622 GROUP PROJECT 3

Jessica Valencia

2023-03-31

```
# Tidyverse and text analysis tools
library(tidyverse)
library(tidytext)

# For decision trees
library(rpart)
library(rpart.plot)
library(rattle)

# For nearest neighbors
library(class)

# For ML evaluation
library(caret)
```

First, we brought in the data file after handcoded sentiment has been conducted.

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.2.2
```

```
datasubset <- read_csv("hw3datasubset.csv")
```

```
## New names:
## Rows: 200 Columns: 30
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (9): date_utc, title, text, subreddit, url, datetime, title_text, Coder... dbl
## (21): ...1, timestamp, comments, WordCount, SentimentGI, NegativityGI, P...
## i Use 'spec()' to retrieve the full column specification for this data. i
## Specify the column types or set 'show_col_types = FALSE' to quiet this message.
## * '' -> '...1'
```

```
#View(hw3datasubset)
```

#REMOVING NEUTRAL SENTIMENT FROM HANDCODED SENTIMENT Then, we removed the observations with a "neutral" hand-coded sentiment from the data set considering that this information is not useful for the purpose of this report.

```
data<-datasubset[datasubset$Handcoded_Sentiment!="Neutral",]

View(data)
```

#CLEANING HANDCODED SENTIMENT We decided to clean the hand-coded sentiment column in order to make sure our results are as optimal and accurate as possible by removing human error.

```
data$Handcoded_Sentiment <- replace(data$Handcoded_Sentiment, data$Handcoded_Sentiment == "Posotive", "

data$Handcoded_Sentiment <- replace(data$Handcoded_Sentiment, data$Handcoded_Sentiment == "Postive", "Po

data$Handcoded_Sentiment <- replace(data$Handcoded_Sentiment, data$Handcoded_Sentiment == "postive", "Po
```