Final Report

Sentiment Analysis of Amazon Fine Food Reviews using Machine Learning

1. Problem Statement and Motivation:

This project explores how machine learning can be applied to predict the sentiment of Amazon Fine Food reviews. Specifically, the goal is to label a review as positive, negative, or neutral based on the textual content. Sentiment analysis is useful as reviews directly influence the decisions consumers make. If a model can accurately determine the tone of a review through text analysis, it would enable businesses to process customer feedback much more efficiently. **Thesis Statement:** Can a machine learn to recognize whether a review is positive, neutral, or negative just from the text?

Challenges addressed:

- **Data Volume:** The full dataset contains over 500,000 reviews. A smaller, representative sample of approximately 10,000 reviews was selected to make the analysis more manageable.
- **Sentiment Imbalance:** Most reviews are positive, which can make it harder to train models to recognize neutral and negative sentiments.
- **Text Preprocessing:** Reviews contain slang, typos, and special characters, requiring thorough cleaning and standardization.
- **Model Selection:** Choosing effective machine learning models (e.g., Logistic Regression, Random Forest) that can handle text data and produce high accuracy.

Final Report

2. Introduction and Data Description:

The Amazon Fine Food Reviews dataset from Kaggle was used for this project. It contains over 500,000 reviews collected between 1999 and 2012. Each review includes the star rating, timestamp, summary, and full review text.

To make the analysis manageable, a smaller sample of approximately 10,000 reviews was selected. Rating distribution was maintained on the scaled dataset.

Step 1: Data Loading and Review:

The dataset was loaded, and the structure of the reviews was analyzed. Missing or duplicate entries were identified and handled.

Step 2: Data Refinement:

Duplicate reviews were removed, missing values were handled, and new features such as word count and review year were created.

Step 3: Sentiment Labeling:

Sentiment labels were assigned based on the Score field (Negative: 1–2 stars, Neutral: 3 stars, Positive: 4–5 stars). The text was then cleaned by converting to lowercase, removing punctuation and numbers, and filtering out very short words.

3. Modeling Approach:

The following machine learning models were applied:

Final Report

- Baseline: Logistic Regression
- Advanced: Random Forest Classifier
- Additional: K-Nearest Neighbors (KNN)

Each model used TF-IDF vectorization for converting text to numerical features.

Logistic Regression Baseline Model:

The logistic regression model was used as a baseline to implement and classify the sentiment of Amazon food reviews. The text data was first transformed into numerical features using TF-IDF, which helps to capture the most important words based on their frequency and relevance across reviews. This transformation was essential for the model to handle high-dimensional text data. Following feature extraction, the model was trained on the data. It leveraged the sentiment labels: negative, neutral, and positive (which were assigned based on the review scores). The training process involved splitting the data into training and testing sets to evaluate model performance. Evaluation metrics such as accuracy and F1 score were used to assess the model's overall performance, with a particular focus on precision and recall for each class. A confusion matrix was generated to provide a deeper understanding of the model's classification performance. Misclassified examples were examined to highlight areas where the model struggled, offering valuable insights for future improvements. Furthermore a results table was created, showcasing precision and recall for each sentiment class. The Logistic Regression model achieved an accuracy of 91.8% and an F1 score of 0.9532, showing strong performance. The confusion matrix indicates that the model performed well in classifying positive reviews but had a lower recall for negative reviews.

Final Report

Random Forest Model:

For the advanced machine learning model, the Random Forest Classifier was used because of its ability to handle both classification and regression tasks. Additionally, it does not need extensive preprocessing of the data and works with high-dimensional data such as our dataset being about reviews from Amazon. While typically one might need to scale features in Logistic Regression, the Random Forest does not require this additional step and can still make predictions that can be useful. In the implementation of the Random Forest Classifier with the dataset of Amazon Food Reviews, it creates sentiment labels based on what the review scores were after ensuring the columns needed are in the dataframe to prevent any issues. This is followed with a label encoder procedure to turn the sentiment labels into numerical values. In this case, anything under 2 was given a negative label, a 3 was neutral and anything above a 4 was positive. These positive, neutral, and negative labels are then also denoted with numbers to be able to be used in the Random Forest machine learning algorithm. By using a TF-IDF Vectorizer, it converts the text data into numerical features and in doing so, limiting less important words. Starting the training and testing, the Random Forest Model utilized 100 trees. For this model, the accuracy was 80%. The five most important words that were deemed as important by the model were “great”, “disappointed”, “bad”, “thought”, and “love”. By using these words, the model was able to determine if the reviews had a sentiment of a positive, negative or neutral review.

K Nearest Neighbors Model:

To further take a look into the data and try to find what words would be most important and key for a machine learning system to identify the sentiment of a review, K Nearest

Neighbors was implemented. In doing so, the goal was to use simple, distance based methods to see how sentimentally similar reviews could be predicted. KNN builds models by classifying reviews on proximity to other reviews in a feature space. This allows for a straightforward approach to visualizing how individual words group reviews together. Essentially, shared words such as “great” or “love” in positive reviews should be naturally grouped together, whereas common negative review words such as “waste” or “bad” would be grouped together. Based on this, distinct clusters should form, allowing data to be separated. The success of the KNN model also explains the data simply. The better it does, the stronger the difference in word choice in positive vs negative reviews. Implementation of this model was simple and based on K = 5. Overall accuracy was found to be roughly 77%, indicating the model was able to capture structure in the data. It is worth noting that the model had a higher relative share of false positives, indicating that the model performed much better for positive reviews than negative or neutral reviews. In conclusion, while the model has reasonable accuracy, alternative models that are better able to handle class imbalance would be preferable.

4. Project Trajectory, Results and Interpretation

Revised Project Question:

Based on what we found in the EDA, such as the imbalance between positive, neutral, and negative reviews, and how review length and word choice differ by sentiment, we updated our thesis to: Can we build a model that accurately predicts review sentiment using the text, while also handling the imbalance to better identify neutral and negative reviews?

Final Report

Results:

The results indicate that the Logistic Regression model performed best. It achieved the highest accuracy at 91.8%, whereas the Random Forest and K Nearest Neighbors models had accuracies of 80% and 77% respectively. The LR model also boasted an impressive F1 score of 0.9532. It should be noted that all three models performed exceptionally well in detecting positive reviews. However, the gaps in model accuracy began to appear when sorting negative reviews. Although the LR model also struggled with false positives, or classifying negative reviews, both the RF and KNN models performed worse in this area. Therefore, Logistic Regression was clearly the most reliable model for sentiment classification in this study.

5. Conclusions and Future Work:

Limitations:

A limitation to consider is the language of the review. If the model is trained and tested in one language, then it may not perform well for another language. For example, if the model is trained in English then tested in French, many words could be misgrouped or ignored despite similar structure. Even worse, if the test data had reviews in a language using a different alphabet, such as Arabic, the model would certainly not know how to classify the review.

Future work:

This study could be used in the future by companies or businesses to determine how well a product is doing in the market or even how the general public feel about certain music, art, politics, and so much more.