Task 3.1

It is necessary to clean and process data as we need to work on the important part of the data that is the data we are concerned about to draw conclusions from the data or to make some prediction. It is necessary while processing we need to remove outliers , inconsistent data and unnecessary information (noise).

This can be done by following steps:-

1) Lower Casing: -
   All the textual data should be normalized by converting it to lowercase. This can be easily done using Python's .lower() function. The reason for this step is to ensure consistency across the dataset. For instance, "Hello World", "hello World", and "hello world" should all be treated as the same text. By converting everything to lowercase, we eliminate any discrepancies caused by different capitalizations, ensuring that words are consistently represented, which improves the quality of text processing and analysis.

2) Data Cleaning: Handling Emoticons, Emojis and Special Characters-
   In this step, we remove unnecessary information, often referred to as "noise," from the text data. Noise can lead to unwanted outputs and hinder the effectiveness of text processing. By removing special characters, emojis, and other non-essential elements, we ensure that the data is focused and clean for analysis. For example, social media data often includes symbols like @, #, and emojis, which may not add value to the text analysis. Removing these elements helps in reducing noise and enhances the overall quality of the data.

3) Stop Words : -
   In this step we remove stop words. These are words which occur the most in any natural language data set like the big data we have in which there are some words like "and" ." the" , "is" etc These words provide little or no semantic meaning . Having these words increase the size of the data set and an additional amount of time is required to analyse data. So after removing it the prediction and analysis time is reduced. Various libraries such as 'Natural Language Toolkit' (NLTK), 'spaCy', and 'Scikit-Learn' can be used to remove stopwords. These libraries have collection of major stop words. But we can also make a collection of our own stop words as the words which are used often normally may be important for us in some analysing process so it depends on data set.

4) Stemming: -
   Stemming means converting words to their root or base form. If the word is automate, automation, automating or the like it will convert it to automat. It basically cuts of endings and extracts the core form of data.There are various algorithms for stemming one is Porter Stemmer algorithm Snowball Stemmer algorithm Lovins Stemmer algorithm.

5) Lemmatization: -
   Lemmatization does the same thing that is shortens the words but the meaning is not lost.In the above eg meaning of automate was lost when it became automat . It has a predefined dictionary it checks the words before removing thee endings and shortening the word . Eg:- “cars” ,”car’s” and similar words are converted to “car”.

6) Tokenization: -
   In this basically all the long strings and sentences are converted to tokens so that we can deal with things in bits and pieces . This step is necessary for further steps like stemming and Lemmatization. Its easier to deal with tokens and perform algorithms on them .

7) Removing Punctuations, Numbers and URLs:-
   It is necessary to remove punctuations , numbers and urls because they are not required for text analysis .Advantages of removing it are same as removing special characters and emojis as they increase the length of the data set and takes time in analysis.