Milestone #3 Model Evaluation

To evaluate the performance of the sentiment classification model, we used a range of metrics to capture how well the model performed across all classes. The goal of this task was to classify reviews into negative, neutral, or positive sentiment, so metrics such as accuracy, precision, recall, and F1-score were chosen to evaluate performance. These metrics provide a balanced view of both overall performance and how the model handles each sentiment class individually, and this is particularly important in the presence of class imbalance.

Two models were tested: Logistic Regression and Naïve Bayes. Both were trained using TF-IDF vectorization with unigrams and bigrams, and oversampling was applied using SMOTE to address the class imbalance in the training data (there were many more positive reviews than neutral and negative). Since the original dataset was heavily skewed toward positive sentiment, balancing the classes helped ensure that the model had enough examples from the smaller classes (neutral and negative) to learn from. After applying SMOTE, each class had an equal number of examples, which made the training process more balanced.

Logistic Regression performed better overall, achieving an accuracy of 87.81%. It showed strong performance in classifying positive reviews, with precision and recall both above 90%, and an F1-score of 0.94. However, performance on the neutral and negative classes was significantly lower, especially for neutral, where the F1-score dropped to 0.25. Naïve Bayes had also performed well on the positive class but struggled even more with the smaller classes, ending with a slightly lower overall accuracy of 83.01% and a macro F1-score of 0.50. There were several different runs conducted of these metrics and the accuracy, precision, recall, and F1 had very slight variations but remained consistent over time.

```
Logistic Regression Accuracy: 0.8781
Logistic Regression Classification Report:
              precision    recall  f1-score   support

    negative       0.30      0.59      0.40       162
     neutral       0.19      0.35      0.25       300
    positive       0.97      0.91      0.94      6464

    accuracy                           0.88      6926
   macro avg       0.49      0.62      0.53      6926
weighted avg       0.92      0.88      0.90      6926
```

```
Naïve Bayes Accuracy: 0.8301
Naïve Bayes Classification Report:
              precision    recall  f1-score   support

    negative       0.24      0.59      0.34       162
     neutral       0.16      0.50      0.25       300
    positive       0.98      0.85      0.91      6464

    accuracy                           0.83      6926
   macro avg       0.46      0.65      0.50      6926
weighted avg       0.93      0.83      0.87      6926
```

The results tell us that while the model does well in identifying positive sentiment, it struggles with neutral and negative reviews. This could be due to the ambiguous nature of neutral reviews, which often include both positive and negative aspects, making them harder to classify. Additionally, since the test set remained imbalanced, the models had difficulty generalizing to these underrepresented sentiments. Even with oversampling using SMOTE during training, the class imbalance in real-world data is a challenge.

We wanted to explore whether more sophisticated models could boost performance, especially in cases where the baseline models might oversimplify the decision boundary. We used **Support Vector Machine (SVM)** model, which is great for high-dimensional spaces such as text data through TF-IDF or word embeddings. SVM's margin-maximization also helps create more distinct boundaries between closely related classes, which is especially valuable when reviews may contain both positive and negative signals (e.g., "We like this product, but the delivery was later than expected date").

In addition, we also added an **ensemble model** using majority voting to combine predictions from Naive Bayes, Logistic Regression, and SVM. The reason behind ensembling is that different models make different types of errors—Naive Bayes might possibly over-rely on word frequency, Logistic Regression may under fit certain patterns, and SVM might be sensitive to tuning of parameters. By aggregating their predictions, the ensemble can smooth out individual weaknesses of each of the models while amplifying their strengths.

```
SVM Performance:
Accuracy: 0.9337
Precision: 0.6008
Recall: 0.4355
F1 Score: 0.4759

Ensemble Performance:
Accuracy: 0.9370
Precision: 0.7110
Recall: 0.4018
F1 Score: 0.4381
```

We see that SVM and Ensemble both had higher levels of accuracy compared to naïve bayes and logistic regression. SVM had the best F1 score out of the models. However, we still saw varying levels of difficulty in neutral reviews being correctly identified as neutral, which builds on our earlier hypotheses of these reviews having both positive and negative sentiment which make them difficult to classify.

Finally, there can be further analysis such as ablation studies, which could involve testing the result of removing n-grams or skipping SMOTE and this could possibly provide deeper insight into what components of the pipeline are most impactful. While the current models performs well for its dominant class, there is clear room for improvement in handling less frequent sentiments of neutral and negative reviews, specifically neutral reviews.