Task 3.1

Preprocessing textual data is an essential step before performing any analysis or modeling. Here are the steps I would follow to preprocess the data:

**Step 1: Tokenization** Tokenization involves breaking down the text into individual words or tokens. This can be done using the **unnest\_tokens()** function from the **tidytext** package in R.

*library(tidytext)*

*library(tidyverse)*

*data <- data %>%*

*unnest\_tokens(output = token, input = text)*

**Step 2: Remove Stop Words** Stop words are common words like "the", "and", "a", etc. that do not add much value to the analysis. These can be removed using the **stopwords** package in R.

*library(stopwords)*

*stop\_words <- stopwords(language = "en")*

*data <- data %>%*

*filter(!token %in% stop\_words)*

**Step 3: Remove Punctuation and Special Characters** Punctuation and special characters can be removed using the **str\_replace\_all()** function from the **stringr** package in R.

*library(stringr)*

*data <- data %>%*

*mutate(token = str\_replace\_all(token, "[[:punct:]]", ""))*

**Step 4: Remove Digits** Digits can be removed using the **str\_replace\_all()** function from the **stringr** package in R.

*data <- data %>%*

*mutate(token = str\_replace\_all(token, "[0-9]", ""))*

**Step 5: Stemming or Lemmatization** Stemming or lemmatization involves reducing words to their base form. This can be done using the **wordStem()** function from the **SnowballC** package in R.

*library(SnowballC)*

*data <- data %>%*

*mutate(token = wordStem(token, language = "en"))*

**Step 6: Remove Infrequent Words** Infrequent words can be removed to reduce the dimensionality of the data. This can be done by filtering out words that appear less than a certain frequency.

*data <- data %>%*

*group\_by(token) %>%*

*filter(n() > 5)*

These are the basic steps involved in preprocessing textual data. The specific steps and techniques may vary depending on the nature of the data and the analysis being performed.