**Problem Statement**

Build, implement, and evaluate an End-to-End Text Classification for Customer Feedback Analysis or Sentiment Analysis in Social Media or Topic Modeling for Document Clustering using spaCy and other Python libraries. The goal of the project is to develop a system that can

* Classify customer feedback into categories like complaints, suggestions, compliments, etc.,

**OR**

* Determine the sentiment (positive, negative, neutral) of social media posts

**OR**

* Identify different topics within a collection of documents and cluster similar documents together.

**Dataset description**

For feedback analysis, gather customer feedback data, which might be labeled or unlabeled.

**OR**

For sentiment analysis, you would need text data labeled with sentiments.

**OR**

For topic modeling, a large collection of unlabeled documents is required.

**Tasks to be completed**:

**Marks: 50**

**1.** **Data Collection and Preprocessing (****2+3+1 = 6 marks)**

* Collect relevant datasets.
* Preprocess the text data that will involve different NLP tasks where needed.
* Split the dataset into a training set and a test set.

**2. Feature Extraction and Feature Engineering (3+3 = 6 marks)**

* Extract and select numerical features from text data using NLP techniques - Bag of Words (BoW), TF-IDF, or word embeddings (Word2Vec, GloVe).

**3. Model Selection (6 marks)**

* Feedback/Sentiment Analysis: Choose a classification model like Naive Bayes, Logistic Regression, Random Forest, or neural networks.
* Topic Modeling: Use unsupervised models like clustering, LDA, or other unsupervised methods.

**4.** **Model Training and Evaluation (2 + 2 + 2 = 6 marks)**

* Train the model on the training dataset.
* Evaluate the model using appropriate metrics:
  + Classification: Accuracy, Precision, Recall, F1-Score.
  + Topic Modeling: Perplexity, Topic Coherence.
* Fine-tune model parameters for optimal performance.

**5. Model Refinement and Optimization (3+3 = 6 marks)**

* Experiment with different autoencoder architectures, hyperparameters, and reconstruction error thresholds to fine-tune the model's performance.
* Apply techniques such as learning rate scheduling, regularization, and dropout to improve the model's generalization and robustness.

**6. Documentation (10 marks) and 15 mins Presentation (10 marks)**

Document the entire project, including the dataset description, preprocessing steps, model architecture, hyperparameter settings, and evaluation results.

Create a presentation summarizing the project's objectives, methods, findings, and recommendations.