingface Pipeline would be the best choice. If it is more important to have an overall balanced result, then Roberta is the better choice.

How do the findings relate to the problem statement?
Amazon can harness Roberta's balanced performance to distinguish consumer preferences accurately. which directly impacts market demand. Positive sentiment improves product visibility and negative sentiment indicates areas for improvement. These insights help to make data-driven pricing and promotional decisions. Amazon can refine its marketplace offerings to maximize sales by using the HuggingFace Pipeline, with perfect recall, which may guarantee no positive sentiment and focus on popular products.

For small businesses, models like Roberta can assist in understanding products that resonate most with customers, leading to more educated choices around promoting successful items and addressing those with negative feedback. Additionally, sentiment analysis can inform pricing strategies by highlighting the rate of consumer satisfaction. If she decides to sell on Amazon, she can use these insights to position her products better, improve visibility, make informed and strategic choices giving her a competitive advantage in the market.

6.23 Limitations of the methods and suggestions for improvement

Model	Limitations	Suggestions for Improvement
VADER	• Relies on predefined lexicons and rules • Limited understanding of context, sarcasm, and nuanced expressions	• Context-aware lexicons or a rule-based layer can be added. • This will account for sarcasm, irony, and mixed sentiments. • Domain-specific sentiment dictionaries can be incorporated.
	• Misinterprets complex or mixed sentiment sentences • Interprets uppercase words as more emotionally intense • Not always be accurate	• The model needs to be trained with bakery and confectionery-related data • This shall improve detection of food industry specific sentiments.
	• It doesn't understand industry-specific terms or jargon.	• VADER can be combined with other models like a simple classifier • This will help to handle more complex or ambiguous cases by leveraging VADER's efficiency for basic sentiment detection.
Roberta	• Roberta requires substantial resources • Less efficient compared to lighter models like VADER.	• Optimization techniques can be applied like pruning, quantization, or distillation (e.g., DistilRoberta) • This will reduce computational time while maintaining performance.
	• Requires large, high-quality labeled data, which might not be available for niche markets.	• Roberta can be fine-tuned with a high-quality, bakery industry-specific dataset • This will improve sentiment interpretation of food products.
	• Due to its size, Roberta may experience slower inference times and latency.	• To prevent overfitting applying regularization or cross validation can be beneficial • This will ensure generalization across different market conditions.
	• More prone to overfitting with insufficient data or improper fine-tuning.	
HuggingFace Pipeline	• HuggingFace models may have lower precision, leading to an overestimation of positive sentiment.	• Sentiment predictions can be refined using post-processing techniques. • Combining sentiment scores with additional features (e.g., product categories) will improve precision.
	• HuggingFace models require significant computational resources • Less feasible for smaller businesses or limited infrastructure.	• HuggingFace modelscan be fine tuned on bakery product-related reviews • This will improve sentiment interpretation and decrease misclassification
	• These models still struggle with sarcasm, idiomatic language, or highly ambiguous sentiment.	• Ensembling with other models like VADER or Roberta • This will balance the strengths (e.g., HuggingFace's recall with VADER's precision) for better predictions.
	• Dependency on fine-tuning to perform optimally in niche markets.	

Table 11

7. Conclusion

In this project, we explored the sentiment analysis of customer reviews of bakery and confectionary products to determine their influence on market strategies, product development, and pricing decisions for large e-commerce platforms like Amazon and small businesses in the food industry. The analysis highlights that consumer sentiment significantly affects market demand and pricing. While traditional sales data lacks qualitative insights, sentiment analysis can be instrumental in providing actionable data to optimize product demand and supply. Among all the models evaluated, Roberta was the most balanced and accurate, while HuggingFace Pipeline had a perfect recall.

VADER model was quick and ideal for short texts. But it struggles with complex sentences and domain-specific terms. On the other hand, state-of-the-art models like Roberta which is effective for capturing complex sentiments was computationally intensive. HuggingFace models had high recall but lower precision. HuggingFace suffered from having false positives in positive sentiment. Roberta stood out as the most accurate and efficient model to extract insights from customer reviews.

These findings prove to be useful for small businesses and larger e-commerce platforms, specifically in the food industry, to drive product improvements, optimize pricing strategies, and enhance market positioning in a competitive space.

On comparing the model performances, it was worth noting that sentiment analysis models emphasize the importance of balancing accuracy with cost-effectiveness which small businesses can leverage to align marketing strategies and improve their customer satisfaction.

However, VADER, Roberta, and HuggingFace Pipeline each had strengths and weaknesses. Limitations include potential inaccuracies in classifying ambiguous reviews and overlooking contextual subtleties like sarcasm.

Future work building on this approach could try to include additional data for fine tuning and training models. This will refine the model techniques and address review biases. One can also try to delve deeper into the negative feedback and process of manual review to collect insights to improve customer satisfaction.

8. References

Anvil Works. (n.d.). Using Hugging Face transformers on the Anvil server. Retrieved March 18, 2025, from https://anvil.works/forum/t/using-hugging-face-transformers-on-the-anvil-server/17264

Analytics Vidhya. (2022, October). Sentiment analysis using VADER. Retrieved March 18, 2025, from https://www.analyticsvidhya.com/blog/2022/10/sentiment-analysis-using-vader/

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148

GeeksforGeeks. (n.d.). How to create a pie chart in Seaborn? Retrieved March 18, 2025, from https://www.geeksforgeeks.org/how-to-create-a-pie-chart-in-seaborn/

Hugging Face. (n.d.). CardiffNLP/twitter-roberta-base-sentiment. Retrieved March 18, 2025, from https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

Kaggle. (n.d.). Amazon Fine Food Reviews dataset. Retrieved March 18, 2025, from https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews

Matplotlib. (n.d.). Matplotlib.pyplot.suptitle. Retrieved March 18, 2025, from https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.suptitle.html

Mulla, R. (2021). Sentiment analysis Python YouTube tutorial. Kaggle. Retrieved March 18, 2025, from https://www.kaggle.com/code/robikscube/sentiment-analysis-python-youtube-tutorial

PrettyTable. (n.d.). PrettyTable (version 3.15.1) [Python package]. PyPI. Retrieved March 21, 2025, from https://pypi.org/project/prettytable/

Scikit-learn. (n.d.). *TfidfVectorizer*. Retrieved March 18, 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Stanford SNAP. (n.d.). Web-FineFoods dataset. Retrieved March 18, 2025, from https://snap.stanford.edu/data/web-FineFoods.html

Stack Overflow. (2018, March 1). How to use tqdm with pandas in a Jupyter notebook? Retrieved March 18, 2025, from https://stackoverflow.com/questions/40476680/how-to-use-tqdm-with-pandas-in-a-jupyter-notebook

Stack Overflow. (2019, April 15). How to add extra stop words in addition to default stopwords in WordCloud? Retrieved March 18, 2025, from https://stackoverflow.com/questions/53997443/how-to-add-extra-stop-words-in-addition-to-default-stopwords-in-wordcloud