

Classifier

```
set.seed(42)

library(caret) # highly correlated features removal

## Loading required package: ggplot2
## Loading required package: lattice

library(tidyverse)

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr    1.5.1
## v lubridate  1.9.3      v tibble     3.2.1
## v purrr      1.0.2      v tidyr      1.3.1

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x purrr::lift()    masks caret::lift()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(tidymodels)

## -- Attaching packages ----- tidymodels 1.2.0 --
## v broom       1.0.5      v rsample     1.2.1
## v dials       1.3.0      v tune        1.2.1
## v infer       1.0.7      v workflows   1.1.4
## v modeldata   1.4.0      v workflowsets 1.1.0
## v parsnip     1.2.1      v yardstick   1.3.2
## v recipes     1.1.0

## -- Conflicts ----- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x purrr::lift()     masks caret::lift()
## x yardstick::precision() masks caret::precision()
## x yardstick::recall() masks caret::recall()
## x yardstick::sensitivity() masks caret::sensitivity()
## x yardstick::spec()   masks readr::spec()
## x yardstick::specificity() masks caret::specificity()
## x recipes::step()     masks stats::step()
## * Use suppressPackageStartupMessages() to eliminate package startup messages

library(e1071)

##
## Attaching package: 'e1071'
##
```

```
## The following object is masked from 'package:tune':
##
##     tune
##
## The following object is masked from 'package:rsample':
##
##     permutations
##
## The following object is masked from 'package:parsonip':
##
##     tune
```

Helpers

```
train_svm <- function(
  training_set,
  testing_set,
  columns,
  kernel = "radial",
  gamma = if (is.vector(training_set)) 1 else 1 / ncol(training_set),
  cost = 1) {
  model <- svm(
    training_set[columns],
    training_set$class,
    kernel = kernel, type = "C-classification",
    gamma = gamma,
    cost = cost,
    probability = TRUE,
    cross = 10
  )

  if (is.null(testing_set)) {
    return(list(
      model = model
    ))
  }

  pred <- predict(model, testing_set[columns], probability = TRUE)
  set_with_preds <- testing_set %>%
    mutate(
      pred = pred,
      prob_good = attr(pred, "probabilities")[, "good"],
      prob_bad = attr(pred, "probabilities")[, "bad"]
    )

  cm <- confusionMatrix(
    set_with_preds$pred, set_with_preds$class,
    mode = "everything",
    positive = "good"
  )

  return(list(
    model = model,
```

```

    prediction_set = set_with_preds,
    cm = cm
  ))
}

train_glm <- function(training_set, testing_set, columns) {
  formula <- reformulate(colnames(training_set[columns]), "class")
  model <- glm(
    formula,
    training_set,
    family = "binomial"
  )
  pred <- predict(model, testing_set[columns], type = "response")
  set_with_preds <- testing_set %>%
    mutate(
      prob_good = pred,
      prob_bad = 1 - pred,
      pred = if_else(pred > .5, "good", "bad") %>%
        factor(levels = c("bad", "good"))
    )

  cm <- confusionMatrix(
    set_with_preds$pred, set_with_preds$class,
    mode = "everything",
    positive = "good"
  )

  return(list(
    model = model,
    prediction_set = set_with_preds,
    cm = cm
  ))
}

get_mismatch_details <- function(data_with_predictions) {
  plot <- data_with_predictions %>%
    ggplot(aes(x = prob_good, y = class, color = subcorpus)) +
    geom_jitter(height = 0.2, width = 0)
  print(plot)

  cat("Confusion matrices by subcorpora:\n")
  data_with_predictions %>%
    select(pred, class, subcorpus) %>%
    table() %>%
    print()

  cat("\n")

  deviations <- data_with_predictions %>%
    filter(pred != class) %>%
    mutate(abs_dev = abs(prob_good - 0.5)) %>%
    arrange(-abs_dev)

```

```

cat("Greatest deviations:\n")
deviations %>%
  select(abs_dev, prob_good, class, subcorpus, FileName) %>%
  mutate(across(c(prob_good, abs_dev), ~ round(.x, 3))) %>%
  print(n = round(nrow(data_with_predictions) / 5))

cat("Names of highest-deviating documents:\n")
highest_deviation_names <- deviations %>%
  filter(abs_dev >= 0.17) %>%
  arrange(-abs_dev) %>%
  pull(FileName)

print(highest_deviation_names)

return(list(
  deviations = deviations,
  highest_deviations = highest_deviation_names,
  plot = plot
))
}

analyze_outlier <- function(doc_name, variable_importances, dataset) {
  important_variables <- sort(variable_importances, decreasing = TRUE) %>%
    head(n = 16)
  varnames <- names(important_variables)

  varscores <- tibble(feats = character(), score = numeric())
  for (v in varnames) {
    vgood <- filter(dataset, class == "good")[[v]]
    vbad <- filter(dataset, class == "bad")[[v]]
    vdoc <- filter(dataset, FileName == doc_name)[[v]]
    docclass <- filter(dataset, FileName == doc_name)$class

    # so that good values are always greater
    if (mean(vgood) < mean(vbad)) {
      vbad <- -vbad
      vgood <- -vgood
      vdoc <- -vdoc
    }

    qgood <- quantile(vgood, probs = c(.25, .75))
    qbad <- quantile(vbad, probs = c(.25, .75))

    # -2 very bad, -1 bad, 0 medium, +1 good, +2 very good
    vscore <- sum(c(
      vdoc > qbad[[1]], vdoc > qbad[[2]], vdoc > qgood[[1]], vdoc > qgood[[2]]
    )) - 2

    varscores <- varscores %>% add_row(feats = v, score = vscore)
  }

  varscores <- varscores %>%
    mutate(verbose_score = case_when(

```

```

    score == -2 ~ "very bad",
    score == -1 ~ "bad",
    score == 1 ~ "good",
    score == 2 ~ "very good",
    .default = "medium"
  )) %>%
  rowid_to_column("rank") %>%
  select(rank, everything())

cat(paste("class", docclass, "and:\n"))
if (docclass == "good") {
  print(
    varscores %>%
      filter(score < 0) %>%
      select(rank, feat, verbose_score) %>%
      as.data.frame()
  )
} else {
  print(
    varscores %>%
      filter(score > 0) %>%
      select(rank, feat, verbose_score) %>%
      as.data.frame()
  )
}
cat("even though:\n")
if (docclass == "good") {
  print(
    varscores %>%
      filter(score >= 0) %>%
      select(rank, feat, verbose_score) %>%
      as.data.frame()
  )
} else {
  print(
    varscores %>%
      filter(score <= 0) %>%
      select(rank, feat, verbose_score) %>%
      as.data.frame()
  )
}

dmut <- dataset %>%
  select(KUK_ID, FileName, class, all_of(varnames)) %>%
  mutate(across(all_of(varnames), ~ scale(.x))) %>%
  pivot_longer(
    all_of(varnames),
    names_to = "feature", values_to = "value"
  ) %>%
  mutate(across(value, ~ .x[, 1])) %>%
  mutate(across(feature, ~ factor(.x, levels = varnames)))

```

```

cat(
  nrow(dmut %>% filter(value > 5)),
  "observation(s) removed from the plot\n"
)
dmutf <- dmut %>% filter(value <= 5)

plot <- dmutf %>%
  ggplot(aes(x = class, y = value)) +
  facet_wrap(~feature) +
  geom_boxplot() +
  geom_point(
    data = dmut %>% filter(FileName == doc_name), color = "red", size = 5
  ) +
  labs(y = "measurements (scaled)")

return(plot)
}

```

Load and tidy data

```

pretty_names <- read_csv("../feat_name_mapping.csv")

## Rows: 85 Columns: 2
## -- Column specification -----
## Delimiter: ","
## chr (2): name_orig, name_pretty
##
## 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.

prettify_feat_name <- function(x) {
  name <- pull(pretty_names %>%
    filter(name_orig == x), name_pretty)
  if (length(name) == 1) {
    return(name)
  } else {
    return(x)
  }
}

prettify_feat_name_vector <- function(x) {
  map(
    x,
    prettify_feat_name
  ) %>% unlist()
}

data <- read_csv("../measurements/measurements.csv")

## Rows: 753 Columns: 108
## -- Column specification -----
## Delimiter: ","

```

```
## chr (20): fpath, KUK_ID, FileName, FileFormat, FolderPath, subcorpus, Source...
## dbl (85): RuleAbstractNouns, RuleAmbiguousRegards, RuleAnaphoricReferences, ...
## lgl (3): ClarityPursuit, SyllogismBased, Bindingness
##
## 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.
```

```
.firstnonmetacolumn <- 18

data_no_nas <- data %>%
  select(!c(
    fpath,
    # KUK_ID,
    # FileName,
    FolderPath,
    # subcorpus,
    DocumentTitle,
    ClarityPursuit,
    # Readability,
    SyllogismBased,
    SourceDB
  )) %>%
  # replace -1s in variation coefficients with NAs
  mutate(across(c(
    `RuleDoubleAdpos.max_allowable_distance.v`,
    `RuleTooManyNegations.max_negation_frac.v`,
    `RuleTooManyNegations.max_allowable_negations.v`,
    `RuleTooManyNominalConstructions.max_noun_frac.v`,
    `RuleTooManyNominalConstructions.max_allowable_nouns.v`,
    `RuleCaseRepetition.max_repetition_count.v`,
    `RuleCaseRepetition.max_repetition_frac.v`,
    `RulePredSubjDistance.max_distance.v`,
    `RulePredObjDistance.max_distance.v`,
    `RuleInfVerbDistance.max_distance.v`,
    `RuleMultiPartVerbs.max_distance.v`,
    `RuleLongSentences.max_length.v`,
    `RulePredAtClauseBeginning.max_order.v`,
    `mattr.v`,
    `maentropy.v`
  ), ~ na_if(.x, -1))) %>%
  # replace NAs with 0s
  replace_na(list(
    RuleGPcoordovs = 0,
    RuleGPdeverbaddr = 0,
    RuleGPpatinstr = 0,
    RuleGPdeverbsubj = 0,
    RuleGPadjective = 0,
    RuleGPpatbenperson = 0,
    RuleGPwordorder = 0,
    RuleDoubleAdpos = 0,
    RuleDoubleAdpos.max_allowable_distance.v = 0,
    RuleAmbiguousRegards = 0,
    RuleReflexivePassWithAnimSubj = 0,
    RuleTooManyNegations = 0,
```

```

RuleTooManyNegations.max_negation_frac.v = 0,
RuleTooManyNegations.max_allowable_negations.v = 0,
RuleTooManyNominalConstructions.max_noun_frac.v = 0,
RuleTooManyNominalConstructions.max_allowable_nouns.v = 0,
RuleFunctionWordRepetition = 0,
RuleCaseRepetition.max_repetition_count.v = 0,
RuleCaseRepetition.max_repetition_frac.v = 0,
RuleWeakMeaningWords = 0,
RuleAbstractNouns = 0,
RuleRelativisticExpressions = 0,
RuleConfirmationExpressions = 0,
RuleRedundantExpressions = 0,
RuleTooLongExpressions = 0,
RuleAnaphoricReferences = 0,
RuleLiteraryStyle = 0,
RulePassive = 0,
RulePredSubjDistance = 0,
RulePredSubjDistance.max_distance.v = 0,
RulePredObjDistance = 0,
RulePredObjDistance.max_distance.v = 0,
RuleInfVerbDistance = 0,
RuleInfVerbDistance.max_distance.v = 0,
RuleMultiPartVerbs = 0,
RuleMultiPartVerbs.max_distance.v = 0,
RuleLongSentences.max_length.v = 0,
RulePredAtClauseBeginning.max_order.v = 0,
RuleVerbalNouns = 0,
RuleDoubleComparison = 0,
RuleWrongValencyCase = 0,
RuleWrongVerbominalCase = 0,
RuleIncompleteConjunction = 0
)) %>%
# replace NAs with medians
mutate(across(c(
  RuleDoubleAdpos.max_allowable_distance,
  RuleTooManyNegations.max_negation_frac,
  RuleTooManyNegations.max_allowable_negations,
  RulePredSubjDistance.max_distance,
  RulePredObjDistance.max_distance,
  RuleInfVerbDistance.max_distance,
  RuleMultiPartVerbs.max_distance
), ~ coalesce(., median(., na.rm = TRUE)))) %>%
# merge GPs
mutate(
  GPs = RuleGPcoordovs +
    RuleGPdeverbaddr +
    RuleGPpatinstr +
    RuleGPdeverbsubj +
    RuleGPadjective +
    RuleGPpatbenperson +
    RuleGPwordorder
) %>%
select(!c(

```



```

    RuleGPcoordovs,
    RuleGPdeverbaddr,
    RuleGPpatinstr,
    RuleGPdeverbsubj,
    RuleGPadjective,
    RuleGPpatbenperson,
    RuleGPwordorder
  ))

data_clean <- data_no_nas %>%
  # norm data expected to correlate with text length
  mutate(across(c(
    GPs,
    RuleDoubleAdpos,
    RuleAmbiguousRegards,
    RuleFunctionWordRepetition,
    RuleWeakMeaningWords,
    RuleAbstractNouns,
    RuleRelativisticExpressions,
    RuleConfirmationExpressions,
    RuleRedundantExpressions,
    RuleTooLongExpressions,
    RuleAnaphoricReferences,
    RuleLiteraryStyle,
    RulePassive,
    RuleVerbalNouns,
    RuleDoubleComparison,
    RuleWrongValencyCase,
    RuleWrongVerbominalCase,
    RuleIncompleteConjunction,
    num_hapax,
    RuleReflexivePassWithAnimSubj,
    RuleTooManyNominalConstructions,
    RulePredSubjDistance,
    RuleMultiPartVerbs,
    RulePredAtClauseBeginning
  ), ~ .x / word_count)) %>%
  mutate(across(c(
    RuleTooFewVerbs,
    RuleTooManyNegations,
    RuleCaseRepetition,
    RuleLongSentences,
    RulePredObjDistance,
    RuleInfVerbDistance
  ), ~ .x / sent_count)) %>%
  # remove variables identified as text-length dependent
  select(!c(
    RuleTooFewVerbs,
    RuleTooManyNegations,
    RuleTooManyNominalConstructions,
    RuleCaseRepetition,
    RuleLongSentences,
    RulePredAtClauseBeginning,

```

```

    syllab_count,
    char_count
  )) %>%
  # remove variables identified as unreliable
  select(!c(
    RuleAmbiguousRegards,
    RuleFunctionWordRepetition,
    RuleDoubleComparison,
    RuleWrongValencyCase,
    RuleWrongVerbonominalCase
  )) %>%
  # remove further variables belonging to the 'acceptability' category
  select(!c(RuleIncompleteConjunction)) %>%
  # remove artificially limited variables
  select(!c(
    RuleCaseRepetition.max_repetition_frac,
    RuleCaseRepetition.max_repetition_frac.v
  )) %>%
  # remove variables with too many NAs
  select(!c(
    RuleDoubleAdpos.max_allowable_distance,
    RuleDoubleAdpos.max_allowable_distance.v
  )) %>%
  mutate(across(c(
    class,
    FileFormat,
    subcorpus,
    DocumentVersion,
    LegalActType,
    Objectivity,
    AuthorType,
    RecipientType,
    RecipientIndividuation,
    Anonymized
  ), ~ as.factor(.x)))

# no NAs should be present now
data_clean[!complete.cases(data_clean[,firstnonmetacolumn:ncol(data_clean)]), ]

## # A tibble: 0 x 78
## # i 78 variables: KUK_ID <chr>, FileName <chr>, FileFormat <fct>,
## #   subcorpus <fct>, SourceID <chr>, Readability <chr>, DocumentVersion <fct>,
## #   ParentDocumentID <chr>, LegalActType <fct>, Objectivity <fct>,
## #   Bindingness <lgl>, AuthorType <fct>, RecipientType <fct>,
## #   RecipientIndividuation <fct>, Anonymized <fct>, Recipient Type <chr>,
## #   class <fct>, RuleAbstractNouns <dbl>, RuleAnaphoricReferences <dbl>,
## #   RuleCaseRepetition.max_repetition_count <dbl>, ...

colnames(data_clean) <- prettify_feat_name_vector(colnames(data_clean))

data_scaled <- data_clean %>%
  mutate(across(all_of(firstnonmetacolumn:ncol(data_clean)), ~ scale(.x)[, 1]))

data_stratified <- data_scaled %>%

```

```
unite("strata", c("class", "subcorpus"), remove = FALSE)
```

Important features identification

```
feature_importances <- read_csv("../importance_measures/featcomp.csv")

## Rows: 61 Columns: 21
## -- Column specification -----
## Delimiter: ","
## chr  (2): Variable, Sign
## dbl (15): Importance, p_value, estimate, wilcox_p, wilcox_r, kw_p, kw_chi2, ...
## lgl  (4): selected_pval, wilcox_sel, kw_sel, selected_reg
##
## 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.

selected_features_names <- feature_importances %>%
  filter(kw_sel) %>%
  pull(Variable)

variable_importances <- feature_importances %>%
  filter(kw_sel) %>%
  pull(kw_epsilon2)
names(variable_importances) <- selected_features_names
```

Formulas

```
columns_all <- colnames(data_stratified)[
  (.firstnonmetacolumn + 1):ncol(data_stratified)
]
columns_cleaned <- columns_all[!(columns_all %in% c("atl", "cli"))]
columns_readabilty_forms <- c("ari", "fkgl", "fre", "gf", "smog")
```

Hyperparameters

```
colside <- c("all", "cleaned", "readforms")
colsets <- list(columns_all, columns_cleaned, columns_readabilty_forms)
```

Splits and folds

```
.splitprop <- 3 / 4

split <- initial_split(data_stratified, .splitprop, strata = strata)

training_set <- training(split)
testing_set <- testing(split)

training_set %>%
```

```

select(class) %>%
table()

## class
## bad good
## 310 253

testing_set %>%
select(class) %>%
table()

## class
## bad good
## 104 86

training_set %>%
select(subcorpus, class) %>%
table()

##          class
## subcorpus bad good
## CzCDC      157  0
## FrBo       56 171
## KUKY       64  82
## LiFRLaw    3   0
## OmbuFlyers 30   0

testing_set %>%
select(subcorpus, class) %>%
table()

##          class
## subcorpus bad good
## CzCDC      54  0
## FrBo       22 58
## KUKY       20 28
## LiFRLaw    0  0
## OmbuFlyers 8   0

```

Tune

```

tune_res <- tibble(
  columns = character(),
  kernel = character(),
  gamma = numeric(),
  cost = numeric(),
  error = numeric(),
  dispersion = numeric()
)

# commented out to reduce computation time

# for (coli in seq_along(colsets)) {
#   colsid <- colsids[coli]
#   columns <- colsets[[coli]]
# }

```

```

# message("tune linear on ", colsid)
# tune_linear <- tune.svm(training_set[columns], training_set$class,
#   cost = 10^(-3:1),
#   kernel = "linear"
# )
# tune_res <- tune_res %>%
#   bind_rows(tune_linear$performances %>%
#     mutate(kernel = "linear", columns = colsid, gamma = 0))

# message("tune radial on ", colsid)
# tune_radial <- tune.svm(training_set[columns], training_set$class,
#   gamma = 10^(-3:3),
#   cost = c(0.01, 0.1, 1, 10, 100, 1000),
#   kernel = "radial"
# )
# tune_res <- tune_res %>%
#   bind_rows(tune_radial$performances %>%
#     mutate(kernel = "radial", columns = colsid))

# message("tune polynomial3 on ", colsid)
# tune_polynomial <- tune.svm(training_set[columns], training_set$class,
#   gamma = 10^(-3:0),
#   degree = 3,
#   cost = 10^(-3:1),
#   kernel = "polynomial"
# )
# tune_res <- tune_res %>%
#   bind_rows(tune_polynomial$performances %>%
#     mutate(kernel = "polynomial3", columns = colsid))

# message("tune polynomial4 on ", colsid)
# tune_polynomial <- tune.svm(training_set[columns], training_set$class,
#   gamma = 10^(-3:-1),
#   degree = 4,
#   cost = 10^(-3:1),
#   kernel = "polynomial"
# )
# tune_res <- tune_res %>%
#   bind_rows(tune_polynomial$performances %>%
#     mutate(kernel = "polynomial4", columns = colsid))

# message("tune polynomial5 on ", colsid)
# tune_polynomial <- tune.svm(training_set[columns], training_set$class,
#   gamma = 10^(-3:-1),
#   degree = 5,
#   cost = 10^(-3:0),
#   kernel = "polynomial"
# )
# tune_res <- tune_res %>%
#   bind_rows(tune_polynomial$performances %>%
#     mutate(kernel = "polynomial5", columns = colsid))

```

```
# message("tune sigmoid on ", colsid)
# tune_sigmoid <- tune.sum(training_set[columns], training_set$class,
#   gamma = 10^(-3:3),
#   cost = 10^(-3:3),
#   kernel = "sigmoid"
# )
# tune_res <- tune_res %>%
#   bind_rows(tune_sigmoid$performances %>%
#     mutate(kernel = "sigmoid", columns = colsid))
# }

# tune_res %>% write_csv("tune_results.csv")
tune_res <- read_csv("tune_results.csv")

## Rows: 429 Columns: 7
## -- Column specification -----
## Delimiter: ","
## chr (2): columns, kernel
## dbl (5): gamma, cost, error, dispersion, degree
##
## 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.
```

```
tune_res %>%
  arrange(error, -dispersion)
```

```
## # A tibble: 429 x 7
##   columns kernel  gamma  cost error dispersion degree
##   <chr>   <chr>   <dbl> <dbl> <dbl>   <dbl> <dbl>
## 1 cleaned radial  0.01    1   0.201   0.0425    NA
## 2 cleaned radial  0.001 100   0.202   0.0444    NA
## 3 all      radial  0.001 100   0.206   0.0576    NA
## 4 cleaned radial  0.001 10   0.206   0.0421    NA
## 5 all      radial  0.01    1   0.208   0.0440    NA
## 6 all      sigmoid 0.001 10   0.210   0.0530    NA
## 7 all      radial  0.001 10   0.216   0.0679    NA
## 8 cleaned linear  0      0.1  0.217   0.0285    NA
## 9 all      sigmoid 0.001 100   0.222   0.0759    NA
## 10 cleaned linear 0      0.01 0.222   0.0383    NA
## # i 419 more rows
```

```
tune_res %>%
  arrange(error + dispersion)
```

```
## # A tibble: 429 x 7
##   columns kernel  gamma  cost error dispersion degree
##   <chr>   <chr>   <dbl> <dbl> <dbl>   <dbl> <dbl>
## 1 cleaned radial  0.01    1   0.201   0.0425    NA
## 2 cleaned linear  0      0.1  0.217   0.0285    NA
## 3 cleaned radial  0.001 100   0.202   0.0444    NA
## 4 cleaned radial  0.001 10   0.206   0.0421    NA
## 5 all      radial  0.01    1   0.208   0.0440    NA
## 6 cleaned radial  0.01    10   0.240   0.0199    NA
## 7 cleaned linear  0      0.01 0.222   0.0383    NA
```

```
## 8 cleaned sigmoid 0.001 10 0.222 0.0389 NA
## 9 all sigmoid 0.001 10 0.210 0.0530 NA
## 10 all radial 0.001 100 0.206 0.0576 NA
## # i 419 more rows
```

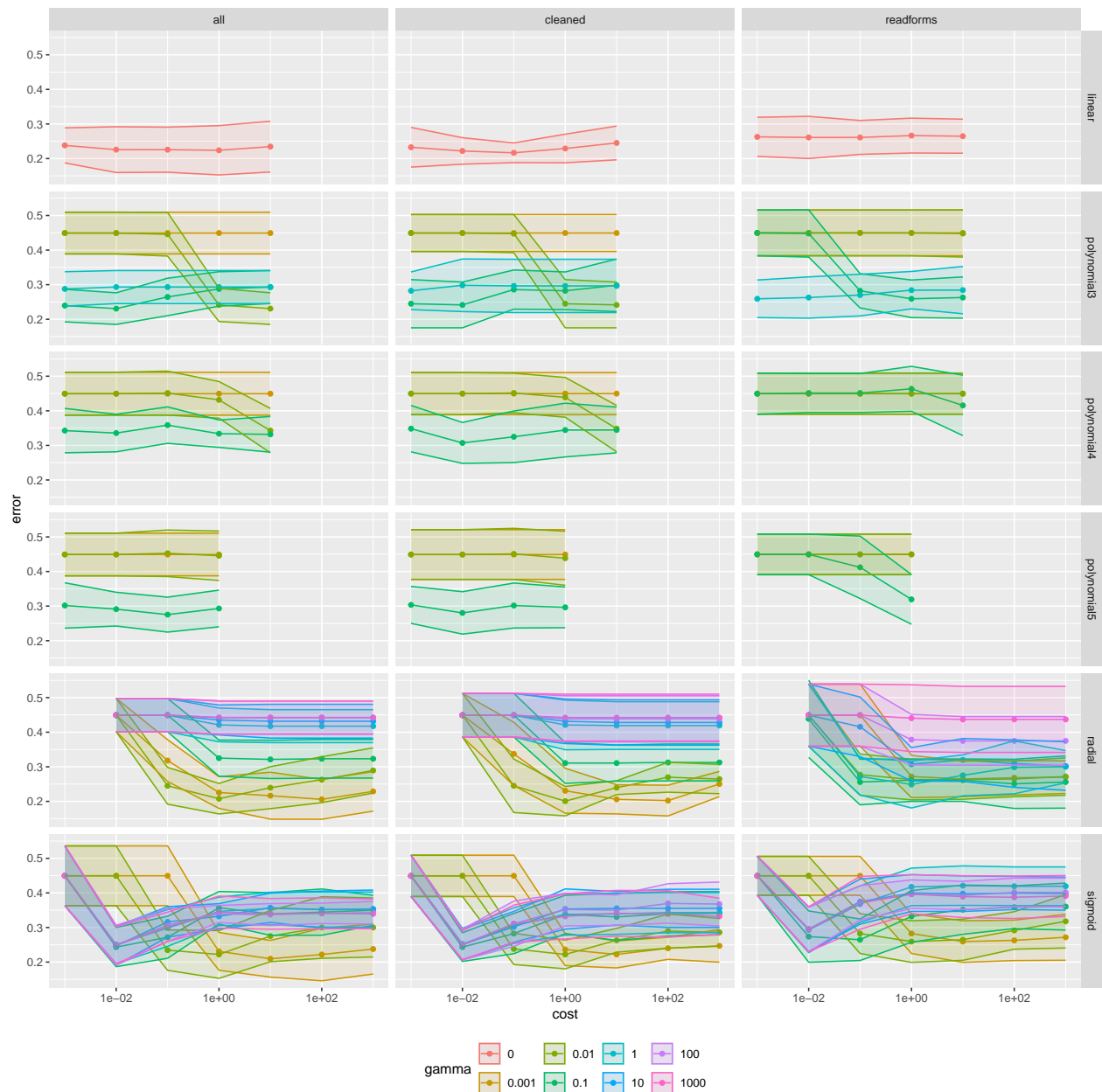
```
tune_res %>%
  filter(columns == "all") %>%
  arrange(error, -dispersion)
```

```
## # A tibble: 143 x 7
##   columns kernel gamma cost error dispersion degree
##   <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 all radial 0.001 100 0.206 0.0576 NA
## 2 all radial 0.01 1 0.208 0.0440 NA
## 3 all sigmoid 0.001 10 0.210 0.0530 NA
## 4 all radial 0.001 10 0.216 0.0679 NA
## 5 all sigmoid 0.001 100 0.222 0.0759 NA
## 6 all sigmoid 0.01 1 0.222 0.0690 NA
## 7 all linear 0 1 0.224 0.0712 NA
## 8 all radial 0.001 1 0.226 0.0458 NA
## 9 all linear 0 0.1 0.226 0.0652 NA
## 10 all linear 0 0.01 0.226 0.0662 NA
## # i 133 more rows
```

```
tune_res %>%
  filter(str_detect(columns, "notcorrelating.*")) %>%
  arrange(error, -dispersion)
```

```
## # A tibble: 0 x 7
## # i 7 variables: columns <chr>, kernel <chr>, gamma <dbl>, cost <dbl>,
## # error <dbl>, dispersion <dbl>, degree <dbl>
```

```
tune_res %>%
  mutate(across(gamma, as.factor)) %>%
  ggplot(aes(
    x = cost, y = error, ymin = error - dispersion,
    ymax = error + dispersion, color = gamma, fill = gamma
  )) +
  geom_point() +
  geom_line() +
  geom_ribbon(alpha = 0.1) +
  scale_x_log10() +
  facet_grid(kernel ~ columns) +
  theme(legend.position = "bottom")
```



SVM cleaned

```
set.seed(42)

model_cleaned <- train_svm(
  training_set, testing_set, columns_cleaned, "radial",
  gamma = 0.01, cost = 1
)
model_cleaned$cm
```

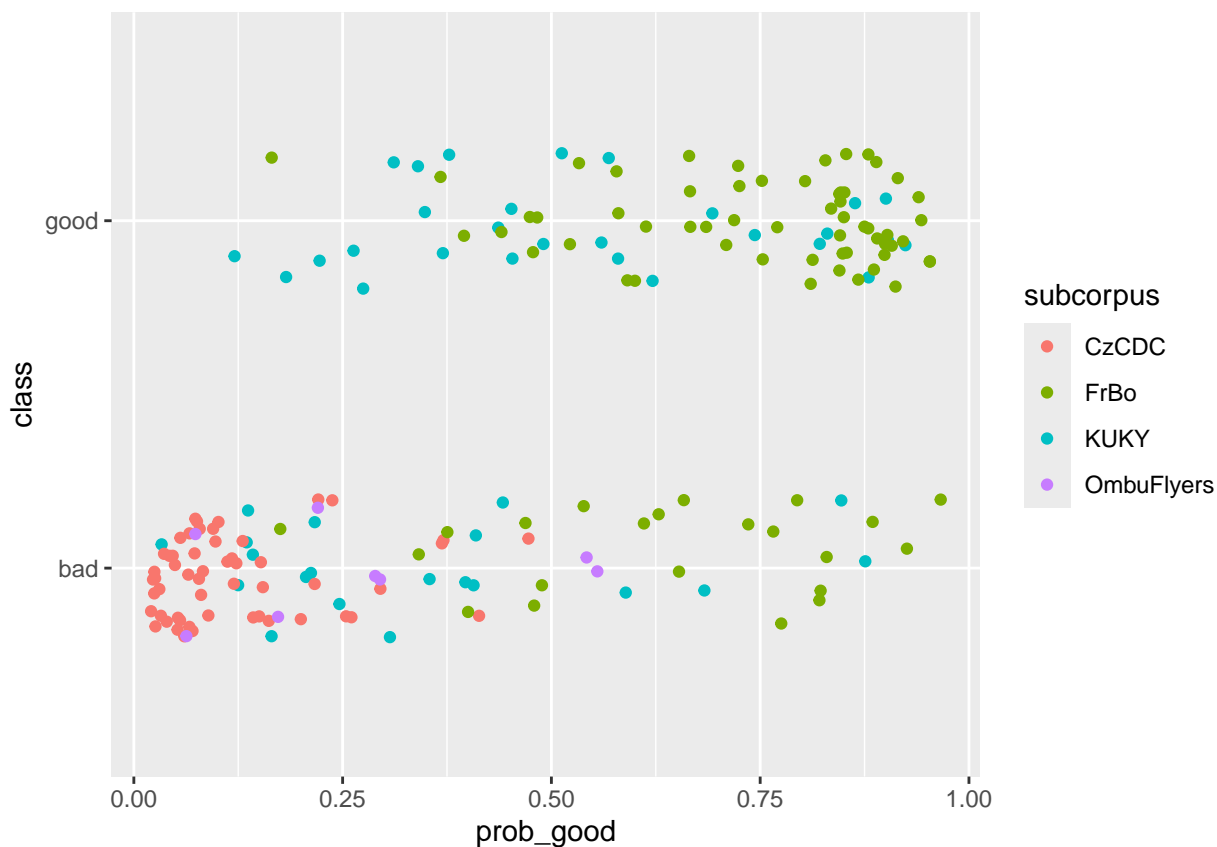
```
## Confusion Matrix and Statistics
##
##          Reference
```



```

## Prediction bad good
##      bad   83   21
##      good  21   65
##
##              Accuracy : 0.7789
##              95% CI : (0.7132, 0.8358)
##      No Information Rate : 0.5474
##      P-Value [Acc > NIR] : 2.6e-11
##
##              Kappa : 0.5539
##
##      McNemar's Test P-Value : 1
##
##              Sensitivity : 0.7558
##              Specificity : 0.7981
##              Pos Pred Value : 0.7558
##              Neg Pred Value : 0.7981
##              Precision : 0.7558
##              Recall : 0.7558
##              F1 : 0.7558
##              Prevalence : 0.4526
##              Detection Rate : 0.3421
##              Detection Prevalence : 0.4526
##              Balanced Accuracy : 0.7769
##
##      'Positive' Class : good
##
mismatches_cleaned <- get_mismatch_details(model_cleaned$prediction_set)

```



```
## Confusion matrices by subcorpora:
```

```
## , , subcorpus = CzCDC
```

```
##
```

```
##      class
## pred  bad good
## bad   54   0
## good   0   0
```

```
##
```

```
## , , subcorpus = FrBo
```

```
##
```

```
##      class
## pred  bad good
## bad    7   7
## good  15  51
```

```
##
```

```
## , , subcorpus = KUKY
```

```
##
```

```
##      class
## pred  bad good
## bad   16  14
## good   4   14
```

```
##
```

```
## , , subcorpus = LiFRLaw
```

```
##
```

```
##      class
## pred  bad good
## bad    0   0
```

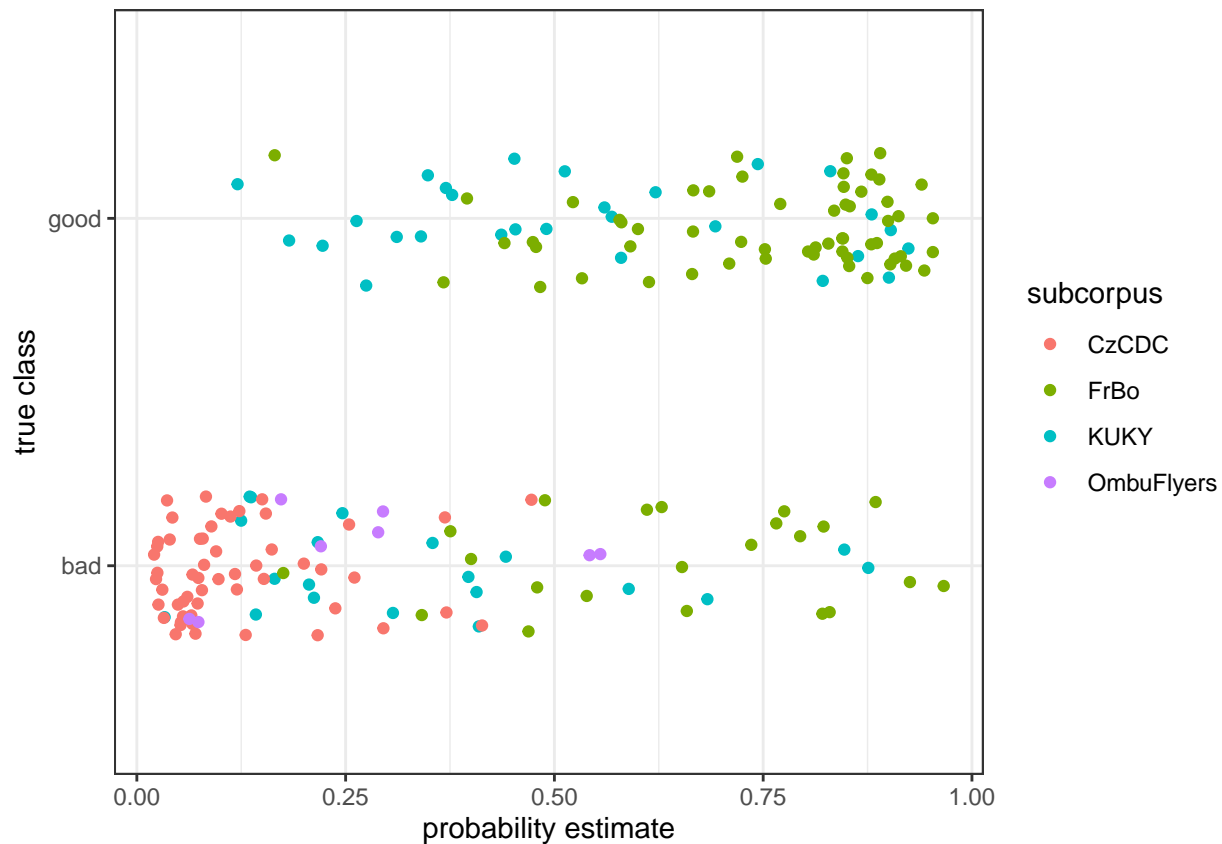
```

##   good    0    0
##
## , , subcorpus = OmbuFlyers
##
##       class
## pred   bad good
##   bad     6    0
##   good    2    0
##
##
## Greatest deviations:
## # A tibble: 42 x 5
##   abs_dev prob_good class subcorpus FileName
##   <dbl>    <dbl> <fct> <fct>    <chr>
## 1  0.466    0.966 bad   FrBo    orig_Jaká pravidla platí pro veřejné zaká-
## 2  0.426    0.926 bad   FrBo    orig_Kompletní průvodce pořizováním nahrá-
## 3  0.385    0.885 bad   FrBo    orig_Jak řešit lavinovitou černou skládku-
## 4  0.379    0.121 good  KUKY    AK_JH_Hroch_ustavni_stiznost
## 5  0.376    0.876 bad   KUKY    PR_Konecny__Miliak
## 6  0.347    0.847 bad   KUKY    PR_Masinova
## 7  0.335    0.165 good  FrBo    red_Certifikáty autorizovaných inspektorů
## 8  0.33     0.83   bad   FrBo    orig_Mohou spolky ve správních žalobách p-
## 9  0.322    0.822 bad   FrBo    orig_Jaké trestné činy mohou souviset s k-
## 10 0.321    0.821 bad   FrBo    orig_Sousedské vztahy
## 11 0.318    0.182 good  KUKY    Mestsky_urad_PRIKAZ
## 12 0.294    0.794 bad   FrBo    28
## 13 0.278    0.222 good  KUKY    Odvolani_proti_rozhodnuti_o_nepovoleni_ka-
## 14 0.275    0.775 bad   FrBo    orig_Jak se bránit neposkytnutí projektov-
## 15 0.266    0.766 bad   FrBo    orig_Jak se bránit obtěžování kouřem a pá-
## 16 0.237    0.263 good  KUKY    Mestsky_urad_usneseni_-_slouceni_pred
## 17 0.236    0.736 bad   FrBo    153
## 18 0.225    0.275 good  KUKY    AK_JH_Podani_US_podpis
## 19 0.189    0.311 good  KUKY    Mestsky_urad_PRIKAZ_REV2
## 20 0.183    0.683 bad   KUKY    043_Plisen-a-zavady-v-byte
## 21 0.16     0.34   good  KUKY    Mestsky_urad_kontrola_po
## 22 0.159    0.659 bad   FrBo    orig_Kompletní průvodce občana obtěžované-
## 23 0.153    0.653 bad   FrBo    176
## 24 0.151    0.349 good  KUKY    6417_2023_VOP
## 25 0.133    0.367 good  FrBo    red_Co je to úřední deska a jak ji využít
## 26 0.13     0.37   good  KUKY    Obecni_urad_rozhodnuti_zadost_dle_106pdf
## 27 0.128    0.628 bad   FrBo    orig_Jak využít svého práva být informová-
## 28 0.122    0.378 good  KUKY    Mestsky_urad_kontrola_pred
## 29 0.111    0.611 bad   FrBo    42
## 30 0.105    0.395 good  FrBo    156
## 31 0.089    0.589 bad   KUKY    Dopis_studentské brigády
## 32 0.064    0.436 good  KUKY    Mestsky_urad__Vyzva_k_odstraneni_trabanta
## 33 0.06     0.44   good  FrBo    red_10_významných práv účastníka správníh-
## 34 0.055    0.555 bad   OmbuFlyers Pozemkove-urady
## 35 0.048    0.452 good  KUKY    6421_2023_VOP
## 36 0.047    0.453 good  KUKY    Mestsky_urad_Nesoucinnost-U_sroz
## 37 0.042    0.542 bad   OmbuFlyers Skolstvi
## 38 0.039    0.539 bad   FrBo    orig_Předcházení ekologické újmě - jak se-
## # i 4 more rows
## Names of highest-deviating documents:

```

```
## [1] "orig_Jaká pravidla platí pro veřejné zakázky malého rozsahu_final"
## [2] "orig_Kompletní průvodce pořizováním nahrávek veřejné správy"
## [3] "orig_Jak řešit lavinovitou černou skládku úprava svépomoc 2021"
## [4] "AK_JH_Hroch_ustavni_stiznost"
## [5] "PR_Konecny__Miliak"
## [6] "PR_Masinova"
## [7] "red_Certifikáty autorizovaných inspektorů"
## [8] "orig_Mohou spolky ve správních žalobách používat věcné argumenty_final"
## [9] "orig_Jaké trestné činy mohou souviset s korupcí"
## [10] "orig_Sousedské vztahy"
## [11] "Mestsky_urad_PRIKAZ"
## [12] "28"
## [13] "Odvolani_proti_rozhodnuti_o_nepovoleni_kaceni"
## [14] "orig_Jak se bránit neposkytnutí projektové dokumentace"
## [15] "orig_Jak se bránit obtěžování kouřem a pálením odpadu"
## [16] "Mestsky_urad_usneseni_-_sloucení_pred"
## [17] "153"
## [18] "AK_JH_Podani_US_podpis"
## [19] "Mestsky_urad_PRIKAZ_REV2"
## [20] "043_Plisen-a-zavady-v-byte"
```

```
mismatches_cleaned$plot +
  theme_bw() +
  labs(y = "true class", x = "probability estimate")
```



```
ggsave("model_cleaned_probabilities.pdf")
```

```
## Saving 6.5 x 4.5 in image
```

```
mismatches_cleaned$deviations %>%
  select(subcorpus, class) %>%
  table()
```

```
##           class
## subcorpus    bad good
##   CzCDC         0   0
##   FrBo         15   7
##   KUKY          4  14
##   LiFRLaw       0   0
##   OmbuFlyers    2   0
```

```
mismatches_cleaned$deviations %>%
  select(subcorpus, Readability) %>%
  table(useNA = "ifany")
```

```
##           Readability
## subcorpus    high low medium <NA>
##   CzCDC         0   0      0     0
##   FrBo          7   0     14     1
##   KUKY         14   2      2     0
##   LiFRLaw       0   0      0     0
##   OmbuFlyers    0   0      0     2
```

```
mismatches_cleaned$deviations %>%
  filter(abs_dev >= 0.17) %>%
  select(subcorpus, class) %>%
  table()
```

```
##           class
## subcorpus    bad good
##   CzCDC         0   0
##   FrBo         10   1
##   KUKY          3   6
##   LiFRLaw       0   0
##   OmbuFlyers    0   0
```

```
mismatches_cleaned$deviations %>%
  filter(abs_dev >= 0.17) %>%
  select(subcorpus, Readability) %>%
  table(useNA = "ifany")
```

```
##           Readability
## subcorpus    high low medium <NA>
##   CzCDC         0   0      0     0
##   FrBo          1   0      9     1
##   KUKY          6   1      2     0
##   LiFRLaw       0   0      0     0
##   OmbuFlyers    0   0      0     0
```

```
variable_importances_cleaned <- variable_importances[
  names(variable_importances) %in% columns_cleaned
]
```

```
for (doc in mismatches_cleaned$highest_deviations) {
  doc_row <- mismatches_cleaned$deviations %>% filter(FileName == doc)
```

```

cat(paste(doc, "/", doc_row["subcorpus"][[1]], "\n"))
cat("KUK_ID:", doc_row["KUK_ID"][[1]], "\n")
cat("dev:", doc_row["abs_dev"][[1]] %>% round(3), "\n")
cat("Readability:", doc_row["Readability"][[1]], "\n")

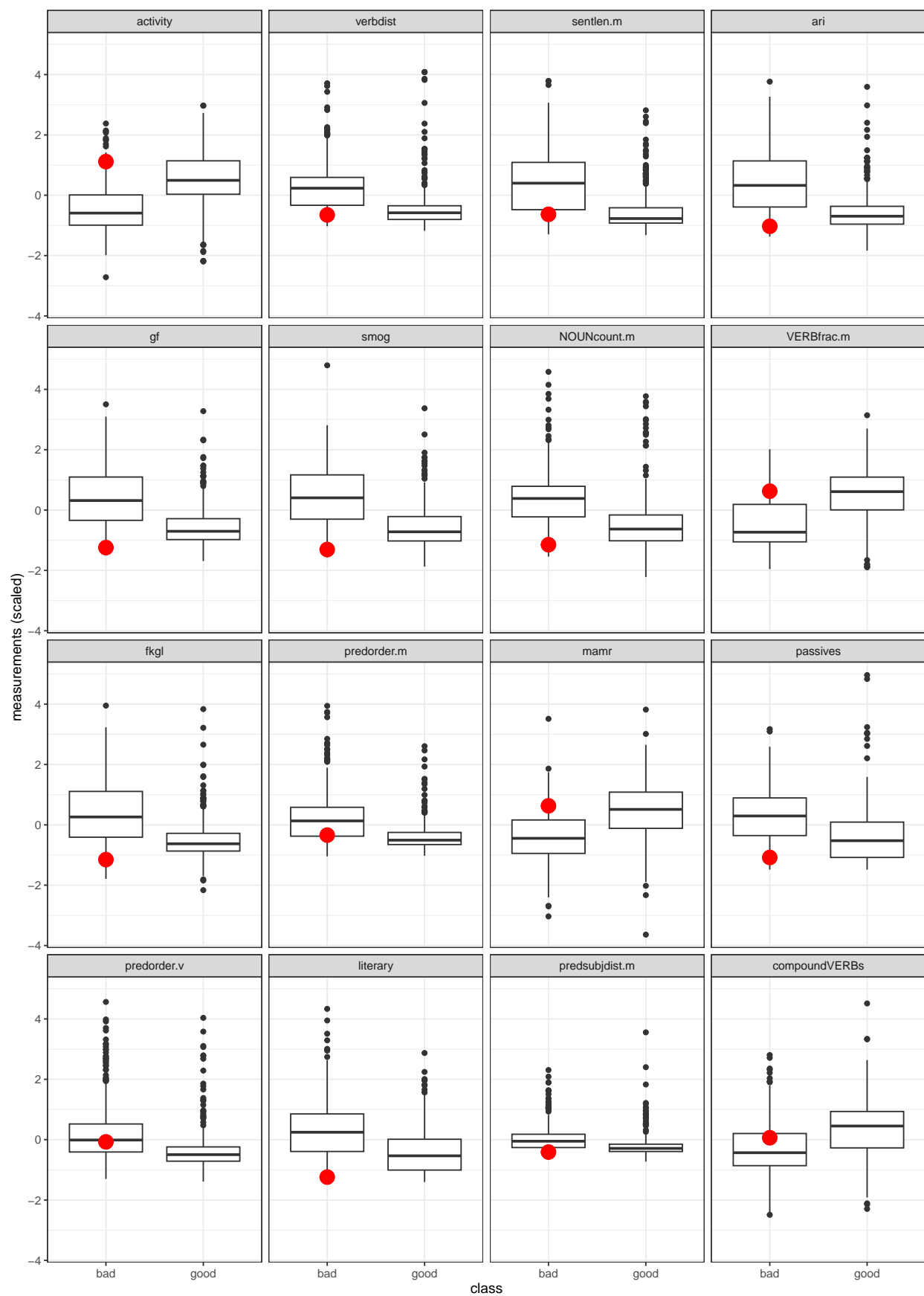
plt <- analyze_outlier(doc, variable_importances_cleaned, data_clean) +
  theme_bw()
print(plt)
ggsave(
  paste(
    c("outlier_cleaned_", doc_row["KUK_ID"][[1]], ".pdf"),
    collapse = ""
  ), plt,
  width = 8,
  height = 8
)
}

```

```

## orig_Jaká pravidla platí pro veřejné zakázky malého rozsahu_final / FrBo
## KUK_ID: Fart_orig_05482
## dev: 0.466
## Readability: medium
## class bad and:
##   rank      feat verbose_score
## 1      1      activity          good
## 2      2      verbdist          good
## 3      3      sentlen.m         good
## 4      4          ari      very good
## 5      5          gf      very good
## 6      6          smog      very good
## 7      7  NOUNcount.m         very good
## 8      8  VERBfrac.m          good
## 9      9          fkg1      very good
## 10     11          mamr          good
## 11     12      passives      very good
## 12     14      literary      very good
## 13     15 predsubjdist.m      very good
## even though:
##   rank      feat verbose_score
## 1     10  predorder.m         medium
## 2     13  predorder.v          bad
## 3     16 compoundVERBs        medium
## 16 observation(s) removed from the plot

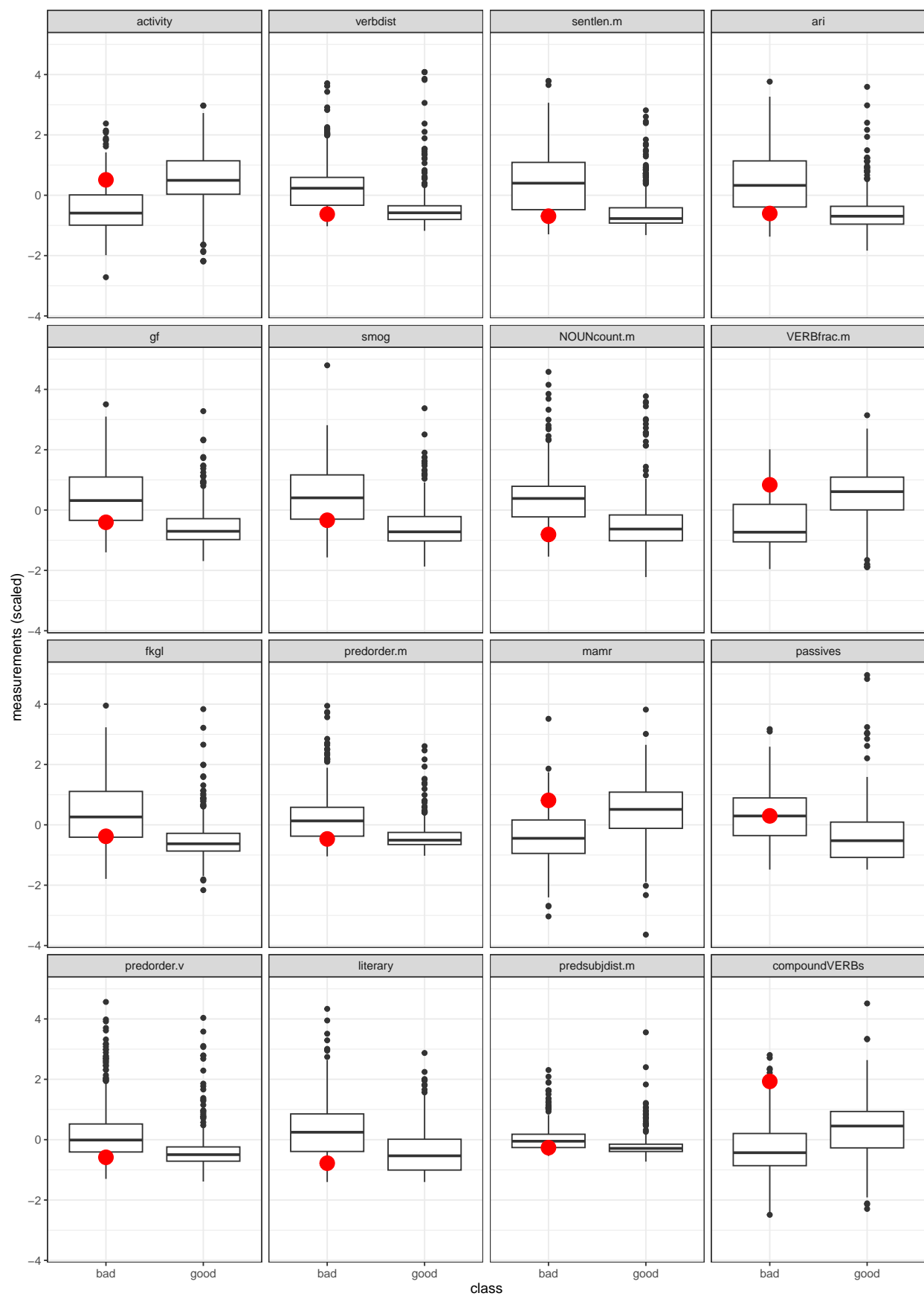
```



```

## orig_Kompletní průvodce pořizováním nahrávek veřejné správy / FrBo
## KUK_ID: Fart_orig_05648
## dev: 0.426
## Readability: medium
## class bad and:
##   rank      feat verbose_score
## 1      1      activity          good
## 2      2      verbdist          good
## 3      3      sentlen.m         good
## 4      4      ari               good
## 5      5      gf                good
## 6      6      smog              good
## 7      7      NOUNcount.m       good
## 8      8      VERBfrac.m        good
## 9     10      predorder.m       good
## 10     11      mamr              good
## 11     13      predorder.v       good
## 12     14      literary          good
## 13     15      predsubjdist.m   good
## 14     16      compoundVERBs    very good
## even though:
##   rank      feat verbose_score
## 1      9      fkg1             medium
## 2     12      passives          bad
## 16 observation(s) removed from the plot

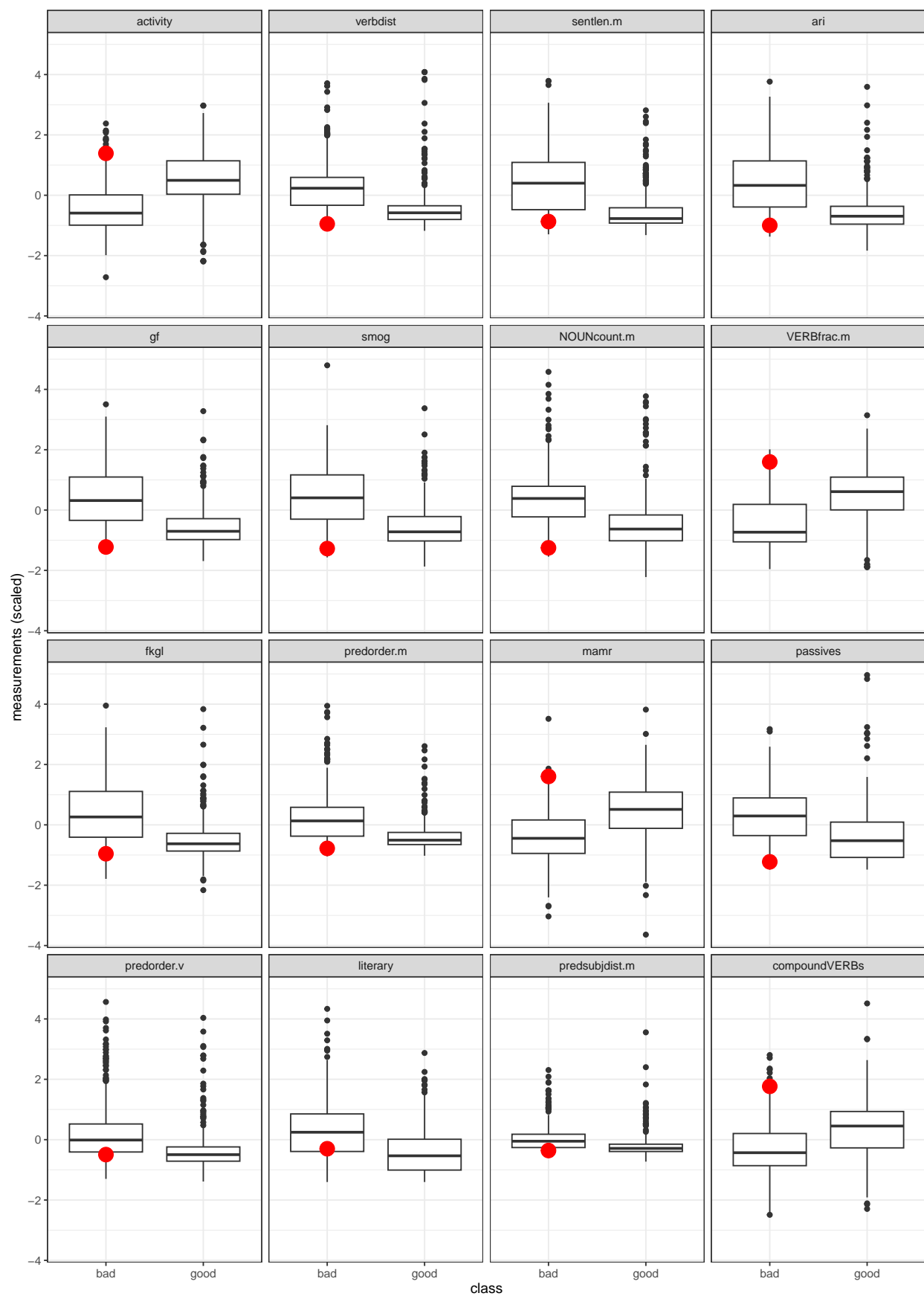
```

```

## orig_Jak řešit lavinovitou černou skládku úprava svépomoc 2021 / FrBo
## KUK_ID: Fart_orig_06078
## dev: 0.385
## Readability: NA
## class bad and:
##   rank      feat verbose_score
## 1      1      activity      very good
## 2      2      verbdist      very good
## 3      3      sentlen.m      good
## 4      4      ari          very good
## 5      5      gf          very good
## 6      6      smog        very good
## 7      7      NOUNcount.m    very good
## 8      8      VERBfrac.m     very good
## 9      9      fkgl        very good
## 10     10     predorder.m     very good
## 11     11     mamr         very good
## 12     12     passives      very good
## 13     13     predorder.v     good
## 14     15     predsubjdist.m  good
## 15     16     compoundVERBs   very good
## even though:
##   rank      feat verbose_score
## 1     14 literary      medium
## 16 observation(s) removed from the plot

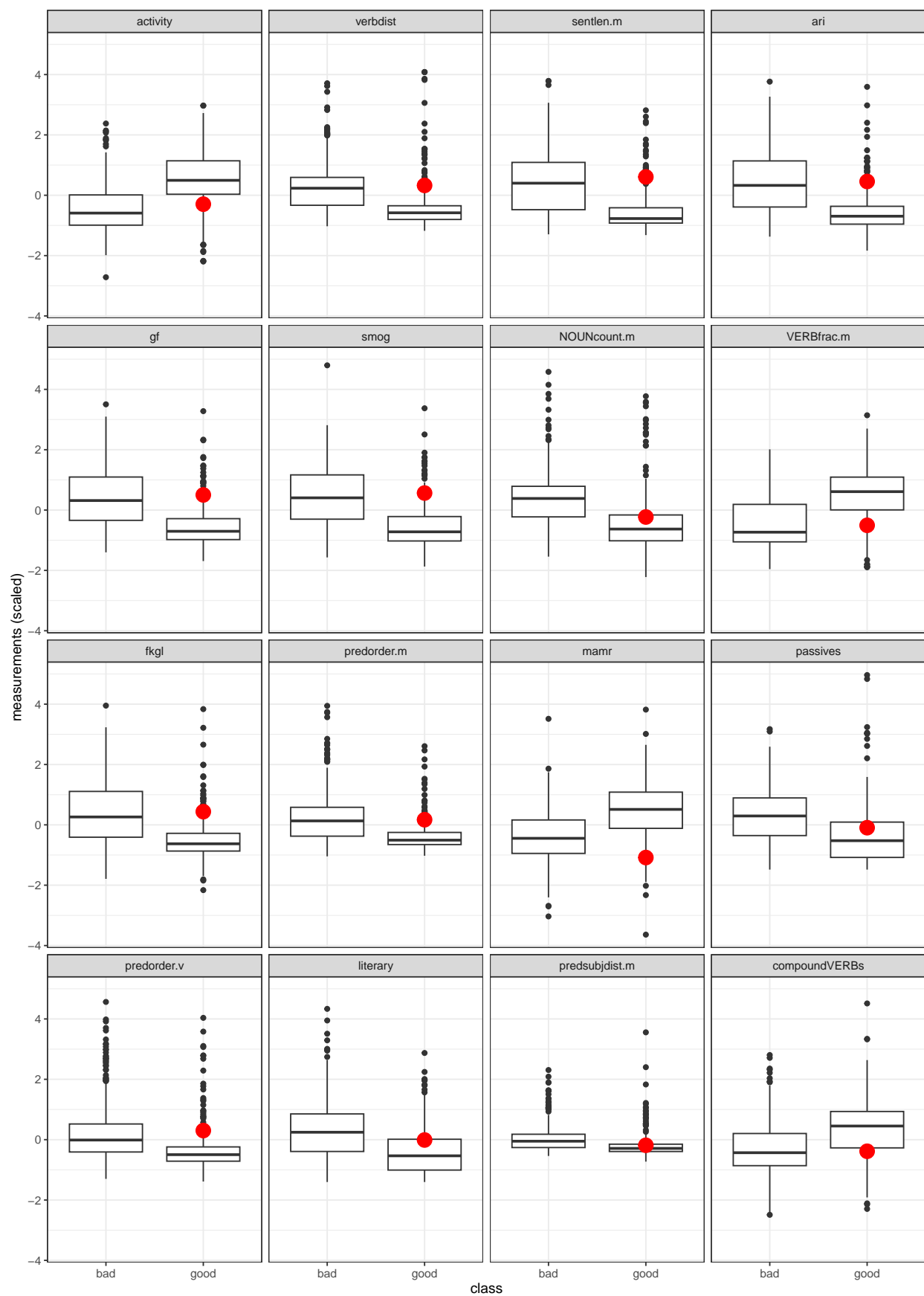
```



```

## AK_JH_Hroch_ustavni_stiznost / KUKY
## KUK_ID: 66f19554c6537d54ff062451
## dev: 0.379
## Readability: high
## class good and:
##      rank      feat verbose_score
## 1      1      activity          bad
## 2      2      verbdist          bad
## 3      3      sentlen.m          bad
## 4      4          ari          bad
## 5      5          gf          bad
## 6      6          smog          bad
## 7      8      VERBfrac.m          bad
## 8      9          fkg1          bad
## 9     10      predorder.m          bad
## 10     11          mamr      very bad
## 11     13      predorder.v          bad
## 12     16      compoundVERBs          bad
## even though:
##      rank      feat verbose_score
## 1      7      NOUNcount.m          good
## 2     12          passives      medium
## 3     14          literary      medium
## 4     15      predsubjdist.m      medium
## 16 observation(s) removed from the plot

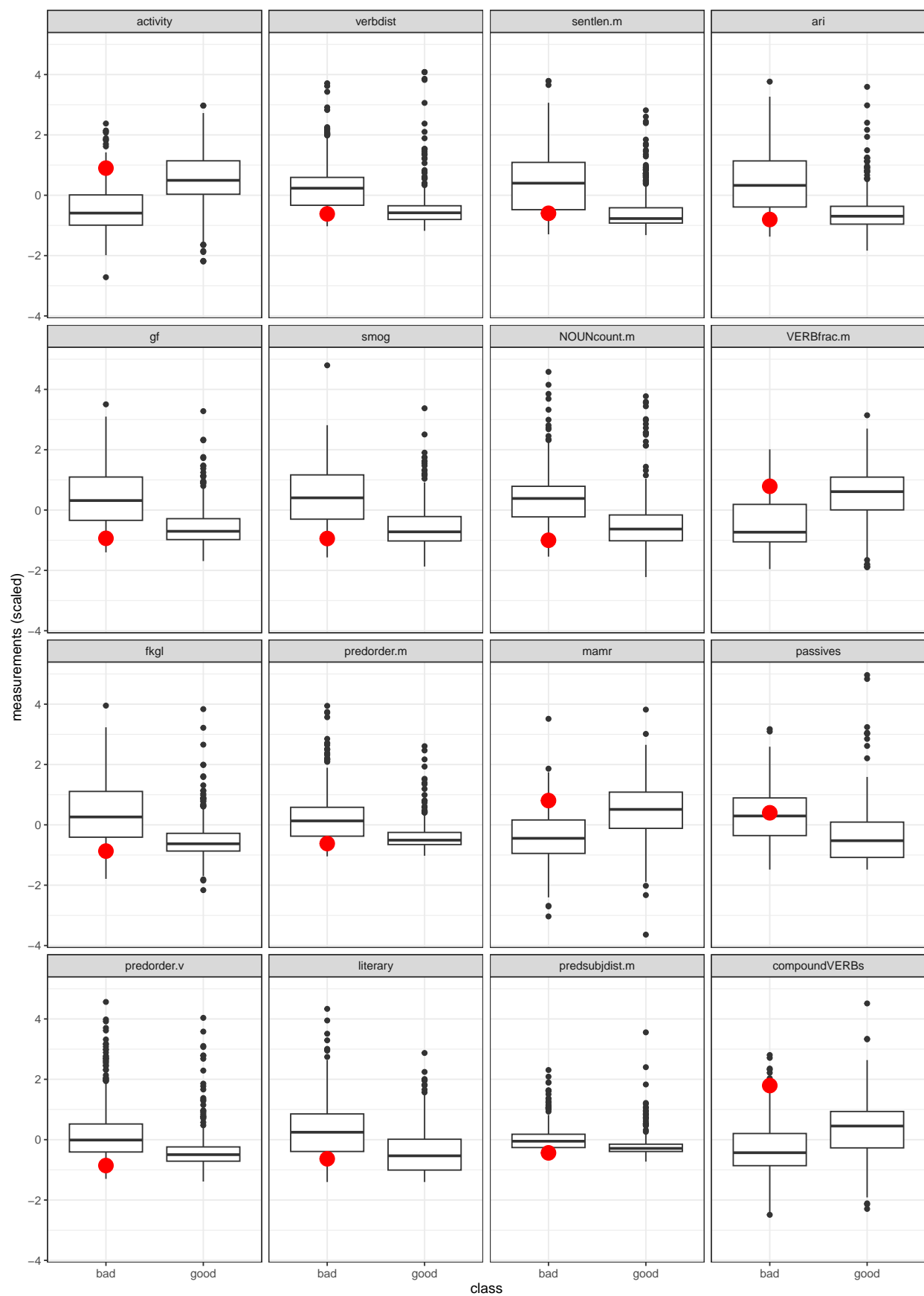
```



```

## PR_Konecny__Miliak / KUKY
## KUK_ID: 66f19554c6537d54ff062454
## dev: 0.376
## Readability: medium
## class bad and:
##   rank      feat verbose_score
## 1      1      activity          good
## 2      2      verbdist          good
## 3      3      sentlen.m         good
## 4      4      ari               good
## 5      5      gf                good
## 6      6      smog              good
## 7      7      NOUNcount.m       good
## 8      8      VERBfrac.m        good
## 9      9      fkgl             very good
## 10     10     predorder.m       good
## 11     11     mamr              good
## 12     13     predorder.v       very good
## 13     14     literary          good
## 14     15     predsubjdist.m    very good
## 15     16     compoundVERBs     very good
## even though:
##   rank      feat verbose_score
## 1      12     passives          bad
## 16 observation(s) removed from the plot

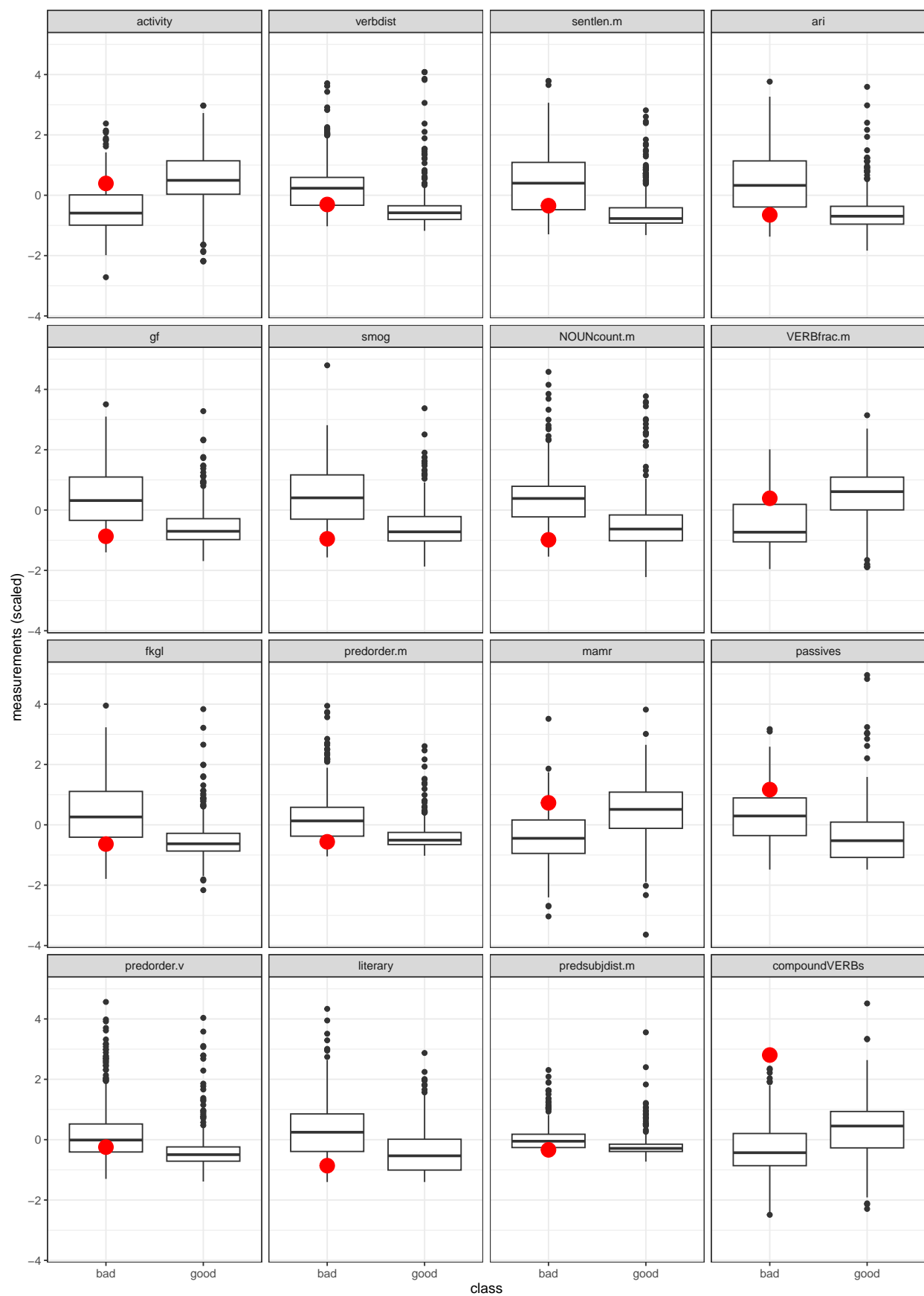
```



```

## PR_Masinova / KUKY
## KUK_ID: 66f19554c6537d54ff06244c
## dev: 0.347
## Readability: medium
## class bad and:
##   rank      feat verbose_score
## 1      1      activity          good
## 2      4          ari          good
## 3      5          gf          good
## 4      6          smog          good
## 5      7  NOUNcount.m          good
## 6      8  VERBfrac.m          good
## 7      9          fkg1          good
## 8     10  predorder.m          good
## 9     11          mamr          good
## 10    14          literary      good
## 11    15  predsubjdist.m        good
## 12    16  compoundVERBs        very good
## even though:
##   rank      feat verbose_score
## 1      2  verbdist              bad
## 2      3  sentlen.m             bad
## 3     12  passives              very bad
## 4     13  predorder.v           medium
## 16 observation(s) removed from the plot

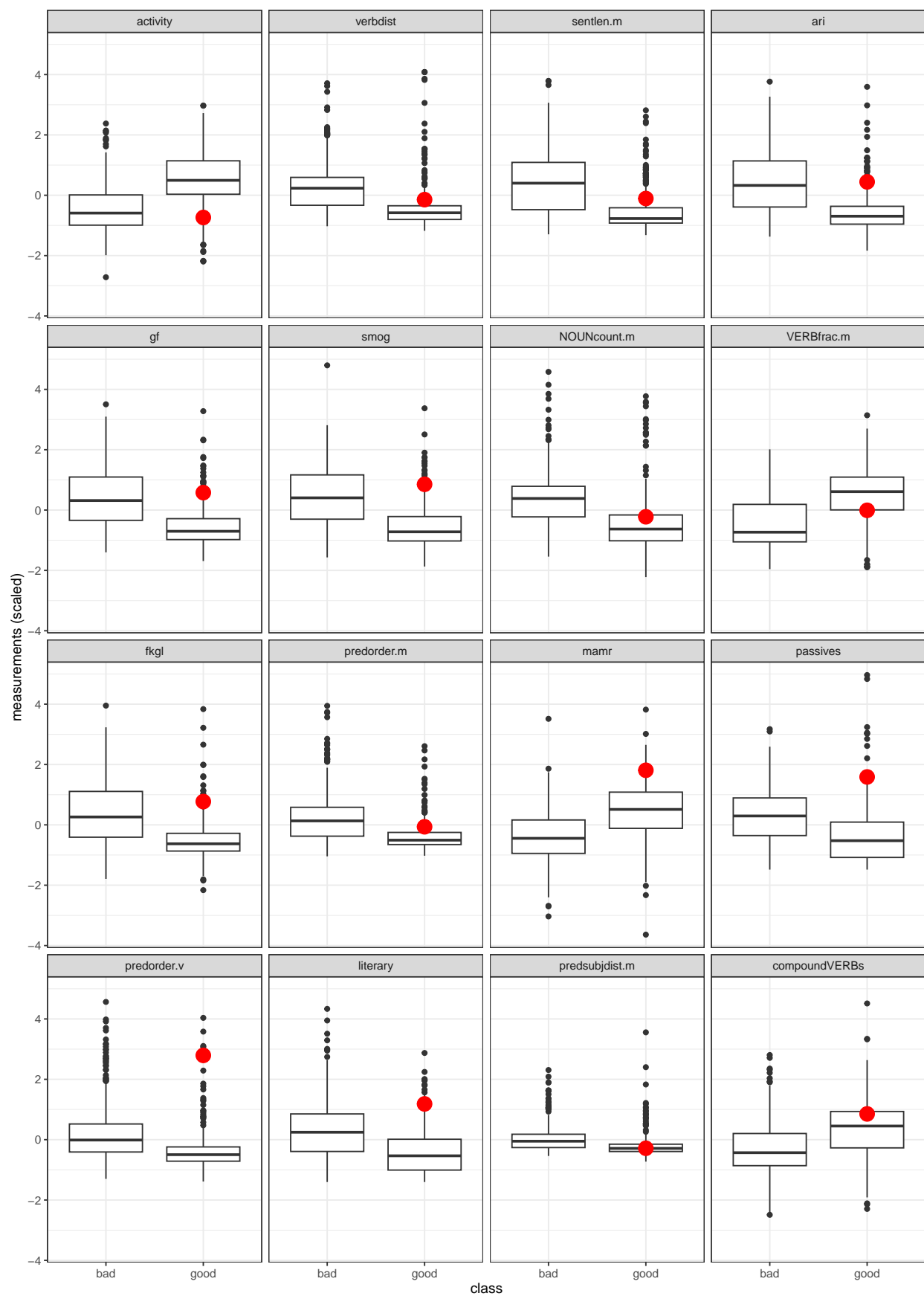
```

```

## red_Certifikáty autorizovaných inspektorů / FrBo
## KUK_ID: Fart_red_00253
## dev: 0.335
## Readability: high
## class good and:
##   rank      feat verbose_score
## 1      1    activity          bad
## 2      2    verbdist          bad
## 3      3    sentlen.m          bad
## 4      4      ari             bad
## 5      5      gf              bad
## 6      6      smog            bad
## 7      8    VERBfrac.m          bad
## 8      9      fkg1            bad
## 9     10 predorder.m          bad
## 10     12    passives        very bad
## 11     13 predorder.v        very bad
## 12     14    literary        very bad
## even though:
##   rank      feat verbose_score
## 1      7    NOUNcount.m        medium
## 2     11      mamr          very good
## 3     15 predsubjdist.m        good
## 4     16 compoundVERBs        good
## 16 observation(s) removed from the plot

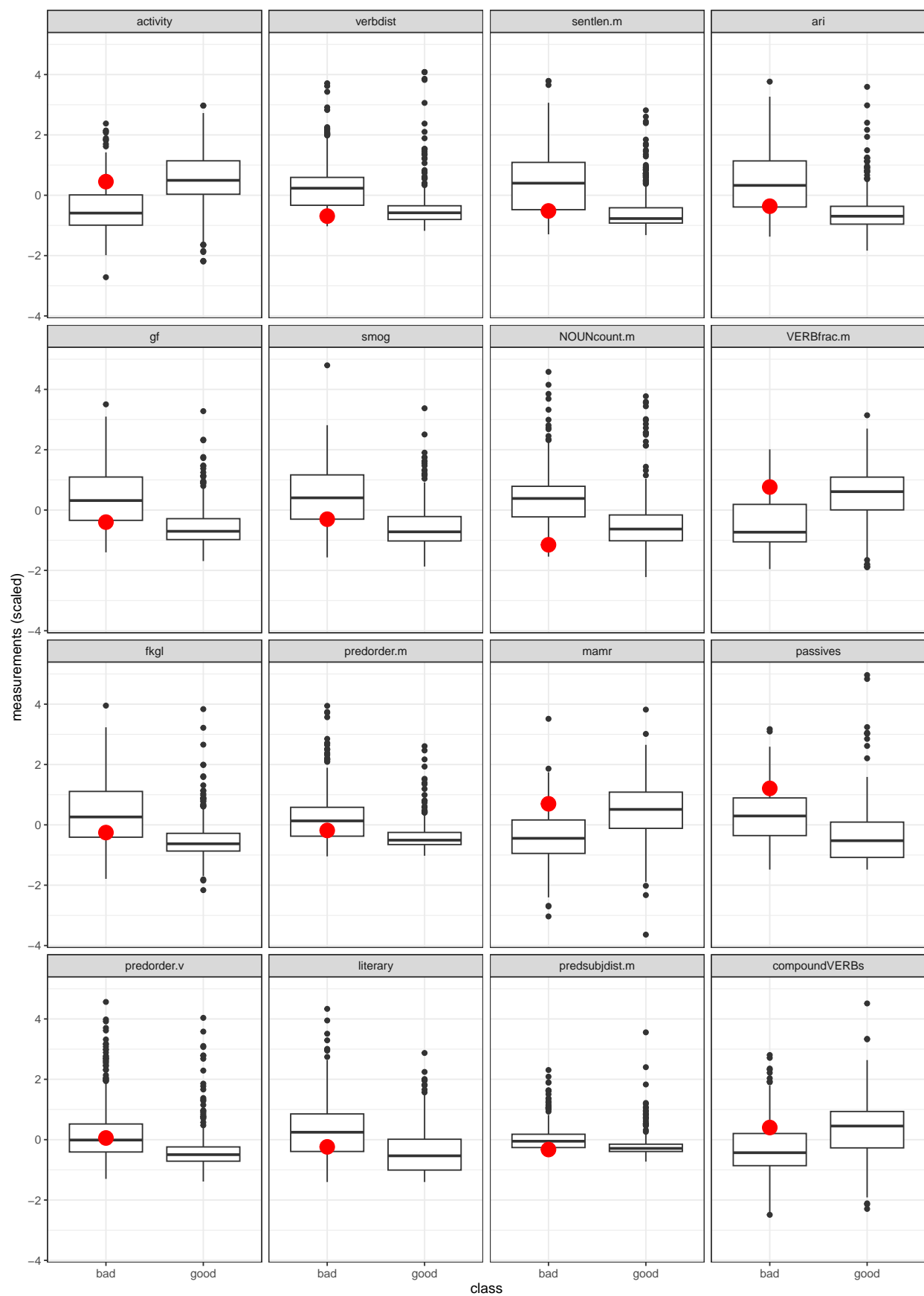
```



```

## orig_Mohou spolky ve správních žalobách používat věcné argumenty_final / FrBo
## KUK_ID: Fart_orig_5938b
## dev: 0.33
## Readability: medium
## class bad and:
##   rank      feat verbose_score
## 1      1      activity          good
## 2      2      verbdist          good
## 3      3      sentlen.m         good
## 4      5      gf                good
## 5      6      smog              good
## 6      7      NOUNcount.m       very good
## 7      8      VERBfrac.m        good
## 8     11      mamr              good
## 9     15      predsubjdist.m    good
## 10    16      compoundVERBs     good
## even though:
##   rank      feat verbose_score
## 1      4      ari              bad
## 2      9      fkg1            bad
## 3     10      predorder.m       bad
## 4     12      passives          very bad
## 5     13      predorder.v       bad
## 6     14      literary          medium
## 16 observation(s) removed from the plot

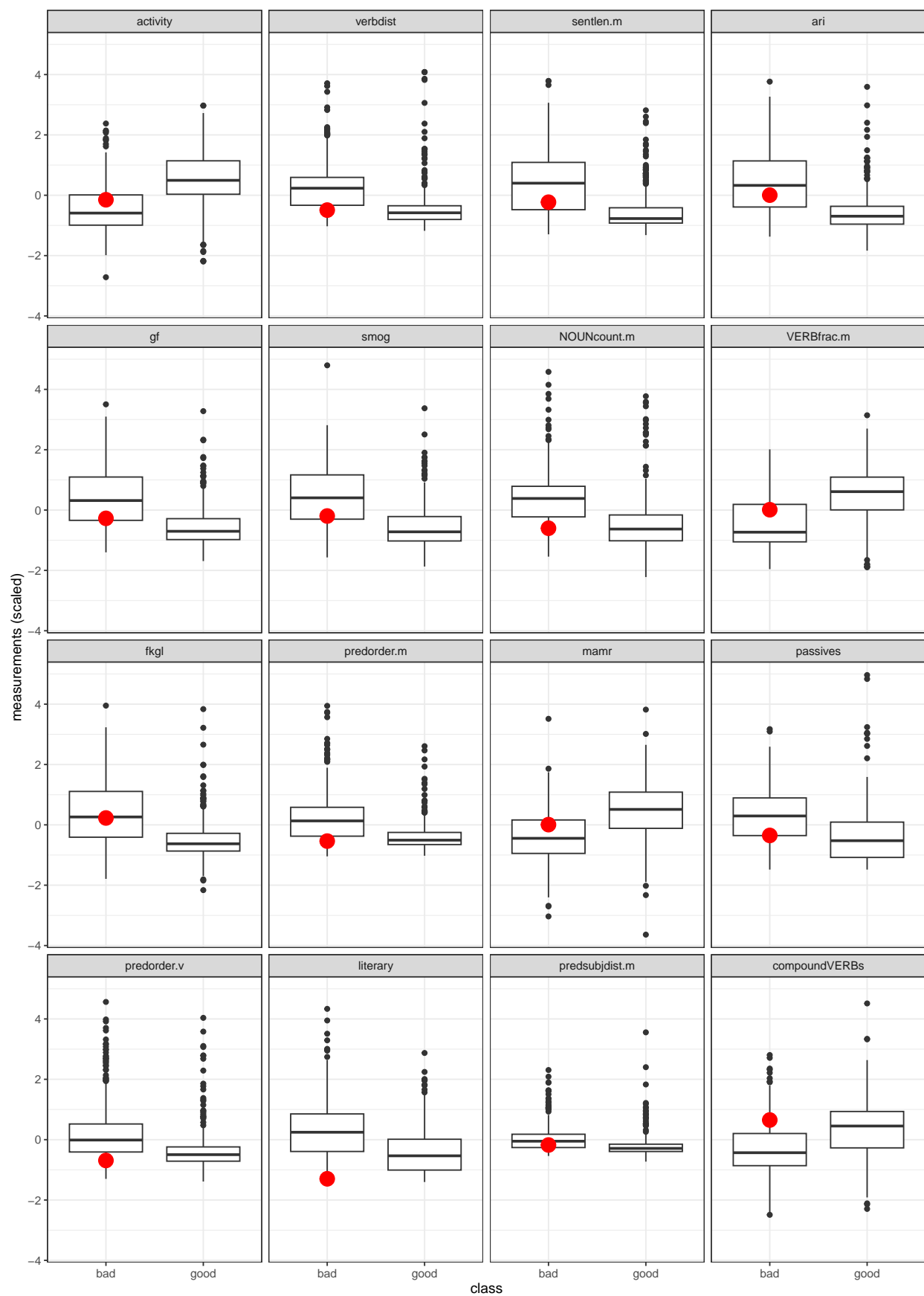
```



```

## orig_Jaké trestné činy mohou souviset s korupcí / FrBo
## KUK_ID: Fart_orig_00285
## dev: 0.322
## Readability: medium
## class bad and:
##   rank      feat verbose_score
## 1    2      verbdist          good
## 2    7  NOUNcount.m          good
## 3   10  predorder.m          good
## 4   13  predorder.v          good
## 5   14      literary    very good
## 6   16 compoundVERBs          good
## even though:
##   rank      feat verbose_score
## 1    1      activity          bad
## 2    3      sentlen.m          bad
## 3    4          ari          bad
## 4    5          gf          bad
## 5    6          smog          bad
## 6    8  VERBfrac.m          medium
## 7    9          fkgl          bad
## 8   11          mamr          medium
## 9   12      passives          medium
## 10  15  predsubjdist.m          medium
## 16 observation(s) removed from the plot

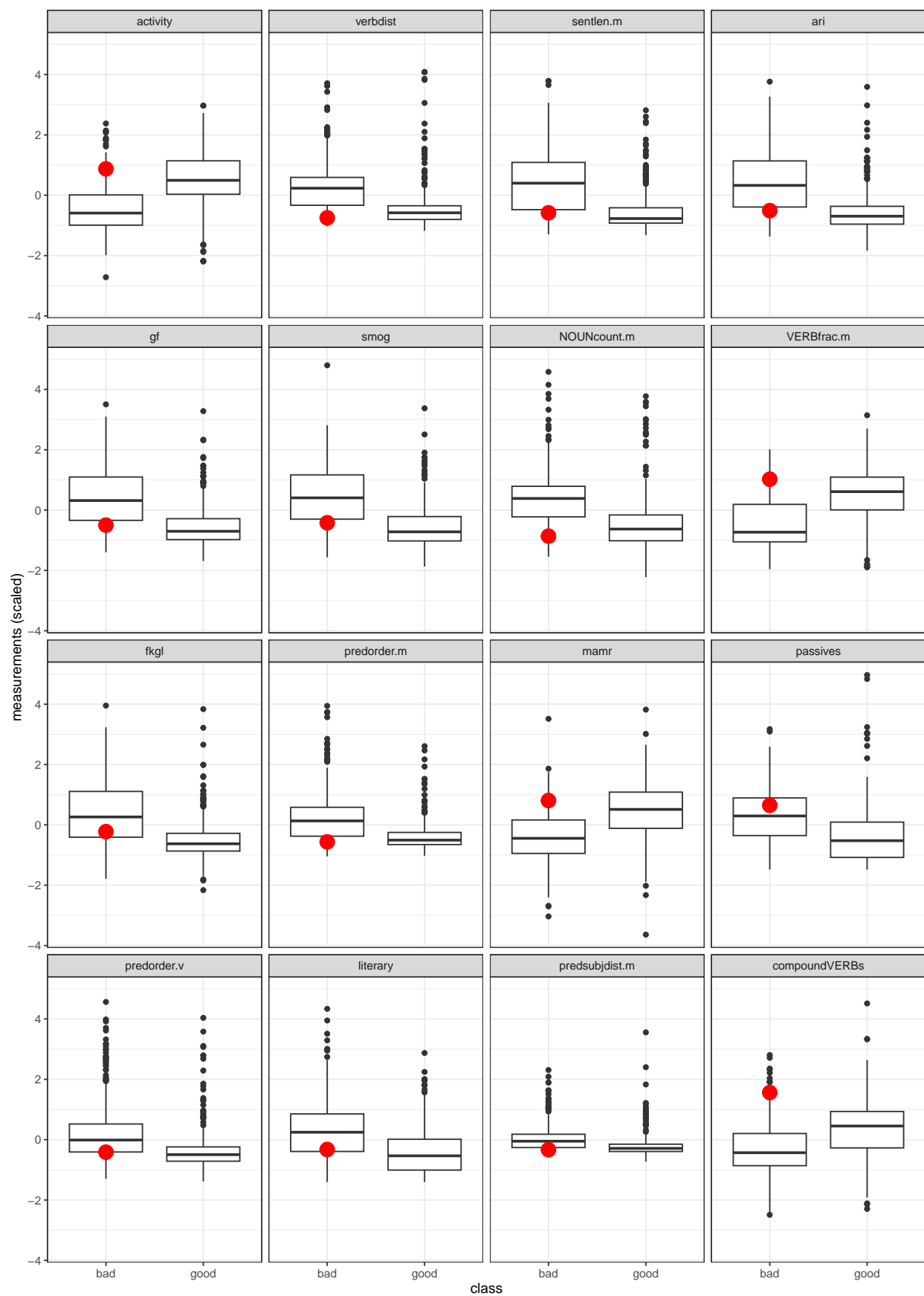
```



```

## orig_Sousedské vztahy / FrBo
## KUK_ID: Fart_orig_0mVfN
## dev: 0.321
## Readability: medium
## class bad and:
##   rank      feat verbose_score
## 1      1      activity          good
## 2      2      verbdist          good
## 3      3      sentlen.m         good
## 4      4      ari               good
## 5      5      gf                good
## 6      6      smog              good
## 7      7      NOUNcount.m       good
## 8      8      VERBfrac.m        good
## 9     10      predorder.m       good
## 10     11      mamr              good
## 11     13      predorder.v       good
## 12     15      predsubjdist.m   good
## 13     16      compoundVERBs    very good
## even though:
##   rank      feat verbose_score
## 1      9      fkg1              bad
## 2     12      passives           bad
## 3     14      literary           medium
## 16 observation(s) removed from the plot

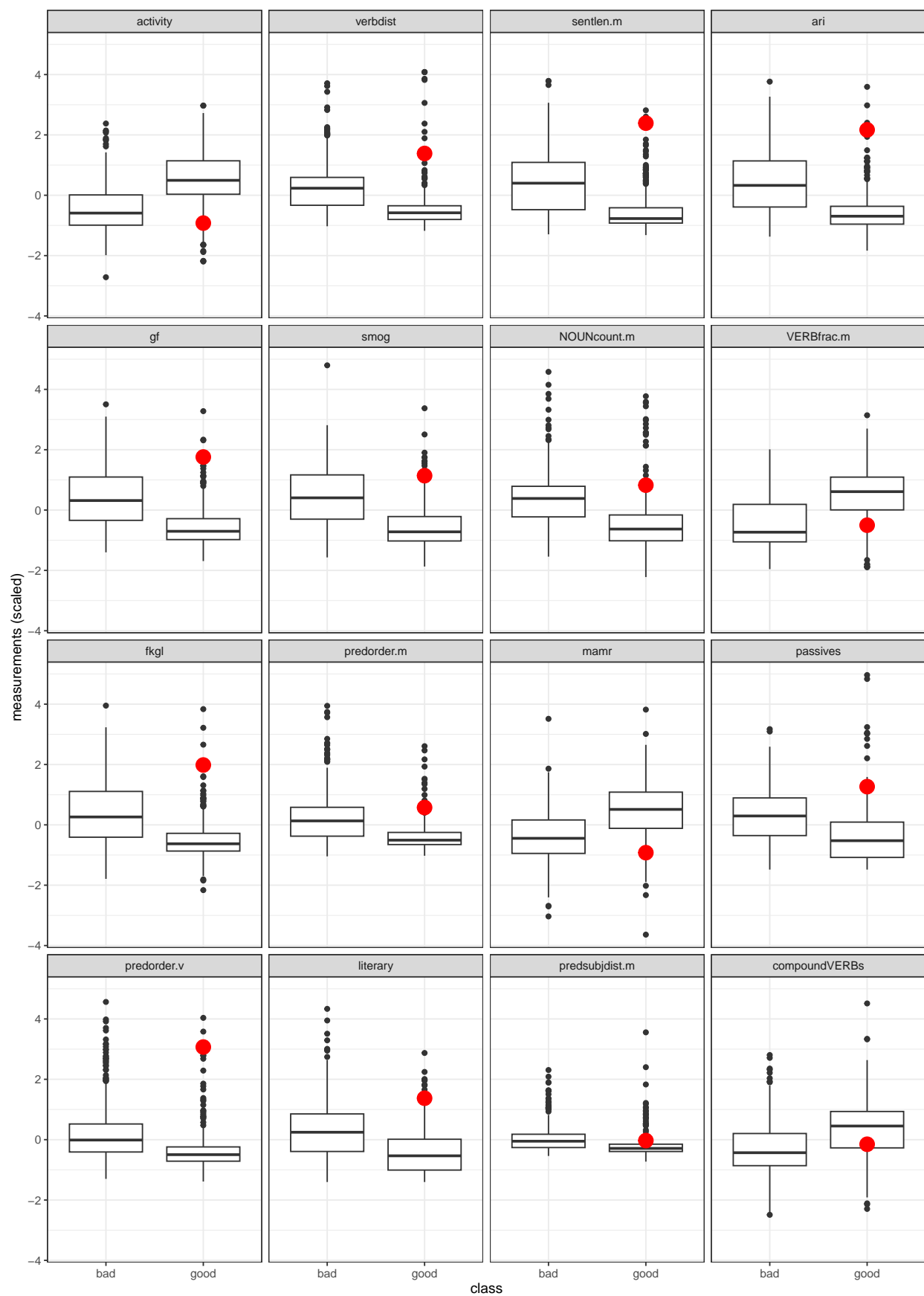
```

```

## Mestsky_urad_PRIKAZ / KUKY
## KUK_ID: 66f1be84c6537d54ff062490
## dev: 0.318
## Readability: high
## class good and:
##      rank      feat verbose_score
## 1      1      activity          bad
## 2      2      verbdist      very bad
## 3      3      sentlen.m      very bad
## 4      4          ari      very bad
## 5      5          gf      very bad
## 6      6      smog          bad
## 7      7  NOUNcount.m      very bad
## 8      8  VERBfrac.m          bad
## 9      9      fkg1      very bad
## 10     10  predorder.m          bad
## 11     11      mamr          bad
## 12     12      passives      very bad
## 13     13  predorder.v      very bad
## 14     14      literary      very bad
## 15     15  predsubjdist.m      bad
## even though:
##      rank      feat verbose_score
## 1      16  compoundVERBs      medium
## 16 observation(s) removed from the plot

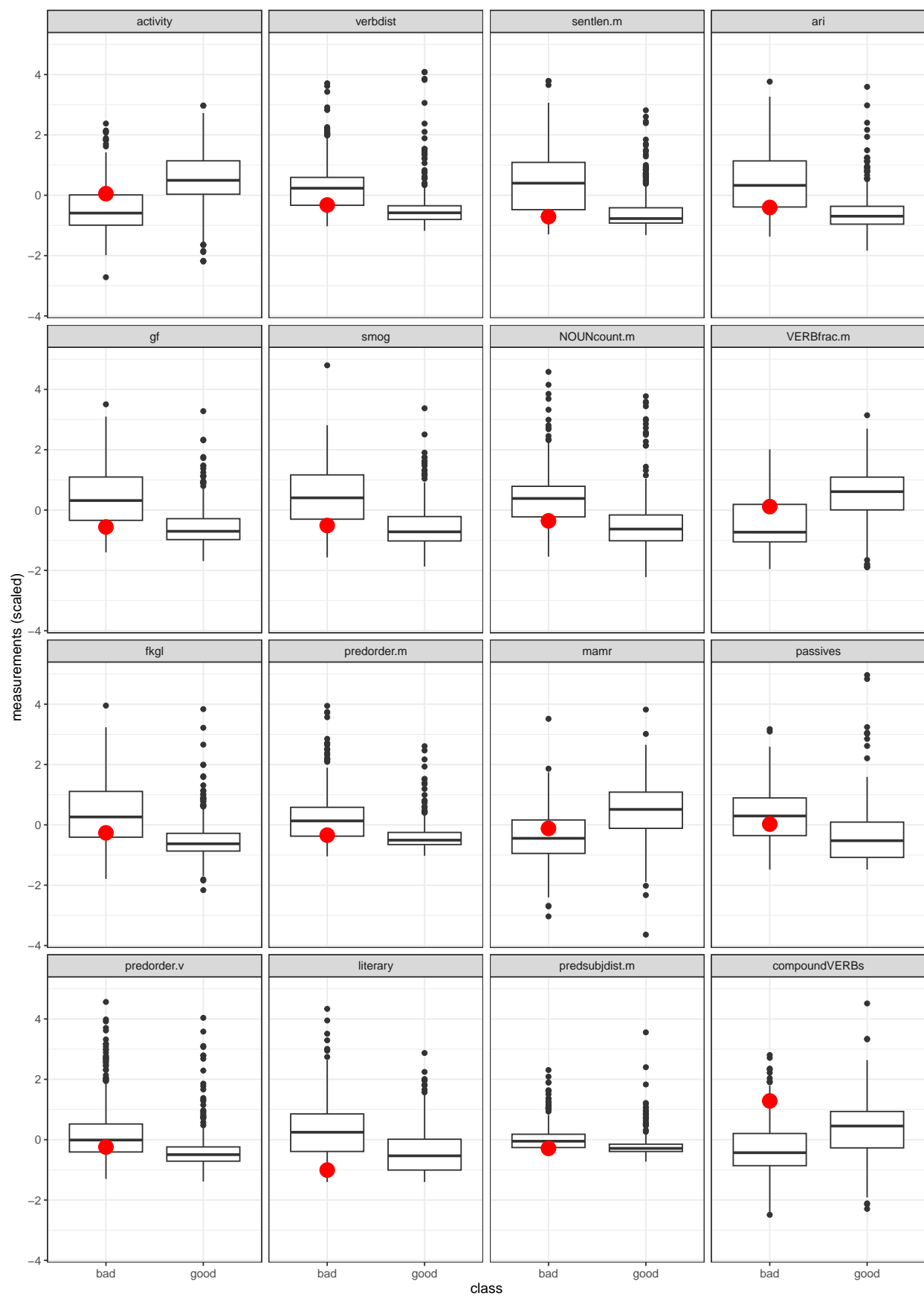
```



```

## 28 / FrBo
## KUK_ID: Fana_00028
## dev: 0.294
## Readability: medium
## class bad and:
##   rank      feat verbose_score
## 1    1      activity          good
## 2    3      sentlen.m          good
## 3    4          ari            good
## 4    5          gf             good
## 5    6          smog           good
## 6    7      NOUNcount.m        good
## 7   14      literary           good
## 8   15 predsubjdist.m          good
## 9   16 compoundVERBs          very good
## even though:
##   rank      feat verbose_score
## 1    2      verbdist           bad
## 2    8      VERBfrac.m         medium
## 3    9          fkg1           bad
## 4   10 predorder.m            medium
## 5   11          mamr           bad
## 6   12      passives           medium
## 7   13 predorder.v            medium
## 16 observation(s) removed from the plot

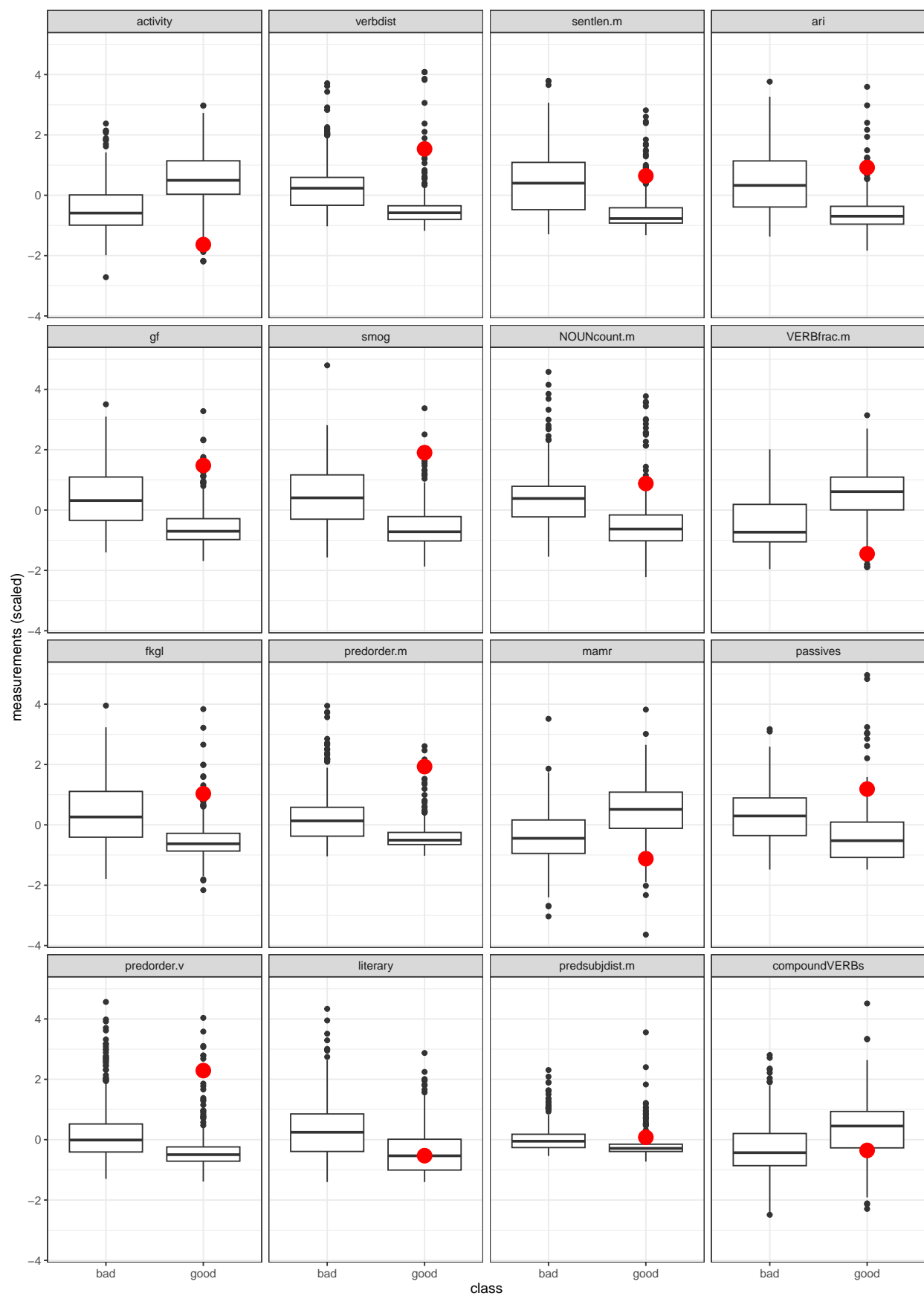
```



```

## Odvolani_proti_rozhodnuti_o_nepovoleni_kaceni / KUKY
## KUK_ID: 6745acb5c6537d54ff0636ca
## dev: 0.278
## Readability: high
## class good and:
##   rank      feat verbose_score
## 1      1      activity      very bad
## 2      2      verbdist      very bad
## 3      3      sentlen.m      bad
## 4      4      ari           bad
## 5      5      gf           very bad
## 6      6      smog          very bad
## 7      7      NOUNcount.m    very bad
## 8      8      VERBfrac.m     very bad
## 9      9      fkg1          bad
## 10     10     predorder.m     very bad
## 11     11     mamr           very bad
## 12     12     passives       very bad
## 13     13     predorder.v     very bad
## 14     15     predsubjdist.m  bad
## 15     16     compoundVERBs   bad
## even though:
##   rank      feat verbose_score
## 1      14     literary        good
## 16 observation(s) removed from the plot

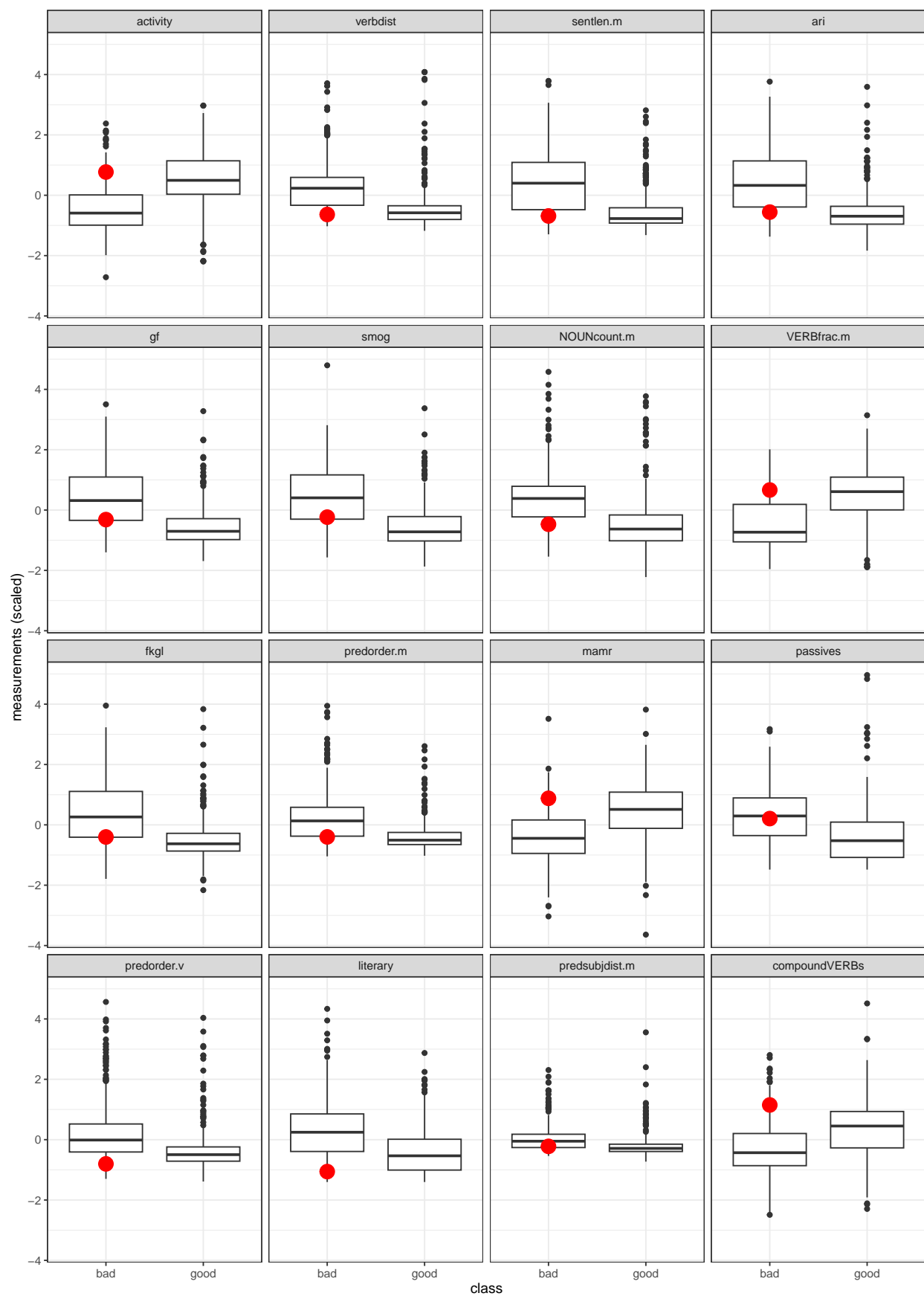
```



```

## orig_Jak se bránit neposkytnutí projektové dokumentace / FrBo
## KUK_ID: Fart_orig_00259
## dev: 0.275
## Readability: medium
## class bad and:
##      rank      feat verbose_score
## 1      1      activity           good
## 2      2      verbdist           good
## 3      3      sentlen.m          good
## 4      4          ari            good
## 5      7  NOUNcount.m            good
## 6      8  VERBfrac.m             good
## 7     10  predorder.m            good
## 8     11      mamr               good
## 9     13  predorder.v           very good
## 10    14      literary           very good
## 11    16  compoundVERBs         very good
## even though:
##      rank      feat verbose_score
## 1      5          gf            medium
## 2      6          smog           medium
## 3      9          fkg1           medium
## 4     12      passives           bad
## 5     15  predsubjdist.m         medium
## 16 observation(s) removed from the plot

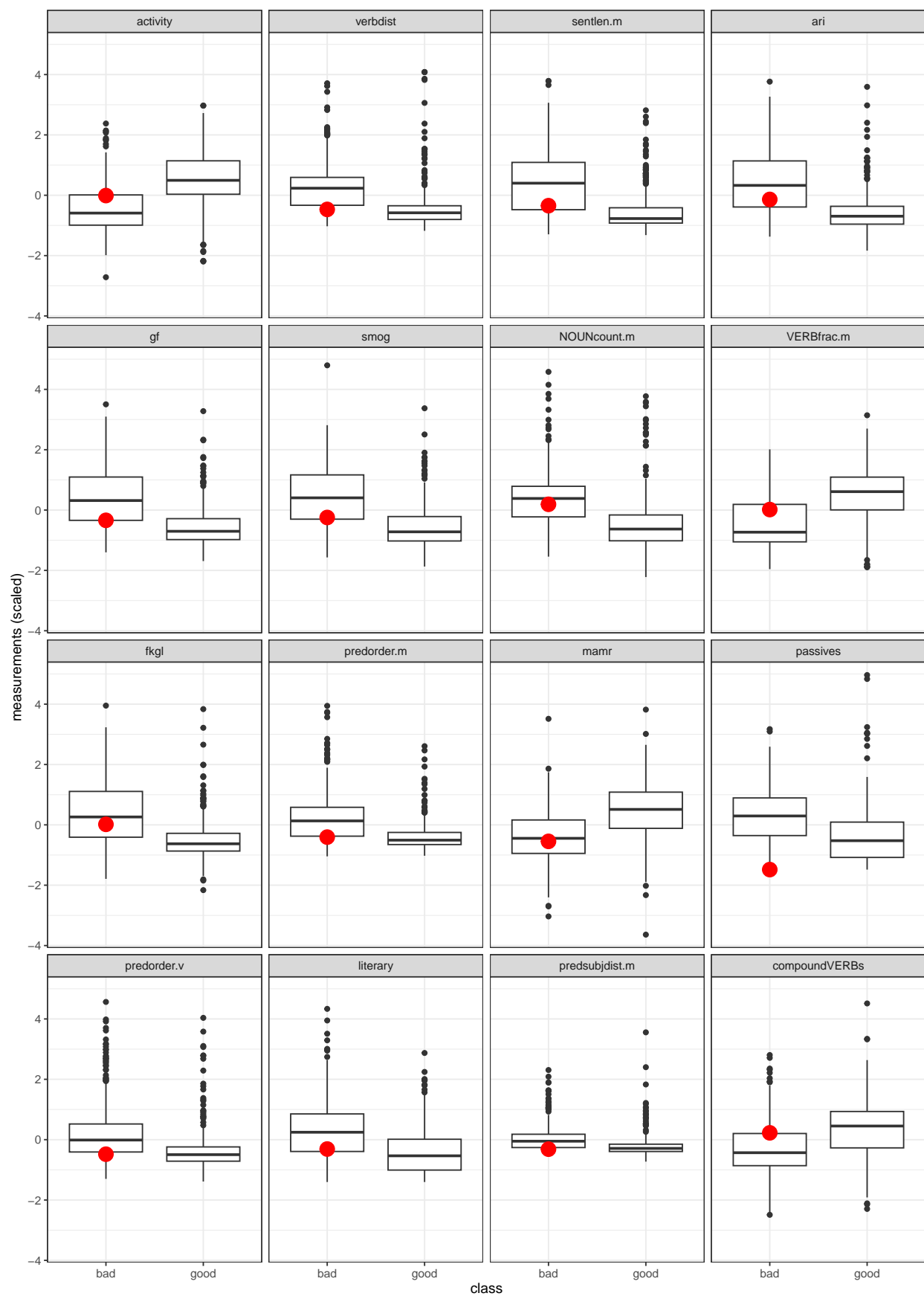
```

```

## orig_Jak se bránit obtěžování kouřem a pálením odpadu / FrBo
## KUK_ID: Fart_orig_05620
## dev: 0.266
## Readability: medium
## class bad and:
##   rank      feat verbose_score
## 1    2      verbdist          good
## 2   10    predorder.m          good
## 3   12      passives    very good
## 4   13    predorder.v          good
## 5   15 predsubjdist.m          good
## 6   16  compoundVERBs          good
## even though:
##   rank      feat verbose_score
## 1    1    activity          bad
## 2    3   sentlen.m          bad
## 3    4        ari          bad
## 4    5         gf    medium
## 5    6        smog    medium
## 6    7 NOUNcount.m          bad
## 7    8  VERBfrac.m    medium
## 8    9        fkg1          bad
## 9   11        mamr          bad
## 10   14   literary    medium
## 16 observation(s) removed from the plot

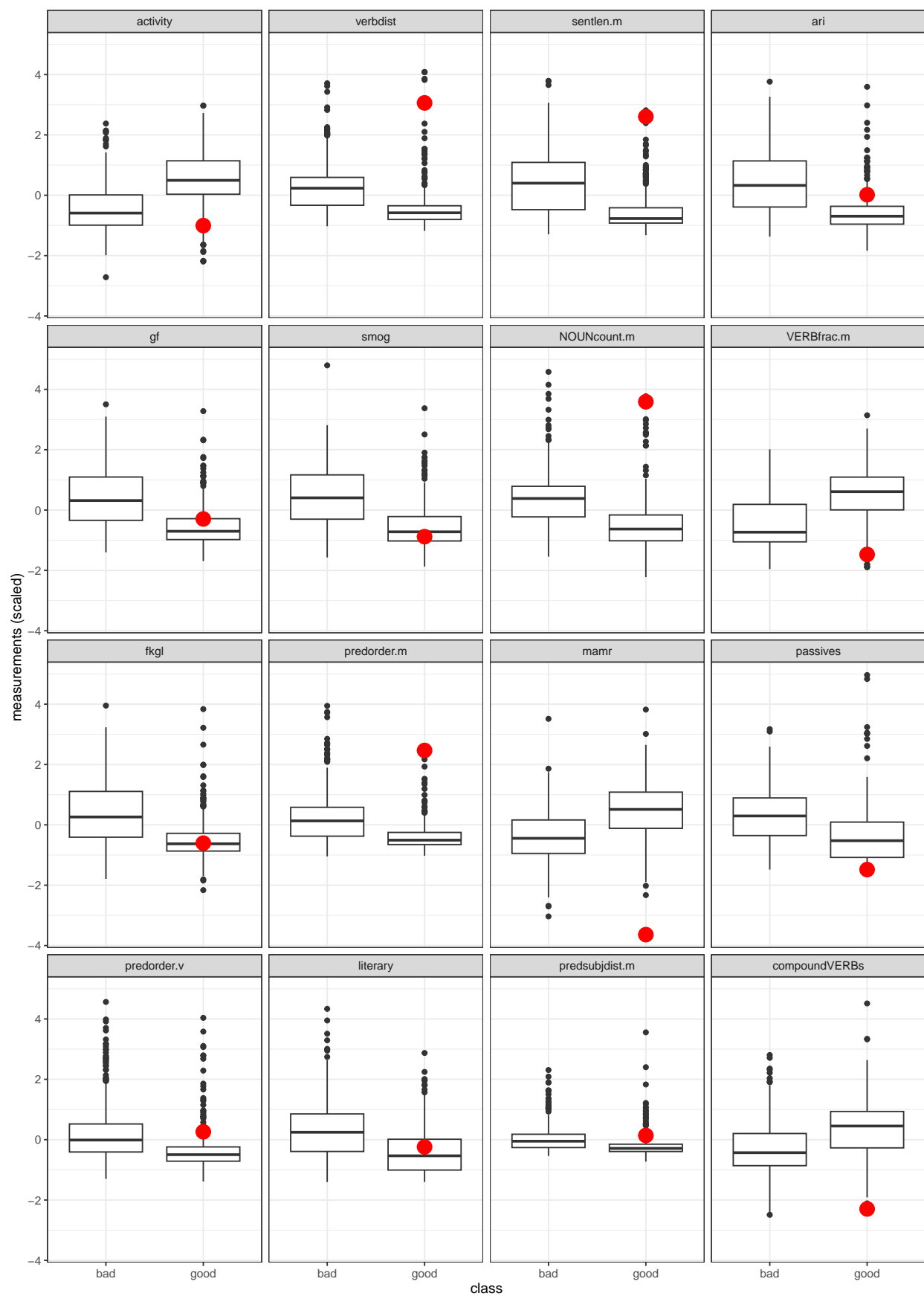
```



```

## Mestsky_urad_usneseni_-_slouceni_pred / KUKY
## KUK_ID: 66f19554c6537d54ff062453
## dev: 0.237
## Readability: high
## class good and:
##   rank      feat verbose_score
## 1      1      activity      very bad
## 2      2      verbdist      very bad
## 3      3      sentlen.m      very bad
## 4      4      ari           bad
## 5      7      NOUNcount.m     very bad
## 6      8      VERBfrac.m      very bad
## 7     10      predorder.m      very bad
## 8     11      mamr           very bad
## 9     13      predorder.v      bad
## 10    15      predsubjdist.m    bad
## 11    16      compoundVERBs     very bad
## even though:
##   rank      feat verbose_score
## 1      5      gf           medium
## 2      6      smog          good
## 3      9      fkg1          good
## 4     12      passives      very good
## 5     14      literary      medium
## 16 observation(s) removed from the plot

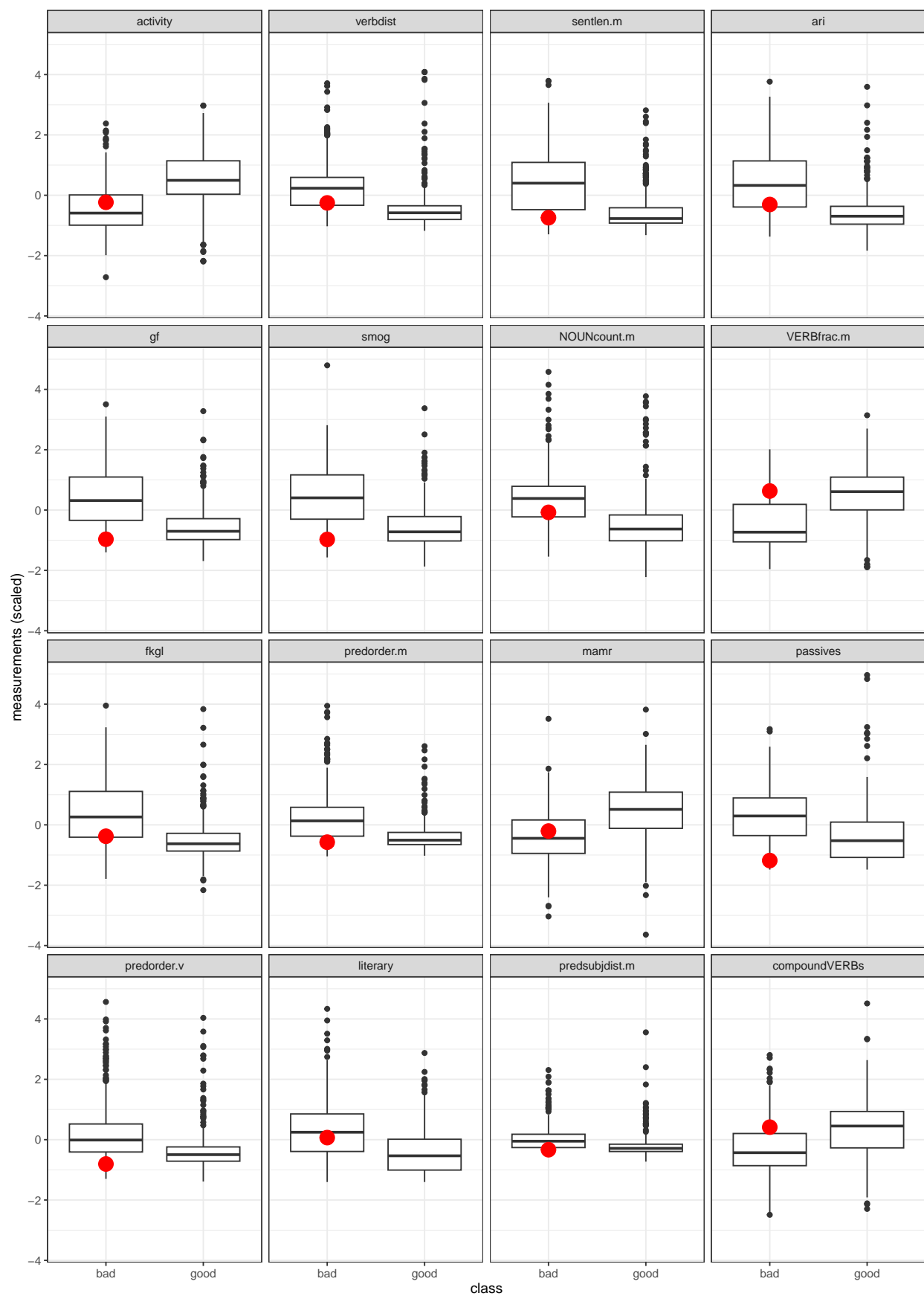
```



```

## 153 / FrBo
## KUK_ID: Fana_00153
## dev: 0.236
## Readability: medium
## class bad and:
##   rank      feat verbose_score
## 1    3      sentlen.m          good
## 2    5          gf            good
## 3    6          smog            good
## 4    8      VERBfrac.m          good
## 5   10    predorder.m          good
## 6   12      passives        very good
## 7   13    predorder.v        very good
## 8   15 predsubjdist.m          good
## 9   16  compoundVERBs          good
## even though:
##   rank      feat verbose_score
## 1    1    activity            bad
## 2    2    verbdist            bad
## 3    4      ari              bad
## 4    7 NOUNcount.m            bad
## 5    9      fkg1             medium
## 6   11      mamr             bad
## 7   14    literary            bad
## 16 observation(s) removed from the plot

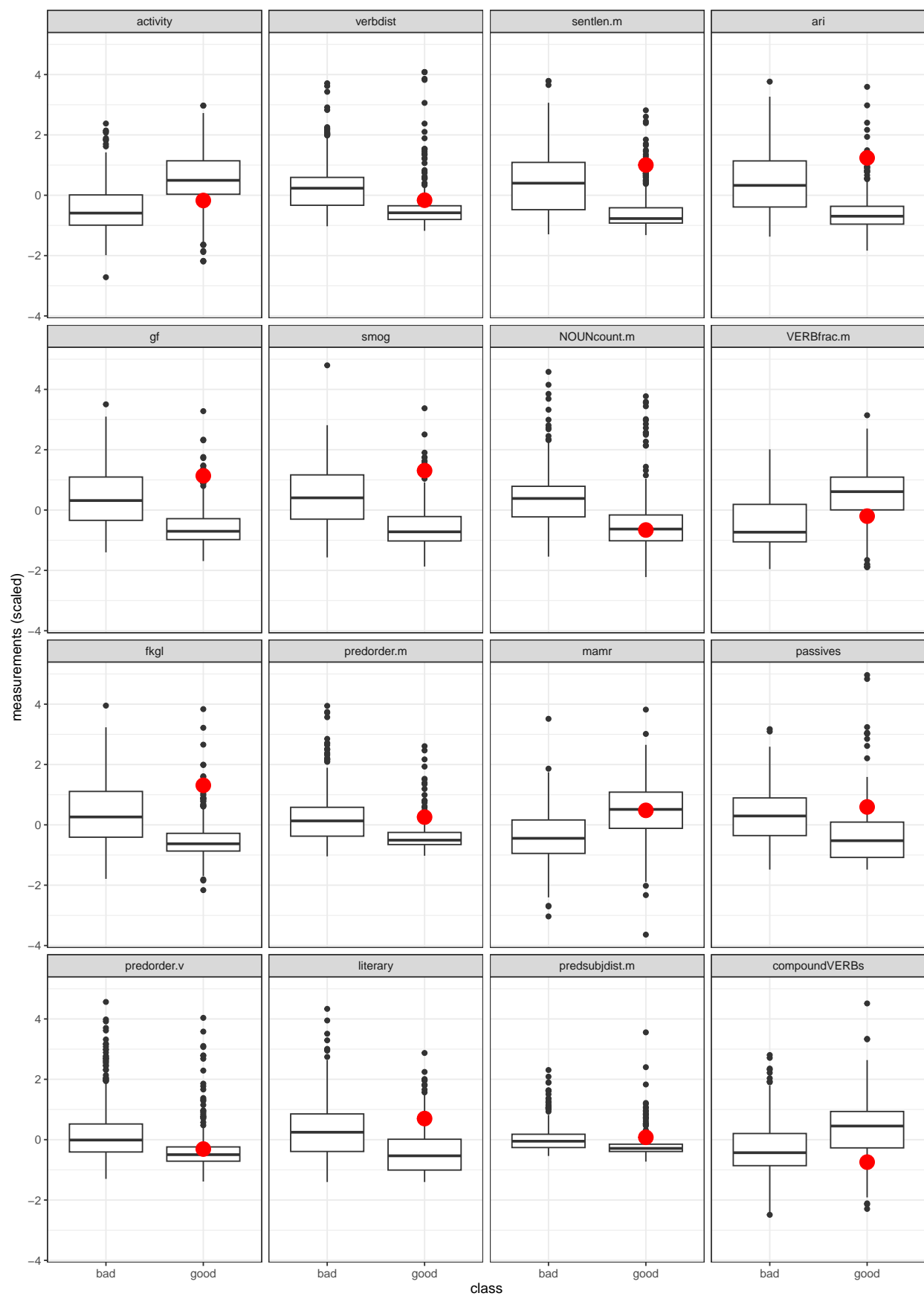
```



```

## AK_JH_Podani_US_podpis / KUKY
## KUK_ID: 66f19554c6537d54ff06244f
## dev: 0.225
## Readability: high
## class good and:
##      rank      feat verbose_score
## 1      1      activity          bad
## 2      2      verbdist          bad
## 3      3      sentlen.m          bad
## 4      4          ari      very bad
## 5      5          gf      very bad
## 6      6      smog      very bad
## 7      8      VERBfrac.m          bad
## 8      9      fkg1      very bad
## 9     10      predorder.m          bad
## 10     12      passives          bad
## 11     14      literary          bad
## 12     15      predsubjdist.m          bad
## 13     16      compoundVERBs          bad
## even though:
##      rank      feat verbose_score
## 1      7 NOUNcount.m          good
## 2     11      mamr          good
## 3     13      predorder.v      medium
## 16 observation(s) removed from the plot

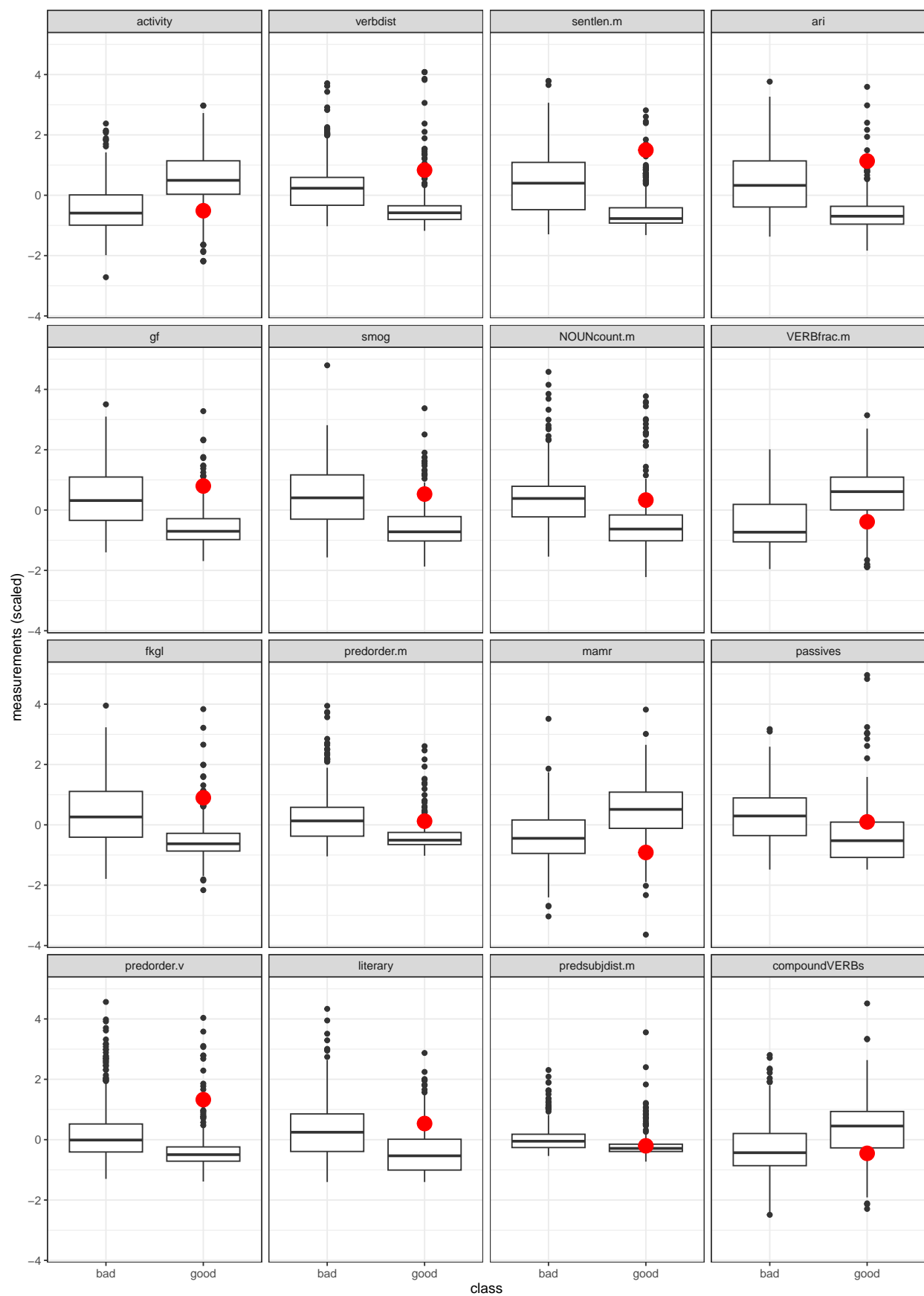
```

```

## Mestsky_urad_PRIKAZ_REV2 / KUKY
## KUK_ID: 66f1be84c6537d54ff062492
## dev: 0.189
## Readability: high
## class good and:
##      rank      feat verbose_score
## 1      1      activity          bad
## 2      2      verbdist      very bad
## 3      3      sentlen.m      very bad
## 4      4          ari          bad
## 5      5          gf          bad
## 6      6          smog          bad
## 7      7  NOUNcount.m          bad
## 8      8  VERBfrac.m          bad
## 9      9          fkg1          bad
## 10     10 predorder.m          bad
## 11     11          mamr          bad
## 12     12      passives          bad
## 13     13 predorder.v      very bad
## 14     14      literary          bad
## 15     16 compoundVERBs          bad
## even though:
##      rank      feat verbose_score
## 1      15 predsubjdist.m      medium
## 16 observation(s) removed from the plot

```



```

## 043_Plisen-a-zavady-v-byte / KUKY
## KUK_ID: 673b7a38c6537d54ff062bb3
## dev: 0.183
## Readability: low
## class bad and:
##      rank      feat verbose_score
## 1      1      activity      good
## 2      2      verbdist      good
## 3      3      sentlen.m      good
## 4      4      ari      very good
## 5      5      gf      very good
## 6      6      smog      very good
## 7      8      VERBfrac.m      very good
## 8      9      fkg1      very good
## 9     10     predorder.m      very good
## 10     11     mamr      good
## 11     12     passives      very good
## 12     13     predorder.v      very good
## 13     15     predsubjdist.m      good
## even though:
##      rank      feat verbose_score
## 1      7     NOUNcount.m      medium
## 2     14     literary      very bad
## 3     16     compoundVERBs      medium
## 16 observation(s) removed from the plot

```

