STA6714_Step2_Term_Project

Benjamin Garcia

2024-03-08

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.1.2
```

```
library(ggplot2)
features2 <- read_excel("features2.xlsx")
```

```
table(features2$JJS, features2$Label)
```

```
##
##     objective subjective
##   0       633        360
##   1         2          1
##   3         0          2
##   4         0          1
##   5         0          1
```

```
table(features2$NNP)
```

```
##
##    0
## 1000
```

```
table(features2$WRB)
```

```
##
##    0
## 1000
```

```
table(features2$exclamationmarks, features2$Label)
```

```
##
##      objective subjective
##   0        597        299
##   1         31         34
##   2          1         13
##   3          3          5
##   4          1          8
##   5          0          1
##   6          1          3
##   7          0          1
##   10         1          0
##   11         0          1
```

```
#table(features2$questionmarks)
#table(features2$semicolon)
table(features2$ellipsis)
```

```
##
##   0   3
## 999   1
```

```
table(features2$TOs, features2$Label)
```

```
##
##     objective subjective
##   0       624        326
##   1        10         33
##   2         1          5
##   3         0          1
```

```
# can remove predictors JJS, NNP, WRB, exclamationmark, TOs, and ellipsis due to sparsity or all zeroes
```

```
table(features2$sentence1st)
```

```
##
##   0   1
##  73 927
```

```
table(features2$sentencelast)
```

```
##
##   0   1
##   5 995
```

```
# can remove these as well
```

```
feat3 <- features2[,-c(1, 2, 15, 19, 31, 42, 51, 60 , 61 )]

# removed TextID, URL, JJS, NNP, TOs, WRB, ellipsis, sentence1st, sentencelast
```

```
# Let 1 = objective and 0 = subjective

feat3$Label <- ifelse(feat3$Label == "objective", 1, 0)
table(feat3$Label)
```

```
##
##   0   1
## 365 635
```

```
round(cor(feat3[,c(1,43:53)]), 2)
```

```
##               Label semicolon colon pronouns1st pronouns2nd pronouns3rd
## Label          1.00     -0.18 -0.11       -0.24       -0.39       -0.44
## semicolon     -0.18      1.00  0.29        0.12        0.23        0.28
## colon         -0.11      0.29  1.00        0.15        0.10        0.23
## pronouns1st   -0.24      0.12  0.15        1.00        0.40        0.40
## pronouns2nd   -0.39      0.23  0.10        0.40        1.00        0.52
## pronouns3rd   -0.44      0.28  0.23        0.40        0.52        1.00
## compsupadjadv -0.47      0.37  0.38        0.35        0.48        0.67
## past          -0.17      0.18  0.23        0.42        0.37        0.68
## imperative    -0.49      0.29  0.31        0.50        0.63        0.69
## present3rd    -0.52      0.36  0.34        0.44        0.57        0.75
## present1st2nd -0.50      0.33  0.33        0.61        0.66        0.66
## txtcomplexity -0.09      0.13  0.04        0.22        0.14        0.14
##               compsupadjadv  past imperative present3rd present1st2nd
## Label                 -0.47 -0.17      -0.49      -0.52         -0.50
## semicolon              0.37  0.18       0.29       0.36          0.33
## colon                  0.38  0.23       0.31       0.34          0.33
## pronouns1st            0.35  0.42       0.50       0.44          0.61
## pronouns2nd            0.48  0.37       0.63       0.57          0.66
## pronouns3rd            0.67  0.68       0.69       0.75          0.66
## compsupadjadv          1.00  0.51       0.68       0.72          0.68
## past                   0.51  1.00       0.49       0.37          0.39
## imperative             0.68  0.49       1.00       0.74          0.76
## present3rd             0.72  0.37       0.74       1.00          0.79
## present1st2nd          0.68  0.39       0.76       0.79          1.00
## txtcomplexity          0.17  0.16       0.12       0.15          0.21
##               txtcomplexity
## Label                 -0.09
## semicolon              0.13
## colon                  0.04
## pronouns1st            0.22
## pronouns2nd            0.14
## pronouns3rd            0.14
## compsupadjadv          0.17
## past                   0.16
## imperative             0.12
## present3rd             0.15
## present1st2nd          0.21
## txtcomplexity          1.00
```

```
round(cor(feat3[,c(1:17)]), 2)
```

```
##                   Label totalWordsCount semanticobjscore semanticsubjscore
## Label              1.00           -0.45            -0.45             -0.49
## totalWordsCount   -0.45            1.00             0.96              0.89
## semanticobjscore  -0.45            0.96             1.00              0.87
## semanticsubjscore -0.49            0.89             0.87              1.00
## CC                -0.08            0.62             0.61              0.40
## CD                -0.46            0.98             0.95              0.86
## DT                -0.41            0.60             0.59              0.58
## EX                -0.34            0.71             0.68              0.68
## FW                -0.44            0.98             0.95              0.86
## INs               -0.42            0.96             0.94              0.87
## JJ                -0.38            0.70             0.70              0.62
## JJR               -0.37            0.70             0.67              0.61
## LS                -0.51            0.79             0.76              0.73
## MD                -0.38            0.97             0.93              0.85
## NN                -0.11            0.46             0.44              0.36
## NNPS              -0.37            0.92             0.88              0.79
## NNS               -0.30            0.42             0.39              0.43
##                      CC    CD    DT    EX    FW   INs    JJ   JJR    LS    MD
## Label             -0.08 -0.46 -0.41 -0.34 -0.44 -0.42 -0.38 -0.37 -0.51 -0.38
## totalWordsCount    0.62  0.98  0.60  0.71  0.98  0.96  0.70  0.70  0.79  0.97
## semanticobjscore   0.61  0.95  0.59  0.68  0.95  0.94  0.70  0.67  0.76  0.93
## semanticsubjscore  0.40  0.86  0.58  0.68  0.86  0.87  0.62  0.61  0.73  0.85
## CC                 1.00  0.59  0.23  0.37  0.64  0.64  0.47  0.49  0.38  0.64
## CD                 0.59  1.00  0.61  0.64  0.96  0.94  0.68  0.68  0.78  0.94
## DT                 0.23  0.61  1.00  0.40  0.60  0.56  0.45  0.40  0.58  0.54
## EX                 0.37  0.64  0.40  1.00  0.66  0.67  0.46  0.48  0.56  0.67
## FW                 0.64  0.96  0.60  0.66  1.00  0.94  0.69  0.68  0.76  0.95
## INs                0.64  0.94  0.56  0.67  0.94  1.00  0.66  0.67  0.74  0.95
## JJ                 0.47  0.68  0.45  0.46  0.69  0.66  1.00  0.52  0.65  0.65
## JJR                0.49  0.68  0.40  0.48  0.68  0.67  0.52  1.00  0.58  0.68
## LS                 0.38  0.78  0.58  0.56  0.76  0.74  0.65  0.58  1.00  0.74
## MD                 0.64  0.94  0.54  0.67  0.95  0.95  0.65  0.68  0.74  1.00
## NN                 0.38  0.44  0.22  0.26  0.46  0.44  0.23  0.33  0.33  0.47
## NNPS               0.69  0.91  0.52  0.56  0.91  0.91  0.65  0.67  0.70  0.91
## NNS                0.08  0.43  0.36  0.36  0.39  0.39  0.30  0.34  0.41  0.37
##                      NN  NNPS   NNS
## Label             -0.11 -0.37 -0.30
## totalWordsCount    0.46  0.92  0.42
## semanticobjscore   0.44  0.88  0.39
## semanticsubjscore  0.36  0.79  0.43
## CC                 0.38  0.69  0.08
## CD                 0.44  0.91  0.43
## DT                 0.22  0.52  0.36
## EX                 0.26  0.56  0.36
## FW                 0.46  0.91  0.39
## INs                0.44  0.91  0.39
## JJ                 0.23  0.65  0.30
## JJR                0.33  0.67  0.34
## LS                 0.33  0.70  0.41
## MD                 0.47  0.91  0.37
## NN                 1.00  0.45  0.13
```

```
## NNPS               0.45  1.00  0.35
## NNS                0.13  0.35  1.00
```

```
round(cor(feat3[,c(1,18:32)]), 2)
```

```
##       Label   PDT   POS   PRP  PRP$    RB   RBR   RBS    RP   SYM    UH    VB
## Label  1.00 -0.22 -0.48 -0.37 -0.52 -0.37 -0.28 -0.28  0.12 -0.44 -0.49 -0.17
## PDT   -0.22  1.00  0.36  0.37  0.42  0.31  0.26  0.29 -0.23  0.42  0.41  0.31
## POS   -0.48  0.36  1.00  0.71  0.85  0.49  0.29  0.61  0.12  0.81  0.76  0.66
## PRP   -0.37  0.37  0.71  1.00  0.79  0.44  0.33  0.60  0.20  0.79  0.71  0.69
## PRP$  -0.52  0.42  0.85  0.79  1.00  0.58  0.36  0.68  0.12  0.87  0.80  0.66
## RB    -0.37  0.31  0.49  0.44  0.58  1.00  0.23  0.39  0.05  0.53  0.47  0.38
## RBR   -0.28  0.26  0.29  0.33  0.36  0.23  1.00  0.23  0.01  0.34  0.32  0.23
## RBS   -0.28  0.29  0.61  0.60  0.68  0.39  0.23  1.00  0.16  0.68  0.60  0.61
## RP     0.12 -0.23  0.12  0.20  0.12  0.05  0.01  0.16  1.00  0.14  0.03  0.38
## SYM   -0.44  0.42  0.81  0.79  0.87  0.53  0.34  0.68  0.14  1.00  0.76  0.68
## UH    -0.49  0.41  0.76  0.71  0.80  0.47  0.32  0.60  0.03  0.76  1.00  0.49
## VB    -0.17  0.31  0.66  0.69  0.66  0.38  0.23  0.61  0.38  0.68  0.49  1.00
## VBD   -0.37  0.48  0.71  0.77  0.82  0.49  0.33  0.68  0.17  0.81  0.70  0.67
## VBG   -0.38  0.37  0.64  0.76  0.79  0.49  0.34  0.56  0.15  0.80  0.63  0.64
## VBN   -0.50  0.35  0.79  0.63  0.82  0.49  0.30  0.54  0.04  0.76  0.76  0.39
## VBP   -0.52  0.45  0.76  0.66  0.80  0.49  0.40  0.56  0.02  0.75  0.75  0.37
##         VBD   VBG   VBN   VBP
## Label -0.37 -0.38 -0.50 -0.52
## PDT    0.48  0.37  0.35  0.45
## POS    0.71  0.64  0.79  0.76
## PRP    0.77  0.76  0.63  0.66
## PRP$   0.82  0.79  0.82  0.80
## RB     0.49  0.49  0.49  0.49
## RBR    0.33  0.34  0.30  0.40
## RBS    0.68  0.56  0.54  0.56
## RP     0.17  0.15  0.04  0.02
## SYM    0.81  0.80  0.76  0.75
## UH     0.70  0.63  0.76  0.75
## VB     0.67  0.64  0.39  0.37
## VBD    1.00  0.73  0.70  0.75
## VBG    0.73  1.00  0.67  0.66
## VBN    0.70  0.67  1.00  0.79
## VBP    0.75  0.66  0.79  1.00
```

```
round(cor(feat3[,c(1,32:42)]), 2)
```

```
##                  Label   VBP   VBZ   WDT    WP   WP$ baseform Quotes
## Label             1.00 -0.52 -0.36 -0.45 -0.10 -0.46    -0.48   0.21
## VBP              -0.52  1.00  0.55  0.65  0.21  0.66     0.79  -0.10
## VBZ              -0.36  0.55  1.00  0.57  0.23  0.51     0.67   0.07
## WDT              -0.45  0.65  0.57  1.00  0.25  0.65     0.71   0.01
## WP               -0.10  0.21  0.23  0.25  1.00  0.13     0.22   0.12
## WP$              -0.46  0.66  0.51  0.65  0.13  1.00     0.73   0.02
## baseform         -0.48  0.79  0.67  0.71  0.22  0.73     1.00   0.01
## Quotes            0.21 -0.10  0.07  0.01  0.12  0.02     0.01   1.00
## questionmarks    -0.42  0.58  0.40  0.53  0.16  0.49     0.56  -0.10
## exclamationmarks -0.17  0.18  0.18  0.18  0.08  0.14     0.18  -0.09
## fullstops        -0.39  0.78  0.56  0.63  0.23  0.64     0.84   0.15
## commas           -0.39  0.70  0.65  0.66  0.21  0.70     0.88   0.04
##                  questionmarks exclamationmarks fullstops commas
## Label                    -0.42            -0.17     -0.39  -0.39
## VBP                       0.58             0.18      0.78   0.70
## VBZ                       0.40             0.18      0.56   0.65
## WDT                       0.53             0.18      0.63   0.66
## WP                        0.16             0.08      0.23   0.21
## WP$                       0.49             0.14      0.64   0.70
## baseform                  0.56             0.18      0.84   0.88
## Quotes                   -0.10            -0.09      0.15   0.04
## questionmarks             1.00             0.29      0.51   0.53
## exclamationmarks          0.29             1.00      0.15   0.16
## fullstops                 0.51             0.15      1.00   0.81
## commas                    0.53             0.16      0.81   1.00
```

```
library(reshape)
```

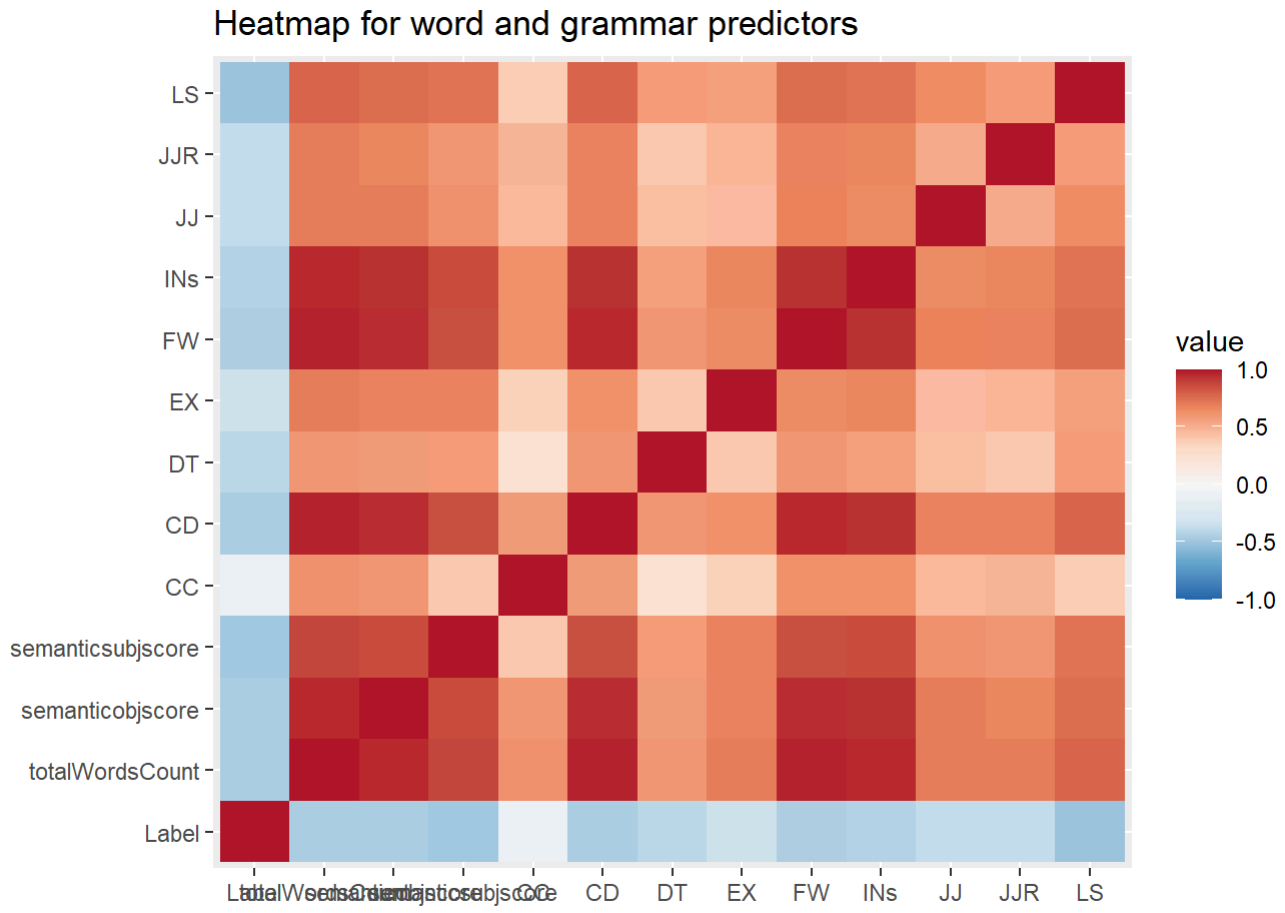
```
## Warning: package 'reshape' was built under R version 4.1.2
```

```
cor.mat2 <- round(cor(feat3[,c(1:13)]), 2)
melted.cor.mat2 <- melt(cor.mat2)
```

```
## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by the
## caller; using TRUE
```

```
## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by the
## caller; using TRUE
```

```
ggplot(melted.cor.mat2, aes(x=X1, y=X2, fill=value)) +
  geom_tile() + xlab("") + ylab("") + ggtitle("Heatmap for word and grammar predictors") +
  scale_fill_distiller(palette="RdBu", limits=c(-1, 1))
```




```
head(feat3)
```

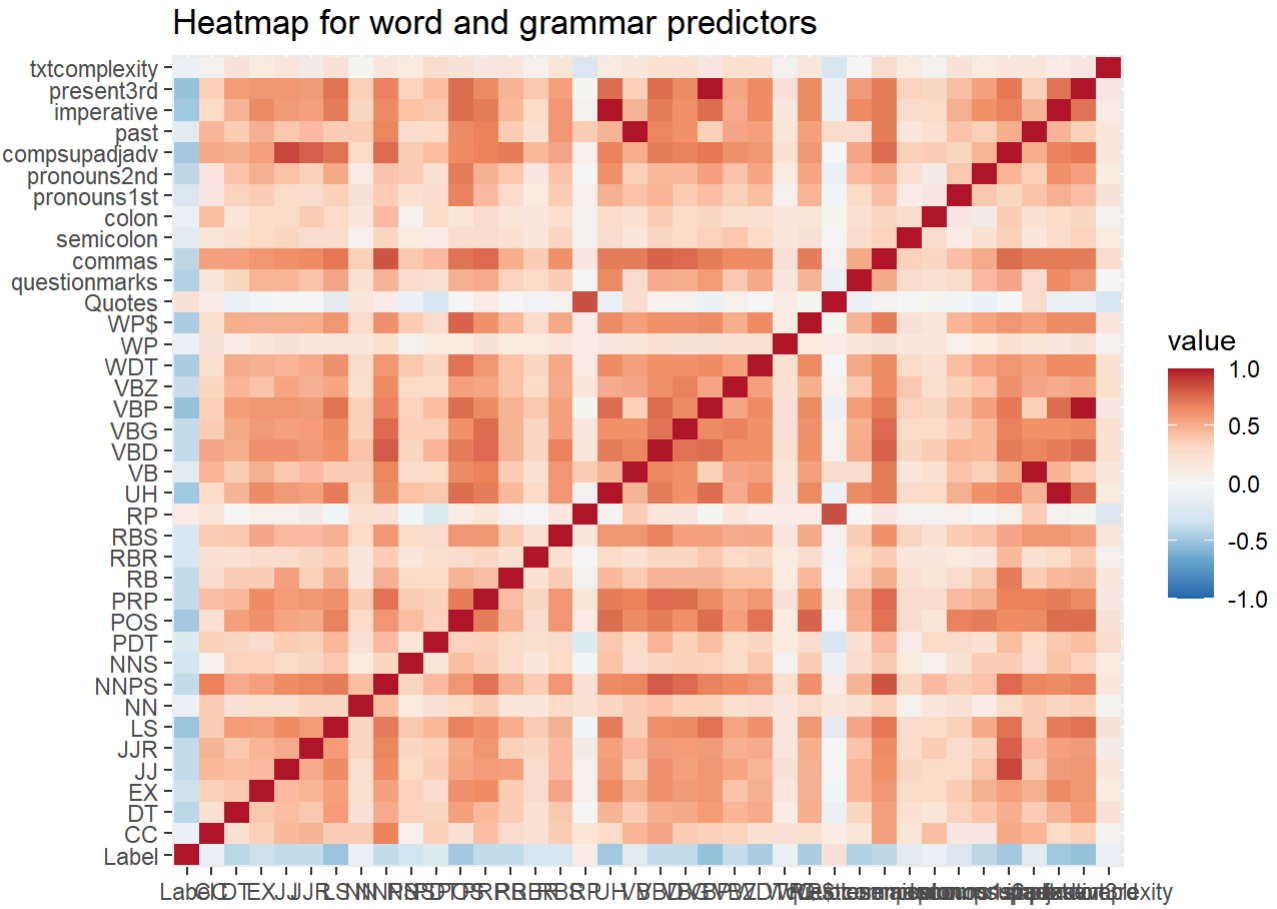
```
## # A tibble: 6 x 53
##   Label totalWordsCount semanticobjscore semanticsubjscore    CC    CD    DT
##   <dbl>           <dbl>            <dbl>              <dbl> <dbl> <dbl> <dbl>
## 1     1             109                0                  1     7     9     0
## 2     1             309               21                  4     1    19     1
## 3     1             149                6                  1     8    14     0
## 4     1             305               18                  5     7    26     0
## 5     1             491               23                  8    33    47     0
## 6     1             314               14                  1    17    17     0
## # i 46 more variables: EX <dbl>, FW <dbl>, INs <dbl>, JJ <dbl>, JJR <dbl>,
## #   LS <dbl>, MD <dbl>, NN <dbl>, NNPS <dbl>, NNS <dbl>, PDT <dbl>, POS <dbl>,
## #   PRP <dbl>, `PRP$` <dbl>, RB <dbl>, RBR <dbl>, RBS <dbl>, RP <dbl>,
## #   SYM <dbl>, UH <dbl>, VB <dbl>, VBD <dbl>, VBG <dbl>, VBN <dbl>, VBP <dbl>,
## #   VBZ <dbl>, WDT <dbl>, WP <dbl>, `WP$` <dbl>, baseform <dbl>, Quotes <dbl>,
## #   questionmarks <dbl>, exclamationmarks <dbl>, fullstops <dbl>, commas <dbl>,
## #   semicolon <dbl>, colon <dbl>, pronouns1st <dbl>, pronouns2nd <dbl>, ...
```

```
# The following variables are being removed due to high correlation with other predictors

feat4 <- feat3[,-c(2,3,4,6,9,10,14,21,26,31,40,41,37,47,52)]
```

```
#round(cor(feat4), 2) > .7

cor.mat3 <- round(cor(feat4[,]), 2)
melted.cor.mat3 <- melt(cor.mat3)
```

```
## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by the
## caller; using TRUE

## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by the
## caller; using TRUE
```

```
ggplot(melted.cor.mat3, aes(x=X1, y=X2, fill=value)) +
  geom_tile() + xlab("") + ylab("") + ggtitle("Heatmap for word and grammar predictors") +
  scale_fill_distiller(palette="RdBu", limits=c(-1, 1))
```

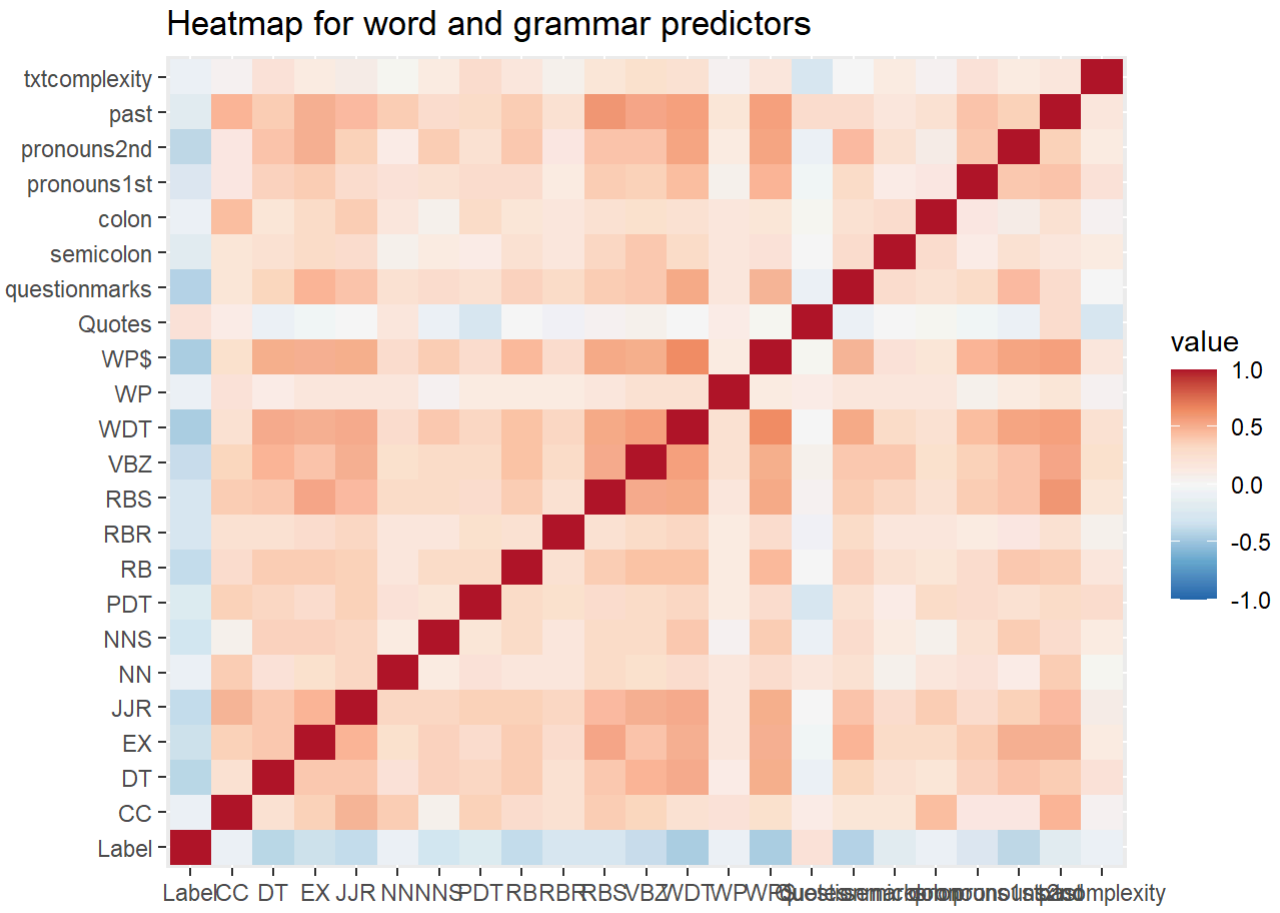



```
#additional removal due to correlation

feat5 <- feat4[,-c(5,7,9,12,13,17,18,19,20,21,22,29,34,36,37)]
```

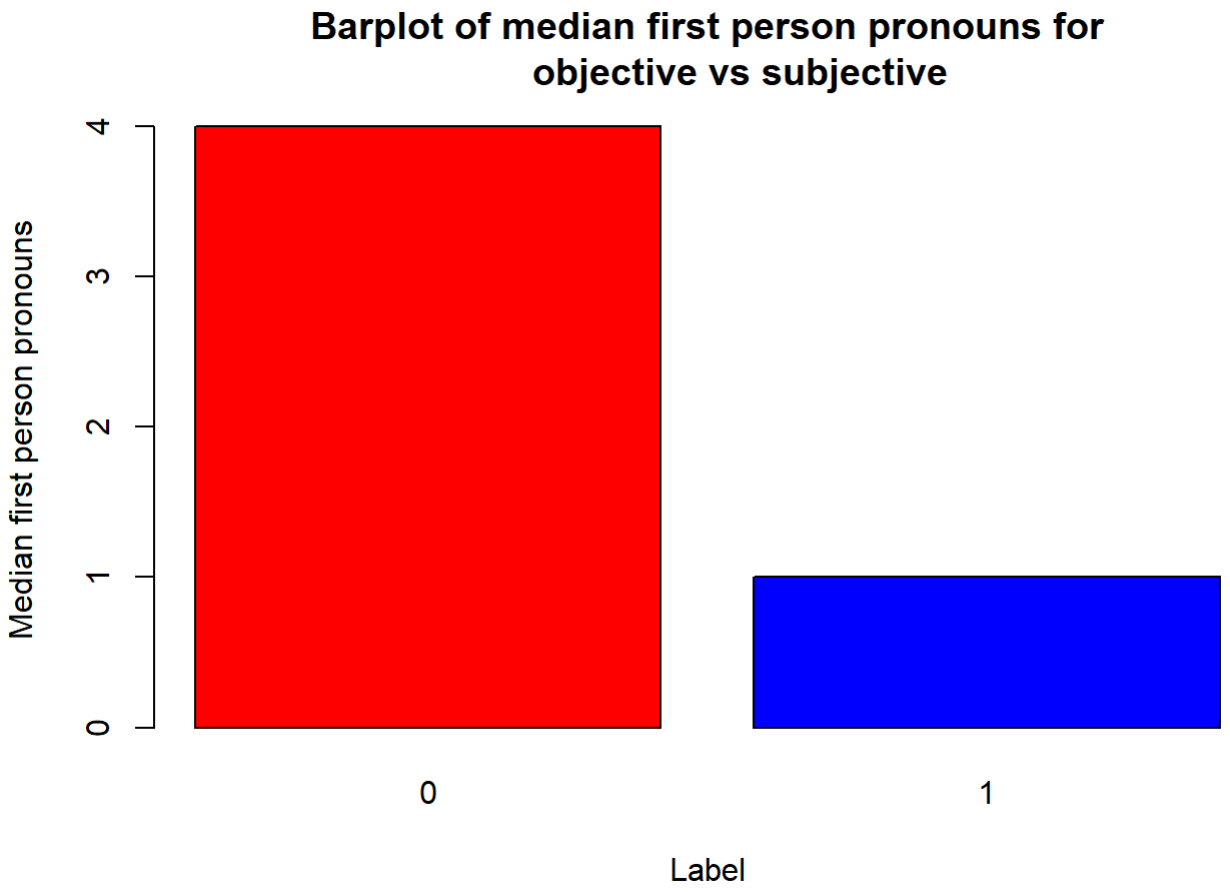
```
cor.mat4 <- round(cor(feat5[,]), 2)
melted.cor.mat4 <- melt(cor.mat4)
```

```
## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by the
## caller; using TRUE

## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by the
## caller; using TRUE
```

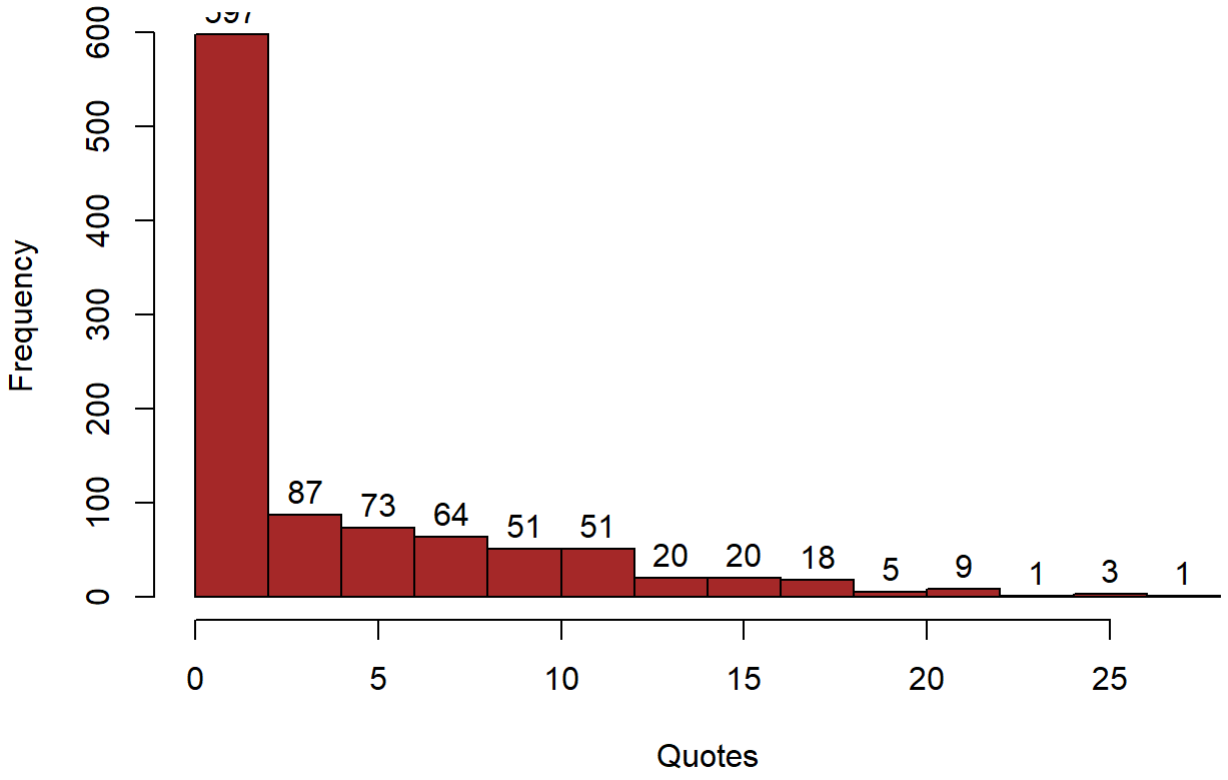
```
ggplot(melted.cor.mat4, aes(x=X1, y=X2, fill=value)) +
  geom_tile() + xlab("") + ylab("") + ggtitle("Heatmap for word and grammar predictors") +
  scale_fill_distiller(palette="RdBu", limits=c(-1, 1))
```




```
# exploratory data analysis

barplot(tapply(feat5$pronouns1st, feat5$Label, median), xlab = "Label",
        ylab = "Median first person pronouns", main = "Barplot of median first person pronouns f
or
      objective vs subjective", col = c("red", "blue"))
```



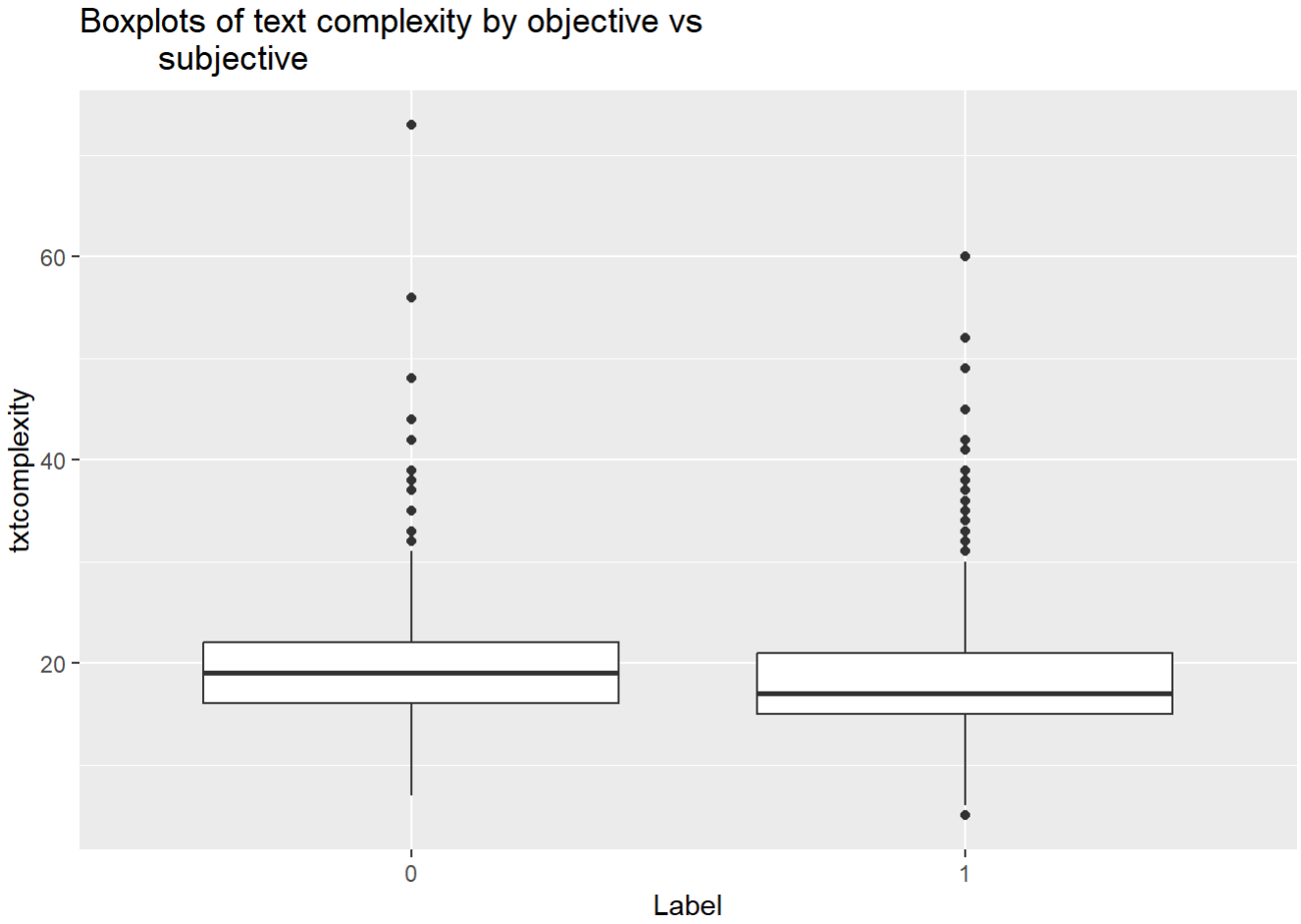

```
g <- hist(feat5$Quotes, main = "Histogram of number of quotes used",
          xlab = "Quotes", col = "brown")
text(g$mids,g$counts,labels=g$counts, adj=c(0.5, -0.5))
```



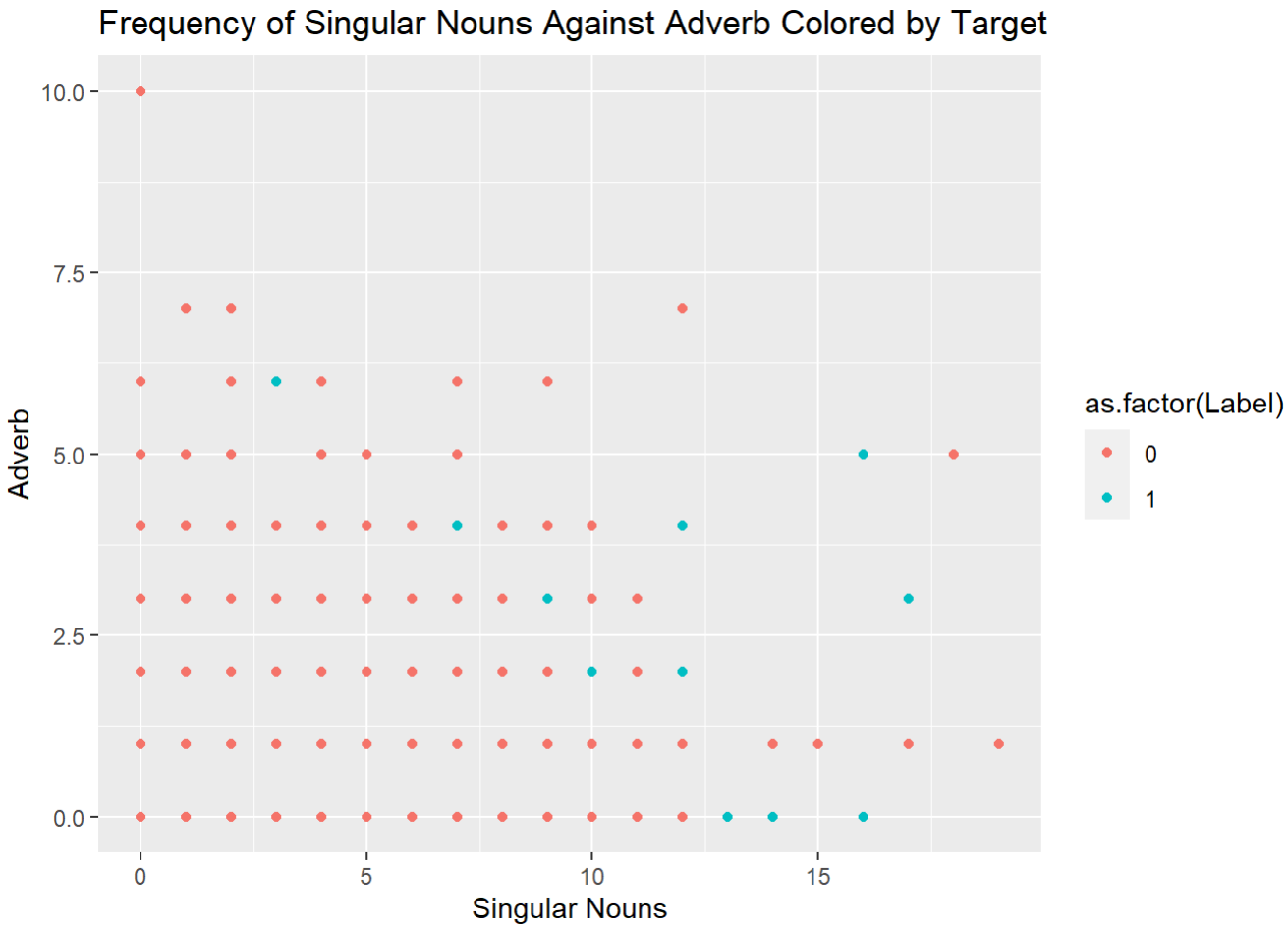

```
table(feat5$Quotes)
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
## 436  93  68  41  46  43  30  31  33  25  26  25  26  13   7  13   7  13   5   4
##  20  21  22  23  25  26  28
##   1   5   4   1   2   1   1
```

```
ggplot(feat5, aes(as.factor(Label), txtcomplexity)) + geom_boxplot() +
  labs(title = "Boxplots of text complexity by objective vs
       subjective", x = "Label")
```




```
ggplot(feat5, aes(x = NN, y = RB, color = as.factor(Label))) + geom_point() +
  labs(title = "Frequency of Singular Nouns Against Adverb Colored by Target",
       x = "Singular Nouns", y = "Adverb")
```




```
# scaling

feat5_scal <- cbind(feat5[,1], scale(feat5[,c(2:23)]))
```

```
# full model
set.seed(7)

ind <- sample(1:1000, 700, replace = F)
train.df <- feat5_scal[ind,]
holdout.df <- feat5_scal[-ind,]

logmod1 <- glm(formula = Label ~., family = binomial(link = "logit"), data = train.df)

summary(logmod1)
```

```
##
## Call:
## glm(formula = Label ~ ., family = binomial(link = "logit"), data = train.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6277  -0.4121   0.4254   0.5840   3.2467
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.394752   0.116714   3.382 0.000719 ***
## CC             0.003034   0.164921   0.018 0.985320
## DT            -0.170513   0.158735  -1.074 0.282735
## EX            -0.261789   0.154454  -1.695 0.090089 .
## JJR           -0.200104   0.160564  -1.246 0.212669
## NN             0.167314   0.131141   1.276 0.202016
## NNS           -0.059208   0.142969  -0.414 0.678778
## PDT            0.235723   0.146393   1.610 0.107353
## RB            -0.398273   0.135691  -2.935 0.003334 **
## RBR           -0.137889   0.125294  -1.101 0.271106
## RBS           -0.167900   0.137292  -1.223 0.221351
## VBZ           -0.173622   0.157528  -1.102 0.270389
## WDT           -0.635211   0.209194  -3.036 0.002394 **
## WP            -0.184189   0.115368  -1.597 0.110369
## `WP$`         -0.864592   0.200164  -4.319 1.56e-05 ***
## Quotes         0.512235   0.141446   3.621 0.000293 ***
## questionmarks -1.030836   0.222102  -4.641 3.46e-06 ***
## semicolon      0.065387   0.116428   0.562 0.574385
## colon          0.143771   0.146563   0.981 0.326618
## pronouns1st    0.091242   0.122161   0.747 0.455123
## pronouns2nd   -0.119301   0.172161  -0.693 0.488332
## past           0.798668   0.194640   4.103 4.07e-05 ***
## txtcomplexity  0.099222   0.119094   0.833 0.404765
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 921.47  on 699  degrees of freedom
## Residual deviance: 557.89  on 677  degrees of freedom
## AIC: 603.89
##
## Number of Fisher Scoring iterations: 6
```

```
# Step-wise or bidirection regression applied to full model

step_mod_both <- MASS::stepAIC(
  object = logmod1,
  direction = "both"
)
```

```
## Start:  AIC=603.89
## Label ~ CC + DT + EX + JJR + NN + NNS + PDT + RB + RBR + RBS +
##     VBZ + WDT + WP + `WP$` + Quotes + questionmarks + semicolon +
##     colon + pronouns1st + pronouns2nd + past + txtcomplexity
##
##                 Df Deviance    AIC
## - CC             1   557.89 601.89
## - NNS            1   558.06 602.06
## - semicolon      1   558.21 602.21
## - pronouns2nd    1   558.37 602.37
## - pronouns1st    1   558.45 602.45
## - txtcomplexity  1   558.60 602.60
## - colon          1   558.82 602.82
## - DT             1   559.05 603.05
## - VBZ            1   559.11 603.11
## - RBR            1   559.12 603.12
## - RBS            1   559.35 603.35
## - JJR            1   559.44 603.44
## - NN             1   559.54 603.54
## <none>               557.89 603.89
## - WP             1   560.34 604.34
## - PDT            1   560.44 604.44
## - EX             1   560.70 604.70
## - RB             1   566.67 610.67
## - WDT            1   567.61 611.61
## - Quotes         1   572.44 616.44
## - past           1   575.64 619.64
## - `WP$`          1   577.78 621.78
## - questionmarks  1   584.23 628.23
##
## Step:  AIC=601.89
## Label ~ DT + EX + JJR + NN + NNS + PDT + RB + RBR + RBS + VBZ +
##     WDT + WP + `WP$` + Quotes + questionmarks + semicolon + colon +
##     pronouns1st + pronouns2nd + past + txtcomplexity
##
##                 Df Deviance    AIC
## - NNS            1   558.06 600.06
## - semicolon      1   558.21 600.21
## - pronouns2nd    1   558.38 600.38
## - pronouns1st    1   558.45 600.45
## - txtcomplexity  1   558.60 600.60
## - colon          1   558.87 600.87
## - DT             1   559.06 601.06
## - VBZ            1   559.11 601.11
## - RBR            1   559.12 601.12
## - RBS            1   559.35 601.35
## - JJR            1   559.55 601.55
## - NN             1   559.61 601.61
## <none>               557.89 601.89
## - WP             1   560.35 602.35
## - PDT            1   560.48 602.48
## - EX             1   560.72 602.72
```

```
## + CC             1   557.89 603.89
## - RB             1   566.72 608.72
## - WDT            1   568.10 610.10
## - Quotes         1   572.49 614.49
## - past           1   576.94 618.94
## - `WP$`          1   577.85 619.85
## - questionmarks  1   584.29 626.29
##
## Step:  AIC=600.06
## Label ~ DT + EX + JJR + NN + PDT + RB + RBR + RBS + VBZ + WDT +
##     WP + `WP$` + Quotes + questionmarks + semicolon + colon +
##     pronouns1st + pronouns2nd + past + txtcomplexity
##
##                 Df Deviance    AIC
## - semicolon      1   558.41 598.41
## - pronouns1st    1   558.65 598.65
## - pronouns2nd    1   558.68 598.68
## - txtcomplexity  1   558.79 598.79
## - colon          1   559.12 599.12
## - RBR            1   559.27 599.27
## - DT             1   559.29 599.29
## - VBZ            1   559.37 599.37
## - RBS            1   559.52 599.52
## - NN             1   559.78 599.78
## - JJR            1   559.85 599.85
## <none>               558.06 600.06
## - WP             1   560.45 600.45
## - PDT            1   560.68 600.68
## - EX             1   560.97 600.97
## + NNS            1   557.89 601.89
## + CC             1   558.06 602.06
## - RB             1   567.04 607.04
## - WDT            1   569.03 609.03
## - Quotes         1   572.67 612.67
## - past           1   577.52 617.52
## - `WP$`          1   578.07 618.07
## - questionmarks  1   584.51 624.51
##
## Step:  AIC=598.41
## Label ~ DT + EX + JJR + NN + PDT + RB + RBR + RBS + VBZ + WDT +
##     WP + `WP$` + Quotes + questionmarks + colon + pronouns1st +
##     pronouns2nd + past + txtcomplexity
##
##                 Df Deviance    AIC
## - pronouns1st    1   559.03 597.03
## - pronouns2nd    1   559.06 597.06
## - txtcomplexity  1   559.19 597.19
## - RBR            1   559.54 597.54
## - VBZ            1   559.57 597.57
## - RBS            1   559.61 597.61
## - colon          1   559.69 597.69
## - DT             1   559.73 597.73
```

```
## - JJR            1   560.06 598.06
## - NN             1   560.10 598.10
## <none>               558.41 598.41
## - WP             1   560.96 598.96
## - PDT            1   561.11 599.11
## - EX             1   561.19 599.19
## + semicolon      1   558.06 600.06
## + NNS            1   558.21 600.21
## + CC             1   558.41 600.41
## - RB             1   567.29 605.29
## - WDT            1   569.94 607.94
## - Quotes         1   573.22 611.22
## - past           1   577.53 615.53
## - `WP$`          1   578.45 616.45
## - questionmarks  1   585.04 623.04
##
## Step:  AIC=597.03
## Label ~ DT + EX + JJR + NN + PDT + RB + RBR + RBS + VBZ + WDT +
##     WP + `WP$` + Quotes + questionmarks + colon + pronouns2nd +
##     past + txtcomplexity
##
##                 Df Deviance    AIC
## - pronouns2nd    1   559.42 595.42
## - txtcomplexity  1   559.95 595.95
## - VBZ            1   560.01 596.01
## - RBS            1   560.13 596.13
## - RBR            1   560.22 596.22
## - DT             1   560.37 596.37
## - colon          1   560.66 596.66
## - NN             1   560.75 596.75
## <none>               559.03 597.03
## - JJR            1   561.07 597.07
## - PDT            1   561.67 597.67
## - EX             1   561.71 597.71
## - WP             1   561.86 597.86
## + pronouns1st    1   558.41 598.41
## + semicolon      1   558.65 598.65
## + NNS            1   558.80 598.80
## + CC             1   559.03 599.03
## - RB             1   568.19 604.19
## - WDT            1   570.83 606.83
## - Quotes         1   573.53 609.53
## - `WP$`          1   578.55 614.55
## - past           1   580.79 616.79
## - questionmarks  1   585.93 621.93
##
## Step:  AIC=595.42
## Label ~ DT + EX + JJR + NN + PDT + RB + RBR + RBS + VBZ + WDT +
##     WP + `WP$` + Quotes + questionmarks + colon + past + txtcomplexity
##
##                 Df Deviance    AIC
## - txtcomplexity  1   560.40 594.40
```

```
## - RBR            1   560.49 594.49
## - VBZ            1   560.59 594.59
## - RBS            1   560.62 594.62
## - colon          1   561.13 595.13
## - DT             1   561.19 595.19
## - NN             1   561.32 595.32
## - JJR            1   561.41 595.41
## <none>               559.42 595.42
## - PDT            1   561.94 595.94
## - WP             1   562.15 596.15
## - EX             1   562.66 596.66
## + semicolon      1   559.02 597.02
## + pronouns2nd    1   559.03 597.03
## + pronouns1st    1   559.06 597.06
## + NNS            1   559.09 597.09
## + CC             1   559.40 597.40
## - RB             1   568.79 602.79
## - WDT            1   571.70 605.70
## - Quotes         1   574.42 608.42
## - `WP$`          1   580.79 614.79
## - past           1   581.28 615.28
## - questionmarks  1   587.37 621.37
##
## Step:  AIC=594.4
## Label ~ DT + EX + JJR + NN + PDT + RB + RBR + RBS + VBZ + WDT +
##     WP + `WP$` + Quotes + questionmarks + colon + past
##
##                 Df Deviance    AIC
## - VBZ            1   561.28 593.28
## - RBR            1   561.44 593.44
## - RBS            1   561.52 593.52
## - colon          1   561.88 593.88
## - DT             1   562.04 594.04
## - NN             1   562.17 594.17
## <none>               560.40 594.40
## - JJR            1   562.77 594.77
## - WP             1   563.08 595.08
## - PDT            1   563.36 595.36
## + txtcomplexity  1   559.42 595.42
## - EX             1   563.54 595.54
## + pronouns1st    1   559.94 595.94
## + semicolon      1   559.95 595.95
## + pronouns2nd    1   559.95 595.95
## + NNS            1   560.04 596.04
## + CC             1   560.39 596.39
## - RB             1   569.43 601.43
## - WDT            1   572.27 604.27
## - Quotes         1   574.42 606.42
## - `WP$`          1   581.51 613.51
## - past           1   582.84 614.84
## - questionmarks  1   592.72 624.72
##
```

```
## Step:  AIC=593.28
## Label ~ DT + EX + JJR + NN + PDT + RB + RBR + RBS + WDT + WP +
##     `WP$` + Quotes + questionmarks + colon + past
##
##                 Df Deviance    AIC
## - RBR            1   562.60 592.60
## - RBS            1   562.71 592.71
## - colon          1   562.76 592.76
## - NN             1   562.98 592.98
## <none>               561.28 593.28
## - DT             1   563.47 593.47
## - WP             1   563.92 593.92
## - PDT            1   564.04 594.04
## - JJR            1   564.07 594.07
## - EX             1   564.39 594.39
## + VBZ            1   560.40 594.40
## + txtcomplexity  1   560.59 594.59
## + pronouns2nd    1   560.67 594.67
## + NNS            1   560.82 594.82
## + semicolon      1   561.01 595.01
## + pronouns1st    1   561.02 595.02
## + CC             1   561.28 595.28
## - RB             1   570.76 600.76
## - WDT            1   574.16 604.16
## - Quotes         1   575.19 605.19
## - past           1   582.94 612.94
## - `WP$`          1   583.47 613.47
## - questionmarks  1   593.47 623.47
##
## Step:  AIC=592.6
## Label ~ DT + EX + JJR + NN + PDT + RB + RBS + WDT + WP + `WP$` +
##     Quotes + questionmarks + colon + past
##
##                 Df Deviance    AIC
## - RBS            1   563.89 591.89
## - colon          1   564.12 592.12
## - NN             1   564.37 592.37
## <none>               562.60 592.60
## - DT             1   564.88 592.88
## - PDT            1   564.95 592.95
## + RBR            1   561.28 593.28
## + VBZ            1   561.44 593.44
## - WP             1   565.44 593.44
## + txtcomplexity  1   561.98 593.98
## + pronouns2nd    1   562.13 594.13
## - JJR            1   566.20 594.20
## + NNS            1   562.20 594.20
## - EX             1   566.25 594.25
## + pronouns1st    1   562.29 594.29
## + semicolon      1   562.42 594.42
## + CC             1   562.60 594.60
## - RB             1   572.17 600.17
```

```
## - Quotes         1   576.35 604.35
## - WDT            1   576.95 604.95
## - `WP$`          1   584.75 612.75
## - past           1   584.88 612.88
## - questionmarks  1   595.95 623.95
##
## Step:  AIC=591.89
## Label ~ DT + EX + JJR + NN + PDT + RB + WDT + WP + `WP$` + Quotes +
##     questionmarks + colon + past
##
##                 Df Deviance    AIC
## - colon          1   565.23 591.23
## - NN             1   565.45 591.45
## <none>               563.89 591.89
## - DT             1   566.26 592.26
## - PDT            1   566.31 592.31
## + VBZ            1   562.42 592.42
## + RBS            1   562.60 592.60
## + RBR            1   562.71 592.71
## - WP             1   566.82 592.82
## + pronouns2nd    1   563.28 593.28
## + txtcomplexity  1   563.36 593.36
## + NNS            1   563.49 593.49
## - JJR            1   567.50 593.50
## + pronouns1st    1   563.68 593.68
## + semicolon      1   563.88 593.88
## + CC             1   563.89 593.89
## - EX             1   568.62 594.62
## - RB             1   573.83 599.83
## - Quotes         1   578.57 604.57
## - WDT            1   579.09 605.09
## - past           1   584.97 610.97
## - `WP$`          1   586.65 612.65
## - questionmarks  1   597.10 623.10
##
## Step:  AIC=591.23
## Label ~ DT + EX + JJR + NN + PDT + RB + WDT + WP + `WP$` + Quotes +
##     questionmarks + past
##
##                 Df Deviance    AIC
## - NN             1   566.60 590.60
## <none>               565.23 591.23
## - DT             1   567.34 591.34
## - WP             1   567.65 591.65
## + VBZ            1   563.75 591.75
## + colon          1   563.89 591.89
## + RBR            1   564.01 592.01
## + RBS            1   564.12 592.12
## - JJR            1   568.24 592.24
## + pronouns2nd    1   564.55 592.55
## - PDT            1   568.59 592.59
## + NNS            1   564.66 592.66
```

```
## + txtcomplexity  1   564.83 592.83
## + pronouns1st    1   564.85 592.85
## + semicolon      1   565.15 593.15
## + CC             1   565.16 593.16
## - EX             1   569.29 593.29
## - RB             1   575.20 599.20
## - Quotes         1   581.07 605.07
## - WDT            1   581.08 605.08
## - past           1   585.83 609.83
## - `WP$`          1   588.04 612.04
## - questionmarks  1   597.59 621.59
##
## Step:  AIC=590.6
## Label ~ DT + EX + JJR + PDT + RB + WDT + WP + `WP$` + Quotes +
##     questionmarks + past
##
##                 Df Deviance    AIC
## - DT             1   568.57 590.57
## <none>               566.60 590.60
## - WP             1   569.00 591.00
## - JJR            1   569.22 591.22
## + VBZ            1   565.22 591.22
## + NN             1   565.23 591.23
## + RBR            1   565.30 591.30
## + colon          1   565.45 591.45
## + RBS            1   565.66 591.66
## + pronouns2nd    1   565.75 591.75
## + NNS            1   566.02 592.02
## + pronouns1st    1   566.24 592.24
## + txtcomplexity  1   566.26 592.26
## + CC             1   566.37 592.37
## - PDT            1   570.53 592.53
## + semicolon      1   566.54 592.54
## - EX             1   570.76 592.76
## - RB             1   576.91 598.91
## - WDT            1   581.96 603.96
## - Quotes         1   583.73 605.73
## - `WP$`          1   588.70 610.70
## - past           1   589.17 611.17
## - questionmarks  1   599.06 621.06
##
## Step:  AIC=590.57
## Label ~ EX + JJR + PDT + RB + WDT + WP + `WP$` + Quotes + questionmarks +
##     past
##
##                 Df Deviance    AIC
## <none>               568.57 590.57
## + VBZ            1   566.57 590.57
## + DT             1   566.60 590.60
## - WP             1   570.81 590.81
## + pronouns2nd    1   567.05 591.05
## + RBR            1   567.23 591.23
```

```
## - JJR            1   571.33 591.33
## + NN             1   567.34 591.34
## + RBS            1   567.54 591.54
## + colon          1   567.63 591.63
## + NNS            1   567.75 591.75
## + txtcomplexity  1   568.34 592.34
## + pronouns1st    1   568.35 592.35
## - PDT            1   572.40 592.40
## + CC             1   568.41 592.41
## + semicolon      1   568.49 592.49
## - EX             1   573.56 593.56
## - RB             1   579.67 599.67
## - WDT            1   586.59 606.59
## - Quotes         1   588.45 608.45
## - past           1   590.51 610.51
## - `WP$`          1   594.46 614.46
## - questionmarks  1   602.63 622.63
```

```
step_mod_both
```

```
##
## Call:  glm(formula = Label ~ EX + JJR + PDT + RB + WDT + WP + `WP$` +
##     Quotes + questionmarks + past, family = binomial(link = "logit"),
##     data = train.df)
##
## Coefficients:
##   (Intercept)             EX            JJR            PDT             RB
##        0.3852        -0.3130        -0.2342         0.2769        -0.4280
##           WDT             WP          `WP$`         Quotes  questionmarks
##       -0.7664        -0.1713        -0.9035         0.5551        -1.1131
##          past
##        0.7542
##
## Degrees of Freedom: 699 Total (i.e. Null);  689 Residual
## Null Deviance:       921.5
## Residual Deviance: 568.6     AIC: 590.6
```

```
anova(logmod1, step_mod_both, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Label ~ CC + DT + EX + JJR + NN + NNS + PDT + RB + RBR + RBS +
##     VBZ + WDT + WP + `WP$` + Quotes + questionmarks + semicolon +
##     colon + pronouns1st + pronouns2nd + past + txtcomplexity
## Model 2: Label ~ EX + JJR + PDT + RB + WDT + WP + `WP$` + Quotes + questionmarks +
##     past
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1       677     557.89
## 2       689     568.57 -12  -10.683   0.5563
```

```
# predict

pred <- predict(step_mod_both, holdout.df[,2:23])

prob.predictions <- 1 / (1 + exp(-pred))
```

```
# confusion matrix for .5

caret::confusionMatrix(factor(ifelse(prob.predictions > .5, 1, 0)), factor(holdout.df$Label))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  74  22
##          1  33 171
##
##                Accuracy : 0.8167
##                  95% CI : (0.7682, 0.8588)
##     No Information Rate : 0.6433
##     P-Value [Acc > NIR] : 3.207e-11
##
##                   Kappa : 0.5911
##
##  Mcnemar's Test P-Value : 0.1775
##
##             Sensitivity : 0.6916
##             Specificity : 0.8860
##          Pos Pred Value : 0.7708
##          Neg Pred Value : 0.8382
##              Prevalence : 0.3567
##          Detection Rate : 0.2467
##    Detection Prevalence : 0.3200
##       Balanced Accuracy : 0.7888
##
##        'Positive' Class : 0
##
```

```
# confusion matrix for .25
caret::confusionMatrix(factor(ifelse(prob.predictions > .25, 1, 0)), factor(holdout.df$Label))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  56  13
##          1  51 180
##
##                Accuracy : 0.7867
##                  95% CI : (0.7359, 0.8317)
##     No Information Rate : 0.6433
##     P-Value [Acc > NIR] : 5.029e-08
##
##                   Kappa : 0.4952
##
##  Mcnemar's Test P-Value : 3.746e-06
##
##             Sensitivity : 0.5234
##             Specificity : 0.9326
##          Pos Pred Value : 0.8116
##          Neg Pred Value : 0.7792
##              Prevalence : 0.3567
##          Detection Rate : 0.1867
##    Detection Prevalence : 0.2300
##       Balanced Accuracy : 0.7280
##
##        'Positive' Class : 0
##
```

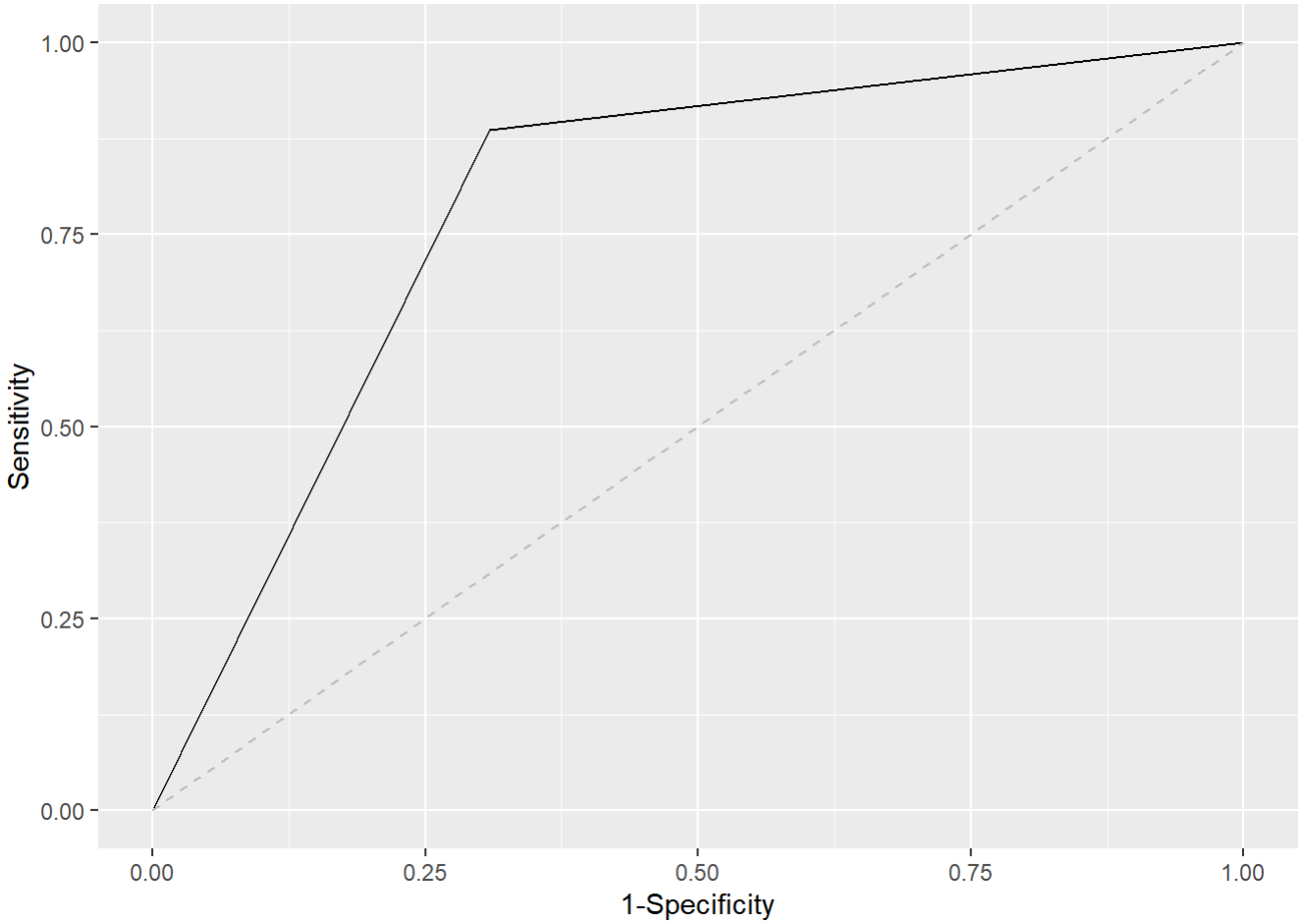
```
# confusion matrix for .75

caret::confusionMatrix(factor(ifelse(prob.predictions > .75, 1, 0)), factor(holdout.df$Label))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  86  41
##          1  21 152
##
##                Accuracy : 0.7933
##                  95% CI : (0.743, 0.8377)
##     No Information Rate : 0.6433
##     P-Value [Acc > NIR] : 1.13e-08
##
##                   Kappa : 0.5677
##
##  Mcnemar's Test P-Value : 0.01582
##
##             Sensitivity : 0.8037
##             Specificity : 0.7876
##          Pos Pred Value : 0.6772
##          Neg Pred Value : 0.8786
##              Prevalence : 0.3567
##          Detection Rate : 0.2867
##    Detection Prevalence : 0.4233
##       Balanced Accuracy : 0.7957
##
##        'Positive' Class : 0
##
```

```
# ROC curve
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 4.1.3
```

```
predob <- prediction(ifelse(prob.predictions > .5, 1, 0), holdout.df$Label)
perf <- performance(predob, "tpr", "fpr")
perf.df <- data.frame(tpr = perf@x.values[[1]],
                      fpr = perf@y.values[[1]])
ggplot2::ggplot(perf.df, aes(x = tpr, y = fpr))+
  geom_line()+
  geom_segment(aes(x=0, y=0, xend=1, yend=1), color = "gray", linetype = "dashed")+
  labs(x = "1-Specificity", y = "Sensitivity")
```




```
performance(predob, measure = "auc")@y.values[[1]]
```

```
## [1] 0.7887996
```