Sentiment Analysis Combined with Predictive Modeling for Sales and Price Impact

DSC 412 Final Project | Jiyun Yi

Abstract

This project focuses on analyzing the relationship of Starbucks' annual price, consumer sentiment, and revenue from 2020 to 2023. By utilizing web-scraped data from platforms such as Reddit, the study aims to compute sentiment scores for public reactions. The analysis pipeline includes data augmentation, model training, and visualization. This study aims to go beyond analyzing sentiment changes regarding Starbucks' price adjustments by exploring the actual revenue impact. By doing so, it seeks to provide actionable insights that can contribute to shaping effective business strategies.

---

1. Background

Starbucks, as a global leader in the coffeehouse industry, has consistently adjusted its pricing strategy to navigate market dynamics, cost fluctuations, and consumer demand. These price changes have significant implications not only on consumer sentiment but also on the company's revenue streams. Understanding the dual impact of price adjustments on both public perception and financial performance is crucial for devising effective business strategies. While previous studies have primarily focused on the sentiment analysis of consumer feedback in response to price hikes, this research extends the scope by concurrently examining the actual revenue changes associated with these price adjustments. By integrating sentiment analysis with revenue data, this study aims to provide a comprehensive view of how pricing strategies influence both customer attitudes and the company's financial health.

2. Data Analysis

The analysis began by collecting consumer opinions from Reddit's Starbucks-related discussions, with a focus on posts and comments reacting to price changes. A total of 2,500 data points (500 per year) were extracted, capturing diverse consumer sentiments. Also, after looking up the data, I changed the The data was preprocessed to remove noise such as stopwords, irrelevant punctuation, and special characters.

1) Data Augmentation

In the process of building a custom sentiment analysis model, I encountered a challenge in obtaining sentiment data related specifically to cafe menu price increases. Despite searching for relevant data, I was unable to find sufficient datasets directly addressing consumer sentiment regarding Starbucks' price hikes.

Although sentiment data related to stock price increases was available, I considered this data to be too different in nature from the focus of my analysis. The stock market is different type compared to the coffee menu.

To address this gap, I decided to generate relevant sentiment data through ChatGPT. I prompted the model to create opinions and sentiment scores specifically related to Starbucks price increases. These opinions were then classified into three categories: negative (-1), neutral (0), and positive (1). This approach allowed me to generate a tailored dataset that was aligned with the scope of my research, ensuring the relevance and accuracy of the sentiment analysis for Starbucks price hikes.

2) Model Selection

Three sentiment analysis models were applied: a Trained Model by ChatGPT-generated sentiment data (based on Naïve Bayes Model), TextBlob and VADER.

Trained Naive Bayes Model: Selected for its strong performance in classifying structured text data, it was trained on labeled examples to ensure accurate sentiment predictions tailored to Starbucks-related discussions.
TextBlob: Incorporated for its simplicity and ease of use, it provided a quick polarity analysis, leveraging a lexicon-based approach for general sentiment evaluation.

VADER: Specifically designed for social media text, VADER excelled in handling informal language, emojis, and nuanced expressions, making it highly suited to analyzing Reddit posts and comments.

These models quantified sentiment on a scale from -1 (negative) to 1 (positive).

3) Training Methodology

To train the sentiment analysis model, I utilized a Multinomial Naive Bayes classifier, which is particularly effective for text classification tasks. I first gathered a dataset of opinions on Starbucks price increases, which had been generated through ChatGPT and labeled with sentiment scores: negative (-1), neutral (0), and positive (1).

The data was then processed using the CountVectorizer from the scikit-learn library, which converts the text data into numerical features. This step involved transforming the raw text into a matrix of token counts, while excluding common stopwords (like "the," "and," etc.) using the NLTK stopwords corpus to ensure more meaningful features for model training.

For the training process, I used the MultinomialNB classifier, which was trained on the vectorized text data (X) and the corresponding sentiment labels (y). During training, the model learned to associate specific word patterns with sentiment classes, adjusting its internal parameters to minimize prediction errors.

Regarding hyperparameter tuning, the primary parameter that was adjusted was the stopwords filtering. While the default CountVectorizer settings were sufficient for the initial training, additional experimentation with different n-gram ranges (unigrams, bigrams) and adjusting the alpha parameter for the Naive Bayes model could improve performance. However, this was not done within the current setup.

To prevent overfitting, I ensured that the dataset was sufficiently large and diverse to represent various types of opinions. I also avoided the use of overly complex models or too many features that could cause the model to memorize the training data. Additionally, I performed cross-validation to check for generalization issues.

The model was validated using the dataset by testing it on previously unseen opinion files for each year (2020-2024). The output of the model, in the form of sentiment scores, was then aggregated to calculate average sentiment scores for each year, which were used to assess the overall sentiment towards Starbucks price increases over time.

3. Results

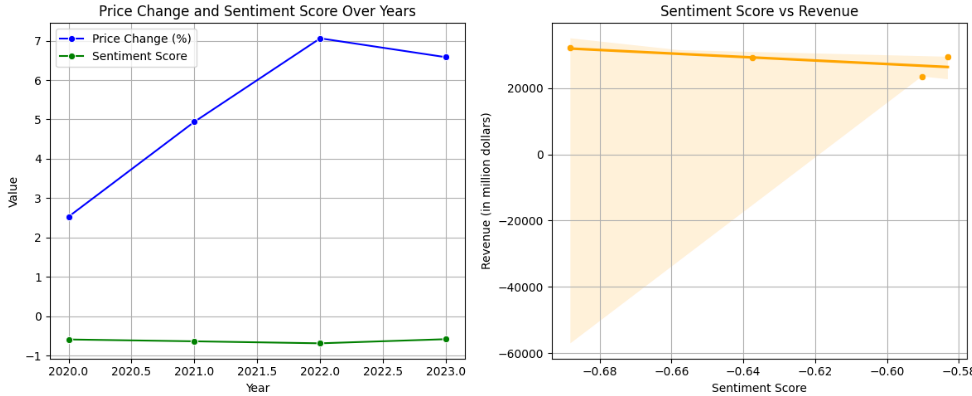

Figure 1Trained Naive Bayes Model

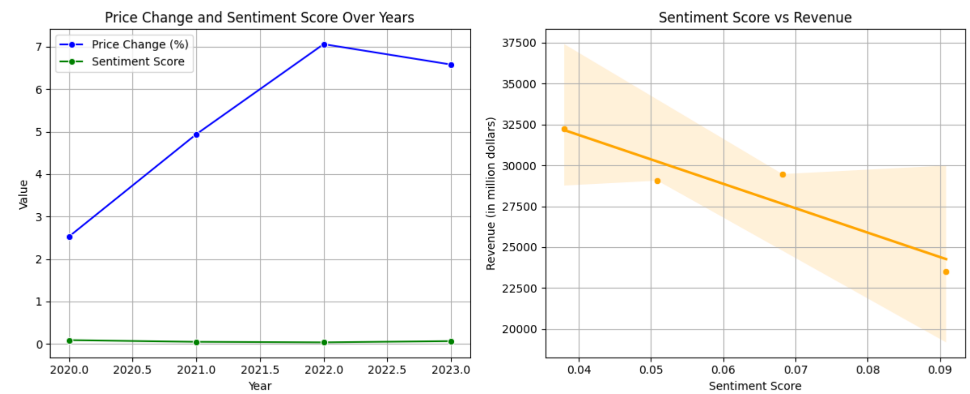

Figure 2 TextBlob

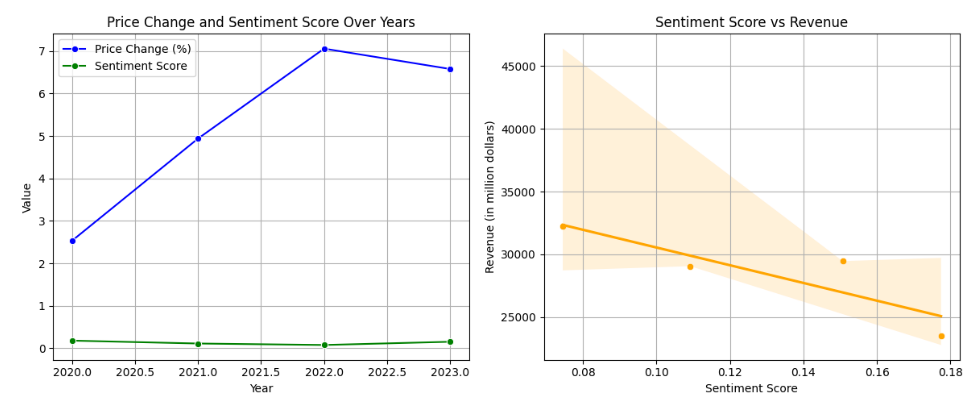

Figure 3 VADER

Trained Naïve Bayes Model	TextBlob	VADER
0.25	0.617	0.493

Figure 4 Sentiment Score ~ Price Change R-squared

1) **Trained Naïve Bayes Model**: The R-squared value of 0.25 suggests that the Naïve Bayes model has a weak correlation between sentiment scores and price changes. This indicates that other factors, aside from the sentiment captured by the model, may be influencing the relationship between sentiment and Starbucks price changes. While the Naïve Bayes model was trained on data derived from ChatGPT-generated sentiment scores, the model may need further fine-tuning or additional features to improve its predictive power.

2) **TextBlob**: With an R-squared value of 0.617, TextBlob demonstrates a moderate to strong correlation between sentiment and price changes. This suggests that TextBlob is relatively better at capturing sentiment trends that align with the price change data. TextBlob's reliance on pre-trained sentiment lexicons may have contributed to its better performance in this context, making it more suitable for this analysis than the Naïve Bayes model.

3) **VADER**: The R-squared value of 0.493 indicates a moderate correlation between sentiment and price changes. VADER's performance lies between the Naïve Bayes model and TextBlob, which suggests that while it captures sentiment effectively, it may not be as well-suited for the specific task of correlating sentiment with price changes as TextBlob.

4. Future Work

Limited Data Availability: Due to the limited data available, covering only the years 2020 to 2024 from Reddit, a comprehensive analysis became challenging. The lack of a broader dataset hindered the ability to accurately analyze the relationship between sentiment scores and revenue. As a result, I was unable to calculate the R-squared value for the correlation between sentiment and revenue, as the available data was insufficient to draw definitive conclusions. Also, as the dataset was based on sentiment data generated through ChatGPT, the results may not fully capture the diverse range of public opinions. Moreover, a deeper analysis involving sentiment scores in relation to Starbucks' revenue was not feasible due to the limited dataset, which hindered a comprehensive exploration of the broader business context.

Impact of External Factors (e.g., COVID-19): The period analyzed includes the COVID-19 pandemic, which introduced significant market fluctuations, making it difficult to isolate the effect of price changes alone. I used the data from 2020 to 2023, that might be influenced by COVID-19. The pandemic likely had a profound influence on consumer behavior and sentiment, and this should be taken into account in future studies. Other external factors, such as seasonal events or global economic conditions, should also be considered to better understand the relationship between price changes, sentiment, and revenue.

Linear Analysis Approach: In this study, sentiment scores were analyzed in a linear fashion, assuming a direct correlation with price changes. However, this approach may oversimplify the complex relationship between sentiment and price changes. In future work, it would be valuable to explore non-linear models that can better capture potential interactions between sentiment and other variables.

Incorporating Multiple Variables: The analysis focused solely on sentiment scores, but this is just one of many factors that could influence consumer behavior and business performance. Future research should aim to incorporate additional parameters, such as customer demographics, social media engagement, or even competitor pricing, to provide a more holistic view of the factors affecting Starbucks' pricing strategy and revenue.

5. Stakeholder Acknowledgements

The primary stakeholders in this project are Starbucks, consumers, and researchers in the field of business strategy and consumer sentiment analysis. Starbucks, as a company, could benefit from this model by gaining insights into how price changes influence consumer sentiment, potentially guiding their pricing strategy to optimize customer satisfaction and revenue. Understanding these sentiments could enable Starbucks to make data-driven decisions regarding price adjustments, promotions, or other business strategies.

However, the model could also have a negative impact if the sentiment analysis results are misinterpreted or overemphasized. If the model indicates strong negative sentiment, for example, Starbucks may be prompted to make premature decisions, such as lowering prices, without fully considering other factors like overall market trends, product quality, or other competitive forces.

The stakeholders involved, including Starbucks management, may be able to test and apply the model to assess customer feedback and sentiments related to pricing decisions. Sharing the results of the sentiment analysis with these stakeholders could allow for more informed decision-making. However, due to the limitations in data and the complexity of external factors, it is essential for stakeholders to use the model results in conjunction with other business insights.

6. Conclusion

This project aimed to explore the relationship between Starbucks' price increases and public sentiment, with the goal of providing valuable insights for business strategy development. By leveraging sentiment analysis tools like the Naïve Bayes Model, TextBlob, and VADER, the sentiment scores of customer opinions regarding Starbucks' price changes were analyzed across data from 2020 to 2024. Despite the challenges posed by limited data and external factors such as the COVID-19 pandemic, the analysis was able to identify patterns in public sentiment, although the correlation between sentiment and revenue could not be fully assessed due to the lack of revenue data.

Through the application of sentiment analysis models, the project revealed that while public sentiment fluctuated in response to price changes, there were limitations in capturing the full scope of factors influencing customer sentiment, such as broader economic conditions and social trends. The performance of the models, as measured by

R-squared values, indicated varying degrees of accuracy, but the results showed room for improvement in both data collection and model performance.

Future work should focus on acquiring more comprehensive and varied data, considering additional factors like regional differences, competition, and other external influences. Moreover, exploring different models and refining hyperparameters could lead to better prediction accuracy and more robust insights. The inclusion of revenue data and expanding the scope of the sentiment analysis to include a broader range of factors will be essential to fully understanding the dynamics between price changes and customer sentiment.

Ultimately, this project serves as a foundation for further research and analysis in the field of business strategy, offering a starting point for a deeper understanding of how price changes influence consumer behavior and how companies can leverage sentiment analysis in decision-making.

7. Citations

1) Public APIs - Reddit API: PublicAPIs.io. "Reddit API." *Public APIs*, [Online]. Available: https://publicapis.io/reddit-api. [Accessed: Nov. 18, 2024].

2) Hugging Face Blog - Sentiment Analysis with Python: Hugging Face. "Sentiment Analysis with Python." *Hugging Face Blog*, [Online]. Available: https://huggingface.co/blog/sentiment-analysis-python. [Accessed: Nov. 18, 2024].

3) Medium - Sentiment Analysis Project Using TextBlob: Qudrohbidemi, "Sentiment Analysis Project Using TextBlob," *Medium*, [Online]. Available: https://medium.com/@qudrohbidemi/sentiment-analysis-project-using-textblob-216d3fe119fc. [Accessed: Nov. 18, 2024].

4) Medium - VADER: A Comprehensive Guide to Sentiment Analysis in Python: R. Slavanyageetha, "VADER: A Comprehensive Guide to Sentiment Analysis in Python," *Medium*, [Online]. Available: https://medium.com/@rslavanyageetha/vader-a-comprehensive-guide-to-sentiment-analysis-in-python-c4f1868b0d2e. [Accessed: Nov. 18, 2024].