Sentiment Analysis for Mental Health - PREPARING DATASET FOR DistilBERT DEEP LEARNING MODEL

2024-11-27

Loading the original dataset

```
# Set the file path for the CSV file
file_path2 <- "C:/Users/jivko/Documents/Data Analytics, Big Data, and Predictive Analytics/Pe
rsonal Project/Sentiment Analysis for Mental Health/Combined Data.csv"


# Read the CSV file into a dataframe
sentiment_analysis <- read.csv(file_path2, header = TRUE)
```

Printing the first few rows of the dataframe

```
print(head(sentiment_analysis))
```

```
##   X
## 1 0
## 2 1
## 3 2
## 4 3
## 5 4
## 6 5
##                                                                  statement
## 1                                                                oh my gosh
## 2               trouble sleeping, confused mind, restless heart. All out of tune
## 3 All wrong, back off dear, forward doubt. Stay in a restless and restless place
## 4                  I've shifted my focus to something else but I'm still worried
## 5       I'm restless and restless, it's been a month now, boy. What do you mean?
## 6   every break, you must be nervous, like something is wrong, but what the heck
##    status
## 1 Anxiety
## 2 Anxiety
## 3 Anxiety
## 4 Anxiety
## 5 Anxiety
## 6 Anxiety
```

Removing redundant X column

```
sentiment_analysis_use <- sentiment_analysis[, !names(sentiment_analysis) %in% c("X")]
```

```
print(head(sentiment_analysis_use))
```

```
##                                                                 statement
## 1                                                               oh my gosh
## 2               trouble sleeping, confused mind, restless heart. All out of tune
## 3 All wrong, back off dear, forward doubt. Stay in a restless and restless place
## 4                  I've shifted my focus to something else but I'm still worried
## 5       I'm restless and restless, it's been a month now, boy. What do you mean?
## 6   every break, you must be nervous, like something is wrong, but what the heck
##    status
## 1 Anxiety
## 2 Anxiety
## 3 Anxiety
## 4 Anxiety
## 5 Anxiety
## 6 Anxiety
```

Structure of dataset

```
str(sentiment_analysis_use)
```

```
## 'data.frame':    53043 obs. of  2 variables:
##  $ statement: chr  "oh my gosh" "trouble sleeping, confused mind, restless heart. All out
of tune" "All wrong, back off dear, forward doubt. Stay in a restless and restless place" "I'
ve shifted my focus to something else but I'm still worried" ...
##  $ status   : chr  "Anxiety" "Anxiety" "Anxiety" "Anxiety" ...
```

Check for missing values

```
# Check for missing values in each column
colSums(is.na(sentiment_analysis_use))
```

```
## statement    status
##         0         0
```

Distribution of mental health statuses

```
status_counts <- table(sentiment_analysis_use$status)
print(status_counts)
```

```
##
##              Anxiety              Bipolar           Depression
##                 3888                 2877                15404
##               Normal Personality disorder               Stress
##                16351                 1201                 2669
##             Suicidal
##                10653
```

Matching each status count to median (3888)

```
# Load the libraries
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.2
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.4.2
```

```
## Loading required package: lattice
```

```
# Assuming your dataset is called sentiment_analysis
# Get the distribution of the classes in the sentiment_analysis dataset
class_counts <- table(sentiment_analysis_use$status)

# Initialize an empty list to hold the balanced dataset
balanced_data <- list()

# Loop through each class
for (class in names(class_counts)) {
  # Subset the data for the current class
  class_data <- sentiment_analysis_use %>% filter(status == class)

  # If the class has fewer than 3888 samples, oversample
  if (nrow(class_data) < 3888) {
    # Oversample with replacement
    class_data <- class_data[sample(1:nrow(class_data), 3888, replace = TRUE), ]
  }

  # If the class has more than 3888 samples, undersample
  else if (nrow(class_data) > 3888) {
    # Undersample to 3888 samples
    class_data <- class_data[sample(1:nrow(class_data), 3888), ]
  }

  # Add the balanced class data to the list
  balanced_data[[class]] <- class_data
}

# Combine the balanced data
balanced_data <- do.call(rbind, balanced_data)

# Check the distribution of the balanced data
balanced_class_counts <- table(balanced_data$status)
print(balanced_class_counts)
```

```
##
##              Anxiety              Bipolar           Depression
##                 3888                 3888                 3888
##               Normal Personality disorder               Stress
##                 3888                 3888                 3888
##             Suicidal
##                 3888
```

Checking structure of balanced data

```
str(balanced_data)
```

```
## 'data.frame':    27216 obs. of  2 variables:
##  $ statement: chr  "oh my gosh" "trouble sleeping, confused mind, restless heart. All out
of tune" "All wrong, back off dear, forward doubt. Stay in a restless and restless place" "I'
ve shifted my focus to something else but I'm still worried" ...
##  $ status   : chr  "Anxiety" "Anxiety" "Anxiety" "Anxiety" ...
```

Setting a fixed sample size per class (25% of 3888)

```
# Set a fixed sample size per class (e.g., 25% of 3888)
fixed_sample_size <- 972  # Round down to ensure consistency across classes

# Perform stratified sampling
sampled_balanced_data <- do.call(rbind, lapply(split(balanced_data, balanced_data$status), fu
nction(class_data) {
  class_data[sample(1:nrow(class_data), fixed_sample_size), ]
}))

# Check the new distribution
table(sampled_balanced_data$status)
```

```
##
##              Anxiety              Bipolar           Depression
##                  972                  972                  972
##               Normal Personality disorder               Stress
##                  972                  972                  972
##             Suicidal
##                  972
```

Structure of sampled balanced dataset

```
str(sampled_balanced_data)
```

```
## 'data.frame':    6804 obs. of  2 variables:
##  $ statement: chr  "Constant twitching in my hand causing bad anxiety I was having a good
week then this morning I woke up with twi"| __truncated__ "I swear I can't sleep...sleep anyw
ay, but I don't sleep...so restless.." "Already worried about picking up the phone" "Urinalys
is came back weird and I’m freaking out 24(F) - I just went for my routine checkup with the d
octor and "| __truncated__ ...
##  $ status   : chr  "Anxiety" "Anxiety" "Anxiety" "Anxiety" ...
```

Light preprocessing of text data and

exporting as CSV for DistilBERT to use

```
# Load required libraries
library(textclean)  # For replace_contraction and text cleaning
```

```
## Warning: package 'textclean' was built under R version 4.4.2
```

```
library(tm)         # For corpus and text manipulation
```

```
## Warning: package 'tm' was built under R version 4.4.2
```

```
## Loading required package: NLP
```

```
##
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
# Replace stemming with lemmatization
corpus <- Corpus(VectorSource(sampled_balanced_data$statement))

# Apply minimal preprocessing steps
corpus <- tm_map(corpus, content_transformer(tolower))  # Convert to lowercase
```

```
## Warning in tm_map.SimpleCorpus(corpus, content_transformer(tolower)):
## transformation drops documents
```

```
corpus <- tm_map(corpus, stripWhitespace)                # Strip extra whitespaces
```

```
## Warning in tm_map.SimpleCorpus(corpus, stripWhitespace): transformation drops
## documents
```

```
# Remove non-text symbols and replace â€™ with '
corpus <- tm_map(corpus, content_transformer(function(x) gsub("[^[:alnum:] [:space:]]", "",
x)))  # Remove non-alphanumeric symbols
```

```
## Warning in tm_map.SimpleCorpus(corpus, content_transformer(function(x)
## gsub("[^[:alnum:] [:space:]]", : transformation drops documents
```

```
corpus <- tm_map(corpus, content_transformer(function(x) gsub("â€™", "'", x)))  # Fix â€™ to
'
```

```
## Warning in tm_map.SimpleCorpus(corpus, content_transformer(function(x)
## gsub("â€™", : transformation drops documents
```

```
# Convert cleaned corpus to a data frame
cleaned_statements_lots <- data.frame(statement = sapply(corpus, as.character),
                                      status = sampled_balanced_data$status)

# View a sample of the cleaned data
head(cleaned_statements_lots)
```

```
##
statement
## 1
constant twitching in my hand causing bad anxiety i was having a good week then this morning
i woke up with twitching in my one hand that hasnt stopped for two hours usually body twitche
s dont set me off but ive never had them for this long in one spot anyone else experience thi
s im worried about it lasting all day idk how ill get through it
## 2
i swear i cant sleepsleep anyway but i dont sleepso restless
## 3
already worried about picking up the phone
## 4
urinalysis came back weird and im freaking out 24f  i just went for my routine checkup with t
he doctor and had a blood and urine test the blood test came back fine for everything but the
urine test came back with elevated leukocyte erasure wbc bacteria and squamous epithelial cel
ls i doctor googled myself into a panic about cancerkidneyliver disease even though it seems
very likely its a uti the only thing is i have no symptoms of a uti could this be dehydration
i definitely dont drink as much water as i should but the day i took the test i drank a lot s
ince i knew id have to take a urine test could it be an asymptomatic uti if so is there any c
hance i caused kidney damage by not catching it earlier possibly relevant i also have a sinus
infectioncold right now the doctor hasnt called yet to explain the results just got them this
morning i guess i just need talking down
## 5 gerd fears  please help hello been a while since i posted here anyways ive had on and of
f acid reflux since i was 19 or 20 23 now it started when i went to an ent because my singing
voice wasnt quite right i sang a cappella in college and he said oh you probably have gerd ne
ver did a diagnostic test or anything then came summer 2017 i heard that acid reflux can lead
to barretts and then esophageal c coincidentally i also started to feel a lump in my throat w
hich i knew had to be esophageal c so i got an endoscopy it came back clean with a little irr
itation but it was clean great needless to say the lump disappeared ive been on and off meds
for gerd ever since for the past 69 months ive been off them because truthfully my reflux has
nt been bad tums or zantac usually suffices if i need it ppis arent good long term anyways ab
out five days ago the lump in my throat returned coincidentally ive been drinking more coffee
lately and being very liberal with my diet im hoping the lump is one of two things 1 irritati
on from increased reflux due to liberal diet 2 cricopharyngeal spasm essentially the upper sp
hincter in throat being too tight due to anxiety stress etc i am literally praying to god eve
ry night that it isnt anything worse but im still terrified could my clean scope have progres
sed to barretts or worse in a year and a half background  i suffer from chronic ocd anxiety a
nd depression i just started my first real job since graduating college and recently moved in
to a new apartment ive suffered from a sphincter spasm before it was down south not fun when
i take my klonopin it is prescribed it seems to get better additionally my throat feels bette
r when eating drinking or in the morning please help me
## 6
justhave you ever been let down by your best friend to the point of feeling too anxious i don
t want to talk about him here but i feel sorry for myself like girl i deserve better then thi
s
##    status
## 1 Anxiety
## 2 Anxiety
## 3 Anxiety
## 4 Anxiety
## 5 Anxiety
```

```
## 6 Anxiety
```

```
# Optional: Export the cleaned data to CSV
file_path <- "C:/Users/jivko/Documents/Data Analytics, Big Data, and Predictive Analytics/Per
sonal Project/Sentiment Analysis for Mental Health/Cleaned_Statements_try1.csv"
write.csv(cleaned_statements_lots, file = file_path, row.names = FALSE)

# Confirm that the file has been saved
cat("CSV file saved at:", file_path)
```

```
## CSV file saved at: C:/Users/jivko/Documents/Data Analytics, Big Data, and Predictive Analy
tics/Personal Project/Sentiment Analysis for Mental Health/Cleaned_Statements_try1.csv
```