<u>Task 3.1:</u>

Let’s say you are given a large amount of textual data- messages, emails, books, etc. Before performing any operations on this data, it is necessary to clean and preprocess the data (removing unnecessary words or symbols, etc.). Explain how you would go about preprocessing. What different steps would be followed? Why are they necessary?

<u>Answer</u>

Data cleaning and preprocessing is an integral part of the Data Mining process to get ideal results from the data that we provide. While it may be slightly more complex when it comes to textual data rather than numerical, it certainly is possible and incredibly essential.

If I were given a large amount of textual data such as in messages, emails, books, I would go about the data cleaning and preprocessing in the following way: -

- Handling any missing data: I would delete any data which has missing or any inconsistent value. For text especially, I would delete blanks and deal with algorithms to efficiently and quickly get rid of these.
- Removing Duplicates: Duplicates can distort analysis results and lead to biased insights. This can increase the favourability of an outcome in a specific direction which is not ideal. Any excessively appearing text should be deleted.
- Addressing Inconsistencies: Inconsistent textual-data fonts, sizes, languages, or excessive jargon used can create confusion and hinder data analysis. Data cleaning involves standardizing and transforming data to ensure consistency, facilitating seamless integration and analysis.
- Removing Punctuation and Spelling errors: When it comes to textual data, it is crucial for the data to follow a certain grammatical syntax. We can either correct this data or get rid of it altogether. These errors can cause problems later and give us inconsistent results.
- Outlier Detection: Outliers, sudden changes in the data, can significantly impact statistical analysis and modelling. I would employ techniques like statistical methods, clustering, or machine learning algorithms to identify and handle outliers appropriately.

- Quantifying the Data: Since a large amount of textual data has been given, it is crucial to quantify this data by using numbers associated with certain fields. This will make it easier to use the given data set and run it through several algorithms that require a quantified set of data.
- Normalization: Upon quantifying this data, it is important to normalize the entire dataset in order for it to work cohesively with different fields in itself. For example, we can scale the values between 0 and 1 so that there is a common field for the values corresponding to the textual data.

These techniques are crucial for extracting ideal results from the data we provide. These processes can be implemented using various algorithms and codes in languages such as Python. It is also important to familiarise myself with the various libraries and frameworks that do so.