Building A Pre-Processing Function for Cleaning different type of Texts data. This Function will be used to in all of my NLP analysis

```
## Libraries
library(plyr)
library(dplyr)
library(textclean)
library(NLP)
library(tm)
library(textstem)
options(stringsAsFactors = F)
```

```
path <- "/Users/mo/Desktop/Desktop/School/USF/Courses/Fall 2021/NLP/datasets/city_of_SF_tweets.csv"
data <- read.csv(path)
data$Message[4]
```

```
## [1] "The @SFUnified Meal Distribution program will continue serving students throughout the summer. Adults can
pick up meals for children without a child present. Pick up 5 days worth of food every Wednesday! Visit http://sf
usd.edu/schoolfood for locations."
```

```
pre_process_corpus <- function(data, text_col, replace_numbers = FALSE, non_stopwords = NULL,
                               extra_stopwords = NULL, root_gen = NULL, output_corpus = FALSE, replace_emoji = TR
UE, replace_url = TRUE,
                               replace_symbol = TRUE){
  text <- data[, text_col]
  ## replace contraction such as wasn't tp was not
  text <- replace_contraction(data[, text_col])

  ## replace emojies from (<F0><9F><92><9B> Wear a face covering <F0><9F><92><9A>) to Wear a face covering
  if(replace_emoji == T){
    text <- replace_emoji(text)
  }
  ## remove all hyperlinks
  if(replace_url == T){
    text <- replace_url(text)
  }
  ## remove numbers
  if(replace_numbers == T){
    text <- suppressWarnings(replace_number(text))
  }
  ## remove all sym
  if(replace_symbol == T){
    text <- gsub("@", "mention ", text)
    text <- gsub("#", "hashtag ",text)
  }

  text <- tolower(text)
  text <- gsub("[^\001-\177]", '', text, perl = TRUE)
  text <- VCorpus(VectorSource(text))


  stopwords <- stopwords()
  stopwords <- stopwords[which(!stopwords %in% non_stopwords)]
  stopwords <- c(stopwords, extra_stopwords)

  text <- tm_map(text, function(x) {removeWords(x,stopwords)})

  text <- tm_map(text, function(x) {removePunctuation(x)})
  text <- tm_map(text, function(x) {removeNumbers(x)})
  text <- tm_map(text, function(x) {stripWhitespace(x)})


  if(!is.null(root_gen)){
    if(root_gen == 'stem'){
      text <- unlist(lapply(text, function(x) {stem_strings(x$content)}))
      text <- VCorpus(VectorSource(text))
    }

    if(root_gen == 'lemmatize') {
      text <- unlist(lapply(text, function(x) {lemmatize_strings(x$content)}))
      text <- VCorpus(VectorSource(text))
    }
  }
  if(output_corpus == TRUE) {
    return(text)
  } else {return(unlist(lapply(text, function(x) {x$content})))}
}
```

```
clean_text <- pre_process_corpus(data, "Message",
                                 replace_emoji = TRUE,
                                 replace_symbol = T,
                                 replace_numbers = T,
                                 root_gen = "lemmatize",
                                 replace_url = T)
```