Naive Bayes Classification by EJvH

```
##
## 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
```

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

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

```
## Package version: 1.3.14
```

```
## Parallel computing: 2 of 4 threads used.
```

```
## See https://quanteda.io for tutorials and examples.
```

```
##
## Attaching package: 'quanteda'
```

```
## The following object is masked from 'package:utils':
##
##     View
```

Evaluating Models

Precision = TP / (TP + FP), Recall = FP / (TP + FN). F1 score = 2 x (Precision x Recall) / (Precision + Recall).

Charity Primary Purpose

% claiming that as purpose:

```
set.seed(300)
id_train <- sample(1:3446, 1000, replace = FALSE)

dat2$id_numeric <- 1:nrow(dat2)
dat.corpus <- data.frame(lapply(dat2, as.character), stringsAsFactors=FALSE)
dat.corpus <- corpus(dat.corpus,
                     text_field = "Corpus")

dfmat_training <- corpus_subset(dat.corpus, id_numeric %in% id_train) %>%
  dfm(stem = TRUE)

# get test set (documents not in id_train)
dfmat_test <- corpus_subset(dat.corpus, !id_numeric %in% id_train) %>%
  dfm(stem = TRUE)

tmod_nb <- textmodel_nb(dfmat_training, docvars(dfmat_training, "Orgpurposecharitable"))


dfmat_matched <- dfm_select(dfmat_test, pattern=dfmat_training, selection = "keep")

actual_class <- docvars(dfmat_matched, "Orgpurposecharitable")
predicted_class <- predict(tmod_nb, newdata = dfmat_matched)
tab_class <- prop.table(table(actual_class, predicted_class))
sum(dat$Orgpurposecharitable)/nrow(dat)
```

```
## [1] 0.7867092
```

```
confusionMatrix(tab_class, mode = "everything")
```

```
## Confusion Matrix and Statistics
## 
##              predicted_class
## actual_class          0          1
##            0 0.09403107 0.11856092
##            1 0.10139002 0.68601799
## 
##                 Accuracy : 0.78
##                   95% CI : (NA, NA)
##      No Information Rate : NA
##      P-Value [Acc > NIR] : NA
## 
##                    Kappa : 0.3231
##   Mcnemar's Test P-Value : 0.03611
## 
##              Sensitivity : 0.48117
##              Specificity : 0.85264
##           Pos Pred Value : 0.44231
##           Neg Pred Value : 0.87124
##                Precision : 0.44231
##                   Recall : 0.48117
##                       F1 : 0.46092
##               Prevalence : 0.19542
##           Detection Rate : 0.09403
##     Detection Prevalence : 0.21259
##        Balanced Accuracy : 0.66691
## 
##         'Positive' Class : 0
## 
```

Religious Primary Purpose

% claiming that as purpose:

```
## [1] 0.1320371
```

```
## Confusion Matrix and Statistics
## 
##              predicted_class
## actual_class          0          1
##            0 0.82788226 0.04170074
##            1 0.05928046 0.07113655
## 
##                 Accuracy : 0.899
##                   95% CI : (NA, NA)
##      No Information Rate : NA
##      P-Value [Acc > NIR] : NA
## 
##                    Kappa : 0.5277
##   Mcnemar's Test P-Value : 0.001991
## 
##              Sensitivity : 0.9332
##              Specificity : 0.6304
##           Pos Pred Value : 0.9520
##           Neg Pred Value : 0.5455
##                Precision : 0.9520
##                   Recall : 0.9332
##                       F1 : 0.9425
##               Prevalence : 0.8872
##           Detection Rate : 0.8279
##     Detection Prevalence : 0.8696
##        Balanced Accuracy : 0.7818
## 
##         'Positive' Class : 0
## 
```

Education Primary Purpose

% claiming that as purpose:

```
## [1] 0.4358677
## Confusion Matrix and Statistics
## 
##             predicted_class
## actual_class         0         1
##            0 0.4213772 0.1474820
##            1 0.1464543 0.2846865
## 
##                Accuracy : 0.7061
##                  95% CI : (NA, NA)
##     No Information Rate : NA
##     P-Value [Acc > NIR] : NA
## 
##                   Kappa : 0.4009
##  Mcnemar's Test P-Value : 0.06539
## 
##             Sensitivity : 0.7421
##             Specificity : 0.6587
##          Pos Pred Value : 0.7407
##          Neg Pred Value : 0.6603
##               Precision : 0.7407
##                  Recall : 0.7421
##                      F1 : 0.7414
##              Prevalence : 0.5678
##          Detection Rate : 0.4214
##    Detection Prevalence : 0.5689
##       Balanced Accuracy : 0.7004
## 
##        'Positive' Class : 0
## 
```

Scientific Primary Purpose

% claiming that as purpose:

```
## [1] 0.06529309
```

```
## Confusion Matrix and Statistics
## 
##             predicted_class
## actual_class          0          1
##            0 0.90441932 0.03031860
##            1 0.04881809 0.01644399
## 
##                Accuracy : 0.9209
##                  95% CI : (NA, NA)
##     No Information Rate : NA
##     P-Value [Acc > NIR] : NA
## 
##                   Kappa : 0.2529
##  Mcnemar's Test P-Value : 0.0004848
## 
##             Sensitivity : 0.9488
##             Specificity : 0.3516
##          Pos Pred Value : 0.9676
##          Neg Pred Value : 0.2520
##               Precision : 0.9676
##                  Recall : 0.9488
##                      F1 : 0.9581
##              Prevalence : 0.9532
##          Detection Rate : 0.9044
##    Detection Prevalence : 0.9347
##       Balanced Accuracy : 0.6502
## 
##        'Positive' Class : 0
## 
```

Literary Primary Purpose

% claiming that as purpose:

```
## [1] 0.03946605
## Confusion Matrix and Statistics
## 
##             predicted_class
## actual_class          0          1
##            0 0.93833505 0.02209661
##            1 0.03288798 0.00668037
## 
##                Accuracy : 0.945
##                  95% CI : (NA, NA)
##     No Information Rate : NA
##     P-Value [Acc > NIR] : NA
## 
##                   Kappa : 0.1678
##  Mcnemar's Test P-Value : 2.458e-05
## 
##             Sensitivity : 0.9661
##             Specificity : 0.2321
##          Pos Pred Value : 0.9770
##          Neg Pred Value : 0.1688
##               Precision : 0.9770
##                  Recall : 0.9661
##                      F1 : 0.9715
##              Prevalence : 0.9712
##          Detection Rate : 0.9383
##    Detection Prevalence : 0.9604
##       Balanced Accuracy : 0.5991
## 
##        'Positive' Class : 0
## 
```

Public Safety Primary Purpose

% claiming that as purpose:

```
## [1] 0.01160766
```

```
## Confusion Matrix and Statistics
## 
##             predicted_class
## actual_class           0           1
##            0 0.975847893 0.012846865
##            1 0.008735868 0.002569373
## 
##                Accuracy : 0.9784
##                  95% CI : (NA, NA)
##     No Information Rate : NA
##     P-Value [Acc > NIR] : NA
## 
##                   Kappa : 0.1816
##  Mcnemar's Test P-Value : 1.211e-11
## 
##             Sensitivity : 0.9911
##             Specificity : 0.1667
##          Pos Pred Value : 0.9870
##          Neg Pred Value : 0.2273
##               Precision : 0.9870
##                  Recall : 0.9911
##                      F1 : 0.9891
##              Prevalence : 0.9846
##          Detection Rate : 0.9758
##    Detection Prevalence : 0.9887
##       Balanced Accuracy : 0.5789
## 
##        'Positive' Class : 0
## 
```

Sports Primary Purpose

% claiming that as purpose:

```
## [1] 0.06355194
## Confusion Matrix and Statistics
##
##              predicted_class
## actual_class          0          1
##             0 0.89876670 0.03494347
##             1 0.02517986 0.04110997
##
##                Accuracy : 0.9399
##                  95% CI : (NA, NA)
##     No Information Rate : NA
##     P-Value [Acc > NIR] : NA
##
##                   Kappa : 0.5454
##  Mcnemar's Test P-Value : 5.38e-05
##
##             Sensitivity : 0.9727
##             Specificity : 0.5405
##          Pos Pred Value : 0.9626
##          Neg Pred Value : 0.6202
##               Precision : 0.9626
##                  Recall : 0.9727
##                      F1 : 0.9676
##              Prevalence : 0.9239
##          Detection Rate : 0.8988
##    Detection Prevalence : 0.9337
##       Balanced Accuracy : 0.7566
##
##        'Positive' Class : 0
##
```

Cruelty Primary Purpose

% claiming that as purpose:

```
## [1] 0.06326175
```

```
## Confusion Matrix and Statistics
##
##              predicted_class
## actual_class          0          1
##             0 0.90698869 0.02980473
##             1 0.02620761 0.03699897
##
##                Accuracy : 0.944
##                  95% CI : (NA, NA)
##     No Information Rate : NA
##     P-Value [Acc > NIR] : NA
##
##                   Kappa : 0.5392
##  Mcnemar's Test P-Value : 2.552e-05
##
##             Sensitivity : 0.9719
##             Specificity : 0.5538
##          Pos Pred Value : 0.9682
##          Neg Pred Value : 0.5854
##               Precision : 0.9682
##                  Recall : 0.9719
##                      F1 : 0.9700
##              Prevalence : 0.9332
##          Detection Rate : 0.9070
##    Detection Prevalence : 0.9368
##       Balanced Accuracy : 0.7629
##
##        'Positive' Class : 0
##
```