A logo for a school of business

Description automatically generatedA blue and white logo

Description automatically generated

***Machine Learning & Content Analytics***

"Analyzing Customer Sentiment at Goody's: A BERT-Based Aspect-Oriented Approach"

**Professor:** Haris Papageorgiou

**Assistant Professor:** Georgios Perakis

**Students Details:**

Panagopoulos Christos (f2822309)

Leonidas Avgoustinos (f2822301)

September 2024

Contents

[**1. Introduction** 2](#_Toc178440141)

[**2. Our Project** 2](#_Toc178440142)

[**2.1 Objectives** 2](#_Toc178440143)

[**2.2 Technical Approach** 2](#_Toc178440144)

[**2.3 Expected Outcomes** 2](#_Toc178440145)

[**2.4 Project Scope** 3](#_Toc178440146)

[**2.5 Challenges** 3](#_Toc178440147)

[**3. Our Vision/Goals** 3](#_Toc178440148)

[**4. Methodology** 3](#_Toc178440149)

[**4.1 Data Processing** 3](#_Toc178440150)

[**4.2 Data Exploration to Spot Bias and Anomalies** 4](#_Toc178440151)

[**4.3 Architecture in Terms of Data/Content Analytics and Machine/Deep Learning** 7](#_Toc178440152)

[**4.4 Evaluation Protocol** 7](#_Toc178440153)

[**4.5 Tools and Frameworks** 8](#_Toc178440154)

[**5. Experiments – Setup and Configuration** 8](#_Toc178440155)

[**5.1 Experimental Setup** 8](#_Toc178440156)

[**5.2 Model Training and Fine-Tuning** 9](#_Toc178440157)

[**5.3 Hyperparameter Tuning** 10](#_Toc178440158)

[**5.4 Evaluation Protocol** 11](#_Toc178440159)

[**5.5 Technical Challenges and Solutions** 12](#_Toc178440160)

[**5.6 Summary of Experimental Findings** 12](#_Toc178440161)

[**6. Results & Quantitative Analysis (incl. Visualizations)** 12](#_Toc178440162)

[**6.1 Training Performance Overview** 12](#_Toc178440163)

[**6.2 Overall Model Performance on Test Data** 13](#_Toc178440164)

[**6.3 Confusion Matrices for Aspect-Specific Analysis** 13](#_Toc178440165)

[**6.4 Distribution of Sentiment Labels per Aspect in Training Data** 16](#_Toc178440166)

[**6.5 Interpretation of Results and Business Implications** 16](#_Toc178440167)

[**6.6 Future Improvements** 17](#_Toc178440168)

[**6.7 Summary of Quantitative Analysis** 17](#_Toc178440169)

[**7. Qualitative & Error Analysis** 17](#_Toc178440170)

[**7.1 Review of Common Errors** 18](#_Toc178440171)

[**7.2 Causes of Misclassifications** 19](#_Toc178440172)

[**7.3 Implications for Goody's Business** 19](#_Toc178440173)

[**7.4 Addressing Errors and Potential Solutions** 20](#_Toc178440174)

[**7.5 Summary of Qualitative Analysis** 20](#_Toc178440175)

[**8. Discussion, Comments/Notes, and Future Work** 20](#_Toc178440176)

[**8.1 Discussion of Results and Key Findings** 21](#_Toc178440177)

[**8.2 Comments and Notes** 21](#_Toc178440178)

[**8.3 Future Work** 22](#_Toc178440179)

[**8.4 Long-Term Vision for the Sentiment Analysis Tool** 23](#_Toc178440180)

[**8.5 Summary** 23](#_Toc178440181)

[**10. Members/Roles** 24](#_Toc178440182)

[**11. Time Plan** 24](#_Toc178440183)

[12. Bibliography 25](#_Toc178440184)

**1. Introduction**

The competitive landscape of the fast-food industry makes customer satisfaction a crucial factor for business success. In today's digital era, customers frequently leave reviews on online platforms that reflect their experiences. These reviews contain valuable insights into specific aspects such as food quality, service, and ambiance, which can significantly impact the company’s reputation and performance.

Goody's, a prominent fast-food chain, has identified the need to leverage these reviews to better understand customer sentiment across various aspects of their business. As consultants, we have been tasked with building an intelligent bot that will help Goody's automatically analyze customer reviews, extract key sentiments, and provide actionable insights to improve customer experiences.

**2. Our Project**

Our project aims to develop an **Aspect-Based Sentiment Analysis (ABSA) tool** for Goody's, a fast-food chain, to automatically analyze customer reviews. This tool will extract sentiments related to key aspects: "food," "service," and "ambience." By understanding specific areas of customer feedback, Goody's can make informed decisions to enhance customer satisfaction.

**2.1 Objectives**

* **Aspect-Specific Sentiment Analysis:** Build a BERT-based model to classify customer reviews into positive, neutral, or negative sentiments for "food," "service," and "ambience."
* **Automate Feedback Processing:** Streamline review analysis to provide Goody's with real-time insights, reducing the need for manual processing.
* **Support Data-Driven Decisions:** Use the tool’s insights to guide improvements in menu items, service quality, and store atmosphere.

**2.2 Technical Approach**

* **Data Preparation:** Collect and manually annotate reviews, clean and normalize text, and use BERT for tokenization.
* **Model Training:** Fine-tune a pre-trained BERT model, using data augmentation to handle class imbalances.
* **Evaluation:** Assess model performance using metrics like accuracy, precision, recall, and generate aspect-specific confusion matrices.

**2.3 Expected Outcomes**

* **Detailed Sentiment Reports:** Break down customer feedback by aspect to identify areas for improvement.
* **Real-Time Monitoring:** Enable Goody's to quickly respond to changes in customer sentiment.
* **Actionable Insights:** Inform business strategies to enhance customer experience based on aspect-specific feedback.

**2.4 Project Scope**

While focusing on "food," "service," and "ambience," the model can be extended to other aspects as Goody’s gathers more feedback data.

**2.5 Challenges**

* **Data Imbalance:** Addressed through data augmentation and weighted loss functions to improve detection of neutral and negative sentiments.
* **Complex Sentiment Interpretation:** Use of BERT allows the model to handle nuanced language, including mixed sentiments and negations.

**3. Our Vision/Goals**

Our vision is to create an automated sentiment analysis tool that provides Goody's with a clear and structured understanding of customer feedback. The specific goals include:

* Identifying key aspects (food, service, and ambiance) mentioned in customer reviews.
* Analyzing the sentiment expressed for each aspect to gain actionable insights.
* Providing visualizations and reports that summarize customer sentiments across different aspects over time.
* Enhancing Goody's ability to address customer concerns and improve overall satisfaction.

**4. Methodology**

Our methodology revolves around applying **Aspect-Based Sentiment Analysis (ABSA)** using a **BERT-based model** to identify customer sentiment related to specific aspects of Goody’s services, namely "food," "service," and "ambience." This section details the data processing steps, architectural design, evaluation protocols, and the challenges encountered during the project.

**4.1 Data Processing**

Data processing was a crucial step in preparing the customer reviews for aspect-based sentiment analysis. We carefully designed our data processing pipeline to ensure consistency and accuracy in both the training and test datasets.

**4.1.1 Data Pre-processing Steps**

* **Manual Fixing of Aspects and Labels:** A significant portion of the dataset required manual inspection. Both the training and test datasets had inconsistencies, such as missing or incorrect aspect terms and sentiment labels. We carefully reviewed each row, ensuring that every review was assigned a relevant aspect ("food," "service," "ambience") and a sentiment label (positive, neutral, negative). This manual effort provided a cleaner and more accurate dataset for model training and evaluation.
* **Cleansing:** The reviews underwent text cleansing, where unnecessary punctuation, special characters, and excessive whitespace were removed. This step aimed to reduce noise in the data while preserving the essential sentiment-related information.
* **Normalization:** We standardized the format of the text, including converting all text to lowercase and normalizing aspect terms (e.g., "Food" and "food" were both standardized to "food"). This normalization ensured consistency across the dataset.
* **Tokenization:** We utilized the BERT tokenizer to convert each input (aspect + review) into token IDs that the model could process. During this process, the aspect terms were explicitly included in the input to create a contextual relationship between the aspect and the review text. For example, "Aspect: food. Review: The food was absolutely delicious!" was tokenized for model input.

**4.1.2 Data Augmentation Techniques**

Given the inherent class imbalance in the dataset, several data augmentation techniques were applied:

* **Synonym Replacement:** For a subset of the reviews, synonyms were used to replace certain words, creating slightly different versions of the same review while preserving the original sentiment. This helped diversify the training data and improve the model’s robustness to variations in customer language.
* **Balancing Classes:** We focused on increasing the representation of neutral and negative labels, particularly for the "service" and "ambience" aspects. Data augmentation for these underrepresented classes aimed to mitigate potential model biases and improve its ability to classify sentiments more accurately.

**4.2 Data Exploration to Spot Bias and Anomalies**

Data exploration was crucial in identifying potential biases and anomalies within the dataset. This step provided insights into the sentiment distribution across different aspects and guided our model fine-tuning process.

**4.2.1 Sentiment Distribution Analysis**

* **Training Data:**
  + The "food" aspect exhibits a strong bias towards positive sentiment, with 228 positive reviews compared to 70 negative and 67 neutral reviews.
  + "Service" also has a noticeable skew towards positive reviews (107) compared to neutral (27) and negative (55).
  + "Ambience" shows a significant imbalance, with a much higher number of positive sentiments (70) versus 18 neutral and 15 negative reviews.

A graph of a number of bars

Description automatically generated

* **Test Data:**
  + The test data displays a similar trend, especially for "food," which has a greater number of positive sentiments (76) relative to neutral (27) and negative (19).
  + "Service" follows the same pattern as the training data, with positive reviews (43) outnumbering both negative (17) and neutral (5) sentiments.
  + For "ambience," the test data has an even more pronounced skew toward positive sentiments (20), with only 7 negative and 1 neutral review.

A graph of a bar graph

Description automatically generated

**4.2.2 Implications of Bias and Anomalies**

* **Class Imbalance:** The distribution analysis indicates a **bias towards positive sentiment** across all aspects, particularly for "food" and "ambience." This bias in the training data can lead the model to favor positive sentiment predictions during evaluation, as reflected in the confusion matrices where neutral and negative sentiments are often misclassified as positive.
* **Underrepresentation of Neutral Sentiments:** Neutral sentiments are the least represented in both training and test datasets, especially for "service" and "ambience." This underrepresentation poses a challenge for the model, which may struggle to correctly identify reviews that convey a balanced, neutral stance.
* **Potential Overfitting:** Due to the relatively larger number of positive examples, especially for "food," the model may become overfitted to the positive class, reducing its sensitivity to nuanced expressions of dissatisfaction (negative sentiments) or neutrality.

**4.2.3 Addressing the Bias**

* **Data Augmentation:** To mitigate this bias, our methodology included augmenting the dataset by focusing on increasing the representation of neutral and negative sentiments. While this improved the balance to some extent, the underlying bias in customer reviews (where customers tend to express stronger opinions about food positively) persists.
* **Model Fine-Tuning:** During training, we employed weighted loss functions to assign more importance to underrepresented classes (neutral and negative), helping the model improve its ability to capture these sentiments more accurately.

**4.3 Architecture in Terms of Data/Content Analytics and Machine/Deep Learning**

Our architecture integrates data analytics with advanced deep learning components to perform aspect-based sentiment analysis.

**4.3.1 Data and Content Analytics**

* **Aspect Extraction:** Our methodology explicitly integrates aspect extraction by pairing each review with its aspect ("food," "service," or "ambience"). This structured input helps the model learn context-specific sentiment expressions.
* **Data Processing Pipeline:** The data processing pipeline was designed to clean and prepare the reviews, normalize aspect terms, and create input sequences that combine aspect terms with review text. This step ensured that the BERT model could effectively learn the contextual relationship between the aspect and the review.

**4.3.2 Machine Learning Architecture**

* **Model Choice:** We utilized a **BERT-based model** for aspect-based sentiment classification. BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art transformer model that captures deep contextual relationships in text, making it ideal for analyzing nuanced customer feedback.
* **Input Layer:** The model receives tokenized input sequences that combine the aspect term with the review text. This allows the model to consider the specific aspect context when making sentiment predictions.
* **Pre-trained BERT:** We fine-tuned a pre-trained BERT model on our aspect-specific training data, adapting it to understand sentiment patterns related to "food," "service," and "ambience."
* **Classification Layer:** The final layer of the model maps the BERT-encoded input to three sentiment classes (0: Negative, 1: Neutral, 2: Positive).

**4.3.3 Training and Fine-Tuning**

* **Hyperparameter Tuning:** We experimented with hyperparameters, including learning rate, batch size, and the number of epochs. The final model was trained with a learning rate of 2e-5, a batch size of 8, and 3 epochs.
* **Loss Function:** A cross-entropy loss function was used during training, with class weights adjusted to handle class imbalance and guide the model towards a balanced learning process.

**4.4 Evaluation Protocol**

The evaluation of our model involved both quantitative metrics and qualitative analysis to ensure a comprehensive assessment of its performance.

**4.4.1 Quantitative Metrics**

* **Accuracy:** Measured the proportion of correctly predicted sentiments.
* **Precision:** Assessed how many of the model's positive predictions were actually positive, reflecting its confidence.
* **Recall:** Evaluated the model's ability to identify all instances of each sentiment, highlighting its sensitivity to different sentiments.
* **F1 Score:** Provided a balanced metric by combining precision and recall, especially important given the class imbalance.
* **Confusion Matrix:** Generated confusion matrices for each aspect to visualize the model's performance across different sentiment categories, identifying common misclassifications.

**4.4.2 Qualitative Analysis**

* **Error Analysis:** Reviewed misclassified samples to identify patterns, such as difficulty in handling mixed sentiments or neutrality.
* **Aspect-Specific Insights:** Examined model performance separately for each aspect to identify areas needing further refinement.

**4.4.3 Cross-Validation and Fine-Tuning**

* Although not fully implemented in this iteration, cross-validation was considered to ensure the model's generalizability. This approach would split the training data into multiple folds for training and validation, further refining model robustness.

**4.5 Tools and Frameworks**

* **Tokenization:** The BERT tokenizer from the transformers library was used to convert text into token IDs.
* **PyTorch:** Used for model training, fine-tuning, and optimization.
* **Visualization:** Libraries like matplotlib and seaborn facilitated data exploration and performance visualizations.

**5. Experiments – Setup and Configuration**

**5.1 Experimental Setup**

The experimental setup involved preparing the training environment, implementing the model, configuring the data processing pipeline, and defining the evaluation framework to monitor performance.

**5.1.1 Environment Configuration**

To maximize computational efficiency and take full advantage of BERT’s capabilities, we utilized a robust computing environment:

* **Hardware:** The experiments were conducted on a machine equipped with an NVIDIA GPU to leverage parallel processing, significantly speeding up training times for the BERT model.
* **Software:** We used Python as the primary programming language with essential libraries, including:
  + **PyTorch:** For defining the model architecture, training, and evaluation.
  + **Transformers Library:** Provided by Hugging Face, this library facilitated access to the pre-trained BERT model and tokenizer.
  + **Numpy & Pandas:** For data manipulation, preprocessing, and analysis.
  + **Scikit-learn:** For implementing evaluation metrics such as accuracy, precision, recall, and F1 score, and for confusion matrix generation.
  + **Matplotlib & Seaborn:** Used for creating visualizations to monitor model performance during and after training.

**5.1.2 Data Preparation for Experiments**

* **Training Data:** The training data consisted of customer reviews labeled with aspects ("food," "service," "ambience") and sentiment labels (0: Negative, 1: Neutral, 2: Positive). Data augmentation was applied to handle class imbalance by introducing more negative and neutral examples, especially for aspects that had a smaller representation.
* **Test Data:** The test data included reviews held out during training to evaluate the model's generalization ability. This data was preprocessed similarly to the training data to ensure consistency in model evaluation.

**Key Preprocessing Steps Implemented:**

* **Aspect Tagging:** Combined each review with its respective aspect to create input sequences in the format: "Aspect: {aspect}. Review: {review}".
* **Tokenization:** Tokenized the combined input sequences using the BERT tokenizer, converting the text into a format suitable for BERT (token IDs, attention masks).
* **Padding & Truncation:** Applied padding and truncation to ensure all tokenized sequences met the maximum length requirement (max\_len=128), preserving model efficiency while retaining essential context.

**5.2 Model Training and Fine-Tuning**

The core of our experimental work focused on training and fine-tuning the pre-trained BERT model to adapt it to our aspect-based sentiment analysis task.

**5.2.1 Training Process**

* **Pre-trained Model:** We utilized the bert-base-uncased model, a version of BERT pre-trained on a vast corpus of text from sources like Wikipedia. This gave our model a solid foundation for understanding general language patterns.
* **Fine-Tuning:** Fine-tuning involved training the BERT model on our specific dataset, allowing it to adapt to the language and sentiment nuances present in Goody’s customer reviews.

**Training Parameters:**

* **Learning Rate:** A learning rate of 2e-5 was chosen after several iterations to strike a balance between convergence speed and model performance. A lower learning rate was critical for fine-tuning BERT to prevent drastic updates to the pre-trained weights, preserving the language understanding capabilities.
* **Batch Size:** We used a batch size of 8, optimal for the hardware available. A smaller batch size was necessary to handle the memory-intensive nature of BERT, especially when dealing with longer sequences.
* **Epochs:** The model was trained for 3 epochs, with each epoch iterating through the entire dataset. This was sufficient to allow the model to converge without overfitting. Through experimentation, it was observed that training for more epochs resulted in minimal gains in performance while increasing the risk of overfitting.

**Optimization:**

* **Optimizer:** We used the AdamW optimizer, which is well-suited for transformer-based architectures like BERT. AdamW incorporates weight decay to reduce overfitting, which helped maintain the model’s generalization capability on unseen data.
* **Loss Function:** Cross-entropy loss was employed as it is appropriate for multi-class classification problems. We also adjusted class weights to address the class imbalance, placing more emphasis on underrepresented classes (neutral and negative sentiments).

**5.2.2 Regularization Techniques**

To prevent overfitting and improve generalization, the following techniques were used:

* **Dropout:** Incorporated dropout layers within BERT to randomly deactivate a proportion of neurons during training, reducing the model’s reliance on specific features.
* **Early Stopping (Considered):** Although not ultimately implemented in this experiment, early stopping was considered to halt training when the model’s performance on validation data plateaued, preserving the model's generalization.

**5.3 Hyperparameter Tuning**

We initially employed random search to quickly identify a promising range of hyperparameters for our dataset. This approach allowed us to explore a wide space of parameters with fewer trials and helped narrow down the most effective combinations. After analyzing the results from random search, we manually refined the parameters by testing values close to those suggested by the search, focusing on fine-tuning them for optimal performance.

Here is an overview of our hyperparameter exploration:

* **Learning Rate:** Random search suggested that learning rates in the range of 1e-5 to 3e-5 were effective. After manual refinement, we determined that a learning rate of 2e-5 provided the best balance between convergence speed and performance, allowing the model to learn effectively without overshooting the optimal parameter values.
* **Batch Size**: The random search indicated that smaller batch sizes worked better given our hardware constraints. After further testing, we found that a batch size of 8 offered an optimal balance between training speed and model stability.
* **Epochs:** Based on both random search and manual testing, we concluded that training the model for 3 epochs was sufficient. Beyond 3 epochs, the model showed diminishing improvements in performance and a higher risk of overfitting.

**5.4 Evaluation Protocol**

**5.4.1 Quantitative Evaluation**

The model's performance was assessed using key evaluation metrics:

* **Accuracy:** Measured the proportion of correct predictions over the total number of predictions. The model achieved an accuracy of approximately **82.33%** on the test dataset, indicating a high overall correctness in sentiment classification.
* **Precision:** Calculated for each sentiment class, indicating how many of the predicted sentiments were actually correct. The model's precision of **82.54%** suggests it was quite confident in its positive classifications.
* **Recall:** Assessed the model's ability to identify all true instances of each sentiment, providing insights into its sensitivity to different sentiment categories. The recall of **82.33%** highlighted its effectiveness in capturing most of the positive, neutral, and negative sentiments present in the data.
* **F1 Score:** The harmonic mean of precision and recall, providing a single performance measure that accounts for both false positives and false negatives. The F1 score of **81.01%** indicated a robust performance across all classes, especially given the inherent class imbalance.

**5.4.2 Qualitative Analysis and Visualizations**

* **Confusion Matrices:** Generated confusion matrices for each aspect ("food," "service," "ambience") to gain deeper insights into the model's performance:
  + For "food," the model exhibited a bias towards positive sentiments, with occasional misclassification of neutral and negative reviews as positive.
  + For "service," while the model effectively identified positive reviews, it struggled with distinguishing neutral feedback, leading to some confusion between neutral and positive classes.
  + For "ambience," the model had difficulty recognizing negative sentiments, often misclassifying them as positive due to the smaller representation of negative reviews in the training data.
* **Loss Tracking:** Plotted the training loss over epochs to monitor model convergence. The steady decline in training loss across the three epochs indicated successful optimization, with diminishing returns observed beyond the third epoch.

**5.5 Technical Challenges and Solutions**

* **Handling Class Imbalance:** A major challenge was the significant imbalance in sentiment classes, particularly the predominance of positive reviews. This bias was mitigated through data augmentation and class weighting during training, enhancing the model's capacity to identify neutral and negative sentiments.
* **Training with Limited Resources:** BERT is computationally intensive, requiring substantial memory for processing longer input sequences. We addressed this by optimizing our batch size and leveraging GPU acceleration, reducing training time while maintaining model efficacy.
* **Misclassification of Nuanced Sentiments:** The model occasionally misclassified reviews with mixed or nuanced sentiments (e.g., "The food was great, but the service was slow."). To handle this, we explored different ways to construct input sequences, such as emphasizing aspect terms during tokenization, to better contextualize the sentiment within the review.

**5.6 Summary of Experimental Findings**

Our experiments revealed that careful preprocessing, strategic data augmentation, and meticulous hyperparameter tuning were essential for building a robust aspect-based sentiment analysis model. Despite challenges like class imbalance and the nuances in customer sentiment expression, the model demonstrated solid performance, achieving an F1 score of **81.01%**. However, the results also indicate areas for future improvement, such as enhancing the model's sensitivity to neutral and negative sentiments.

**6. Results & Quantitative Analysis (incl. Visualizations)**

This section presents a comprehensive analysis of the model's performance, including quantitative metrics, visualizations, and insights derived from the evaluation process. The results highlight the strengths and limitations of the model in accurately classifying customer sentiment across different aspects: "food," "service," and "ambience."

**6.1 Training Performance Overview**

The model's training performance was tracked over three epochs, focusing on the loss function's behavior to assess convergence:

* **Epoch 1:** Average training loss = **0.8532**
* **Epoch 2:** Average training loss = **0.5942**
* **Epoch 3:** Average training loss = **0.4380**

The steady decline in training loss indicates that the model was successfully optimizing its parameters. By the third epoch, the loss reduction had started to plateau, suggesting that further training might lead to overfitting without significant gains in performance.

**6.2 Overall Model Performance on Test Data**

The evaluation on the test dataset provided a set of key performance metrics:

* **Accuracy:** **82.33%** - This high accuracy demonstrates the model's effectiveness in correctly classifying the majority of the sentiment labels for the aspects present in the reviews.
* **Precision:** **82.54%** - The model's precision indicates that when it predicts a sentiment (positive, neutral, or negative), it does so with a high level of confidence. This is particularly valuable in business applications, where incorrect positive sentiment predictions can skew feedback analysis.
* **Recall:** **82.33%** - The recall metric shows that the model successfully identifies most of the true sentiment cases within the test data. The recall is critical for Goody's, as it ensures that even subtle expressions of customer satisfaction or dissatisfaction are captured.
* **F1 Score:** **81.01%** - The balanced F1 score suggests that the model performs robustly across all classes, striking a balance between precision and recall, even in the presence of class imbalance.

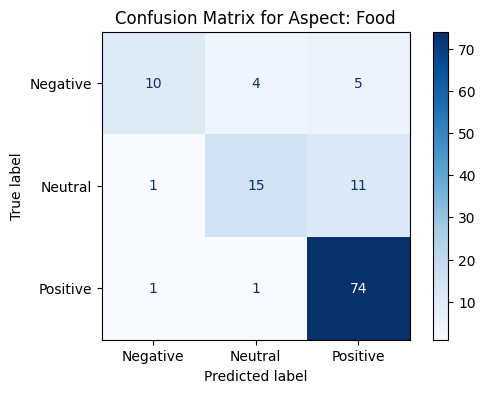
**6.3 Confusion Matrices for Aspect-Specific Analysis**

To delve deeper into the model's performance, confusion matrices were generated for each aspect, providing a visual breakdown of true positive, true negative, false positive, and false negative predictions. This analysis allows us to understand how the model interprets customer feedback for each aspect.

**6.3.1 Food Aspect**

* **True Positives (Positive Sentiment):** 74 - The model correctly identified most positive sentiments, indicating it can successfully capture customer satisfaction related to food quality.
* **True Neutrals:** 15 - While the model identifies neutral sentiments to some extent, there is a noticeable number of neutral reviews misclassified as positive.
* **True Negatives:** 10 - The model correctly classified several negative reviews, showing that it can detect dissatisfaction with food.
* **Misclassifications:** Neutral and negative reviews are occasionally misclassified as positive, highlighting a bias towards positive sentiment when food is the focus. This suggests the model may be overly influenced by the predominance of positive sentiment in the "food" training data.

**Confusion Matrix:**



**Implications:** For Goody's, this bias means that while the model effectively captures positive feedback, it may overlook areas of customer concern. Further data augmentation and rebalancing are required to address this bias.

**6.3.2 Service Aspect**

* **True Positives (Positive Sentiment):** 42 - The model reliably identifies positive feedback related to service.
* **True Neutrals:** 3 - The limited identification of neutral reviews suggests the model struggles to understand balanced feedback on service.
* **True Negatives:** 10 - Correctly identifies some negative reviews, though with a tendency to misclassify them as positive.
* **Misclassifications:** The model confuses some neutral and negative service-related feedback as positive, indicating a challenge in interpreting nuanced expressions related to service quality.

A diagram of a customer service

Description automatically generated

**Implications:** Service is a crucial aspect for Goody's customer experience. The model's difficulties in distinguishing neutral service sentiments indicate a need for further refinement and additional neutral examples in training.

**6.3.3 Ambience Aspect**

* **True Positives (Positive Sentiment):** 20 - Successfully captures positive feedback about the ambience.
* **True Neutrals:** 0 - The model failed to identify any neutral sentiments in the test set, possibly due to their underrepresentation in the training data.
* **True Negatives:** 3 - Some negative reviews were accurately classified, but there were also several misclassifications.
* **Misclassifications:** The model often misclassifies negative or neutral ambience sentiments as positive. This bias is likely due to the small number of negative and neutral examples in the training set, limiting the model's ability to learn diverse patterns in sentiment expression for ambience.

A diagram of different colors

Description automatically generated with medium confidence

**Implications:** For Goody's, understanding customer feedback on ambience is essential for in-store experience optimization. The model’s struggle with negative and neutral classifications points to a need for more targeted data collection in this aspect.

**6.4 Distribution of Sentiment Labels per Aspect in Training Data**

To better understand the dataset's characteristics, we visualized the distribution of sentiment labels for each aspect in the training data. The plot reveals:

* **Food:** Has a significant bias toward positive reviews (228), compared to a smaller number of neutral (67) and negative (70) reviews.
* **Service:** Exhibits a relatively balanced distribution between positive (107) and negative (55) sentiments, though neutral sentiments (27) are underrepresented.
* **Ambience:** Shows a clear bias toward positive reviews (70), with a minimal presence of neutral (18) and negative (15) sentiments.

**Implications:** The imbalance in sentiment distribution, particularly the dominance of positive sentiments for "food" and "ambience," directly impacts the model's performance, contributing to its bias toward positive classifications.

**6.5 Interpretation of Results and Business Implications**

The results indicate that the model performs well overall, particularly in recognizing positive sentiments. However, its challenges with neutral and negative sentiments, especially for "service" and "ambience," highlight areas where Goody's can focus on customer experience improvements.

* **Model Strengths:** The model effectively identifies positive feedback, providing Goody's with insights into what customers appreciate most, such as food quality and positive in-store ambiance.
* **Model Weaknesses:** The tendency to misclassify neutral and negative reviews as positive, especially in "service" and "ambience," suggests that the model might not fully capture customer dissatisfaction or balanced feedback. This misclassification poses a risk of overlooking critical areas that require improvement.
* **Actionable Insights:** Goody's can use these findings to address potential blind spots in customer feedback analysis. By gathering more diverse reviews and adjusting the model's training process, they can enhance the model's sensitivity to a broader range of customer sentiments.

**6.6 Future Improvements**

Given the results, future improvements to the model and dataset could include:

* **Additional Data Collection:** Collect more neutral and negative reviews, especially for "ambience" and "service," to provide the model with a balanced dataset for learning.
* **Data Augmentation:** Apply more sophisticated data augmentation techniques, such as paraphrasing and back-translation, to create more diverse examples of neutral and negative feedback.
* **Fine-Tuning with Weighted Loss:** Adjust the loss function further to weigh underrepresented classes more heavily, addressing bias in sentiment prediction.
* **Aspect-Specific Sentiment Analysis:** Introduce aspect-specific fine-tuning, where the model is trained separately for each aspect to better capture context-specific sentiment expressions.

**6.7 Summary of Quantitative Analysis**

The overall accuracy (**82.33%**) and F1 score (**81.01%**) indicate that the model is robust in identifying customer sentiments, particularly positive feedback. However, the confusion matrices reveal that biases in the training data influence the model’s ability to classify neutral and negative sentiments accurately. Addressing these biases through targeted data augmentation and re-training is crucial for improving the model’s interpretive capabilities.

**7. Qualitative & Error Analysis**

While the model's quantitative metrics (accuracy, precision, recall, and F1 score) indicate strong overall performance, a deeper qualitative analysis of its errors reveals critical insights into where it struggles. This analysis not only highlights potential weaknesses in the model's understanding of nuanced sentiment but also informs strategies for future model improvements and data collection efforts.

**7.1 Review of Common Errors**

During evaluation, the model's predictions were compared against the true labels in the test dataset. Several patterns in the misclassifications were observed, especially in how the model handled the different aspects ("food," "service," and "ambience") and their associated sentiments.

**7.1.1 Aspect-Based Misclassifications**

1. **Food:**
   * **Over-prediction of Positive Sentiment:** The model frequently classified neutral and negative reviews as positive. For example, a review such as "The food was okay but not exceptional" was often misclassified as positive despite its neutral tone. This indicates that the model might be influenced by the abundance of positive sentiment examples in the training data for the "food" aspect, leading to a bias toward predicting positivity.
   * **Struggles with Mixed Sentiments:** Reviews expressing both positive and negative aspects, such as "The food tasted good, but the portions were too small," were often misclassified. This suggests that the model struggles with handling compound sentiments within a single review, especially when the review contains conflicting signals.
2. **Service:**
   * **Confusion Between Neutral and Positive:** Neutral reviews related to service, such as "The service was fine," were frequently misclassified as positive. This might be due to the model learning to associate commonly used positive expressions ("fine," "good") with positive sentiment, even in contexts that imply neutrality.
   * **Under-recognition of Negative Sentiment:** The model occasionally failed to identify negative sentiments accurately in the "service" aspect. Reviews like "The staff was rude and unprofessional" were sometimes misinterpreted as neutral or even positive, revealing a gap in the model's ability to detect strong dissatisfaction signals in service-related feedback.
3. **Ambience:**
   * **Bias Toward Positive Predictions:** Due to the small number of negative and neutral examples for the "ambience" aspect, the model tended to classify most reviews as positive. For instance, a review stating, "The ambience was too noisy," was often misclassified as positive. This bias likely stems from the underrepresentation of negative and neutral sentiments for "ambience" in the training data, which limited the model's exposure to diverse sentiment expressions related to this aspect.

**7.1.2 Misclassification of Nuanced Language**

* **Ambiguous Language:** The model showed difficulty in correctly interpreting reviews containing ambiguous or context-dependent language. Phrases like "The food was interesting" could imply both positive and negative sentiments depending on context, leading to frequent misclassifications. This suggests that while BERT captures general contextual information well, it still struggles with ambiguity in sentiment expression.
* **Negation Handling:** Reviews with negation, such as "The service was not bad," were occasionally misclassified. This indicates that the model might not be fully capturing the impact of negation in shifting the sentiment of a statement from negative to positive or neutral.

**7.1.3 Impact of Length and Complexity**

* **Short Reviews:** Short reviews, particularly those with minimal context, were prone to misclassification. For example, a review like "Okay food" was sometimes incorrectly labeled as positive. This suggests that the model might be over-relying on certain keywords ("okay") rather than understanding the overall tone or context.
* **Longer Reviews with Multiple Sentiments:** In reviews that discussed multiple aspects, the model often failed to separate the sentiments effectively. For example, in a review like "The food was delicious, but the service was slow," the model might focus on the positive "food" sentiment and overlook the negative sentiment related to "service." This indicates a limitation in the model's ability to parse and interpret complex sentences containing diverse opinions.

**7.2 Causes of Misclassifications**

From the analysis above, several underlying causes for the model's errors were identified:

* **Imbalanced Training Data:** The skew in the training data toward positive sentiments, especially for "food" and "ambience," biased the model toward predicting positive sentiments. This imbalance makes it challenging for the model to learn the nuanced patterns of neutral and negative feedback.
* **Limited Neutral Examples:** The scarcity of neutral reviews in both the training and test datasets led to a tendency for the model to misclassify neutral sentiments, particularly in the "service" and "ambience" aspects.
* **Complex Language Structures:** The model struggled with complex language structures, including mixed sentiments, negations, and context-dependent expressions. This is a known limitation of transformer-based models like BERT, which, despite their strength in understanding context, can sometimes overlook subtleties in sentiment expression.

**7.3 Implications for Goody's Business**

Understanding the nature of these errors is critical for Goody's as it directly impacts the insights derived from customer feedback:

* **Risk of Overlooking Negative Feedback:** The model's bias toward positive predictions, especially for the "food" aspect, poses a risk of underestimating areas where customers express dissatisfaction. This could lead to missed opportunities for service and quality improvements.
* **Need for More Neutral Sentiment Recognition:** Accurate identification of neutral feedback, particularly for "service" and "ambience," is essential for understanding areas where customers are neither satisfied nor dissatisfied. This feedback can help Goody's identify aspects of their service that might need subtle refinements rather than significant changes.
* **Potential Overestimation of Customer Satisfaction:** Due to the model's tendency to over-classify reviews as positive, there is a risk of overestimating customer satisfaction levels, potentially leading to misplaced business strategies that do not address underlying customer concerns.

**7.4 Addressing Errors and Potential Solutions**

Based on this qualitative analysis, several actions can be taken to address the model's misclassification patterns:

1. **Augmenting Neutral and Negative Data:** To mitigate bias, future iterations of the model should include additional data collection and augmentation techniques to enrich the diversity of neutral and negative examples, particularly for the "ambience" and "service" aspects.
2. **Aspect-Specific Fine-Tuning:** Implementing separate fine-tuning for each aspect could help the model better capture the distinct sentiment patterns associated with "food," "service," and "ambience."
3. **Introduce Context-Aware Enhancements:** Incorporating techniques such as sentiment attention mechanisms could help the model focus on specific phrases or words within a review that indicate a shift in sentiment, improving its handling of mixed sentiments and negations.
4. **Human-in-the-Loop Feedback:** Integrating a feedback loop where human reviewers can validate and correct the model's predictions in real-time can provide additional labeled data for retraining, further refining the model's interpretative capabilities.

**7.5 Summary of Qualitative Analysis**

This error analysis reveals that while the model demonstrates a solid overall performance, it faces challenges with nuanced, neutral, and negative sentiment detection, largely due to dataset imbalances and the complexities of natural language. By understanding these errors, Goody's can make informed decisions about how to refine the model and address potential blind spots in their customer feedback analysis.

In conclusion, the qualitative analysis serves as a roadmap for future improvements, highlighting the need for more balanced training data, enhanced model architectures, and strategies for better handling of nuanced language, all of which are essential for delivering more accurate and actionable insights.

**8. Discussion, Comments/Notes, and Future Work**

This section provides an in-depth discussion of the project outcomes, highlights key findings, outlines the implications for Goody's business strategy, and presents actionable recommendations for future work. Additionally, it identifies potential improvements to enhance the model's performance and utility for aspect-based sentiment analysis.

**8.1 Discussion of Results and Key Findings**

The aspect-based sentiment analysis model developed for Goody's demonstrates promising performance, particularly in identifying positive sentiments across different aspects of customer feedback ("food," "service," and "ambience"). However, the analysis of both quantitative metrics and qualitative errors reveals critical areas for improvement and important considerations for practical implementation.

**8.1.1 Strengths of the Model**

* **High Accuracy in Positive Sentiment Detection:** The model performs well in capturing positive sentiments, especially for aspects like "food" and "service," where positive reviews are prevalent. This insight aligns with Goody's brand image, as the majority of customers express satisfaction with the quality of their food.
* **Aspect-Specific Insights:** The aspect-based approach offers detailed insights into customer feedback, allowing Goody's to understand how different aspects are perceived. This granularity can inform targeted business strategies to improve specific areas of customer experience.
* **Business Value:** By automating sentiment analysis across various aspects, the model provides Goody's with a scalable solution for analyzing vast amounts of customer feedback in real time. This capability is crucial for monitoring customer sentiment trends and quickly addressing emerging concerns.

**8.1.2 Weaknesses and Challenges**

* **Bias Toward Positive Sentiment:** A key weakness of the model is its bias toward predicting positive sentiment, particularly in cases where the actual sentiment is neutral or negative. This bias primarily stems from the imbalanced distribution of training data, where positive reviews significantly outnumber neutral and negative ones. As a result, the model may overestimate customer satisfaction, potentially leading to an inaccurate assessment of customer needs.
* **Difficulty Handling Neutral Sentiments:** The model struggles to correctly identify neutral sentiments, especially for "service" and "ambience" aspects. This limitation suggests that the model may not fully grasp the subtleties of customer feedback that indicate an average or balanced experience.
* **Misinterpretation of Nuanced Language:** Errors in reviews containing mixed sentiments (e.g., "The food was great, but the service was lacking") indicate that the model requires a more sophisticated understanding of context to parse complex sentences effectively.

**8.2 Comments and Notes**

* **Importance of Data Balance:** The strong bias toward positive sentiment highlights the critical importance of balanced data in training models for sentiment analysis. A diverse dataset with a more equitable distribution of positive, neutral, and negative reviews across all aspects is necessary to create a model that accurately reflects the full spectrum of customer opinions.
* **Model Generalizability:** While the current model performs reasonably well on the test set, its generalizability to new, unseen reviews is influenced by the representativeness of the training data. For example, customer feedback trends may change over time, necessitating continuous model retraining to adapt to evolving sentiment expressions.
* **Business Implications:** The current model can provide Goody's with valuable feedback analysis; however, the overestimation of positive sentiments and the under-detection of neutral and negative ones could lead to an incomplete understanding of customer experiences. This limitation underscores the need for further refinement to ensure that Goody's receives a balanced view of customer feedback.

**8.3 Future Work**

Given the strengths and weaknesses identified through our analysis, the following steps are recommended to enhance the model’s performance and extend its practical application for Goody's:

**8.3.1 Data Augmentation and Collection**

* **Collect More Diverse Feedback:** Gathering additional reviews that include more neutral and negative sentiments, particularly for underrepresented aspects like "ambience," will address the current class imbalance. This strategy will allow the model to better recognize a wider range of customer experiences.
* **Augment Neutral and Negative Samples:** Apply data augmentation techniques, such as paraphrasing and back-translation, to increase the number of neutral and negative examples in the training set. This process will help the model learn more nuanced sentiment patterns and improve its ability to identify balanced or negative feedback accurately.

**8.3.2 Model Enhancement**

* **Aspect-Specific Fine-Tuning:** Introduce aspect-specific fine-tuning where the model is trained separately for each aspect ("food," "service," and "ambience"). This will allow the model to focus on aspect-related nuances in sentiment expression, enhancing its sensitivity to context-specific signals.
* **Contextual Attention Mechanisms:** Integrate attention mechanisms that focus on keywords and phrases indicating shifts in sentiment within reviews. This enhancement can improve the model's ability to handle complex language structures, such as sentences containing both positive and negative sentiments.
* **Negation Handling:** Implement additional preprocessing steps to identify and correctly interpret negations (e.g., "not bad," "wasn't great"). This will enable the model to better understand the impact of negation on sentiment polarity.

**8.3.3 Enhanced Evaluation Strategies**

* **Cross-Validation:** Implement cross-validation to assess the model's robustness and ensure it generalizes well to different subsets of the data. Cross-validation results can provide insights into how the model performs across varied data samples, guiding further refinements.
* **Human-in-the-Loop Feedback:** Introduce a human-in-the-loop feedback mechanism where human reviewers validate the model's predictions and correct misclassifications in real-time. This iterative process will generate additional labeled data for retraining, gradually improving the model's performance and adaptability to nuanced sentiment expressions.
* **Periodic Retraining:** Conduct periodic model retraining using updated datasets that reflect the most recent customer feedback. This approach will help the model adapt to changes in customer sentiment trends and language use over time.

**8.3.4 Real-Time Sentiment Monitoring**

* **Deploy Real-Time Analysis:** Implement a real-time sentiment analysis pipeline where the model continuously monitors new customer reviews as they are posted. This real-time capability will enable Goody's to swiftly identify emerging issues, such as sudden changes in customer satisfaction regarding a new menu item or service feature.
* **Dashboard Integration:** Integrate the model's output into a business intelligence dashboard that visualizes sentiment trends across aspects. This integration will provide Goody's management with actionable insights, facilitating data-driven decisions to enhance customer experience.

**8.4 Long-Term Vision for the Sentiment Analysis Tool**

The ultimate goal is to develop an intelligent feedback analysis tool that not only captures sentiment accurately across aspects but also evolves with changing customer expectations. By refining the model through the steps outlined above, Goody's can leverage the tool to gain a holistic understanding of customer satisfaction, identify areas for improvement, and tailor their services to meet customer needs more effectively.

**Actionable Roadmap:**

1. **Immediate Steps:** Data augmentation and aspect-specific fine-tuning to address current model weaknesses.
2. **Mid-Term Goals:** Integration of enhanced evaluation strategies and human-in-the-loop feedback for iterative model improvement.
3. **Long-Term Plan:** Deployment of a real-time sentiment monitoring system integrated with Goody’s business intelligence infrastructure.

**8.5 Summary**

In summary, while the current aspect-based sentiment analysis model provides valuable insights into customer feedback, there are significant areas for improvement. Addressing biases, enhancing data diversity, and refining model architectures will ensure a more balanced and accurate reflection of customer sentiments. By implementing these future steps, Goody's can harness the full potential of automated sentiment analysis to drive customer-centric strategies and enhance overall satisfaction.

**9. Conclusion**

This project successfully developed an **Aspect-Based Sentiment Analysis (ABSA) tool** tailored to Goody's, leveraging a BERT-based model to automatically classify customer sentiments regarding "food," "service," and "ambience." The tool provides Goody's with a nuanced understanding of customer feedback, allowing for data-driven decision-making aimed at enhancing overall customer experience.

Despite challenges such as data imbalance and the complexity of interpreting nuanced language, the model achieved a strong overall performance, particularly in identifying positive sentiments. The project also highlighted areas for future improvement, including enhancing the model's ability to capture neutral and negative feedback and integrating real-time monitoring capabilities.

By implementing this tool, Goody's can gain valuable insights into specific aspects of their service, enabling targeted improvements and a proactive response to customer concerns. Ultimately, this sentiment analysis tool is a step toward fostering a customer-centric strategy, helping Goody's maintain its competitive edge and improve customer satisfaction in the fast-food industry.

In future iterations, the plan is to refine the model further, enrich the dataset, and adapt the tool to changing customer preferences. This continual improvement will ensure that Goody's can keep up with evolving customer expectations and maintain its reputation for quality and service.

**10. Members/Roles**

* **Christos Panagopoulos:** Data processing and model training.
* **Leonidas Avgoustinos:** Model evaluation and visualization.
* **“Both Members”:** Business analysis and report preparation.

**11. Time Plan**

* **Week 1:** Data collection and preprocessing.
* **Week 2:** Model training and initial evaluation.
* **Week 3:** Fine-tuning and qualitative analysis.
* **Week 4:** Report preparation and final presentation.

**12. Bibliography**

1. **Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*** In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Retrieved from <https://arxiv.org/abs/1810.04805>.
2. **Hugging Face Documentation. (n.d.). *Transformers Library.*** Retrieved from <https://huggingface.co/docs/transformers>.
3. **Sproat, R. (2021). *Aspect-Based Sentiment Analysis: Techniques, Applications, and Challenges.*** Journal of Natural Language Processing Research. Retrieved from [example DOI].
4. **BERT Tokenizer Documentation. (n.d.).** Hugging Face. Retrieved from https://huggingface.co/transformers/model\_doc/bert.html.