Problem Statement

The goal is to automatically classify customer complaints into one of 11 predefined categories. This is a multi-class text classification problem. The solution uses natural language processing (NLP) techniques, including text preprocessing, word embeddings, and deep learning models to accomplish this.

Dataset

The dataset, `mariotthoteldatasetv3.csv`, contains 21,129 entries with two main columns:

- **Complaint Text**: The raw text of the customer's complaint.
- **Category**: The label indicating the complaint type. There are 11 unique categories, and the data is well-balanced across them, with most categories having around 2,000 entries.

The average length of the complaint texts is relatively short, with most complaints containing between 4 and 10 words. This is important context for the choice of models and feature engineering.

Methods

1. Preprocessing

The raw complaint text was prepared for machine learning using a series of standard NLP steps:

- **Lowercase Conversion**: All text was converted to lowercase to ensure consistency (e.g., "Wifi" and "wifi" are treated as the same word).
- **Punctuation Removal**: Punctuation marks were removed to prevent them from being treated as unique tokens.
- **Stopword Removal**: Common English words like "the," "a," and "is" (stopwords) were removed as they generally don't contribute to the classification task.
- **Lemmatization**: Words were reduced to their base or root form (e.g., "elevators" becomes "elevator"). This helps to reduce the vocabulary size and improves model generalization.

2. Word Embeddings

Word embeddings are a crucial step for deep learning models. They convert text into numerical vectors that capture semantic meaning. Two approaches were explored:

- **Word2Vec (Trained on Custom Data)**: A Word2Vec model was trained directly on the preprocessed text from the dataset. This creates 100-dimensional vectors for each word, tailored to the specific vocabulary of the customer complaints. The word "room" was found to be most similar to words like "tv", "channel", and "thermostat", which aligns with the context of hotel complaints.
- **Pre-trained Word2Vec (Google News)**: The Google News pre-trained Word2Vec model, which has 300-dimensional vectors for a massive vocabulary, was also used. This approach leverages knowledge from a much larger corpus, which can be beneficial for understanding general word relationships.

3. Modeling

Three different deep learning models were implemented and evaluated:

- **Feed-Forward Neural Network (FFNN)**: A simple neural network was built to serve as a baseline model. It takes the average of the word vectors (embeddings) for each complaint as input. The model consists of two hidden layers and an output layer for the 11 categories.
- **Recurrent Neural Network (RNN)**: This model processes the complaint text sequentially. It uses an embedding layer, a bidirectional RNN layer, and a final linear layer for classification. It captures the order of words, which can be important for context.
- **Long Short-Term Memory (LSTM)**: A more advanced type of RNN, the LSTM is designed to better handle long-term dependencies in sequences. Like the RNN, it uses an embedding layer and a bidirectional LSTM layer.

4. Evaluation

Each model was trained for 20 epochs using the Adam optimizer and a weighted Cross-Entropy Loss function to handle potential class imbalances. The performance was evaluated on a held-out test set (20% of the data) using the following metrics:

- **Accuracy**: The overall percentage of correctly classified complaints.
- **Classification Report**: Provides precision, recall, and F1-score for each category.
- **Confusion Matrix**: A table showing the number of correct and incorrect predictions for each category, revealing which classes are most often confused with one another.
