Project Report: Automated Customer Review Analysis

1. Introduction

The digital landscape is rich with customer feedback in the form of online reviews. For a business, manually analyzing this vast amount of unstructured data is inefficient and often impractical. The objective of this project was to develop a comprehensive NLP system to automate the analysis of customer reviews, transforming raw feedback into actionable insights. The solution leverages a multi-faceted approach, combining supervised learning for sentiment analysis, unsupervised learning for topic clustering, and generative AI for summarization, all integrated into a live web application.

2. Data Preprocessing

The project utilized a comprehensive dataset of customer reviews. The initial data exploration revealed the presence of missing values, which were addressed by dropping any rows with null entries to ensure a clean dataset for subsequent analysis.

A critical challenge identified was the imbalanced distribution of sentiment classes (positive, neutral, negative). To mitigate this, the RandomOverSampler technique was applied to the training data. This process synthesized new examples for the minority classes, creating a more balanced dataset that prevented the model from being biased towards the majority class and significantly improved its ability to accurately classify all sentiment types.

3. Methodology

3.1 Review Classification (Supervised Learning)

For sentiment classification, a pre-trained distilbert-base-uncased model was fine-tuned on the preprocessed dataset. This approach was chosen to leverage the powerful and nuanced language representations learned by the model during its pre-training phase, allowing for high accuracy without the need for extensive training from scratch. The fine-tuning process adapted the model specifically to the task of sentiment classification for customer reviews.

3.2 Customer Review Clustering (Unsupervised Learning)

To identify recurring themes and topics within the reviews, an unsupervised learning approach was employed. The texts were clustered based on content similarity, grouping related reviews together. This method eliminates the need for manual labeling and provides a high-level overview of common customer issues and feedback, making it easier to identify emerging trends.

3.3 Review Summarization (Generative AI)

The project utilized generative AI to transform clustered reviews into concise, readable summaries. This was achieved using a two-step process:

Zero-Shot Classification: Reviews were first categorized into specific themes using a zero-shot classification model, which can classify text into categories it has not been explicitly trained on.

Text Summarization: A generative AI summarization model was then used to create a short, coherent summary for each category. This provides stakeholders with immediate, actionable insights without having to read through thousands of individual reviews.

4. Results & Analysis

The fine-tuned DistilBERT model for sentiment analysis achieved strong performance metrics on the validation dataset. The model's accuracy, precision, recall, and F1-score demonstrated its effectiveness in correctly identifying positive, neutral, and negative sentiment. The clustering component successfully grouped similar reviews, with a qualitative analysis of the clusters showing clear thematic cohesion. The generative AI-powered summarization provided insightful and accurate summaries, demonstrating the model's ability to extract key information from a large corpus of text.

5. Conclusion & Future Work

This project successfully demonstrated the power of a comprehensive NLP pipeline for automating customer review analysis. The combination of supervised, unsupervised, and generative AI techniques provides a robust system for converting unstructured text into valuable business intelligence.

For future work, the project could be improved by exploring different advanced transformer models for sentiment analysis, using a more sophisticated clustering algorithm, and incorporating entity recognition to identify specific products and features mentioned in the reviews.

Additionally, the Gradio application could be expanded to include more interactive features and a dashboard for a more complete user experience.