Sentiment Analysis of Food Product Reviews on Amazon: Predictive Modeling and Trend Forecasting

By

Claudia Dispinzeri

Instructor: Dr. Zhuojun Gu

BITM 609 (8865) - Topics in Business Analytics and Text Mining

University of Albany

Master of Science in Business Analytics

May 2025

ABSTRACT

TThis data-driven project explores two main business questions related to customer sentiment toward food products on Amazon. The first goal is to automatically predict whether customer reviews express positive or negative sentiment and to detect possible drops in perceived product value (as a proxy) over time. The second goal is to evaluate whether the available data is strong enough to create reliable sentiment-based forecasts that can support decisions in marketing and brand reputation. To address these questions, we applied Natural Language Processing (NLP) techniques, supervised classification models, and time series forecasting methods to turn unstructured text into useful insights. were applied as classification techniques of sentiment in the whole dataset. Additionally, we applied a sample of 50,000 records transformer as an additional technique of sentiment analysis. We also used time series models, including ARIMA, SARIMA, Holt-Winters, and Linear Regression, to analyze and forecast the trend of the negative review rate (NegativeRate). These combined methods help connect theory with practice, allowing us to identify patterns in customer behavior and anticipate important shifts in customer experience.

TABLE OF CONTENTS

INTRODUCTION

In this project, we explored data analysis and machine learning techniques to address two strategic problems related to customer sentiment and experience in food product reviews on Amazon. The dataset used, "Amazon Fine Food Reviews" is publicly available on the Kaggle platform (*https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews*) and contains over 500,000 customer reviews, including textual content, rating scores, timestamps, helpfulness votes, and product identifiers. This large volume of unstructured data represents a opportunity to generate insights that support decision-making in areas such as marketing, customer service, and product development.

The first business problem focuses on how to automatically predict the sentiment (positive or negative) expressed in customer reviews and identify potential future declines in customer perception. To accomplish this, we applied a Natural Language Processing (NLP) pipeline involving text cleaning, tokenization, TF-IDF vectorization, and sentiment classification using supervised models such as Logistic Regression, Decision Trees, Random Forest, Naive Bayes, and Neural Networks. The performance of these models was assessed using metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

We also applied a pre-trained Transformer-based model as a complementary technique to Business Problem 1, using a sample of 50,000 reviews, which represents approximately 9.51% of the full dataset. This technique was an additional initiative proposed by our team, inspired by concepts studied in another course within the master's program, where we explored advanced NLP methods. In this step, we analyzed not only the main review text but also the Summary field as an additional input for classification, aiming to assess whether this secondary textual component could enhance sentiment prediction performance. The sample was selected using a stratified random sampling approach to preserve the distribution of sentiment labels, ensuring that the results remained representative. Although the data originates from a time series, the goal of this analysis was to evaluate text classification performance independently of temporal order, making the use of a large static sample statistically appropriate.

The second business problem investigates whether the available data is sufficient to generate reliable forecasts that can support strategic decision-making. For this, we constructed a monthly time series of the proportion of negative reviews (NegativeRate), covering 109 consecutive months, ranging from October 2003 to October 2012. This period was selected based on data completeness, starting from the first month without missing values in the NegativeRate metric. Earlier months were excluded due to data gaps that could compromise the consistency of time series forecasting models. Although the period analyzed ends in 2012 and is not recent enough to predict current trends, the goal was to demonstrate that the dataset has the potential to generate reliable forecasts, as long as the information is monitored continuously. We applied four forecasting methods: ARIMA, SARIMA, Holt-Winters, and Linear Regression (with dummy variables for month and time), to predict customer sentiment trends for the next 12 months. The accuracy of the models was evaluated using MAE, RMSE, and MAPE, both on the test set and over the most recent 12 months of historical data.

In addition, data preparation involved several key steps to ensure analytical quality and consistency. We began by cleaning the review texts, removing punctuation, stopwords, and irrelevant terms, and applying normalization techniques such as lowercasing. To support a binary classification approach, we excluded reviews with a score of 3, which are typically considered neutral. This allowed for a clearer distinction between positive sentiment (scores above 3) and negative sentiment (scores below 3), and based on this transformation, we created a new column called "Sentiment", where positive reviews were labeled as 1 and negative ones as 0. Although this exclusion may introduce some bias by removing neutral perspectives, our focus was to train a binary model with a clear separation between positive and negative sentiment. Future iterations could explore multi-class or ordinal classification to incorporate more nuanced sentiment levels.

During this preparation stage, we also took care to avoid data leakage. To ensure this, the TF-IDF vectorizer was fit exclusively on the training data so that the test data did not influence the learned parameters. After fitting, the test data was only transformed based on the trained vectorizer, preserving the integrity of the evaluation and the model's generalization ability.

Given the class imbalance in the dataset, approximately 85% positive and 15% negative, we adopted different balancing strategies depending on the model. For algorithms like Logistic Regression, Decision Trees, and Random Forest, we used the parameter class_weight = 'balanced', which automatically adjusts class weights based on data distribution to reduce classification bias. For the Neural Network (MLP), we applied the SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic examples of the minority class to balance the training data. For the Naive Bayes model, we manually calculated the class proportions in the training set and used them as priors, allowing the algorithm to properly reflect class imbalance during training.

This report is organized to present the methodology applied, the main visualizations produced, and the strategic recommendations derived from our analyses. By combining sentiment classification with time series forecasting, the project offers a clear view of how customer perception (as a proxy) evolves, supporting Amazon's customer-centric approach. The final results, including the most effective models for each analytical task, are detailed in the following sections.

A. DATASET USED FOR THE FINAL PROJECT

Regarding the dataset structure, we have:

Original columns:

1. **Id, ProductId, UserId, ProfileName:** identifiers for the reviews, products, and users;
2. **HelpfulnessNumerator, HelpfulnessDenominator:** number of votes indicating whether the review was helpful;
3. **Score:** rating assigned to the product (on a scale from 1 to 5);
4. **Time:** review date in timestamp format;
5. **Summary:** brief summary of the review;
6. **Text:** full text of the review.

During preprocessing and analysis, new columns were created in the main dataset called data:

7. **Sentiment:** a binary classification of sentiment based on the Score, where ratings of 4 and 5 were considered positive, and ratings of 1 and 2 were considered negative. Score 3 reviews were removed from the classification task to improve model clarity and polarity separation, as they often represent ambiguous or mixed feedback. While this may introduce some bias by excluding neutral opinions, our focus was to train a binary model with a clear distinction between positive and negative sentiment. Future iterations could explore multi-class or ordinal classification to incorporate more nuanced sentiment levels;
8. **Date:** conversion of the timestamp into a readable date format;
9. **Cleaned_Text:** a cleaned version of the review text, with punctuation, stopwords, and irrelevant elements removed.

To explore an alternative sentiment analysis approach based on pre-trained neural language models, we selected a random sample of 50,000 records (9.51% of the full dataset), referred to as the sample dataset. This strategy was adopted due to the high computational cost associated with transformer-based methods, specifically the use of the "distilbert-base-uncased-finetuned-sst-2-english" model available through Hugging Face. The technique was originally introduced in another course within the master's program, focused on Natural Language Processing (NLP), and we chose to apply it here as an educational extension to reinforce our learning and expand the comparison across modeling strategies.

Through this sample, additional columns were created for comparison purposes:

10. **Cleaned_Summary:** a cleaned version of the review summary;
11. **Text_Label, Cleaned_Label:** sentiment classifications generated by the model using the original and cleaned review texts, respectively;
12. **Summary_Label, Cleaned_Summary_Label:** sentiment classifications based on the original and cleaned summaries.

There are two Python files attached to this report:

- **FinalProject_TeamClaudiaDispinzeri.ipynb =** the main script where we applied 98% of all the techniques.
- **FinalProject_Ref_AutoArima_TeamClaudiaDispinzeri.ipynb =** a script focused specifically on Auto ARIMA, which we used to model the values detailed in section 4.2.
- **Reviews.csv** = dataset from Amazon, downloaded by Kaggle.
- **negative_rate_series.csv** = file we exported from FinalProject_TeamClaudiaDispinzeri.ipynb during code execution, and later read in FinalProject_Ref_AutoArima_TeamClaudiaDispinzeri.ipynb to analyze Auto ARIMA.

The goal was to compare the consistency of the sentiment predicted by these model-based approaches with the sentiment derived from review scores (Sentiment). This comparison allowed us to evaluate the effectiveness of different text sources (full reviews vs. summaries), the impact of text cleaning on model performance, and the feasibility of scaling this alternative technique to larger datasets in the future.

B. REPORT FOR DIRECTORS

1. Introduction

This report presents the results and insights derived from our comprehensive analysis of product reviews data, conducted to address key business questions related to customer sentiment and perception. Our primary goal was to explore how Natural Language Processing (NLP) techniques, combined with traditional scoring data, can be leveraged to enhance decision-making in areas such as customer experience, product feedback, and brand reputation.

Using machine learning models and text-based classification techniques, we aimed to uncover patterns in customer sentiment that go beyond numerical ratings. These insights can enable the organization to monitor shifts in consumer perception, prioritize improvements in product offerings, and optimize customer communication strategies.

Throughout this project, we systematically processed, analyzed, and modeled customer reviews to validate both conventional and advanced sentiment detection approaches. The following sections outline our objectives, methodology, key findings, and recommendations for leveraging sentiment analysis as a strategic business tool.

2. Business Problems

Our work was guided by critical questions that reflect current challenges in understanding and anticipating customer sentiment across large volumes of user-generated content. In the competitive landscape of e-commerce, the ability to identify signs of dissatisfaction early, especially through textual reviews, can be a valuable differentiator for product and customer service teams.

To address these challenges, we structured the project around two core business questions, each designed to provide practical insights and scalable strategies:

- **How can we automatically classify customer sentiment from review text and use this information as a proxy to anticipate potential future declines in perceived value?**
- **Is the available data sufficient to generate reliable sentiment-based forecasts that support strategic decisions in areas such as marketing and brand reputation?**

In this analysis, customer sentiment was used as a proxy for perceived value, under the assumption that shifts in emotional tone expressed in reviews reflect broader changes in how customers experience and evaluate the product.

Given the high computational cost of language models, this first question examines whether preprocessing or input simplification (e.g., using summaries instead of full reviews) can preserve classification accuracy while improving processing efficiency.

These business questions were developed to address both immediate analytical needs and long-term strategic applications. The report evaluates the quality and consistency of various sentiment analysis strategies and explores their potential integration into operational workflows for ongoing customer feedback monitoring.

Although the dataset we used spans from 2003 to 2012 and does not reflect current customer behavior, our goal was not to make real-time predictions, but rather to test the hypothesis that textual review data, when consistently monitored, can provide meaningful insights and reliable forecasts. The project demonstrates that, even with historical data, the proposed approaches are capable of supporting strategic business questions and can be applied to more recent datasets with regular updates.

Each business problem is presented with a focus on two key aspects:

1. **Methodology** – a technical explanation of the approach used to address the problem and how the results were achieved;
2. **Results, Conclusions, and Recommendations** – the key insights, summarized to support decision-making at the leadership level.

3. Answering the Business Problem 1: Sentiment Classification & Trend Analysis

3.1. Methodology and Techniques

To address the first business question "**How can we automatically classify customer sentiment from review text and use this information as a proxy to anticipate potential future declines in perceived value?**", our team implemented a structured and thoughtful approach that combined text analysis with predictive modeling techniques. The primary goal was to automatically classify customer sentiment based on written product reviews and use this information to identify early signals of declining perceived value.

We began by exploring the dataset and simplifying the sentiment into two categories: positive and negative. Neutral reviews were excluded to ensure more meaningful interpretation of patterns. We then prepared the review text through a process known as text cleaning, removing irrelevant characters, simplifying words to their core meaning, and filtering out common terms that don't contribute to sentiment detection. This step ensured that our analysis focused on content that truly reflected the customer's emotional tone.

To make the textual data usable by algorithms, we applied a technique called TF-IDF, which converts each review into a weighted list of words based on how relevant each term is within the context of the entire dataset. This transformation allowed us to turn thousands of unstructured reviews into structured numerical data that could be analyzed computationally.

Using this structured data, we trained several predictive models to classify the sentiment behind each review. These included logistic regression, which predicts the probability of a review being positive or negative; decision trees and random forests, which simulate decision-making pathways; naive Bayes, a probabilistic model well-suited for text; and a neural network model capable of detecting complex patterns. Each model's performance was measured using metrics such as accuracy, precision, recall, and F1-score, that are common metrics used to help us understand their strengths and trade-offs.

Because the dataset contained far more positive reviews than negative ones, we used a balancing method called SMOTE. This technique generates synthetic examples of the underrepresented class, improving the models' ability to detect negative sentiment without bias.

In addition to our primary modeling pipeline, we explored a complementary approach using a pre-trained Transformer model, a cutting-edge tool in natural language understanding. Applied to a random sample of 50,000 reviews, this model served two purposes. First, it worked as a benchmark, helping us validate the effectiveness of our own models. Second, it allowed us to test the potential of using the "Summary" field, which contains a shorter version of each review, as an alternative input. The results were promising: while summaries are more concise, they often capture the core of the customer's message. We found that analyzing summaries required significantly less processing time and still delivered reliable results, suggesting that this could be a viable and efficient option for future applications, particularly in real-time scenarios.

In summary, our methodology combined rigorous analysis with practical innovation. It enabled us not only to classify customer sentiment with strong predictive accuracy, but also to explore smarter and more scalable ways to monitor perception and detect early warning signs that can drive proactive decision-making.

3.2. Results and Recommendations

The results obtained from the sentiment classification models confirmed the analytical pipeline's effectiveness and the preprocessing stages' consistency. Among the models tested, the Neural Network (MLPClassifier) delivered the best overall performance, with 94.9% accuracy, 96.8% precision, and 97.1% recall, resulting in an F1-score of 96.9% and an AUC of 96.5%. This balance highlights the model's ability to accurately classify both positive and negative sentiments, making it ideal for supporting customer engagement strategies.

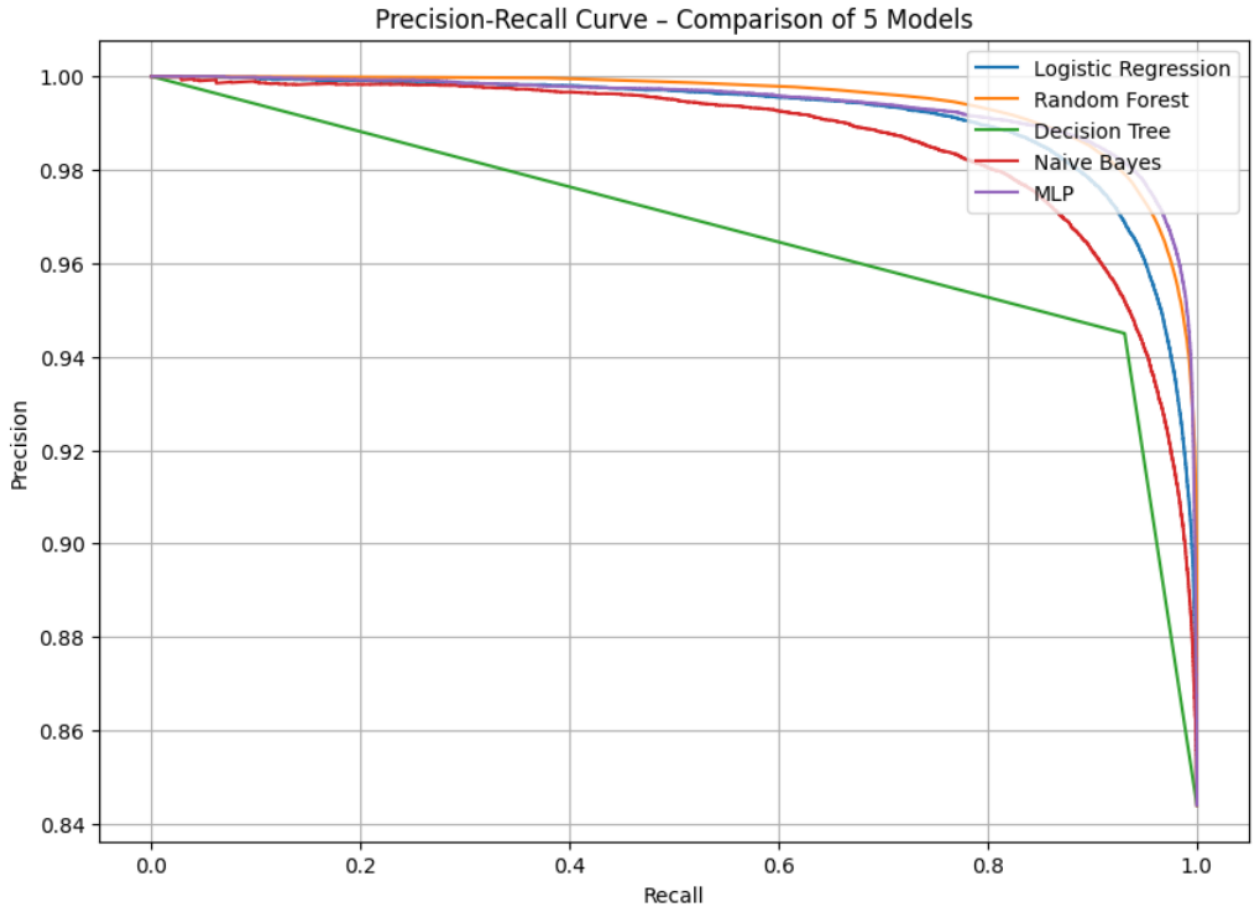
	Accuracy	Precision	Recall	F1 Score	ROC AUC	True Negatives	False Positives	False Negatives	True Positives
Logistic Regression	0.894316	0.978739	0.894211	0.934567	0.958637	14677.0	1724.0	9389.0	79363.0
Decision Tree	0.895952	0.944995	0.930909	0.937899	0.818839	11592.0	4809.0	6132.0	82620.0
Neural Network	0.948998	0.968420	0.971246	0.969831	0.971980	13590.0	2811.0	2552.0	86200.0
Naive Bayes	0.880973	0.879359	0.995561	0.933859	0.936676	4279.0	12122.0	394.0	88358.0
Random Forest	0.939117	0.938508	0.992924	0.964949	0.973678	10627.0	5774.0	628.0	88124.0

The Random Forest model also produced strong results, with 93.9% accuracy, 99.3% recall, and an AUC of 95.9%, indicating high sensitivity to negative reviews. However, its slightly lower precision (91.7%) compared to the neural network suggests a higher rate of false positives. Logistic Regression also performed competitively, with 97.8% precision, an F1-score of 92.9%, and a leading AUC of 95.8%, making it a reliable choice for binary classification with lower computational complexity.

Although Naive Bayes achieved the highest recall (99.5%) and an AUC of 88.1%, its lower precision (87.9%) and weaker true negative detection may lead to misleading interpretations of customer sentiment, potentially impacting strategic decisions based on an inflated perception of dissatisfaction.

The Decision Tree model, while interpretable, showed the lowest AUC among the tested models (84.7%), reflecting limited generalization capability under class imbalance.

The precision-recall curve, which measures the trade-off between precision (how many predicted positives are correct) and recall (how many actual positives are captured), further reinforced these findings, as shown in the figure below. The Neural Network and Random Forest models maintained curves that stayed close to the top-right corner of the plot, indicating consistently high precision even as recall increased. This suggests strong robustness and reliability in identifying negative sentiment. In contrast, the Naive Bayes and Decision Tree models exhibited sharp declines in precision as recall rose, indicating reduced stability and a higher likelihood of false positives under more permissive thresholds.



To complement the sentiment classification evaluation, we experimented using a random sample of 50,000 reviews and applied a pre-trained Transformer model to four input variations: full review text, cleaned full text, summary, and cleaned summary. While the original review text produced results that were relatively aligned with the score-based sentiment distribution (approximately 15.4% negative), the cleaned full-text version showed a substantially higher negative sentiment rate of 50.31%. This difference suggests that common preprocessing steps, such as removing punctuation, stopwords, and converting text to lowercase, may eliminate important linguistic cues that the Transformer model relies on, since it was trained on raw and unaltered language. In contrast, the cleaned summaries yielded a 22.25% negative rate, which is much closer to the expected distribution and was achieved with significantly lower computational cost. These findings indicate that cleaned summaries may serve as an efficient and practical solution for real-time or resource-constrained applications, while also highlighting the importance of aligning preprocessing strategies with the architecture and assumptions of the model being used.

	By Score	By Text	By Cleaned Text	By Summary	By Cleaned Summary
Negative	15.4	27.63	50.31	24.36	22.25
Positive	84.6	72.37	49.69	75.64	77.75

The word cloud visualizations below also revealed insights. In negative reviews, words like “product”, “flavor”, and “taste” were among the most frequent, reflecting dissatisfaction related to sensory and functional experiences. In contrast, positive reviews featured terms like “recommend,” “great,” and “store,” highlighting appreciation for product quality, convenience, and retail trust.

Based on these findings, our business analytics team proposes the following strategic recommendations to answer business problem 1:

- We recommend adopting Neural Networks or Random Forest models for operational deployment, as they demonstrated high accuracy and consistent performance even under challenging conditions. These models are well-suited for real-time sentiment monitoring and can be integrated into intelligent customer service systems or dynamic campaign evaluation tools.
- To enable fast and resource-efficient sentiment analysis, we suggest leveraging the Review Summary field, especially in its cleaned version. This approach delivers results comparable to full-text models while reducing processing time and computational cost, making it a practical solution for agile or resource-constrained environments.
- Sentiment classification outputs should be integrated into company dashboards to enable continuous trend monitoring and early detection of shifts in customer perception. This supports more timely and informed adjustments in marketing campaigns, product messaging, and communication strategies.
- Marketing and product teams should use these insights to prioritize improvements based on frequently cited negative themes, such as “taste” and “use,” while reinforcing positive attributes like “recommend” and “great” that drive customer satisfaction and loyalty.

- We also recommend continuous monitoring of customer reactions following major events such as product launches, pricing changes, or promotional campaigns. By applying classification models to review summaries in near real time, it is possible to detect early signs of dissatisfaction and act before brand sentiment is negatively impacted.
- Sentiment data can further support more personalized CRM and remarketing strategies. Classifying customers by emotional tone enables differentiated communication, such as proactive recovery messages for dissatisfied users and loyalty campaigns for satisfied ones. This builds stronger relationships and reduces churn.
- Aggregated sentiment insights should inform campaign planning and positioning. Analyzing the most frequent expressions in positive and negative reviews can guide the tone of marketing messages, improve alignment with customer expectations, and avoid language that previously triggered negative reactions.

4. Answering Business Problem 2: Sentiment Forecasting for Marketing and Brand Strategy

4.1. Methodology and Techniques

To answer the second business question "**Is the available data sufficient to generate reliable sentiment-based forecasts that support strategic decisions in areas such as marketing and brand reputation?"** our data analytics team developed a time series analysis focused on the monthly trend of negative sentiment in product reviews.

We started by transforming the unstructured text data into a structured time series. To do this, we grouped the sentiment classification results by month and calculated the proportion of negative reviews out of the total reviews published each month. During this process, we found some months with missing values (NaNs), which were removed to ensure data quality. After cleaning the data, we had a consistent and reliable time series with 109 monthly records, starting in October 2003 and ending in October 2012.

Before applying any forecasting models, our team conducted an exploratory analysis to better understand the behavior of the data. First, we used a time series decomposition method to break the series into three main components: trend, which shows the long-term direction; seasonality, which captures recurring patterns across the years; and residual (noise), which represents the random variations not explained by the other components. This helped us choose the most appropriate forecasting models based on the structure of the data.

As part of this exploration, we also used three simple but powerful visual trend analysis techniques. We plotted a linear trend line to observe the overall direction of the data over time. We then applied a six-month moving average to smooth short-term fluctuations and highlight longer-term shifts in sentiment. Finally, we calculated the month-to-month percentage change, which allowed us to spot periods of unusual variation, often linked to external or seasonal events. These steps gave us a clear picture of how customer sentiment evolved over time and helped us prepare for the modeling stage.

We then moved into the modeling phase, which was done in two key steps for each forecasting method. In the first step, we split the time series into two parts: a training set (about 70% of the earliest months) and a test set (the remaining 30%). Each model was trained using only the training set and then tested to see how well it could

predict the most recent months. This approach simulates how real-world forecasting works when future data is still unknown.

In the second step, we retrained each model using the full 109-month dataset and generated forecasts for the next 12 months. Even though these forecasted months are now in the past, this step helped us evaluate how each model performs in projecting future trends. It also allowed us to generate confidence intervals around each forecast point. Narrow intervals indicate higher prediction confidence, while wider ones show greater uncertainty, something decision-makers should keep in mind when interpreting the results.

We applied four forecasting techniques, each with different strengths. The ARIMA model uses past values and errors to forecast future data points. The SARIMA model extends ARIMA by including seasonal effects, which was useful given the yearly cycles we observed. The Holt-Winters method applies exponential smoothing to adapt to trend and seasonality in real time. Finally, we used a linear regression model that includes monthly patterns (called dummy variables) to represent both time trends and seasonal effects in a simple and easy-to-understand format.

One important point we would like to highlight is that, before applying the ARIMA and SARIMA models, we tested whether the time series was stationary, meaning it had a stable average and variance over time, a key requirement for these forecasting methods. To do this, we used two widely accepted statistical tests: the Augmented Dickey-Fuller (ADF) test and the KPSS test. While the ADF test checks whether the series is stationary, the KPSS test evaluates the opposite, whether the series is non-stationary. Using both tests gave us a more complete understanding of the data's behavior.

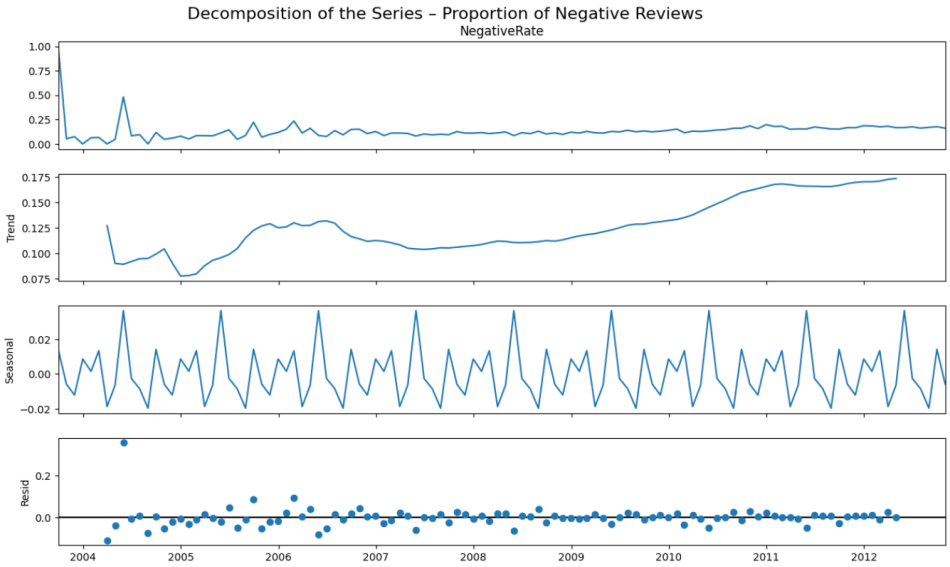
As part of this analysis, we also applied a technique called differencing, which transforms the data by removing trends or seasonal patterns until a stable structure is achieved. This process was repeated iteratively until the statistical conditions for stationarity were met, ensuring that our forecasting models were built on consistent data and capable of delivering more reliable results.

To evaluate how well each model performed, we used three common error metrics. The first was MAE (Mean Absolute Error), which shows the average error in the predictions. The second was RMSE (Root Mean Square Error), which gives more weight to larger errors. The third was MAPE (Mean Absolute Percentage Error), which expresses the error as a percentage, making it easier to compare across models. These metrics were used to assess both the accuracy during the test period and the reliability of the full 12-month forecasts.

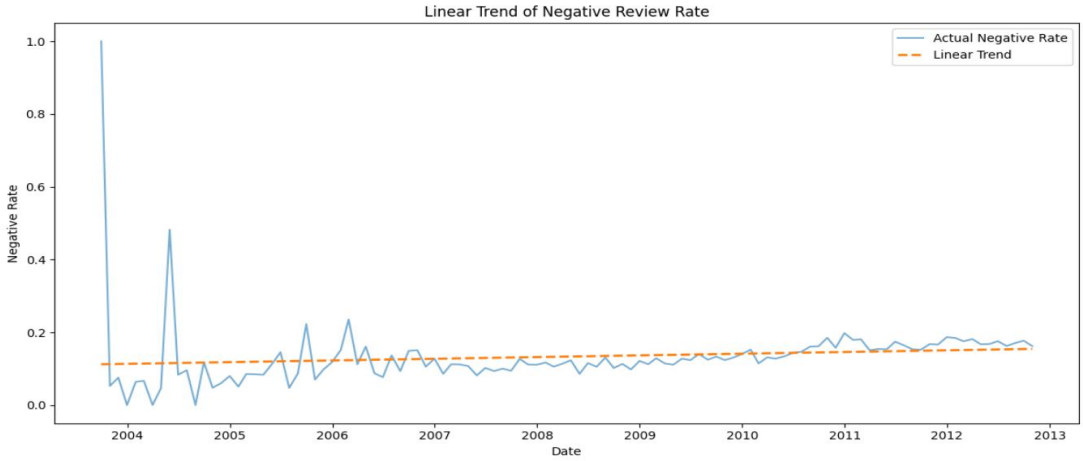
In summary, our analytics team used a combination of data cleaning, trend exploration, seasonal decomposition, statistical forecasting models, and error evaluation techniques to assess whether the data can support reliable sentiment-based predictions. The results from this process will be presented in the next session.

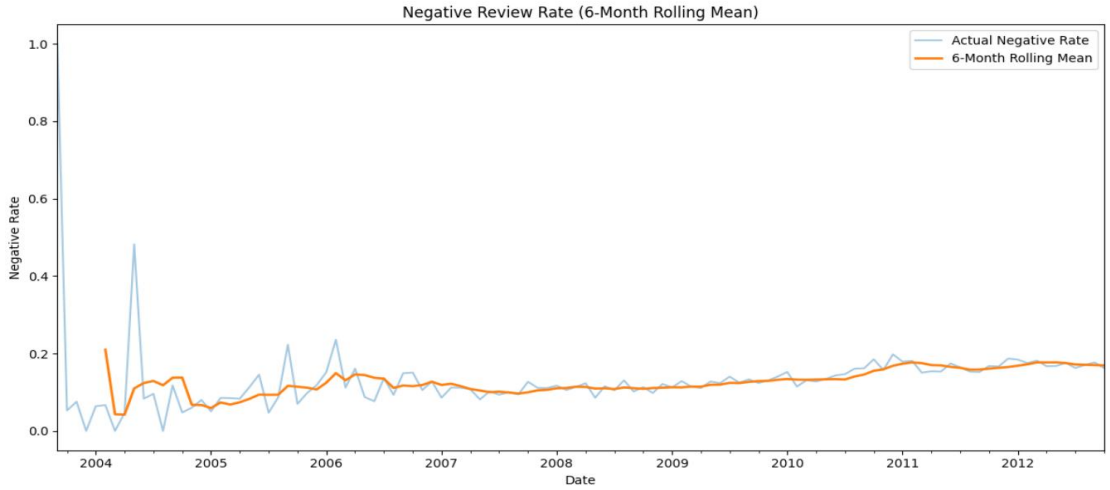
4.2. Results and Recommendations

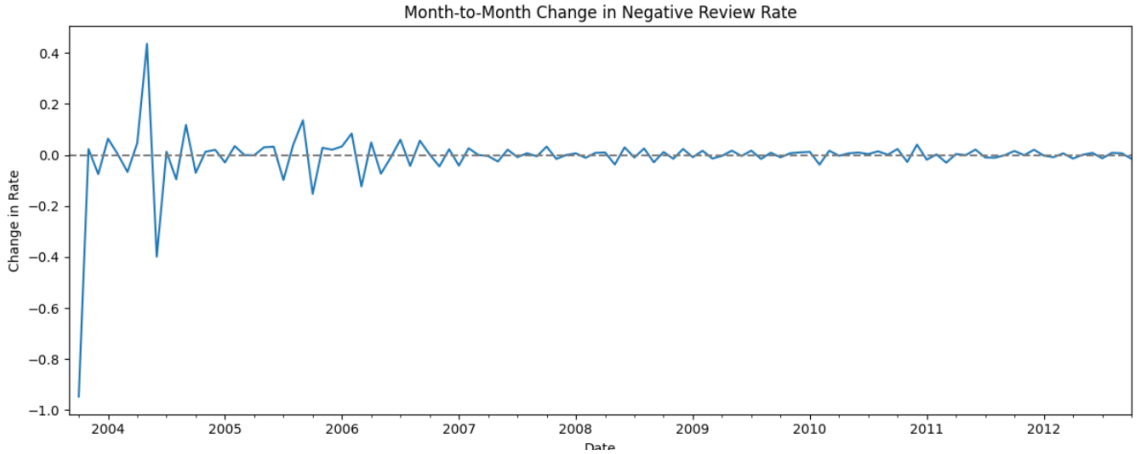
Our analytics team identified a clear upward trend in the proportion of negative reviews over time, particularly after 2009, as shown in the image below. While the variation occurred on a small scale, the pattern is consistent and suggests a gradual decline in how customers perceive their experiences. We also observed well-defined seasonal behavior, with recurring peaks in specific months, which may reflect common consumer dynamics during promotional periods or commercial events. The residual (noise) component remained relatively stable, with no significant random fluctuations that could compromise the predictability of the series.



The three trend analysis techniques confirmed this pattern. The linear trend line showed a steady increase in negative reviews. The six-month moving average highlighted consistent shifts in customer sentiment, with fewer sharp changes in recent years. The month-to-month percentage change revealed moments of instability, possibly linked to external events or operational issues.







Before presenting the results of forecasting, it is important to clarify the structure of the ARIMA and SARIMA models used and explain the rationale behind our model selection process.

Models like ARIMA and SARIMA are widely used to understand the behavior of time series and to make forecasts based on historical data. These models help identify patterns over time and project future trends, which is essential for supporting strategic decision-making.

The ARIMA model works by combining three key elements: the past values of the series itself (such as last month's sales), the changes between periods (used to remove growth or decline trends), and the errors from previous forecasts (used to adjust future predictions). This combination is represented by the notation (p, d, q), where:

- “p” refers to how many past values are taken into account,
- “d” shows how many times the data had to be adjusted to remove trends,
- “q” refers to how many past forecast errors are used to refine the next prediction.

When the data shows seasonality, meaning recurring patterns, like increased sales every December, a more complete model is needed: SARIMA. In addition to everything ARIMA includes, SARIMA adds parameters to capture those seasonal patterns. The notation becomes (p, d, q)(P, D, Q, s), where the seasonal terms follow the

same logic, and s represents the length of the seasonal cycle, for example, s = 12 for monthly data with yearly repetition.

During modeling, it was necessary to make some simple adjustments to the data to make it more stable over time. These adjustments, called *differencing*, were applied to remove long-term trends (using d = 1) and to stabilize seasonal patterns (with D = 1 and s = 12, which reflects yearly seasonality). These values were chosen based on exploratory analysis of the data and are common when working with economic or operational time series.

After that, different model configurations were tested, varying the number of past periods included and the level of complexity, part automatic, which we called auto Arima, and part manual, as we noticed which one performed better. The most effective models were selected based on three main criteria:

1. AIC: a statistical score that compares models and penalizes overly complex ones that don't improve performance;
2. Residuals – the forecast errors were analyzed to ensure the model did not leave important patterns unexplained. To do this, we applied the Ljung-Box test, which is a test that checks whether the residuals show significant autocorrelation. A high p-value in this test indicates that the residuals behave like random noise, suggesting that the model successfully captured the structure of the time series;
3. Forecast accuracy: metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were calculated by comparing the forecasts to actual values in a test period.

Although the auto Arima function was used as a helpful tool to suggest initial models automatically, it has limitations. The algorithm only tests a limited number of combinations, favors simpler models that run faster, and relies solely on AIC to make decisions. As a result, it may overlook more complex models that produce better forecasts. For this reason, we went beyond the automatic suggestions and manually tested additional model combinations, which allowed us to compare more robust options based on their real forecasting performance. Look at the figure below the results:

```
Best model:  ARIMA(0,1,4)(0,0,0)[0] intercept
Total fit time: 5.280 seconds
Best ARIMA model:


Best model:  ARIMA(2,1,3)(0,1,0)[12]
Total fit time: 98.003 seconds
Best SARIMA model:
```

	AIC	MAE	MSE	RMSE	Ljung-Box p-value
ARIMA(0,1,4) [auto_arima]	-190.189321	0.046910	0.002498	0.049980	0.008063
ARIMA(1,1,1) [manual]	-200.526056	0.045042	0.002353	0.048508	0.021486
SARIMA(2,1,3)(0,1,0,12) [auto_sarima]	-135.675459	0.028918	0.001041	0.032262	0.052163
SARIMA(2,1,2)(1,1,1,12) [manual]	-181.723202	0.016572	0.000533	0.023079	0.131858

Based on this comprehensive approach, our analyst team could conclude that the SARIMA(2,1,2)(1,1,1,12) model delivered the best results among the seasonal models, providing highly accurate forecasts while also producing the most stable and statistically reliable prediction errors. The ARIMA(1,1,1) model was also the second most

effective (the first among the non-seasonal models) as it offered the best balance between simplicity and performance.

```
============================================================
Model Performance on Test Set (70/30 Split)
============================================================
```

	MAE	RMSE	MAPE (%)
ARIMA	0.04	0.05	25.74
SARIMA	0.02	0.02	9.75
Holt-Winters	0.07	0.08	43.46
Linear Regression	0.07	0.08	43.41

```
============================================================
Recent Model Performance – Last 12 Months Before Forecast
============================================================
```

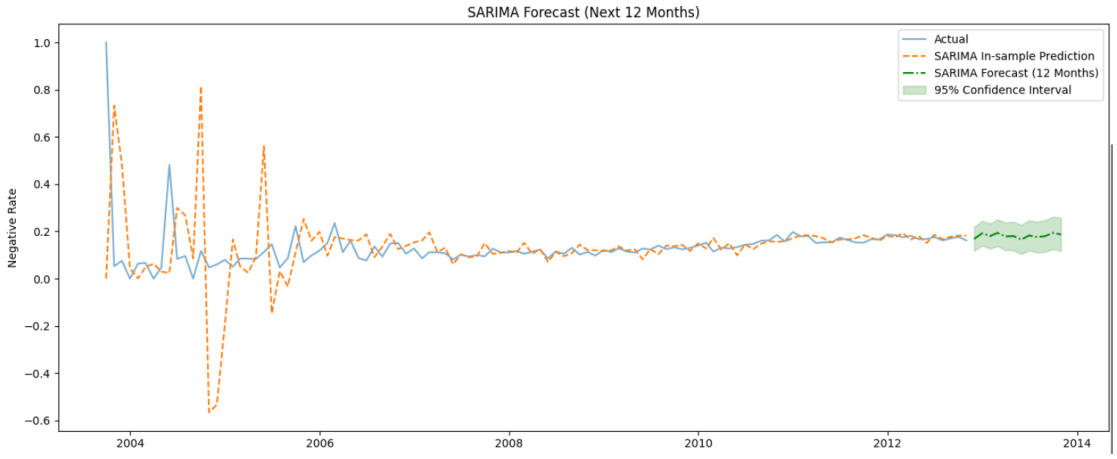
	MAE	RMSE	MAPE (%)
ARIMA	0.02	0.02	11.16
SARIMA	0.01	0.01	6.22
Holt-Winters	0.02	0.04	13.82
Linear Regression	0.04	0.04	20.39

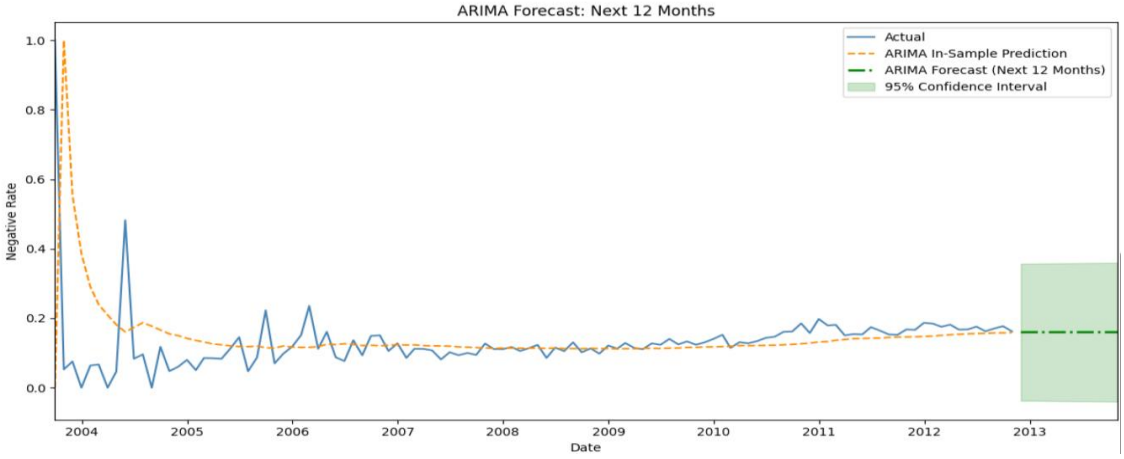
As we can observe in the figure below, among all the models tested, SARIMA and ARIMA stood out for their superior accuracy and consistency. SARIMA demonstrated excellent performance, with a MAPE of 9.75% during validation and an even better 6,22% in the 12-month forecast, confirming its ability to effectively capture seasonal dynamics. While ARIMA does not explicitly model seasonality, it still produced reliable results, reaching 25.74% MAPE in the test set and 11.16% in the forecast period.

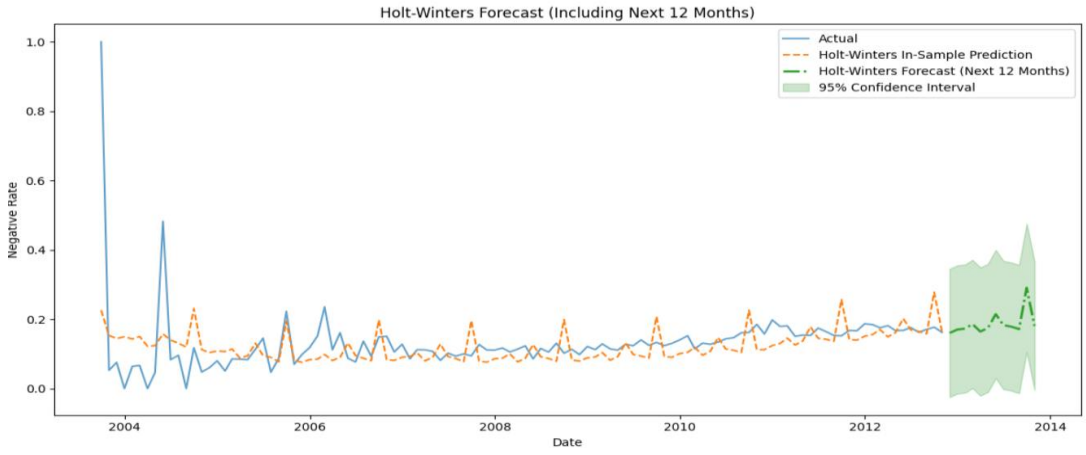
In contrast, Holt-Winters and Linear Regression with monthly dummies showed much higher error rates during validation, with MAPE values exceeding 40%, likely due to their limited ability to adapt to the underlying seasonal and trend structures in the data. Although both models improved somewhat in the forecast phase (with MAPE decreasing to 13.8% for Holt-Winters and 20.4% for Linear Regression), they continued to exhibit wider confidence intervals and greater volatility.

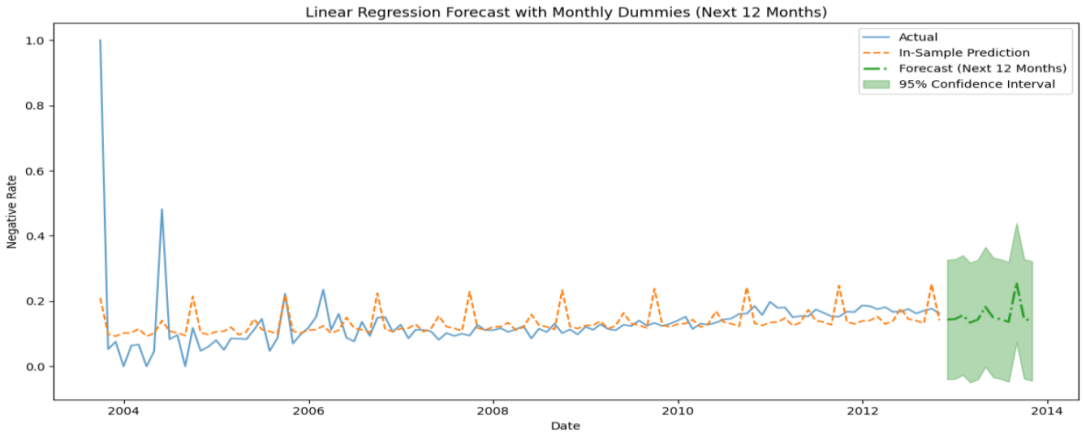
Because of this, our team believes that SARIMA and ARIMA are more suitable for applications where forecast stability and precision are critical, especially in guiding long-term strategic decisions.

The 95% confidence intervals reinforced these conclusions. SARIMA and ARIMA generated tighter intervals, suggesting more dependable forecasts. Holt-Winters and Linear Regression produced wider intervals, especially toward the end of the forecast period, signaling more uncertainty.


SARIMA Forecast (Next 12 Months)
Actual
SARIMA In-sample Prediction
SARIMA Forecast (12 Months)
95% Confidence Interval
Negative Rate
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
2004
2006
2008
2010
2012
2014


ARIMA Forecast: Next 12 Months
Actual
ARIMA In-Sample Prediction
ARIMA Forecast (Next 12 Months)
95% Confidence Interval
Negative Rate
1.0
0.8
0.6
0.4
0.2
0.0
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Date


Holt-Winters Forecast (Including Next 12 Months)
Actual
Holt-Winters In-Sample Prediction
Holt-Winters Forecast (Next 12 Months)
95% Confidence Interval
Negative Rate
1.0
0.8
0.6
0.4
0.2
0.0
2004
2006
2008
2010
2012
2014
Date



Based on these results, in this scenario, our analyst team recommends to answer business problem 2:

- Adopt SARIMA as the primary forecasting model to monitor negative sentiment trends, particularly in scenarios where seasonality plays a significant role. ARIMA can also be used as a simpler alternative for time series without seasonal patterns. Both models are suitable for building early warning systems, enabling the business to detect dissatisfaction trends in advance and support proactive decisions in areas such as customer retention, communication strategy, seasonal campaign timing, and brand reputation management.
- Forecasted increases in negative sentiment allow for proactive brand response. By anticipating periods of growing dissatisfaction, marketing teams can adjust messaging, reposition offerings, or reinforce service delivery before public perception deteriorates. This not only helps contain reputational risk, but also opens opportunities to strengthen trust and brand credibility through timely interventions.
- Incorporating sentiment forecasts into campaign planning improves precision and timing. Seasonal insights revealed by SARIMA can guide marketing teams in selecting the most appropriate periods to launch campaigns, avoiding historically sensitive windows and aligning tone with expected customer mood. This leads to greater message resonance and higher campaign effectiveness.
- Sentiment forecasting acts as a strategic layer of risk prevention. When models predict potential sentiment declines, teams can investigate possible causes early and implement corrective actions, whether through messaging adjustments, customer outreach, or product refinement, before negative sentiment becomes visible in public forums or sales data. This shifts reputation management from a reactive approach to a predictive and preventive one.
- To ensure ongoing reliability, forecasting models should be reviewed and recalibrated periodically. As customer behavior, expectations, and market conditions evolve, model performance may decline over time. Regular updates help keep forecasts aligned with current trends and ensure continued support for strategic decisions.

5. REPORT CONCLUSION

This report explored how sentiment analysis of food product reviews on Amazon can support strategic decisions related to customer experience, brand reputation, and marketing efforts. We structured our work into two main

parts: the first focused on building a sentiment classification model using the review texts, and the second on forecasting trends in negative sentiment over time through time series techniques.

In the first part, we found that supervised models such as Neural Networks and Random Forest achieved consistent results, even with imbalanced data, making them suitable for real-world operational use. Cleaned review summaries proved to be a lighter and more efficient alternative, particularly useful for systems that require speed and lower computational cost.

In the second part, we built a monthly time series covering 109 months to monitor the proportion of negative reviews. Our forecasting analysis showed that SARIMA performed best, effectively capturing seasonal patterns and producing low error rates. This confirmed the potential of using sentiment trends as an early warning tool to guide proactive decisions before customer dissatisfaction impacts brand perception.

The strategic recommendations and detailed answers to the two business problems are presented in Sections 3.2 (Business Problem 1) and 4.2 (Business Problem 2). Throughout the project, our team focused on applying analytical techniques in a practical and contextualized way. More than answering the business questions, we observed how combining text analytics, machine learning, and forecasting models can turn customer feedback into actionable insights, supporting better decisions in marketing, communication, and reputation strategy.

CONCLUSION OF PROJECT

Throughout this project, we applied the knowledge gained during our Master's program to a real-world problem involving unstructured data, customer sentiment, and business decision-making. Working with a large dataset of over 500,000 Amazon food product reviews allowed us to apply key concepts from machine learning, natural language processing, and time series forecasting.

In the first part of the project, we built and compared multiple classification models to detect sentiment in customer reviews. This helped us deepen our understanding of evaluation metrics, the impact of preprocessing and feature engineering, and techniques for addressing class imbalance. We also explored the role of pre-trained language models like Transformers, integrating them as an additional benchmark to validate our results.

In the second part, we developed a time series based on monthly sentiment trends and applied forecasting models to project potential shifts in customer dissatisfaction. This phase challenged us to manage issues like data gaps, stationarity, and model evaluation using error metrics. From this, we identified the most reliable models and translated our findings into actionable recommendations.

Beyond the technical learning, the project strengthened our ability to communicate analytical outcomes in a clear and strategic way. We created visualizations, wrote executive summaries, and translated data findings into business-oriented insights.

Overall, this project was a valuable learning experience that blended theory with real-world application. It gave us the opportunity to simulate the role of data analysts in a business context and demonstrated how analytics can directly contribute to enhancing customer experience and supporting strategy. The skills we developed here will continue to shape our professional journey ahead.