Fake_News_Data_Cleaning

Natália Tosi

8/3/2021

Importing Data and Labels

```
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

```
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

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

Summary Statistics

```
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, P22, 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_19 = if_else(P19 == "Yes", 1, 0),
         unnoticed_fake_news = if_else(P21 == "Yes", 1, 0)) %>%
  rename(c(pol_orientation = P4,
           frequency_fake_news = P20,
           reaction_fake_news = P22,
           resp_population = P23_1,
           resp_gov = P23_2,
           resp_politicians = P23_3,
           resp_press = P23_4,
           resp_social_media = P23_5))


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")),
         pol_orientation = factor(pol_orientation, 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(race_adj = fct_collapse(race,
                                 White = c("White"),
                                 Black = c("Black", "Pardo (brown)"),
                                 Other = c("Indigenous", "Yellow", "Other")))
```

#DEMOGRAPHICS

Sex

```
data_fake_news_dem %>%
  count(sex) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

sex	n	share
Men	942	47.1
Women	1058	52.9

Region

```
data_fake_news_dem %>%
  count(region) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

region	n	share
North	150	7.5
Northeast	538	26.9
Center-West	158	7.9
Southeast	858	42.9
South	296	14.8

City type

```
data_fake_news_dem %>%
  count(type) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

type	n	share
Capital	538	26.90
Metropolitan region	365	18.25
Countryside	1097	54.85

Age

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

mean	median	sd
43.104	42	15.52091

Political Orientation

```
data_fake_news_dem %>%
  count(pol_orientation) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

pol_orientation	n	share
Right/Center-Right	433	21.65
Center	193	9.65
Left/Center-Left	451	22.55
I no longer have a defined political orientation	200	10.00
I never had a political orientation	648	32.40
Unsure	75	3.75

Government Approval Rating

```
data_fake_news_dem %>%
  count(approval) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

approval	n	share
Approves	562	28.1

approval	n	share
Neither approves nor disapproves	302	15.1
Disapproves	1106	55.3
Unsure	30	1.5

Race

```
data_fake_news_dem %>%
  count(race_adj) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

race_adj	n	share
White	857	42.85
Black	1109	55.45
Other	34	1.70

Education

```
data_fake_news_dem %>%
  count(education) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

education	n	share
No education	211	10.55
Elementary School	611	30.55
High School	842	42.10
Higher Education	336	16.80

Class

```
data_fake_news_dem %>%
  count(class) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

class	n	share
A/B	615	30.75
C	921	46.05

class	n	share
D/E	408	20.40
DN/DA	56	2.80

Religion

```
data_fake_news_dem %>%
  count(religion) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

religion	n	share
Catholic	996	49.8
Evangelicals	618	30.9
Other religion	154	7.7
No religion	232	11.6

RESULTS

P19 - Have you ever shared a political news story online that you thought at the time was made up? (Single Answer)

```
data_fake_news_dem %>%
  count(shared_fake_news_19) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

shared_fake_news_19	n	share
0	1589	79.45
1	411	20.55

P20 - How often do you come across news stories online that you think are almost completely made up? (Single Answer)

```
data_fake_news_dem %>%
  count(frequency_fake_news) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

frequency_fake_news	n	share
Hardly ever	325	16.25
Never	111	5.55
Often	1022	51.10
Sometimes	458	22.90
Unsure	84	4.20

P21 - Have you ever shared a political news story online that you later found out was made up? (Single Answer)

```
data_fake_news_dem %>%
  count(unnoticed_fake_news) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

unnoticed_fake_news	n	share
0	1495	74.75
1	505	25.25

P22 - [IS ANSWERED YES] What was your reaction when you found out that the information shared was not true? If it happened more than once, select the option that was most frequent (Single Answer)

```
data_fake_news_dem %>%
  count(reaction_fake_news) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

reaction_fake_news	n	share
I didn't send a warning, but I also didn't share the same information anymore	196	9.80
I just sent a message warning that the information was not true	148	7.40
I kept sharing the information	26	1.30
I sent a message warning that the information was not true along with the correct information	121	6.05
Unsure	14	0.70
NA	1495	74.75

P23 - How much responsibility does each of the following have in trying to prevent made up (fake news) stories from gaining attention". (Single Answer per category)

```
resp_population <- data_fake_news_dem %>%
  count(resp_population) %>%
  mutate(population = (n/sum(n))*100) %>%
  select(-n)

resp_gov <- data_fake_news_dem %>%
  count(resp_gov) %>%
  mutate(governments = (n/sum(n))*100) %>%
  select(-n)

resp_politicians <- data_fake_news_dem %>%
  count(resp_politicians) %>%
  mutate(politicians = (n/sum(n))*100) %>%
  select(-n)

resp_press <- data_fake_news_dem %>%
  count(resp_press) %>%
  mutate(press = (n/sum(n))*100) %>%
  select(-n)

resp_social_media <- data_fake_news_dem %>%
  count(resp_social_media) %>%
  mutate(social_media = (n/sum(n))*100) %>%
  select(-n)

responsibility <- resp_population %>%
  left_join(resp_gov, by = c("resp_population" = "resp_gov")) %>%
  left_join(resp_politicians, by = c("resp_population" = "resp_politicians")) %>%
  left_join(resp_press, by = c("resp_population" = "resp_press")) %>%
  left_join(resp_social_media, by = c("resp_population" = "resp_social_media")) %>%
  mutate(resp_population = factor(resp_population,
  levels = c("A great deal of responsibility", "A fair amount of responsibility",
             "Not much responsibility", "No responsibility at all", "Unsure"))) %>%
  arrange(resp_population) %>%
  rename(c(answer = resp_population))

kable(responsibility)
```

answer	population	governments	politicians	press	social_media
A great deal of responsibility	54.30	62.10	62.15	63.85	56.00
A fair amount of responsibility	21.10	14.15	12.00	14.25	18.05
Not much responsibility	13.40	12.15	13.05	13.20	14.45
No responsibility at all	9.25	9.10	10.35	6.70	7.95
Unsure	1.95	2.50	2.45	2.00	3.55

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"))
```

SHARED FAKE NEWS - SUSPECTING

```
data_fake_news_dem %>%
  count(shared_fake_news_19) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

shared_fake_news_19	n	share
0	1589	79.45
1	411	20.55

Model 1 - Only demographics

```
model_19_1 <- glm(shared_fake_news_19 ~ sex_men + age_full + race_is_white +
                  education_high + income_low + class_c,
                 family = binomial(link = 'logit'),
                 data = data_fake_news_dem)
```

Model 2 - Demographics + Political Orientation

```
model_19_2 <- glm(shared_fake_news_19 ~ 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)
```

Model 3 - Demographics + Political Orientation + City and Region

```
model_19_3 <- glm(shared_fake_news_19 ~ 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)
```

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

```
model_19_4 <- glm(shared_fake_news_19 ~ 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)
```

```
stargazer(model_19_1, model_19_2, model_19_3, model_19_4,
          title = "Logit Models Comparison - Suspected",
          type = "latex",
          digits = 3,
          no.space = TRUE,
          model.numbers = FALSE,
          header = FALSE,
          column.sep.width = "-15pt")
```

Table 17: Logit Models Comparison - Suspected

	Dependent variable:			
	shared_fake_news_19			
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

SHARED FAKE NEWS - UNNOTICED

```
data_fake_news_dem %>%
  count(unnoticed_fake_news) %>%
  mutate(share = (n/sum(n))*100) %>%
  kable()
```

unnoticed_fake_news	n	share
0	1495	74.75
1	505	25.25

Model 1 - Only demographics

```
model_21_1 <- glm(unnoticed_fake_news ~ sex_men + age_full + race_is_white +
                    education_high + income_low + class_c,
                  family = binomial(link = 'logit'),
                  data = data_fake_news_dem)
```

Model 2 - Demographics + Political Orientation

```
model_21_2 <- glm(unnoticed_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)
```

Model 3 - Demographics + Political Orientation + City and Region

```
model_21_3 <- glm(unnoticed_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)
```

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

```r
model_21_4 <- glm(unnoticed_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)
```

```r
stargazer(model_21_1, model_21_2, model_21_3, model_21_4,
          title = "Logit Models Comparison - Unnoticed",
          type = "latex",
          digits = 3,
          no.space = TRUE,
          model.numbers = FALSE,
          header = FALSE,
          column.sep.width = "-15pt")
```

Table 19: Logit Models Comparison - Unnoticed

	Dependent variable:			
	unnoticed_fake_news			
sex_men	−0.0003	−0.068	−0.062	−0.058
	(0.103)	(0.107)	(0.108)	(0.108)
age_full	0.003	0.003	0.003	0.003
	(0.003)	(0.003)	(0.003)	(0.003)
race_is_white	−0.026	−0.035	−0.038	−0.035
	(0.107)	(0.108)	(0.108)	(0.109)
education_high	−0.043	−0.034	−0.043	−0.054
	(0.116)	(0.117)	(0.117)	(0.117)
income_low	−0.412**	−0.391**	−0.419**	−0.429***
	(0.162)	(0.162)	(0.164)	(0.166)
class_c	0.425***	0.425***	0.435***	0.449***
	(0.144)	(0.144)	(0.145)	(0.147)
pol_orientation_right		0.298**	0.295**	0.299**
		(0.138)	(0.142)	(0.142)
pol_orientation_center		0.509***	0.418**	0.418**
		(0.177)	(0.181)	(0.181)
pol_orientation_left		0.291**	0.248*	0.236*
		(0.134)	(0.137)	(0.138)
region_North			0.752***	0.756***
			(0.248)	(0.248)
region_Northeast			0.531***	0.532***
			(0.192)	(0.193)
‘region_Center-West‘			0.699***	0.682***
			(0.243)	(0.244)
region_Southeast			0.783***	0.789***
			(0.179)	(0.179)
capital_metrop			−0.154	−0.153
			(0.106)	(0.107)
religion_Catholic				−0.106
				(0.172)
religion_Evangelicals				−0.059
				(0.182)
‘religion_Other religion‘				0.207
				(0.235)
Constant	−1.077***	−1.256***	−1.756***	−1.707***
	(0.218)	(0.228)	(0.276)	(0.309)
Observations	2,000	2,000	2,000	2,000
Log Likelihood	−1,124.790	−1,119.125	−1,107.227	−1,105.881
Akaike Inf. Crit.	2,263.580	2,258.251	2,244.453	2,247.762
Note:				*p<0.1; **p<0.05; ***p<0.01