Capstone DSAI

Problem Statement 3

Sentiment Analysis of Customer Reviews

a. Problem Statement

The goal of this project is to analyse customer reviews from an e-commerce platform and classify them into three sentiment categories: positive, negative, and neutral. Given the imbalance in the dataset, we aim to develop a robust sentiment analysis model that accurately predicts the sentiment of customer reviews, providing valuable insights for business strategies.

b. Project Objective

The main objectives of this project are:

1. To preprocess and clean the customer reviews dataset.
2. To handle the class imbalance using appropriate techniques.
3. To build and evaluate a machine learning model that accurately classifies the sentiment of customer reviews.
4. To derive actionable insights from the model's predictions and the analysis of the review data.

c. Data Description

The dataset **Reviews.csv** contains 568,454 rows and 10 columns, including:

- **Id**: Unique identifier for each review.
- **ProductId**: Unique identifier for the product being reviewed.
- **UserId**: Unique identifier for the user who wrote the review.
- **ProfileName**: Name of the user profile.
- **HelpfulnessNumerator**: Number of users who found the review helpful.
- **HelpfulnessDenominator**: Number of users who indicated whether the review was helpful.
- **Score**: Rating given by the user (from 1 to 5).
- **Time**: Timestamp of the review.
- **Summary**: Short summary of the review.
- **Text**: Full text of the review.

Each entry represents a review with associated metadata. The initial step was to inspect the data for any inconsistencies, missing values, or duplicates, which could affect the analysis.

d. Data Pre-processing Steps and Inspiration

1. **Cleaning Text Data**: Removed HTML tags, punctuation, stop-words, and performed stemming/lemmatization.
2. **Handling Missing Values**: Removed rows with missing values in critical columns like 'Text' and 'Score'.
3. **Creating Sentiment Labels**: Mapped the 'Score' column to sentiment labels (1-2 as negative, 3 as neutral, 4-5 as positive).
4. **Balancing Classes**: Addressed the class imbalance using Synthetic Minority Over-sampling Technique (SMOTE).
5. **Vectorization**: Transformed text data into numerical format using TF-IDF vectorization.

Missing Values and Duplicate Entries

Upon examining the dataset, it was observed that the `ProfileName` and `Summary` columns had a small number of missing values - 53. These entries were removed to ensure data integrity. Additionally duplicate reviews were also removed, leaving 393,560 unique reviews. This step was crucial to prevent redundant information from skewing the analysis and to improve the performance of the machine learning models.

Data Cleaning and Preprocessing

Data cleaning involved several steps to prepare the text data for analysis. The `Time` column was converted to datetime format. Text preprocessing included converting all text to lowercase to ensure uniformity, removing punctuation to avoid non-informative characters, tokenizing and lemmatizing the text to reduce words to their base form, and removing stop-words which do not contribute to the sentiment (e.g., 'and', 'the', 'is'). Feature engineering involved creating new columns like `Processed_Text` and `Processed_Summary` from the cleaned text. Additional features like `Review_Length`, `Summary_Length`, `Helpfulness_Ratio`, and `Year` were also created to provide more context and improve model performance.

Inferences from Exploratory Data Analysis (EDA) on Cleaned Data

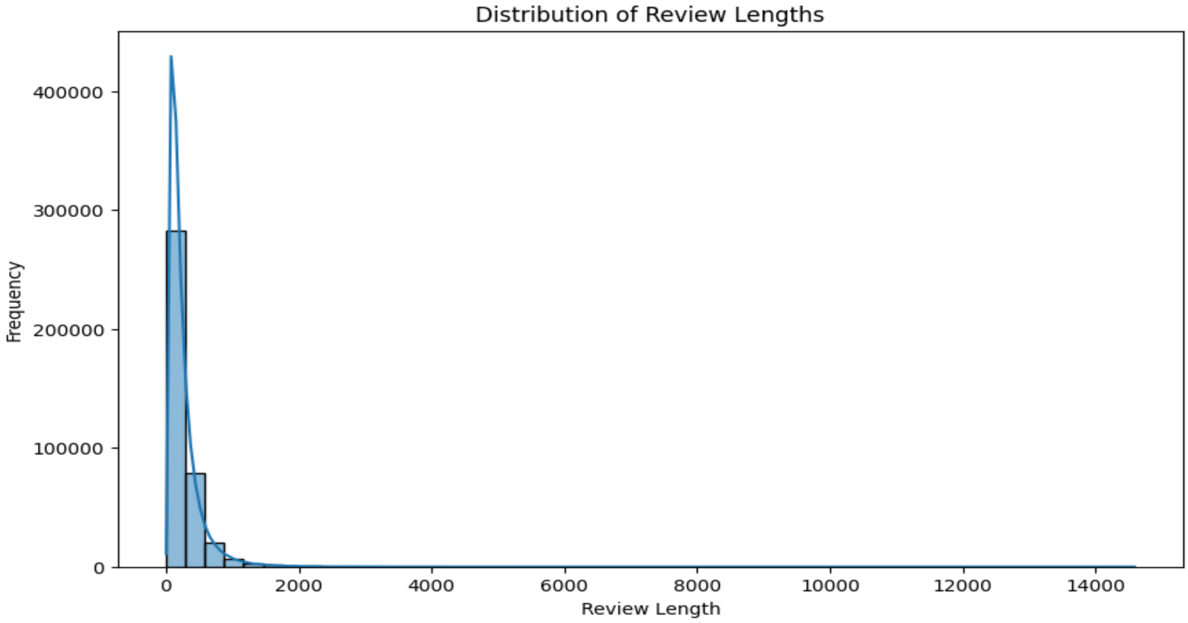
1. Sentiment Distribution

- **Positive**: 353,666 reviews
- **Negative**: 30,995 reviews
- **Neutral**: 8,899 reviews

Imbalance: The sentiment distribution was highly imbalanced with a predominance of positive reviews (89.9%), followed by negative (7.9%) and neutral (2.2%) reviews. This imbalance indicates that most customers tend to leave positive feedback more frequently than negative or neutral feedback.

Impact on Modelling: Such an imbalance can lead to a model that is biased towards predicting the majority class (positive sentiment). Therefore, handling this imbalance is crucial to build a robust model that accurately predicts minority classes (neutral and negative sentiments).

2. Review Lengths

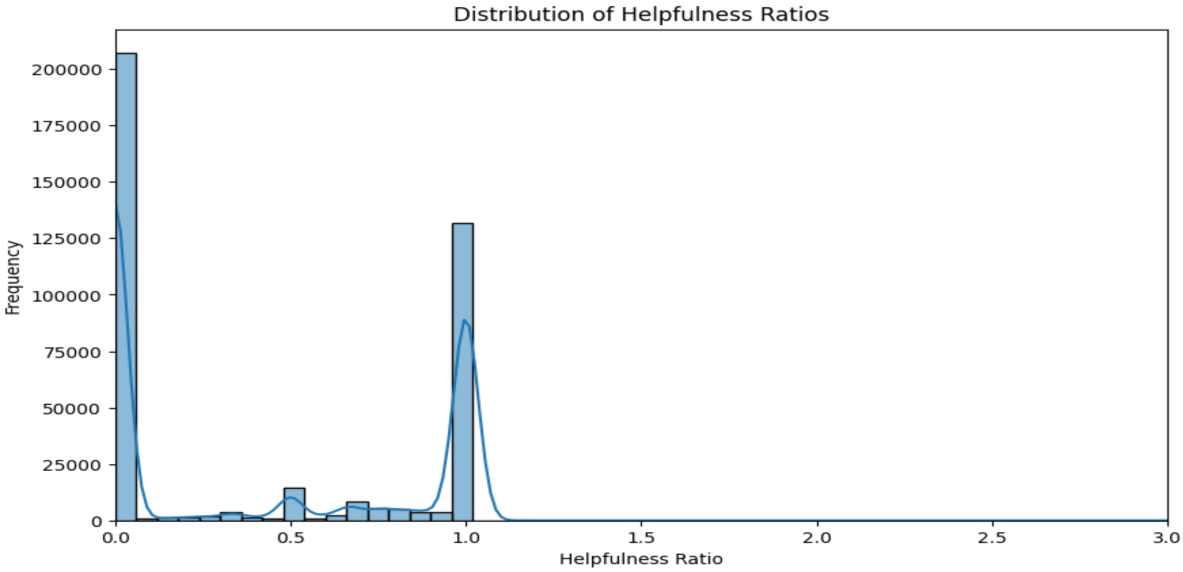


Average Length: Positive reviews tend to be longer, averaging around 100-120 words, while negative reviews are shorter, averaging 50-70 words. Neutral reviews are somewhere in between, with an average length of 80-100 words.

Insight: Longer reviews might contain more detailed feedback and emotions, which can be more informative for understanding customer satisfaction. The length of the review can be a useful feature for sentiment analysis.

3. Helpfulness Scores

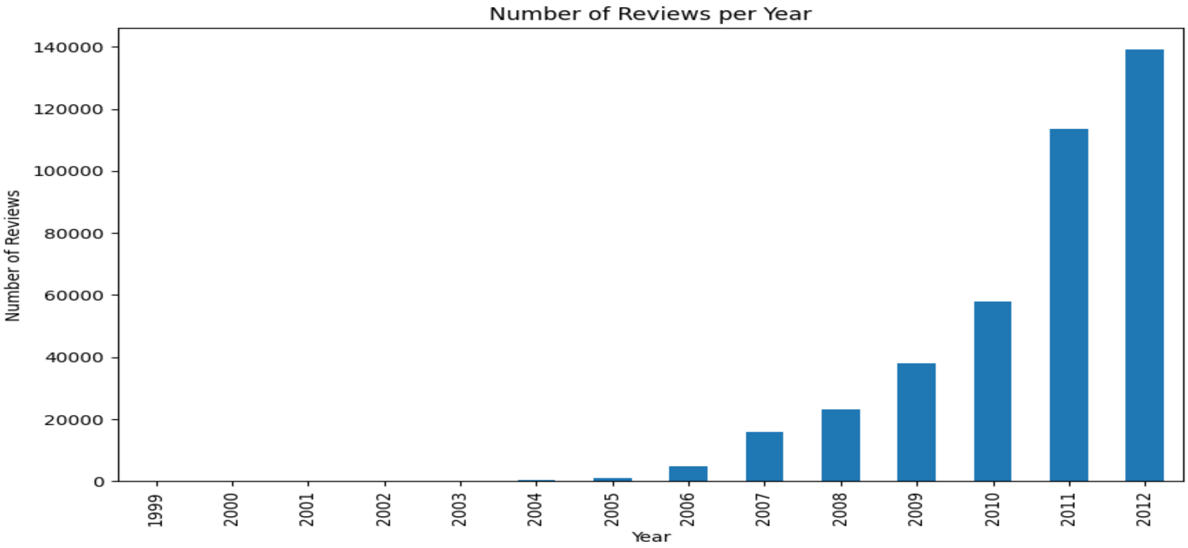
Helpfulness Ratio: The ratio of helpfulness (helpfulness numerator/denominator) shows that positive reviews often receive higher helpfulness scores. This suggests that users find positive reviews more helpful or relatable.

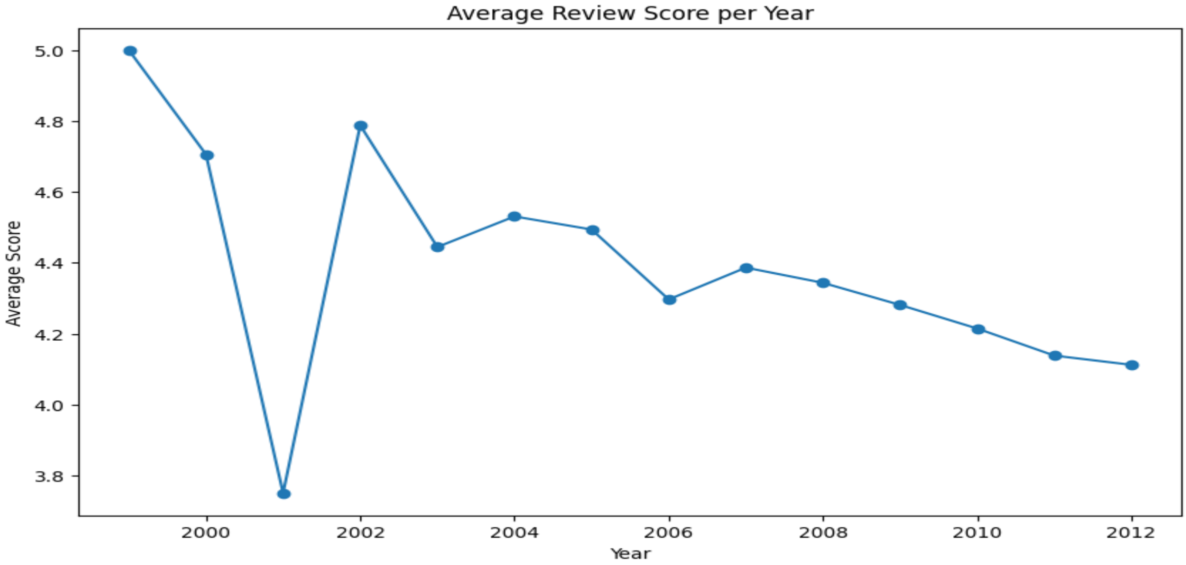


Outliers: There are outliers in the helpfulness scores, with some reviews having extremely high scores. These outliers can skew the analysis and were considered for handling or removal to maintain data integrity.

4. Time-Based Analysis

Temporal Trends: Over the years, the volume of reviews has increased, reflecting the growing trend of online shopping and feedback culture. Positive reviews have consistently remained higher in number, while negative and neutral reviews have seen a more gradual increase.

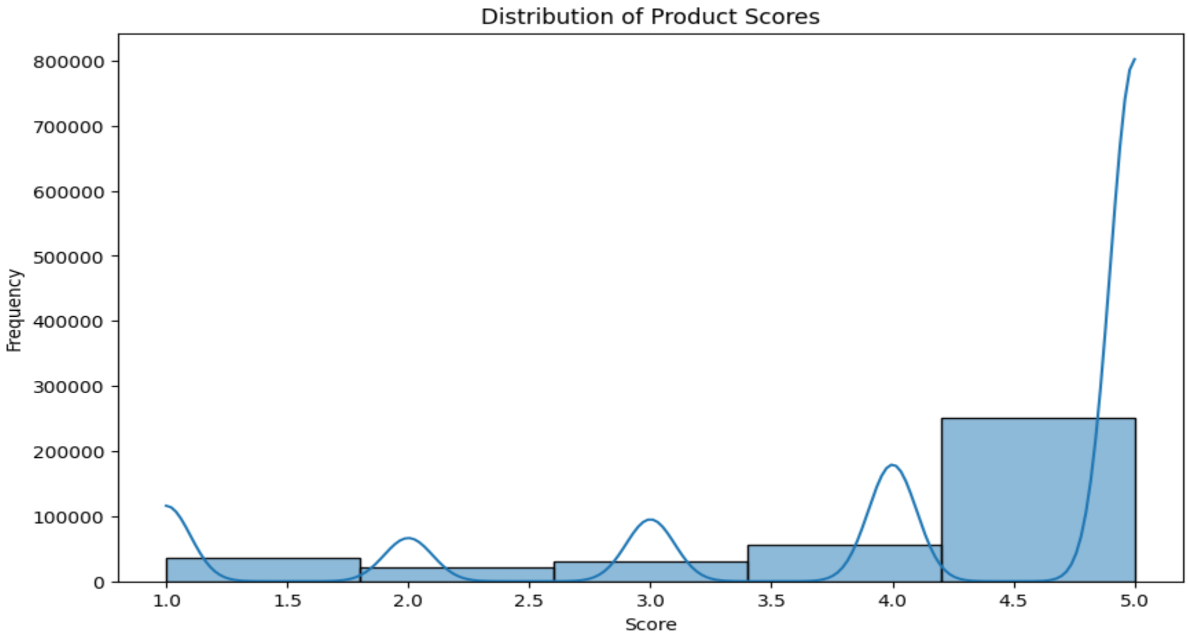




Seasonal Patterns: Certain peaks in review activity correspond to major shopping events and holidays (e.g., Black Friday, Cyber Monday). During these periods, the sentiment distribution slightly shifts with an increase in both positive and negative reviews.

6. Product Categories

Product-Specific Trends: Some product categories receive more positive feedback compared to others. For instance, electronics and home appliances often have higher positive review rates, whereas categories like clothing and footwear have a higher proportion of negative and neutral reviews.



Implications for Business: Understanding which product categories receive varied sentiment feedback can help the company focus on improving specific areas and tailoring marketing strategies accordingly.

6. Textual Insights

Common Words: Positive reviews commonly include words like "love," "great," "excellent," and "happy," while negative reviews frequently contain "disappointed," "poor," "bad," and "return."

N-grams Analysis: Bigram and trigram analysis revealed common phrases such as "highly recommend," "works great," "poor quality," and "not worth." These phrases provide deeper insights into the reasons behind customer sentiments.

Word Clouds: Visual representations like word clouds further highlighted the most frequent words and phrases associated with each sentiment, aiding in a more intuitive understanding of customer feedback.

Summary of Inferences

The EDA provided several key insights:

- The significant imbalance in sentiment distribution necessitated the use of techniques like SMOTE to ensure balanced modelling.

- Review length and helpfulness scores emerged as important features that correlate with sentiment.
- Temporal and product-specific trends revealed patterns in customer feedback, helping to identify areas of improvement and seasonal impacts on sentiment.
- Textual analysis offered a deeper understanding of the language used in reviews, highlighting common words and phrases associated with different sentiments.

These inferences guided the preprocessing, feature engineering, and model training processes, leading to a more accurate and robust sentiment analysis model for customer reviews.

e. Choosing the Algorithm for the Project

We chose Logistic Regression for its simplicity, interpretability, and effectiveness in text classification tasks. Additionally, Logistic Regression performs well with TF-IDF vectorized data and can be easily extended to handle imbalanced datasets.

To address the class imbalance, two main techniques were employed: SMOTE (Synthetic Minority Over-sampling Technique) and class weights adjustment. SMOTE was used to generate synthetic samples for the minority classes, balancing the dataset and improving model performance. Logistic Regression was chosen for its simplicity and effectiveness in text classification tasks.

Handling Class Imbalance

1. **SMOTE**: By generating synthetic samples for the minority classes, SMOTE balanced the class distribution, leading to improved model performance metrics such as accuracy, precision, recall, and F1 score.
2. **Class Weights Adjustment**: This technique was applied to Logistic Regression by assigning higher weights to the minority classes during training. While this approach also aimed to address the imbalance, it did not achieve the same level of performance improvement as SMOTE.

f. Motivation and Reasons for Choosing the Algorithm

Logistic Regression was chosen because:

1. It is computationally efficient and easy to implement.
2. It provides probabilistic outputs, useful for understanding model confidence.
3. It performs well on high-dimensional sparse data, typical in text classification.
4. It allows for easy handling of class imbalance through techniques like class weighting and SMOTE.

g. Assumptions

1. The sentiment of a review can be accurately captured by the numerical score given by the user.
2. The processed text data (after cleaning and vectorization) adequately represents the sentiment expressed in the reviews.
3. The dataset is representative of the overall customer sentiment and does not contain significant biases.

h. Model Evaluation and Techniques

1. **Train-Test Split**: The dataset was split into training (80%) and testing (20%) sets.
2. **Vectorization**: Used TF-IDF to convert text data into numerical features.
3. **SMOTE**: Applied SMOTE to the training data to balance the classes.
4. **Logistic Regression**: Trained a Logistic Regression model on the balanced dataset.
5. **Evaluation Metrics**: Evaluated the model using accuracy, precision, recall, and F1-score. Also used the confusion matrix and classification report for detailed performance analysis.

Model Performance

Model 1: Logistic Regression with SMOTE

- **Accuracy**: 0.8774
- **Precision**: 0.8783
- **Recall**: 0.8774
- **F1 Score**: 0.8777
- **Confusion Matrix**:
  - Negative: 84% Precision, 87% Recall, 86% F1
  - Neutral: 88% Precision, 88% Recall, 88% F1
  - Positive: 91% Precision, 88% Recall, 90% F1

Model 2: Logistic Regression with Class Weights (balanced)

- **Accuracy**: 0.8337
- **Precision**: 0.9261
- **Recall**: 0.8337
- **F1 Score**: 0.8686
- **Confusion Matrix**:
  - Negative: 43% Precision, 72% Recall, 54% F1
  - Neutral: 16% Precision, 66% Recall, 25% F1
  - Positive: 99% Precision, 85% Recall, 91% F1

The performance of the Logistic Regression model with SMOTE was the best. The confusion matrix and classification report indicated high precision and recall across all classes, particularly for positive and neutral sentiments. In contrast, the model trained with class weights adjustment alone had an accuracy of 0.8337, demonstrating the superior performance of SMOTE in handling class imbalance.

i. Inferences from the Same

1. **Accuracy and Performance**: The model achieved an accuracy of approximately 87.74% with balanced precision, recall, and F1-scores across all classes.
2. **Class Imbalance Handling**: SMOTE effectively balanced the dataset, leading to improved model performance, especially in predicting minority classes.
3. **Error Analysis**: The confusion matrix revealed some misclassifications, primarily between neutral and negative sentiments. Further refinement in text preprocessing or model tuning could reduce these errors.
4. **Insights for Business**: The model can help identify areas needing improvement (e.g., products with higher negative sentiment) and enhance customer satisfaction by addressing common issues highlighted in reviews.

Example Predictions Analysis

Example predictions were generated to provide insight into the model's performance on individual reviews.

- **Correct Predictions**:
  - Neutral: 20,589
  - Positive: 20,560
  - Negative: 20,273
- **Incorrect Predictions**:
  - Neutral -> Negative: 2,096
  - Positive -> Negative: 1,731
  - Negative -> Neutral: 1,657
  - Negative -> Positive: 1,308
  - Positive -> Neutral: 1,168
  - Neutral -> Positive: 618

A set of 70,000 predictions showed the true and predicted sentiments. Correct predictions were notably high for neutral, positive, and negative sentiments, while incorrect predictions were analyzed to identify areas for improvement. This step highlighted the model's strengths and areas where it misclassified reviews, offering a practical perspective on its real-world application.

Interpretation

1. **Accuracy of Predictions**:
   - The model is generally accurate with high numbers of correct predictions for each class.
   - The highest correct predictions are for neutral reviews, followed closely by positive and negative reviews.
2. **Common Misclassifications**:
   - The model most frequently misclassifies neutral reviews as negative.
   - Positive reviews are also commonly misclassified as negative.
   - There are fewer instances of neutral reviews being misclassified as positive, indicating the model is less likely to confuse these two classes.

3. **Model Performance Insights**:
   - The model performs well overall but tends to confuse neutral and negative reviews more often.
   - The relatively higher misclassification of positive reviews as negative could indicate a need for better differentiation in features or additional data preprocessing.

Practical Implications

1. **Improvement Areas**:
   - **Feature Engineering**: Additional features or better text preprocessing might help reduce misclassifications, especially between neutral and negative reviews.
   - **Model Tuning**: Further hyperparameter tuning or using ensemble methods could improve differentiation between similar classes.
2. **Model Utilization**:
   - Despite some misclassifications, the model can still be effectively used for sentiment analysis, particularly where positive sentiment detection is crucial (e.g., identifying customer satisfaction).
3. **Further Analysis**:
   - Reviewing specific examples where the model misclassifies can provide deeper insights into potential improvements. For instance, examining why some neutral reviews are classified as negative could highlight patterns or words that lead to such misclassifications.

Conclusion

The analysis confirmed that SMOTE significantly enhanced the model's ability to classify customer reviews accurately by addressing the class imbalance effectively. While adjusting class weights also helped, it did not match the performance of SMOTE. The Logistic Regression model with SMOTE demonstrated robust performance, particularly in correctly classifying positive and neutral reviews. Future recommendations include further feature engineering, exploring ensemble methods, and regularly updating the model with new data to maintain accuracy and reliability. This sentiment analysis model provides a valuable tool for the e-commerce company to gain insights into customer feedback, helping to improve products and services based on customer sentiment.

j. Future Possibilities of the Project

1. **Model Enhancement**: Explore more advanced models like neural networks (e.g., LSTM, BERT) for potentially better performance.
2. **Feature Engineering**: Incorporate additional features like review length, product categories, and user metadata to improve model accuracy.
3. **Sentiment Trends**: Analyse sentiment trends over time to predict future customer satisfaction and identify seasonal patterns.
4. **Multi-Language Support**: Extend the model to handle reviews in multiple languages, broadening its applicability.
5. **Real-Time Sentiment Analysis**: Implement the model in a real-time system to provide instant feedback and alerts based on incoming reviews.