Project NLP | Business Case: Automated Customer Reviews

1. Introduction

This report documents the implementation of an NLP-based sentiment analysis system for customer reviews. The objective is to classify reviews into positive, neutral, or negative sentiments, as well as summarize reviews by rating and product category. The project was conducted in two main parts:

1. **Sentiment Analysis**:
   - Traditional NLP & Machine Learning.
   - LSTM-based Sequence-to-Sequence Modeling.
   - Transformer-based Modeling using Hugging Face API.
2. **Review Summarization**:
   - Summarize grouped reviews by rating (0–5) and product category.
   - Handle large numbers of categories by focusing on the top-K categories.

---

2. Data Preprocessing

The data was preprocessed to ensure consistency and readiness for modeling. The steps included:

- Cleaning text by converting to lowercase, removing special characters, and stripping whitespace.
- Handling missing values by removing rows with null critical fields.
- Balancing class distribution using SMOTE to oversample minority classes.
- Tokenizing and vectorizing text data using TF-IDF and Tokenizer for LSTM.

The cleaned dataset was saved for downstream processing.

---

3. Modeling Approaches

Three different approaches were implemented:

3.1 Traditional NLP & Machine Learning

- **Models**: Logistic Regression, Random Forest, and Naive Bayes.
- **Text Vectorization**: Text was vectorized using TF-IDF.
- **Class Imbalance Handling**: SMOTE was applied to address class imbalance.

3.2 LSTM-based Sequence-to-Sequence Modeling

- **Architecture**: Bidirectional LSTM layers with embedding and dropout were used.
- **Text Processing**: Tokenized text sequences were padded to a fixed length.
- **Class Imbalance Handling**: Class weights were adjusted during training.

3.3 Transformer-based Modeling using Hugging Face API

- **Model**: A pre-trained BERT model (bert-base-uncased) was fine-tuned for sentiment classification.
- **Text Processing**: Tokenization and padding were performed using the BERT tokenizer.
- **Evaluation Metrics**: Metrics including accuracy, precision, recall, and F1-score were computed.

---

4. Review Summarization Using Generative AI

Objective

Summarize reviews by:

1. **Rating (1-5)**.
2. **Product Category** (e.g., top-K categories).

Methodology

1. **Dataset Preparation**:
    - Ensured the dataset included categories, reviews.rating, and reviews.text columns.
    - Filtered missing or empty reviews.
2. **Top-K Category Selection**:
    - Selected the top 10 categories based on review count.

    - Grouped reviews by categories and reviews.rating.

3. **Summarization with Generative AI**:
    - Utilized the **T5-based model** (google/flan-t5-base) for text summarization.
    - Generated concise summaries for grouped reviews.

4. **Output**:
    - Saved summarized reviews in a CSV file for further analysis.

---

5. Results and Analysis

Model Performances

- **Random Forest**: 98.8% accuracy (best overall).
- **BERT**: 94.8% accuracy with strong minority class handling.
- **Logistic Regression**: 90.5% accuracy.
- **Naive Bayes**: 83.7% accuracy.
- **LSTM**: 71% accuracy with balanced metrics.

---

6. Web Application for Sentiment Prediction

A web-based application was developed using **Streamlit** to allow users to upload a CSV file containing customer reviews and predict their sentiment using the trained Random Forest model. Below is a summary of the application features and functionality:

Application Features

- **Model Integration**: The trained Random Forest model and TF-IDF vectorizer were hosted and integrated into the application.
- **CSV File Upload**: Users can upload a CSV file containing a reviews.text column with the customer reviews to analyze.
- **Preprocessing**: Reviews are preprocessed (e.g., filling missing values and transforming text into TF-IDF vectors) before prediction.
- **Prediction**: The model predicts the sentiment of each review as **Positive**, **Neutral**, or **Negative**.

- **Summary Table**: A summary of the sentiment distribution is displayed as a table.
- **Download Option**: Users can download the CSV file containing the original reviews and their predicted sentiments.

---

7. Key Observations

Insights

- **Random Forest**: Best for accuracy and scalability.
- **BERT**: Effective for complex text patterns and minority classes.
- **LSTM**: Requires further tuning for better results.
- **Naive Bayes**: Limited performance with imbalanced data.
- **Summarization**: Generative AI effectively condenses grouped reviews by category and rating.

---

8. Conclusion

The project demonstrated the strengths and limitations of different modeling approaches for sentiment analysis:

- **Random Forest**: Delivered the highest accuracy (98.8%) and performed well across all classes, making it suitable for datasets where computational efficiency is not a concern.
- **Transformer-based Model (BERT)**: Achieved a high accuracy of 94.8% and excelled in handling class imbalance, making it ideal for nuanced text analysis.
- **LSTM**: Overall accuracy (71%) was lower, indicating the need for possible further tuning.
- **Summarization**: Leveraging Generative AI effectively summarized reviews, providing actionable insights grouped by category and rating.