**Subtheme Sentiment Analysis for Customer Reviews**

**Objective**

The objective of this assignment is to develop an approach to identify subthemes and their respective sentiments within customer reviews. Subthemes represent aspects or problems mentioned in the reviews, with associated sentiments (positive, negative, or neutral).

**Approach**

Our approach involves several steps to preprocess the text data, perform sentiment analysis, and identify subthemes within the reviews. The detailed steps are outlined below:

1. **Data Loading**:
   * The dataset is loaded from a CSV file using the Pandas library.
2. **Column Renaming**:
   * The first column, which contains the review text, is renamed to review to standardize its reference.
3. **Text Preprocessing**:
   * Tokenization: The text is split into individual words using the word\_tokenize function from the NLTK library.
   * Stop Words Removal: Common stop words (e.g., "and", "the", "is") are removed from the text to focus on meaningful words.
4. **Sentiment Analysis**:
   * The sentiment of each review is analyzed using the TextBlob library. Reviews are classified as positive, negative, or neutral based on their polarity scores.
5. **Subtheme Identification**:
   * Specific subthemes (e.g., garage service, wait time, incorrect tyres) are identified within each review based on the presence of relevant keywords.
   * The sentiment for each identified subtheme is determined using the TextBlob analysis.
6. **Output Generation**:
   * The processed data, including the original review, the overall sentiment, and the identified subthemes with their sentiments, is saved to a new CSV file (processed\_reviews.csv).

### Code Implementation

### The code implementation for the above steps is attached in the file “ subtheme\_sentiment\_analysis.py”.

### Improvements and Possible Problems

1. **Improvements**:
   * **Expand Subtheme Keywords**: The list of keywords for each subtheme can be expanded to capture more variations and synonyms.
   * **Use Advanced NLP Techniques**: Leveraging more advanced NLP models (e.g., BERT, GPT) could improve the accuracy of sentiment analysis and subtheme identification.
   * **Handle Negations**: Incorporating rules to handle negations (e.g., "not good" should be negative) can improve sentiment accuracy.
2. **Possible Problems**:
   * **Ambiguity in Text**: Reviews with ambiguous language or sarcasm may lead to incorrect sentiment classification.
   * **Limited Keyword List**: The current keyword list for subthemes may not cover all possible variations, leading to incomplete subtheme identification.

**Conclusion**

The provided approach successfully preprocesses customer reviews, performs sentiment analysis, and identifies subthemes with their respective sentiments. The final processed data is saved for further analysis or reporting. This approach can be further refined by expanding keyword lists and leveraging advanced NLP models to improve accuracy and coverage.