Twitter Covid19 Tweet Analysis

Minh Nguyen

3/12/2020

library(tidyverse)  
library(keyring)  
library(tibble)  
library(purrr)  
library(devtools)  
library(usethis)  
library(twitteR)  
library(stringr)  
library(lubridate)  
library(tree)

##Read tidy covid19 tweet data to R

# John Hopkins COVID19 DATA:  
# https://github.com/CSSEGISandData/COVID-19  
  
#Read cleaned covid19 tweet data into R (setwd first)  
covid19tidy <- read.csv("C:/Users/mnguyen/OneDrive/Documents/Personal/School/3 Spring 2020/Predictive Analytics (ITEC 621-002)/Project/twittercovid19/WIP/twittercovid19/Data/covid19mar110.csv")

## Word Mining

covid19tidy %>%  
 mutate(ishealth = str\_count(text, "health"),  
 ispandemic = str\_count(text, "pandemic"),  
 isvirus = str\_count(text, "virus"),  
 isemergency = str\_count(text, "emergency"),  
 isdeaths = str\_count(str\_to\_sentence(text), c("dead","death")),  
 iswho = str\_count(str\_to\_sentence(text), c("who", "wolrd health organization")),  
 iscdc = str\_count(str\_to\_sentence(text), c("cdc", "centers for disease control")),  
 isnih = str\_count(str\_to\_sentence(text), c("nih", "national institutes of health")),  
 isdisease = str\_count(str\_to\_sentence(text), "disease"),   
 isquarantine = str\_count(str\_to\_sentence(text), "quarantine"),   
 isrecover = str\_count(str\_to\_sentence(text), "recover"),  
 isban = str\_count(str\_to\_sentence(text), "ban"),   
 iscoronavirus = str\_count(str\_to\_sentence(text), "coronavirus"),  
 iscovid19 = str\_count(str\_to\_sentence(text), "covid19"),   
 iswash = str\_count(str\_to\_sentence(text), "wash"),   
 isracist = str\_count(str\_to\_sentence(text), c("racist","racism")),   
 isasian = str\_count(str\_to\_sentence(text), "asian"),   
 ischinese = str\_count(str\_to\_sentence(text), c("chinese", "china")),   
 isinfectious = str\_count(str\_to\_sentence(text), c("infectious", "infections"))) -> covid19tidy

##OLS, GLM, Regression Trees, and Classification Trees

#Ordinary Least Squares (OLS) - retweets and favorites total  
retweets.ols <- lm(retweets ~ length + favorites + ishealth + ispandemic +   
 isvirus + isemergency + isdeaths + iswho + iscdc + isnih +   
 isdisease + isquarantine + isrecover + isban + iscoronavirus +   
 iscovid19 + iswash + isracist + isasian + ischinese + isinfectious,   
 data = covid19tidy)  
summary(retweets.ols)

##   
## Call:  
## lm(formula = retweets ~ length + favorites + ishealth + ispandemic +   
## isvirus + isemergency + isdeaths + iswho + iscdc + isnih +   
## isdisease + isquarantine + isrecover + isban + iscoronavirus +   
## iscovid19 + iswash + isracist + isasian + ischinese + isinfectious,   
## data = covid19tidy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7769 -3067 -2318 -813 42703   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -973.194 385.741 -2.523 0.011644 \*   
## length 30.699 2.775 11.064 < 2e-16 \*\*\*  
## favorites -5.343 1.690 -3.161 0.001572 \*\*   
## ishealth 1730.359 195.683 8.843 < 2e-16 \*\*\*  
## ispandemic -2086.144 476.170 -4.381 1.18e-05 \*\*\*  
## isvirus 1237.355 143.103 8.647 < 2e-16 \*\*\*  
## isemergency -2882.574 587.781 -4.904 9.43e-07 \*\*\*  
## isdeaths -1680.407 319.951 -5.252 1.51e-07 \*\*\*  
## iswho -1108.226 214.629 -5.163 2.44e-07 \*\*\*  
## iscdc -1915.588 375.958 -5.095 3.50e-07 \*\*\*  
## isnih -1377.321 908.460 -1.516 0.129503   
## isdisease -1737.685 444.095 -3.913 9.14e-05 \*\*\*  
## isquarantine 1417.425 312.499 4.536 5.76e-06 \*\*\*  
## isrecover -1380.636 674.941 -2.046 0.040808 \*   
## isban -983.339 319.945 -3.073 0.002118 \*\*   
## iscoronavirus -1985.846 161.336 -12.309 < 2e-16 \*\*\*  
## iscovid19 -252.214 84.915 -2.970 0.002978 \*\*   
## iswash -2258.696 346.332 -6.522 7.06e-11 \*\*\*  
## isracist -1620.868 1741.104 -0.931 0.351891   
## isasian -3035.902 1120.743 -2.709 0.006756 \*\*   
## ischinese 952.299 276.389 3.446 0.000571 \*\*\*  
## isinfectious -624.482 852.016 -0.733 0.463597   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7427 on 29978 degrees of freedom  
## Multiple R-squared: 0.0207, Adjusted R-squared: 0.02002   
## F-statistic: 30.18 on 21 and 29978 DF, p-value: < 2.2e-16

favorites.ols <- lm(favorites ~ length + retweets + ishealth + ispandemic +   
 isvirus + isemergency + isdeaths + iswho + iscdc + isnih +   
 isdisease + isquarantine + isrecover + isban + iscoronavirus +   
 iscovid19 + iswash + isracist + isasian + ischinese + isinfectious,   
 data = covid19tidy)  
summary(favorites.ols)

##   
## Call:  
## lm(formula = favorites ~ length + retweets + ishealth + ispandemic +   
## isvirus + isemergency + isdeaths + iswho + iscdc + isnih +   
## isdisease + isquarantine + isrecover + isban + iscoronavirus +   
## iscovid19 + iswash + isracist + isasian + ischinese + isinfectious,   
## data = covid19tidy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.84 -1.68 -1.42 -1.04 1932.49   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.806e+00 1.318e+00 2.129 0.03324 \*   
## length -7.980e-03 9.499e-03 -0.840 0.40086   
## retweets -6.237e-05 1.973e-05 -3.161 0.00157 \*\*  
## ishealth 3.425e-01 6.695e-01 0.512 0.60889   
## ispandemic 4.027e-01 1.627e+00 0.247 0.80455   
## isvirus -4.540e-01 4.895e-01 -0.927 0.35372   
## isemergency -8.042e-01 2.009e+00 -0.400 0.68895   
## isdeaths 3.085e-01 1.094e+00 0.282 0.77789   
## iswho 1.570e+00 7.336e-01 2.141 0.03231 \*   
## iscdc -1.085e+00 1.285e+00 -0.844 0.39865   
## isnih -8.054e-01 3.104e+00 -0.259 0.79527   
## isdisease -1.103e+00 1.518e+00 -0.727 0.46721   
## isquarantine -7.804e-01 1.068e+00 -0.731 0.46498   
## isrecover -7.748e-01 2.306e+00 -0.336 0.73691   
## isban -1.081e+00 1.093e+00 -0.989 0.32283   
## iscoronavirus 3.394e-01 5.526e-01 0.614 0.53914   
## iscovid19 -2.654e-01 2.902e-01 -0.915 0.36032   
## iswash 9.172e-01 1.184e+00 0.775 0.43861   
## isracist -8.789e-01 5.949e+00 -0.148 0.88255   
## isasian 1.183e+00 3.830e+00 0.309 0.75742   
## ischinese -9.072e-01 9.445e-01 -0.960 0.33681   
## isinfectious -8.152e-01 2.911e+00 -0.280 0.77946   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 25.38 on 29978 degrees of freedom  
## Multiple R-squared: 0.0007517, Adjusted R-squared: 5.172e-05   
## F-statistic: 1.074 on 21 and 29978 DF, p-value: 0.3684

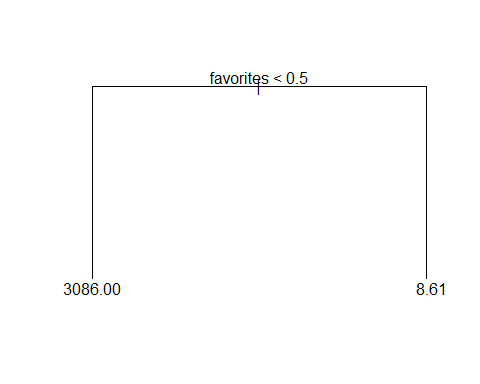
#Generalized Linear Models (GLM) - is it a retweet?  
isretweeted.glm <- glm(isretweeted ~ length + favorites + retweets + ishealth + ispandemic +   
 isvirus + isemergency + isdeaths + iswho + iscdc + isnih +   
 isdisease + isquarantine + isrecover + isban + iscoronavirus +   
 iscovid19 + iswash + isracist + isasian + ischinese + isinfectious,   
 data = covid19tidy)  
summary(isretweeted.glm)

##   
## Call:  
## glm(formula = isretweeted ~ length + favorites + retweets + ishealth +   
## ispandemic + isvirus + isemergency + isdeaths + iswho + iscdc +   
## isnih + isdisease + isquarantine + isrecover + isban + iscoronavirus +   
## iscovid19 + iswash + isracist + isasian + ischinese + isinfectious,   
## data = covid19tidy)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.98504 0.06864 0.11158 0.15052 2.54008   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.922e-02 1.731e-02 -3.998 6.41e-05 \*\*\*  
## length 6.558e-03 1.248e-04 52.554 < 2e-16 \*\*\*  
## favorites -1.754e-03 7.587e-05 -23.120 < 2e-16 \*\*\*  
## retweets 6.502e-06 2.592e-07 25.085 < 2e-16 \*\*\*  
## ishealth 2.256e-02 8.794e-03 2.565 0.01032 \*   
## ispandemic -2.664e-02 2.138e-02 -1.246 0.21265   
## isvirus -4.624e-02 6.430e-03 -7.191 6.59e-13 \*\*\*  
## isemergency -2.901e-02 2.639e-02 -1.099 0.27166   
## isdeaths -1.268e-02 1.437e-02 -0.883 0.37745   
## iswho 2.224e-02 9.637e-03 2.308 0.02102 \*   
## iscdc 3.294e-02 1.688e-02 1.951 0.05105 .   
## isnih 3.202e-02 4.077e-02 0.785 0.43230   
## isdisease 2.847e-02 1.994e-02 1.428 0.15323   
## isquarantine 3.729e-02 1.403e-02 2.658 0.00786 \*\*   
## isrecover -7.404e-02 3.029e-02 -2.444 0.01453 \*   
## isban -2.252e-02 1.436e-02 -1.568 0.11679   
## iscoronavirus 4.315e-02 7.259e-03 5.945 2.80e-09 \*\*\*  
## iscovid19 3.955e-02 3.812e-03 10.376 < 2e-16 \*\*\*  
## iswash -3.034e-02 1.555e-02 -1.951 0.05111 .   
## isracist -5.110e-02 7.814e-02 -0.654 0.51312   
## isasian 8.595e-02 5.031e-02 1.709 0.08755 .   
## ischinese 2.557e-02 1.241e-02 2.061 0.03930 \*   
## isinfectious 2.232e-02 3.824e-02 0.584 0.55951   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.111115)  
##   
## Null deviance: 3813.1 on 29999 degrees of freedom  
## Residual deviance: 3330.9 on 29977 degrees of freedom  
## AIC: 19246  
##   
## Number of Fisher Scoring iterations: 2

#Regression Tree - Retweets total  
tree.retweets <- tree(retweets ~ length + favorites + ishealth + ispandemic +   
 isvirus + isemergency + isdeaths + iswho + iscdc + isnih +   
 isdisease + isquarantine + isrecover + isban + iscoronavirus +   
 iscovid19 + iswash + isracist + isasian + ischinese + isinfectious,   
 data = covid19tidy)  
summary(tree.retweets)

##   
## Regression tree:  
## tree(formula = retweets ~ length + favorites + ishealth + ispandemic +   
## isvirus + isemergency + isdeaths + iswho + iscdc + isnih +   
## isdisease + isquarantine + isrecover + isban + iscoronavirus +   
## iscovid19 + iswash + isracist + isasian + ischinese + isinfectious,   
## data = covid19tidy)  
## Variables actually used in tree construction:  
## [1] "favorites"  
## Number of terminal nodes: 2   
## Residual mean deviance: 55590000 = 1.668e+12 / 30000   
## Distribution of residuals:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -3086.0 -3056.0 -2744.0 0.0 -562.1 42910.0

plot(tree.retweets)  
text(tree.retweets)



#Classification Tree - Is it a retweet?  
tree.isretweeted <- tree(isretweeted ~ length + favorites + retweets + ishealth + ispandemic +   
 isvirus + isemergency + isdeaths + iswho + iscdc + isnih +   
 isdisease + isquarantine + isrecover + isban + iscoronavirus +   
 iscovid19 + iswash + isracist + isasian + ischinese + isinfectious,   
 data = covid19tidy)  
summary(tree.isretweeted)

##   
## Regression tree:  
## tree(formula = isretweeted ~ length + favorites + retweets +   
## ishealth + ispandemic + isvirus + isemergency + isdeaths +   
## iswho + iscdc + isnih + isdisease + isquarantine + isrecover +   
## isban + iscoronavirus + iscovid19 + iswash + isracist + isasian +   
## ischinese + isinfectious, data = covid19tidy)  
## Variables actually used in tree construction:  
## [1] "retweets" "favorites"  
## Number of terminal nodes: 5   
## Residual mean deviance: 0.01433 = 429.8 / 30000   
## Distribution of residuals:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -0.98920 0.01077 0.01077 0.00000 0.01077 0.99300

plot(tree.isretweeted)  
text(tree.isretweeted)

