Fake_News_Data_Cleaning

Natália Tosi

8/3/2021

Importing Data and Labels

```r
data_raw_port <- read_csv(
  "BANCO_NACIONAL_FAKENEWS_2021-08-03_CLEAN - label.csv")

variable_names <- read_csv(
  "BANCO_NACIONAL_FAKENEWS_2021-08-03_CLEAN - variable_names.csv")

answers_labels <- read_csv(
  "BANCO_NACIONAL_FAKENEWS_2021-08-03_CLEAN - answers_translated.csv")
```

Translating Document

```r
data_clean <- as_tibble(data_raw_port)
colnames(data_clean) <- as_vector(variable_names[,2])

data_eng <- data_clean
data_eng[-c(1,2,3,7)] <- lapply(data_clean[-c(1,2,3,7)],
      function(x) answers_labels$answers_eng[match(x,
                                        answers_labels$answers_port)])
```

Save as new CSV file

```r
write.csv(data_eng,'fake_news_db_english.csv')
```

Summary Statistics

```r
data_eng <- data_eng %>%
  mutate(
    evaluation = case_when(
    P1 %in% c("Excellent", "Good") ~ "Excellent/Good",
    P1 %in% c("Bad", "Terrible") ~ "Bad/Terrible",
    TRUE ~ P1),
    approval = case_when(
      P2 %in% c("Strongly approves", "Approves") ~ "Approves",
      P2 %in% c("Strongly disapproves", "Disapproves") ~ "Disapproves",
      TRUE ~ P2))


data_fake_news_dem <- data_eng %>%
  select(idInterview, state, region, type, sex, age_full, age_60, evaluation,
         approval, P4, P19, P20, P21, P23_1, P23_2, P23_3, P23_4, P23_5,
         education_full, race, religion_full, income_full, class_full, age_50,
         education, income, class, religion) %>%
  mutate(shared_fake_news = if_else(P19 == "Yes", 1, 0))


data_fake_news_dem <- data_fake_news_dem %>%
  mutate(sex = factor(sex, levels = c("Men", "Women")),
         region = factor(region, levels = c("North", "Northeast", "Center-West",
                                   "Southeast", "South")),
         type = factor(type, levels = c("Capital", "Metropolitan region",
                                        "Countryside")),
         evaluation = factor(evaluation, levels = c("Excellent/Good", "Regular",
                                        "Bad/Terrible", "Unsure")),
         approval = factor(approval, levels = c("Approves",
                                     "Neither approves nor disapproves",
                                      "Disapproves", "Unsure")),
         P4 = factor(P4, levels = c("Right/Center-Right", "Center",
                                    "Left/Center-Left",
                            "I no longer have a defined political orientation",
                            "I never had a political orientation", "Unsure")),
         race = factor(race, levels = c("White", "Black", "Pardo (brown)",
                                        "Indigenous", "Yellow", "Other")),
         education = factor(education, levels = c("No education", "Elementary School",
                                       "High School", "Higher Education")),
         income = factor(income, levels = c("Up to 1 MW", "1 to 3 MWs", "3 to 6 MWs",
                                    "More than 6 MWs", "Did not answer")),
         class = factor(class, levels = c("A/B", "C", "D/E", "DN/DA")),
         religion = factor(religion, levels = c("Catholic", "Evangelicals",
                                      "Other religion", "No religion")))


#data_fake_news_dem <- data_fake_news_dem %>%
 # mutate_if(is.character, as.factor) %>%
  #dummy_cols(select_columns = c("region", "type", "sex", "evaluation",
   #      "approval", "P4", "P19", "P20", "P21", "P23_1", "P23_2", "P23_3",
    #     "P23_4", "P23_5", "race", "education", "income", "class", "religion"))
```

#DEMOGRAPHICS

Sex

```
data_fake_news_dem %>%
  count(sex) %>%
  mutate(share = n/sum(n))
```

```
## # A tibble: 2 x 3
##   sex       n share
##   <fct> <int> <dbl>
## 1 Men     942 0.471
## 2 Women  1058 0.529
```

Region

```
data_fake_news_dem %>%
  count(region) %>%
  mutate(share = n/sum(n))
```

```
## # A tibble: 5 x 3
##   region          n share
##   <fct>       <int> <dbl>
## 1 North         150 0.075
## 2 Northeast     538 0.269
## 3 Center-West   158 0.079
## 4 Southeast     858 0.429
## 5 South         296 0.148
```

City type

```
data_fake_news_dem %>%
  count(type) %>%
  mutate(share = n/sum(n))
```

```
## # A tibble: 3 x 3
##   type                    n share
##   <fct>               <int> <dbl>
## 1 Capital               538 0.269
## 2 Metropolitan region   365 0.182
## 3 Countryside          1097 0.548
```

Age

```
data_fake_news_dem %>%
   summarise(mean = mean(age_full),
             median = median(age_full),
             sd = sd(age_full))
```

```
## # A tibble: 1 x 3
##    mean median    sd
##   <dbl>  <dbl> <dbl>
## 1  43.1     42  15.5
```

Political Orientation

```
data_fake_news_dem <- rename(data_fake_news_dem, pol_orientation = P4)

data_fake_news_dem %>%
  count(pol_orientation) %>%
  mutate(share = n/sum(n))
```

```
## # A tibble: 6 x 3
##   pol_orientation                                      n  share
##   <fct>                                            <int>  <dbl>
## 1 Right/Center-Right                                 433 0.216
## 2 Center                                             193 0.0965
## 3 Left/Center-Left                                   451 0.226
## 4 I no longer have a defined political orientation   200 0.1
## 5 I never had a political orientation                648 0.324
## 6 Unsure                                              75 0.0375
```

Government Approval Rating

```
data_fake_news_dem %>%
  count(approval) %>%
  mutate(share = n/sum(n))
```

```
## # A tibble: 4 x 3
##   approval                             n share
##   <fct>                            <int> <dbl>
## 1 Approves                           562 0.281
## 2 Neither approves nor disapproves   302 0.151
## 3 Disapproves                       1106 0.553
## 4 Unsure                              30 0.015
```

Race

```
data_fake_news_dem %>%
  count(race) %>%
  mutate(share = n/sum(n))
```

```
## # A tibble: 6 x 3
##   race              n  share
##   <fct>         <int>  <dbl>
## 1 White           857 0.428
## 2 Black           199 0.0995
## 3 Pardo (brown)   910 0.455
## 4 Indigenous        2 0.001
## 5 Yellow           17 0.0085
## 6 Other            15 0.0075
```

Education

```
data_fake_news_dem %>%
  count(education) %>%
  mutate(share = n/sum(n))
```

```
## # A tibble: 4 x 3
##   education              n share
##   <fct>              <int> <dbl>
## 1 No education         211 0.106
## 2 Elementary School    611 0.306
## 3 High School          842 0.421
## 4 Higher Education     336 0.168
```

Class

```
data_fake_news_dem %>%
  count(class) %>%
  mutate(share = n/sum(n))
```

```
## # A tibble: 4 x 3
##   class     n share
##   <fct> <int> <dbl>
## 1 A/B     615 0.308
## 2 C       921 0.460
## 3 D/E     408 0.204
## 4 DN/DA    56 0.028
```

Religion

```
data_fake_news_dem %>%
  count(religion) %>%
  mutate(share = n/sum(n))
```

```
## # A tibble: 4 x 3
##   religion           n share
##   <fct>          <int> <dbl>
## 1 Catholic         996 0.498
## 2 Evangelicals     618 0.309
## 3 Other religion   154 0.077
## 4 No religion      232 0.116
```

DUMMIES

sex_men: 1 Men, 0 Women; region: 5 levels; capital_metrop: 1 Capital and Metropolitan region, 0 Countryside; approvaes_gov: 1 Approves, 0 Neither approves nor disapproves, Disapproves, Unsure; pol_orientation: Right/Center-Right, Center, Left/Center-Left, No orientation (I no longer have a defined political orientation, I never had a political orientation, Unsure); race_is_white: 1 White, 0 Black, Pardo (brown), Indigenous, Yellow, Other; education_high: 1 High School and Higher Education, 0 No education and Elementary School, income_low: 1 Up to 1 MW, 1 to 3 MWs, and Did not answer, 0 3 to 6 MWs and More than 6 MWs; class: 3 levels: A/B, C, D/E and DN/DA; religion: 4 levels: Catholic, Evangelicals, Other religion, No religion

```
data_fake_news_dem <- data_fake_news_dem %>%
  mutate(sex_men = if_else(sex == "Men", 1, 0)) %>%
  dummy_cols(select_columns = c("region")) %>%
  mutate(capital_metrop = if_else(type %in% c("Capital", "Metropolitan region"), 1, 0),
         approves_gov = if_else(approval == "Approves", 1, 0),
         pol_orientation_right = if_else(pol_orientation == "Right/Center-Right", 1, 0),
         pol_orientation_center = if_else(pol_orientation == "Center", 1, 0),
         pol_orientation_left = if_else(pol_orientation == "Left/Center-Left", 1, 0),
         pol_orientation_none = if_else(pol_orientation %in% c(
           "I no longer have a defined   political orientation",
           "I never had a political orientation", "Unsure"), 1, 0),
         race_is_white = if_else(race == "White", 1, 0),
         education_high = if_else(education %in% c("High School", "Higher Education"), 1, 0),
         income_low = if_else(income %in% c("Up to 1 MW", "1 to 3 MWs", "Did not answer"),
                              1, 0),
         class_ab = if_else(class == "A/B", 1, 0),
         class_c = if_else(class == "C", 1, 0),
         class_de = if_else(class %in% c("D/E", "DN/DA"), 1, 0)) %>%
  dummy_cols(select_columns = c("religion"))
```

LIKELIHOOD OF SHARING FAKE NEWS

```
data_fake_news_dem %>%
  count(shared_fake_news) %>%
  mutate(share = n/sum(n))
```

```
## # A tibble: 2 x 3
##   shared_fake_news     n share
##              <dbl> <int> <dbl>
## 1                0  1589 0.794
## 2                1   411 0.206
```

```
sapply(data_fake_news_dem,function(x) sum(is.na(x)))
```

```
##             idInterview                   state                  region
##                       0                       0                       0
##                    type                     sex                age_full
##                       0                       0                       0
##                  age_60              evaluation                approval
##                       0                       0                       0
##         pol_orientation                     P19                     P20
##                       0                       0                       0
##                     P21                   P23_1                   P23_2
##                       0                       0                       0
##                   P23_3                   P23_4                   P23_5
##                       0                       0                       0
##          education_full                    race           religion_full
##                       0                       0                       0
##             income_full              class_full                  age_50
##                       0                       0                       0
##               education                  income                   class
##                       0                       0                       0
##                religion        shared_fake_news                 sex_men
##                       0                       0                       0
##            region_North        region_Northeast      region_Center-West
##                       0                       0                       0
##        region_Southeast            region_South          capital_metrop
##                       0                       0                       0
##            approves_gov   pol_orientation_right  pol_orientation_center
##                       0                       0                       0
##    pol_orientation_left    pol_orientation_none           race_is_white
##                       0                       0                       0
##          education_high              income_low                class_ab
##                       0                       0                       0
##                 class_c                class_de       religion_Catholic
##                       0                       0                       0
##   religion_Evangelicals religion_Other religion    religion_No religion
##                       0                       0                       0
```

```
sapply(data_fake_news_dem, function(x) length(unique(x)))
```

```
##             idInterview                   state                  region
##                    2000                      27                       5
##                    type                     sex                age_full
##                       3                       2                      67
##                  age_60              evaluation                approval
##                       5                       4                       4
##         pol_orientation                     P19                     P20
##                       6                       3                       5
##                     P21                   P23_1                   P23_2
##                       3                       5                       5
##                   P23_3                   P23_4                   P23_5
##                       5                       5                       5
##          education_full                    race           religion_full
##                       5                       6                       9
##             income_full              class_full                  age_50
##                       8                       8                       4
##               education                  income                   class
##                       4                       5                       4
##                religion        shared_fake_news                 sex_men
##                       4                       2                       2
##            region_North        region_Northeast      region_Center-West
##                       2                       2                       2
##        region_Southeast            region_South          capital_metrop
##                       2                       2                       2
##            approves_gov   pol_orientation_right  pol_orientation_center
##                       2                       2                       2
##    pol_orientation_left    pol_orientation_none           race_is_white
##                       2                       2                       2
##          education_high              income_low                class_ab
##                       2                       2                       2
##                 class_c                class_de       religion_Catholic
##                       2                       2                       2
##   religion_Evangelicals religion_Other religion    religion_No religion
##                       2                       2                       2
```

Model 1 - Only demographics

```
model_1 <- glm(shared_fake_news ~ sex_men + age_full + race_is_white +
                education_high + income_low + class_c,
               family = binomial(link = 'logit'),
               data = data_fake_news_dem)

summary(model_1)
```

```
## 
## Call:
## glm(formula = shared_fake_news ~ sex_men + age_full + race_is_white + 
##     education_high + income_low + class_c, family = binomial(link = "logit"), 
##     data = data_fake_news_dem)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.8023  -0.7056  -0.6631  -0.5874   1.9607  
## 
## Coefficients:
##                 Estimate Std. Error z value        Pr(>|z|)    
## (Intercept)    -1.543677   0.236281  -6.533 0.0000000000644 ***
## sex_men        -0.067751   0.111308  -0.609          0.5427    
## age_full        0.004839   0.003641   1.329          0.1838    
## race_is_white  -0.070164   0.115694  -0.606          0.5442    
## education_high  0.027666   0.124701   0.222          0.8244    
## income_low     -0.222803   0.174627  -1.276          0.2020    
## class_c         0.378472   0.153276   2.469          0.0135 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2031.7  on 1999  degrees of freedom
## Residual deviance: 2022.6  on 1993  degrees of freedom
## AIC: 2036.6
## 
## Number of Fisher Scoring iterations: 4
```

Model 2 - Demographics + Political Orientation

```
model_2 <- glm(shared_fake_news ~ sex_men + age_full + race_is_white +
                education_high + income_low + class_c + pol_orientation_right +
                pol_orientation_center + pol_orientation_left,
               family = binomial(link = 'logit'),
               data = data_fake_news_dem)

summary(model_2)
```

```
## 
```

```
## Call:
## glm(formula = shared_fake_news ~ sex_men + age_full + race_is_white +
##     education_high + income_low + class_c + pol_orientation_right +
##     pol_orientation_center + pol_orientation_left, family = binomial(link = "logit"),
##     data = data_fake_news_dem)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8892  -0.7090  -0.6409  -0.5451   2.0688
##
## Coefficients:
##                         Estimate Std. Error z value         Pr(>|z|)
## (Intercept)            -1.760796   0.247873  -7.104 0.00000000000122 ***
## sex_men                -0.151024   0.115330  -1.310          0.19036
## age_full                0.005011   0.003683   1.361          0.17365
## race_is_white          -0.080674   0.116412  -0.693          0.48830
## education_high          0.037590   0.125390   0.300          0.76434
## income_low             -0.198119   0.175347  -1.130          0.25853
## class_c                 0.378931   0.153659   2.466          0.01366 *
## pol_orientation_right   0.400975   0.147863   2.712          0.00669 **
## pol_orientation_center  0.537941   0.190464   2.824          0.00474 **
## pol_orientation_left    0.352302   0.143705   2.452          0.01422 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2031.7  on 1999  degrees of freedom
## Residual deviance: 2009.3  on 1990  degrees of freedom
## AIC: 2029.3
##
## Number of Fisher Scoring iterations: 4
```

Model 3 - Demographics + Political Orientation + City and Region

```r
model_3 <- glm(shared_fake_news ~ sex_men + age_full + race_is_white +
                education_high + income_low + class_c + pol_orientation_right +
                pol_orientation_center + pol_orientation_left + region_North +
                region_Northeast + `region_Center-West` + region_Southeast +
                 capital_metrop,
               family = binomial(link = 'logit'),
               data = data_fake_news_dem)

summary(model_3)
```

```
##
## Call:
## glm(formula = shared_fake_news ~ sex_men + age_full + race_is_white +
##     education_high + income_low + class_c + pol_orientation_right +
##     pol_orientation_center + pol_orientation_left + region_North +
##     region_Northeast + ‘region_Center-West‘ + region_Southeast +
##     capital_metrop, family = binomial(link = "logit"), data = data_fake_news_dem)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9368  -0.7123  -0.6371  -0.5321   2.1093
##
## Coefficients:
##                         Estimate Std. Error z value       Pr(>|z|)
## (Intercept)            -1.821095   0.285328  -6.382 0.000000000174 ***
## sex_men                -0.147244   0.115534  -1.274        0.20250
## age_full                0.005221   0.003700   1.411        0.15823
## race_is_white          -0.081791   0.116571  -0.702        0.48290
## education_high          0.030738   0.125333   0.245        0.80626
## income_low             -0.220713   0.175846  -1.255        0.20943
## class_c                 0.388343   0.153950   2.523        0.01165 *
## pol_orientation_right   0.397592   0.150899   2.635        0.00842 **
## pol_orientation_center  0.505937   0.193436   2.616        0.00891 **
## pol_orientation_left    0.332507   0.146733   2.266        0.02345 *
## region_North            0.348101   0.250151   1.392        0.16405
## region_Northeast        0.072778   0.189261   0.385        0.70058
## ‘region_Center-West‘   -0.007771   0.259593  -0.030        0.97612
## region_Southeast        0.218488   0.173835   1.257        0.20880
## capital_metrop         -0.155170   0.114134  -1.360        0.17398
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2031.7  on 1999  degrees of freedom
## Residual deviance: 2004.1  on 1985  degrees of freedom
## AIC: 2034.1
##
## Number of Fisher Scoring iterations: 4
```

Model 4 - Demographics + Political Orientation + City and Region + Religion

```
model_4 <- glm(shared_fake_news ~ sex_men + age_full + race_is_white +
                education_high + income_low + class_c + pol_orientation_right +
                pol_orientation_center + pol_orientation_left + region_North +
                region_Northeast + `region_Center-West` + region_Southeast +
                 capital_metrop + religion_Catholic + religion_Evangelicals +
                 `religion_Other religion`,
               family = binomial(link = 'logit'),
               data = data_fake_news_dem)

summary(model_4)
```

```
##
## Call:
## glm(formula = shared_fake_news ~ sex_men + age_full + race_is_white +
##     education_high + income_low + class_c + pol_orientation_right +
##     pol_orientation_center + pol_orientation_left + region_North +
##     region_Northeast + ‘region_Center-West‘ + region_Southeast +
##     capital_metrop + religion_Catholic + religion_Evangelicals +
##     ‘religion_Other religion‘, family = binomial(link = "logit"),
##     data = data_fake_news_dem)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0002  -0.7156  -0.6316  -0.5215   2.1694
##
## Coefficients:
##                              Estimate Std. Error z value     Pr(>|z|)
## (Intercept)                 -1.733033   0.320541  -5.407 0.0000000642 ***
## sex_men                     -0.140821   0.115885  -1.215      0.22430
## age_full                     0.005839   0.003724   1.568      0.11690
## race_is_white               -0.072993   0.116716  -0.625      0.53171
## education_high               0.033582   0.126171   0.266      0.79012
## income_low                  -0.285352   0.178417  -1.599      0.10974
## class_c                      0.441575   0.155864   2.833      0.00461 **
## pol_orientation_right        0.380098   0.151287   2.512      0.01199 *
## pol_orientation_center       0.507084   0.193623   2.619      0.00882 **
## pol_orientation_left         0.326286   0.147442   2.213      0.02690 *
## region_North                 0.351211   0.250863   1.400      0.16151
## region_Northeast             0.074401   0.189662   0.392      0.69485
## `region_Center-West`        -0.022432   0.260399  -0.086      0.93135
## region_Southeast             0.223325   0.174301   1.281      0.20010
## capital_metrop              -0.156662   0.114337  -1.370      0.17063
## religion_Catholic           -0.245938   0.182665  -1.346      0.17818
## religion_Evangelicals        0.060075   0.190379   0.316      0.75234
## `religion_Other religion` -0.008482   0.254432  -0.033      0.97341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2031.7  on 1999  degrees of freedom
## Residual deviance: 1997.9  on 1982  degrees of freedom
## AIC: 2033.9
##
## Number of Fisher Scoring iterations: 4
```

```
stargazer(model_1, model_2, model_3, model_4,
          title = "Logit Models Comparison",
          type = "latex",
          digits = 3,
          no.space = TRUE,
          model.numbers = FALSE,
          header = FALSE,
          column.sep.width = "-15pt")
```

balance_table <- data_fake_news_dem %>% select(29:113) %>% lapply(., function(i) tidy(t.test(i ~ data_fake_news_dem$shared_fake_news))) %>% do.call(rbind, .) %>% rownames_to_column(“variable”) %>% rename(mean_diff = estimate, mean_control = estimate1, mean_treatment = estimate2) %>% select(variable, mean_diff, mean_control, mean_treatment, statistic, p.value)

kable(balance_table, caption = “Balance Table - Observable Characteristics”, digits = 3, align = “c”)

colnames(data_fake_news_dem)

Table 1: Logit Models Comparison

	Dependent variable:			
	shared_fake_news			
sex_men	−0.068	−0.151	−0.147	−0.141
	(0.111)	(0.115)	(0.116)	(0.116)
age_full	0.005	0.005	0.005	0.006
	(0.004)	(0.004)	(0.004)	(0.004)
race_is_white	−0.070	−0.081	−0.082	−0.073
	(0.116)	(0.116)	(0.117)	(0.117)
education_high	0.028	0.038	0.031	0.034
	(0.125)	(0.125)	(0.125)	(0.126)
income_low	−0.223	−0.198	−0.221	−0.285
	(0.175)	(0.175)	(0.176)	(0.178)
class_c	0.378**	0.379**	0.388**	0.442***
	(0.153)	(0.154)	(0.154)	(0.156)
pol_orientation_right		0.401***	0.398***	0.380**
		(0.148)	(0.151)	(0.151)
pol_orientation_center		0.538***	0.506***	0.507***
		(0.190)	(0.193)	(0.194)
pol_orientation_left		0.352**	0.333**	0.326**
		(0.144)	(0.147)	(0.147)
region_North			0.348	0.351
			(0.250)	(0.251)
region_Northeast			0.073	0.074
			(0.189)	(0.190)
‘region_Center-West‘			−0.008	−0.022
			(0.260)	(0.260)
region_Southeast			0.218	0.223
			(0.174)	(0.174)
capital_metrop			−0.155	−0.157
			(0.114)	(0.114)
religion_Catholic				−0.246
				(0.183)
religion_Evangelicals				0.060
				(0.190)
‘religion_Other religion‘				−0.008
				(0.254)
Constant	−1.544***	−1.761***	−1.821***	−1.733***
	(0.236)	(0.248)	(0.285)	(0.321)
Observations	2,000	2,000	2,000	2,000
Log Likelihood	−1,011.298	−1,004.658	−1,002.067	−998.930
Akaike Inf. Crit.	2,036.596	2,029.316	2,034.133	2,033.859
Note:	*p<0.1; **p<0.05; ***p<0.01			