W241 Final Project - Tweets are the new folktales

Arriaga, Hue, Khural, Moran

August 3, 2021

0.1 Libraries

```
library(data.table)
library(foreign)
library(tidyverse)
library(dplyr)
library(readr)

# Calculate power
library(effsize)
library(pwr)

# Correlation checks
library(ggpubr)
library(corrplot)
library(PerformanceAnalytics)

# Variance
library(sandwich)

# Reports
library(stargazer)
```

Let's load the data.

```
df<-data.table::fread("../data/processed/tweets_data.csv")
df$truth_f <- as.factor(df$truth)
df$sentiment_f <- as.factor(df$sentiment)

# Robust standard error
rse <- function(model) {
  sqrt(diag(vcovHC(model)))
  }
```

0.2 Experiment Design

Our main goal is to evaluate if truthfulness is a variable that has a short term impact in memory. We can perform an hypotesis testing comparing all positive tweets to all negative tweers.

Based on the literature, we expect that tweets with fake information are less likely to be remembered. In regards to sentiment, Xie and Zhang (2017) found that negative emotions enchance the visual long-term memory. For the power analysis estimation, we will consider the base group to be the treatment with true and negative emotion content. This will be compared with the treatment with the lowest sample size.

0.3 Power analysis

```
# Base group: TN

table1<- table(df$truth_f,df$sentiment_f)
table1
```

```
##
##        negative positive
##   fact       29       26
##   fake       26       26
```

```
# Base group n=29, the rest all have the same size
# Let's see the least optimistic one: smaller effect

# This line outputs "fact" "fake", meaning fact=0, fake=1
levels(df$truth_f)
```

```
## [1] "fact" "fake"
```

```
# Now for sentiment: negative=0, positive=1
levels(df$sentiment_f)
```

```
## [1] "negative" "positive"
```

```
# The level values are consistent with our literatue (base=TN)
tp<- subset(df, truth_f=="fact" & sentiment_f=="positive")
tn<- subset(df, truth_f=="fact" & sentiment_f=="negative")
fp<- subset(df, truth_f=="fake" & sentiment_f=="positive")
fn<- subset(df, truth_f=="fake" & sentiment_f=="negative")

# Let's compare the coehn's do to see which one has the smaller effect size
coehnd_tn_tp <- cohen.d(tn$total_correct,tp$total_correct)
coehnd_tn_tp
```

```
##
## Cohen's d
##
## d estimate: -0.5153612 (medium)
## 95 percent confidence interval:
##       lower       upper
## -1.06596976  0.03524739
```

```
coehnd_tn_fp <- cohen.d(tn$total_correct,fp$total_correct)
coehnd_tn_fp
```

```
##
## Cohen's d
##
## d estimate: -0.7390862 (medium)
## 95 percent confidence interval:
##      lower      upper
## -1.2989380 -0.1792344
```

```
coehnd_tn_fn <- cohen.d(tn$total_correct,fn$total_correct)
coehnd_tn_fn
```

```
##
## Cohen's d
##
## d estimate: -0.2580201 (small)
## 95 percent confidence interval:
##      lower      upper
## -0.8019787  0.2859385
```

We selected the comparison between treatment TN and FN because they show a small effect size (d=-0.26).

```
# We can perform a power test given the effect sizes, samples
pwr.t2n.test(n1 = length(tn$truth) , n2= length(fn$truth), d = coehnd_tn_fn$estimate, sig.level
 = .95)
```

```
##
##      t test power calculation
##
##              n1 = 29
##              n2 = 26
##               d = 0.2580201
##       sig.level = 0.95
##           power = 0.9683005
##     alternative = two.sided
```

After performing a power test using the pwr package developed by Champely following the calculation as outlined by Cohen, we found a 96% power (n1=29, n2=26, d=0.25, two sided test). This means that the rest of the treatments, who had a larger effect and same sample size will show a larger power.

1 Regression analysis

We performed a regression analysis to evaluate the short-memory retention based on the four treatments. Three models were proposed to analyze the data: reduced model which compared only the potential outcome (total correct responses) to the covariate of interest tweet truthfulness. A second model, includes the sentiment as a second independent variable and finaly, a third model adds the interaction between both covariates.

y: Total correct responses F: Content was fake P: Sentiment was positive

1.1 Model 1: Reduced model

$$y = \beta_0 + \beta_1 F + \epsilon_1 \qquad (1)$$

1.2 Model 2: Extended model

$$y = \beta_0 + \beta_1 F + \beta_2 P + \epsilon_2 \qquad (2)$$

1.3 Model 3: Full model

$$y = \beta_0 + \beta_1 F + \beta_2 P + \beta_3 F \cdot P + \epsilon_3 \qquad (3)$$

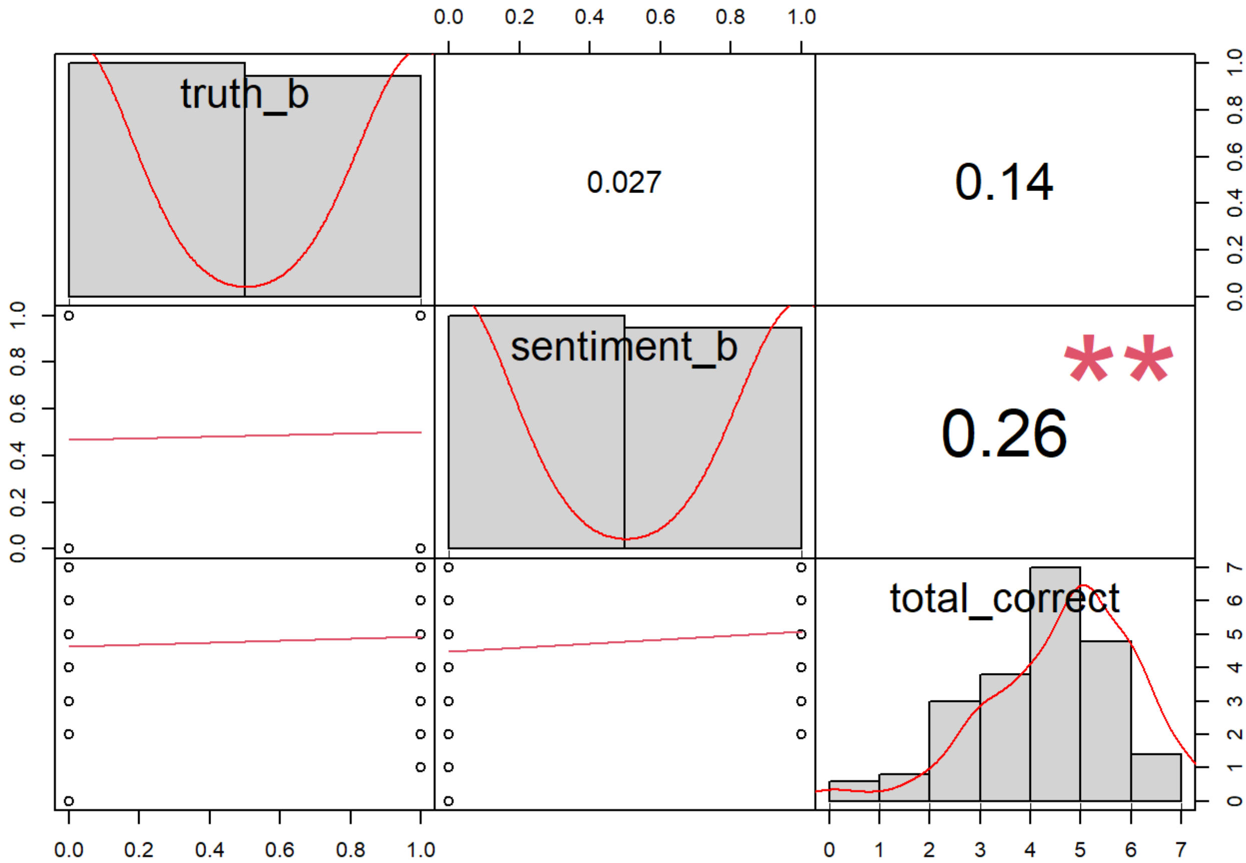
1.4 Correlation check

Before running the analysis we confirmed that our variables are not correlated. We performed a pearson’s correlation to evaluate if the covariates Fake and Positive were indepedent.

```
# Let's check if our correlations are linear

cor(as.integer(df$truth_f),as.integer(df$sentiment_f), method = 'pearson')
```

```
## [1] 0.02727273
```

```
# There seems to be almost no correlation between the variables.

# Let's plot a correlation matrix
df[, truth_b:= ifelse(truth_f=='fake', yes=1,no=0)]
df[, sentiment_b:= ifelse(sentiment_f=='positive', yes=1,no=0)]

df %>% select(c(truth_b,sentiment_b,total_correct)) %>% chart.Correlation()
```



Now we compute the three proposed models.

```
model_1 <- df[, lm(total_correct~truth_f)]
summary(model_1, vcov=vcovHC(model_1))
```

```
##
## Call:
## lm(formula = total_correct ~ truth_f)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4545 -0.8654  0.1346  1.1346  2.5455
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4545     0.1933  23.049   <2e-16 ***
## truth_ffake   0.4108     0.2772   1.482    0.141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.433 on 105 degrees of freedom
## Multiple R-squared:  0.02049,    Adjusted R-squared:  0.01116
## F-statistic: 2.196 on 1 and 105 DF,  p-value: 0.1413
```

Looking at the linear regression, we didn't find a correlation between fake tweets and memory retention. We fail to reject the null hypothesis in our reduced model.

```
model_2 <- df[, lm(total_correct~truth_f+sentiment_f)]
summary(model_2, vcov=vcovHC(model_2))
```

```
##
## Call:
## lm(formula = total_correct ~ truth_f + sentiment_f)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1062 -0.8431  0.1569  0.8938  2.8938
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.1062     0.2266  18.119  < 2e-16 ***
## truth_ffake           0.3907     0.2691   1.452  0.14954
## sentiment_fpositive   0.7369     0.2691   2.738  0.00727 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.391 on 104 degrees of freedom
## Multiple R-squared:  0.08635,    Adjusted R-squared:  0.06878
## F-statistic: 4.915 on 2 and 104 DF,  p-value: 0.009131
```

When running the extended model, we found that the fake covariate coefficient is reduced to 0.39 ($\delta\beta_1 = -0.03$) and the standard error was slightly reduced ($\delta\sigma_f = 0.01$). We found that sentiment had a significant positive impact in short-term memory retention by 0.73 (p-val=0.007, CI=95%), which is closer to remembering the content of one more tweet.

```
model_3 <- df[, lm(total_correct~truth_f+sentiment_f + truth_f:sentiment_f)]
summary(model_3, vcov=vcovHC(model_3))
```

```
##
## Call:
## lm(formula = total_correct ~ truth_f + sentiment_f + truth_f:sentiment_f)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1034 -0.8462  0.1538  0.8966  2.8966
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.10345    0.25953  15.811   <2e-16 ***
## truth_ffake                     0.39655    0.37747   1.051   0.2959
## sentiment_fpositive             0.74271    0.37747   1.968   0.0518 .
## truth_ffake:sentiment_fpositive -0.01194    0.54105  -0.022   0.9824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.398 on 103 degrees of freedom
## Multiple R-squared:  0.08636,    Adjusted R-squared:  0.05974
## F-statistic: 3.245 on 3 and 103 DF,  p-value: 0.02502
```

After running the full model, we found none of the covariates had a significant result.

Let’s organize the results in an organized comparison.

```
# stargazer of all models
stargazer(model_1, model_2, model_3,
  # type="text",
  se = list(rse(model_1),rse(model_2), rse(model_3)),
column.labels = c("F","F+P","F+P+F*P"),
dep.var.labels   = "Correct responses",
covariate.labels = c("Fake information", "Positive sentiment", "Fake:Positive"),
# add.lines = list(c("optimized model", "Yes", "Yes","Yes")),
header=FALSE, type='latex',
title = "Regression results"
)
```

We found that while fake information doesn’t seem to correlate with short-term memory retention, the positive sentiment of tweets does have a significant positive impact in the short-memory retention when explosed to tweets. This is an interesting finding because it’s opposite to what the literature suggested.

1.5 Individual tweet analysis

Will this results be consistent with individual tweets? We ran a linear regression and clustered it by each tweet to see the error depended on specific tweets.

```
model_q_one <- df[ ,lm(bin_georgians~ truth_f+sentiment_f+truth_f:sentiment_f)]
model_q_two <- df[ ,lm(bin_energy~ truth_f+sentiment_f+truth_f:sentiment_f)]
model_q_three <- df[ ,lm(bin_soccer~ truth_f+sentiment_f+truth_f:sentiment_f)]
model_q_four <- df[ ,lm(bin_pollution~ truth_f+sentiment_f+truth_f:sentiment_f)]
model_q_five <- df[ ,lm(bin_fauci~ truth_f+sentiment_f+truth_f:sentiment_f)]
model_q_six <- df[ ,lm(bin_election~ truth_f+sentiment_f+truth_f:sentiment_f)]

# stargazer by question
stargazer(model_q_one, model_q_two, model_q_three,
          model_q_four,model_q_five,model_q_six,
  # type="text",
  se = list(rse(model_q_one),rse(model_q_two), rse(model_q_three),
            rse(model_q_four),rse(model_q_five),rse(model_q_six)),
column.labels = c("Georgians","Energy","Soccer","Pollution","Fauci","Election"),
dep.var.labels =  c("","","","","",""),
covariate.labels = c("Fake information", "Positive sentiment", "Fake:Positive"),
# add.lines = list(c("optimized model", "Yes", "Yes","Yes")),
header=FALSE, type='text',
title = "Regression results"
)
```

```
##
## Regression results
## =====================================================================================
##                                               Dependent variable:
##                                  ----------------------------------------------------
##
##                                  Georgians  Energy   Soccer  Pollution   Fauci   Election
##                                     (1)      (2)      (3)       (4)       (5)      (6)
## -------------------------------------------------------------------------------------
## Fake information                 0.401***   -0.036  0.436***   0.133   -0.605***  -0.028
##                                   (0.121)  (0.132)  (0.116)   (0.138)   (0.110)  (0.122)
##
## Positive sentiment                0.247**   0.080    0.167     0.210    -0.182    0.049
##                                   (0.119)  (0.124)  (0.138)   (0.135)   (0.130)  (0.115)
##
## Fake:Positive                     -0.247    -0.118   -0.206    -0.056   0.682***   0.028
##                                   (0.187)  (0.186)  (0.169)   (0.188)   (0.178)  (0.167)
##
## Constant                          0.138**  0.690*** 0.448*** 0.483***  0.759***  0.759***
##                                   (0.066)  (0.089)  (0.096)   (0.096)   (0.082)  (0.082)
##
## -------------------------------------------------------------------------------------
## Observations                        107      107      107       107       107      107
## R2                                 0.117    0.014    0.152     0.050     0.212    0.006
## Adjusted R2                        0.091    -0.014   0.127     0.022     0.189    -0.023
## Residual Std. Error (df = 103)     0.468    0.471    0.434     0.478     0.451    0.424
## F Statistic (df = 3; 103)        4.556***   0.504  6.130***   1.802   9.246***   0.212
## =====================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

After analyzing the responses by question type in the full model, we found that there is a short-memory retention effect in the principal covariate truthfulness, some cases the sentiment has an effect and also the interaction between fake and positive. This suggests that there could be a relationship between the topic (ommited variable bias) and the short-memory retention.

Entertainment, politics and health topics seem to have a stronger effect than environmental related ones such as pollution or energy. Only the question about statements about georgians during the elections seemed to be impacted positively by the sentiment. Also, we found a strong interaction effect between the Fauci, which mentioned the Covi-19 pandemic, which is a very controversial topic. The fauci tweet had the largest F-statistic from all the cases (F=9.2), it showed a more dramatic response with less variance ($R_{adj} = 0.19$).

1.6 Descriptive statistics

Missing

1.7 Randomization check

Missing

References

Xie, Weizhen, and Weiwei Zhang. 2017. "Negative Emotion Enhances Mnemonic Precision and Subjective Feelings of Remembering in Visual Long-Term Memory." *Cognition* 166: 73–83. https://doi.org/https://doi.org/10.1016/j.cognition.2017.05.025 (https://doi.org/10.1016/j.cognition.2017.05.025).