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

2024-11-25

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
```

```r
# 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

```r
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  "I (F24) don't believe my boyfriend (M24) loves me and it's a me proble
m I have been struggling with worse anxie"| __truncated__ "tips, so you can sleep well, relax
, so you don't get restless.." "let me strum the guitar, believe me.. all your complaints, wo
rries, anxiety, sadness and confusion will disappear instantly" "Health anxiety I’m a type 1
diabetic and in my teenage years it was a rough time where I fainted multiple times"| __trunc
ated__ ...
##  $ status   : chr  "Anxiety" "Anxiety" "Anxiety" "Anxiety" ...
```

Light preprocessing of text data and

exporting as CSV for DistillBERT 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_less <- data.frame(statement = sapply(corpus, as.character),
                                      status = sampled_balanced_data$status)

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

```
##
statement
## 1
i f24 dont believe my boyfriend m24 loves me and its a me problem i have been struggling with
worse anxiety for the past year and recently my anxiety does not let me believe that my boyfr
iend loves me he gives me reassurance everytime i ask but for some reason it is never enough
for instance when he says why he does my thoughts will find complaints with his reasons why i
havent talked to him about this because this is technically not a his issue and i hate burden
ing others with my anxiety he has reassured me everyday and i dont want to ask for more i fee
l like this is something i need to figure out on my own ive had bad experiences with boys in
general so i often time believe that they will say whatever is convenient to them so i cannot
shake this feeling and i do not know what to do ive been talking to a therapist but she has n
ot been giving me effective solutions she listens and validates but i need calming techniques
how do i shake this feeling what should i do
## 2
tips so you can sleep well relax so you dont get restless
## 3
let me strum the guitar believe me all your complaints worries anxiety sadness and confusion
will disappear instantly
## 4 health anxiety im a type 1 diabetic and in my teenage years it was a rough time where i
fainted multiple times because of low blood sugar some of them was so bad where i ended in a
diabetes coma and other my family was pretty sure that i wouldnt make it i could get up in th
e morning and be completely dizzy not knowing what was going around i would look at my watch
and not understand what time it was i got mri scanned and everything was fine while being you
ng and dumb i didnt really care and i thought i was immortal living life with 200 kmp even th
ough my body was screaming after a break a few years later in 21 i caught covid19 a pretty ba
d one and i have got asthma afterwards while being sick with covid19 i realized that i wasnt
immortal at all i realized how vulnerable our body really is the thoughts started to getting
into me from the times i fainted and my covid process i have become afraid of everything and
connects it with heart problems cancer brain tumor ect my body cant be doing anything before
im convinced that im getting the worst news soon im always prepared to say i knew it its driv
ing me insane the last few months it have been brain tumor that is making me going dumb becau
se of a tension headache even though its probably just work related im a teacher and because
i sit and work in different positions that isnt great for either my neck or back i dont recog
nize my self as said before i have become distend from my friends i never seen them or talk w
ith them anymore always finding excuses to not get out of the house i stopped playing soccer
guess why because im thinking i will be getting a cardiac arrest if i do so it have been hard
the last 2 years causing stress depression and healthy anxiety not diagnosed yet only with st
ress i know i need help but how do i get it im not used to getting help i have solved my own
problems ever since i was a kid do i just call the doctor and say hey i think i have a depres
sion and bad health anxiety or what do i even say this really sucks
## 5
thats if youre worriedworried the parno is too much
## 6
anxiety is effecting my school and my life tw hi loves lt33 i 16f am looking for advise on ho
w to go forward for some background i have moved from a small town to the big city after some
majorly traumatising events that i wont get into here but after a year after living here comf
ortably i had an event trigger the traumatising memories from my past after this i have strug
gled with fearing when i leave the house which makes it really hard to go to school i get so
worried that i get physically sick when i have to leave my attendance is getting really bad i
am considering home schooling but my mum thinks i wont be self motivated enough to do it my s
```

```
elf it got so bad that i had to go to the hospital because i tried to end my life my mother w
asnt supportive of me and just is pretending that nothing happened so im starting to get real
ly worried about not being able to finish high school i only have two years left until so i d
ont know if im being dramatic or not i just started medication and they havent helped much an
yway if you got this far thanks for reading any advise would be helpfull  lt33
##    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_try.csv"
write.csv(cleaned_statements_less, 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_try.csv
```