NLP task 1-

Sentiment analysis:

To start with we will take a dataset of sample tweets which have some positive ,some negative and remaining neutral tweets. For this we can dowload the tweet samples from nltk. Then we will simplify the data for processing by splitting the tweets into smaller strings.This splitting will be done at spaces or punctuations. We can use the inbuilt tokenised() method for this.

Next we will remove any part of the text that does not add meaning or information to the tweet from the dataset. Example:

The most common words in a language like “is”, “the”, and “a”. They are generally irrelevant when processing language, unless a specific use case warrants their inclusion. We can import stopword set of English to check and remove all unnecessary words.

Hyperlinks in Twitter are converted to the URL shortener t.co. Therefore, keeping them in the text processing would not add any value to the analysis.

Twitter handles in replies are preceded by a @ symbol, which does not convey any meaning.

Punctuation and special characters often provide context to textual data, this context is often difficult to process. For simplicity, you will remove all punctuation and special characters from tweets

Ater that we have to group together words with the same meaning but different forms. Without normalization, these words would be treated as different words, even though we need them to be treated as the same word. We will use lemmatization, which is a popular technique of normalization.

Now for the sentiment analysis we will give our program the positive and negative tweets as input and train it to identify the sentiment. Then we will test it on the remaining neutral tweets to check their sentiment and the accuracy of the program. For doing this we will use the Naive Bayes Classifier which works on Bayes conditional probability theorem.

We can also use textblob library which gives us the polarity of the cleaned tweet directly where a polarity greater than zero means a positive tweet