1. **Text Preprocessing**: This involves cleaning and formatting the data before we start with analysis. Techniques used are:
   * **Tokenization**: Breaking text into tokens (words, sentences, etc.). In Python, we will start out using spacy library for this but if it doesn’t work, then we can use nltk.
   * **Lowercasing**: Convert all the text to lower case. This is required as python interprets ‘dog’ and ‘DOG’ differently.
   * **Stopwords Removal**: Stopwords are common words that carry less important meaning than keywords. We will start out using spacy library for this but if it doesn’t work, then we can use nltk.
   * **Punctuation Removal**: Remove all punctuation marks from the text. We will start out using spacy library for this but if it doesn’t work, then we can use nltk.
   * **Stemming/Lemmatization**: Reduce words to their root form. For example, ‘running’ to ‘run’. We will start out using spacy library for this but if it doesn’t work, then we can use nltk.
2. **Feature Extraction**: Techniques to extract features from the text:
   * **Bag of Words**: Represents text as a bag of words, disregarding grammar and word order but keeping multiplicity.
   * **TF-IDF**: Stands for Term Frequency-Inverse Document Frequency. It is a statistical measure used to evaluate the importance of a word in a document.
   * **Word Embeddings**: They are a type of word representation that allows words with similar meaning to have similar representation. This is for encoding purposes.
3. **Model Building**: After feature extraction, you can build models to solve your NLP task. You have to train the pretrained LLM model either on your data or other datasets to boost the accuracy of the LLM that is being used. Some common models include:
   * **Naive Bayes Classifier**: It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors.
   * **Logistic Regression**: It is a machine learning algorithm for binary classification.
   * **Neural Networks**: They are a set of algorithms modeled after the human brain, that are designed to recognize patterns.
4. **Evaluation Metrics**: After building your model, you need to evaluate their performance. Some common metrics include:
   * **Precision**
   * **Recall**
   * **F1-Score**
   * **Bleu**
   * **Perplexity**