Twitter Covid19 Tweet Analysis

Minh Nguyen

3/12/2020

library(tidyverse)

## Warning: package 'forcats' was built under R version 3.6.3

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)  
covid19current <- read.csv("C:/Users/mthng/OneDrive/Documents/Personal/School/3 Spring 2020/Predictive Analytics (ITEC 621-002)/Project/twittercovid19/WIP/twittercovid19/Data/covid19march.csv")

## Word Mining

#Was used to create "word" count variables  
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

#USE covid19current  
#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 +  
 ch.confirmed + it.confirmed + ot.confirmed + us.confirmed +  
 ch.deaths + it.deaths + ot.deaths + us.deaths,   
 data = covid19current)  
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 +   
## ch.confirmed + it.confirmed + ot.confirmed + us.confirmed +   
## ch.deaths + it.deaths + ot.deaths + us.deaths, data = covid19current)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -140172 -7127 -2212 2490 188039   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.327e+07 5.828e+05 22.776 < 2e-16 \*\*\*  
## length -9.815e+00 6.683e+00 -1.469 0.141909   
## favorites -9.898e+00 4.193e+00 -2.361 0.018254 \*   
## ishealth -2.611e+02 5.041e+02 -0.518 0.604438   
## ispandemic -1.511e+03 9.188e+02 -1.645 0.100003   
## isvirus 1.127e+03 3.742e+02 3.011 0.002603 \*\*   
## isemergency -4.602e+03 1.308e+03 -3.519 0.000433 \*\*\*  
## isdeaths -3.546e+03 8.718e+02 -4.067 4.77e-05 \*\*\*  
## iswho 2.336e+04 5.155e+02 45.318 < 2e-16 \*\*\*  
## iscdc -3.537e+03 1.014e+03 -3.489 0.000485 \*\*\*  
## isnih -6.725e+03 2.443e+03 -2.753 0.005905 \*\*   
## isdisease -4.478e+03 1.170e+03 -3.829 0.000129 \*\*\*  
## isquarantine -3.605e+03 8.445e+02 -4.268 1.97e-05 \*\*\*  
## isrecover -3.591e+03 1.839e+03 -1.953 0.050846 .   
## isban 1.404e+03 7.507e+02 1.871 0.061362 .   
## iscoronavirus 5.916e+03 4.170e+02 14.188 < 2e-16 \*\*\*  
## iscovid19 -4.350e+03 2.084e+02 -20.872 < 2e-16 \*\*\*  
## iswash -3.012e+03 8.900e+02 -3.385 0.000713 \*\*\*  
## isracist -8.396e+03 3.732e+03 -2.250 0.024472 \*   
## isasian -3.251e+03 2.993e+03 -1.086 0.277442   
## ischinese 4.447e+03 6.769e+02 6.570 5.10e-11 \*\*\*  
## isinfectious -2.552e+03 2.183e+03 -1.169 0.242377   
## ch.confirmed -1.847e+02 8.189e+00 -22.554 < 2e-16 \*\*\*  
## it.confirmed -1.514e+01 6.613e-01 -22.900 < 2e-16 \*\*\*  
## ot.confirmed 1.146e+01 3.742e-01 30.618 < 2e-16 \*\*\*  
## us.confirmed -5.787e+01 4.116e+00 -14.059 < 2e-16 \*\*\*  
## ch.deaths 5.064e+02 2.726e+01 18.576 < 2e-16 \*\*\*  
## it.deaths 3.900e+01 5.396e+00 7.229 4.96e-13 \*\*\*  
## ot.deaths -6.426e+01 1.317e+01 -4.879 1.07e-06 \*\*\*  
## us.deaths -2.665e+02 1.530e+02 -1.741 0.081617 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21530 on 41970 degrees of freedom  
## Multiple R-squared: 0.1797, Adjusted R-squared: 0.1792   
## F-statistic: 317.1 on 29 and 41970 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 + ch.confirmed + it.confirmed + ot.confirmed +  
 us.confirmed + ch.deaths + it.deaths + ot.deaths + us.deaths,   
 data = covid19current)  
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 +   
## ch.confirmed + it.confirmed + ot.confirmed + us.confirmed +   
## ch.deaths + it.deaths + ot.deaths + us.deaths, data = covid19current)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.40 -1.51 -1.18 -0.79 2405.45   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.431e+02 6.826e+02 -0.649 0.5163   
## length -1.030e-02 7.778e-03 -1.325 0.1853   
## retweets -1.341e-05 5.681e-06 -2.361 0.0183 \*  
## ishealth 6.097e-02 5.868e-01 0.104 0.9172   
## ispandemic -2.195e-01 1.069e+00 -0.205 0.8374   
## isvirus -5.422e-01 4.356e-01 -1.245 0.2133   
## isemergency -3.334e-01 1.522e+00 -0.219 0.8266   
## isdeaths 2.896e-01 1.015e+00 0.285 0.7754   
## iswho 1.337e+00 6.145e-01 2.176 0.0295 \*  
## iscdc -9.067e-01 1.180e+00 -0.768 0.4422   
## isnih -6.086e-01 2.843e+00 -0.214 0.8305   
## isdisease -9.777e-01 1.362e+00 -0.718 0.4727   
## isquarantine -8.883e-01 9.832e-01 -0.904 0.3662   
## isrecover -7.754e-01 2.141e+00 -0.362 0.7172   
## isban -8.158e-01 8.738e-01 -0.934 0.3505   
## iscoronavirus 4.523e-01 4.865e-01 0.930 0.3526   
## iscovid19 -3.868e-01 2.438e-01 -1.586 0.1126   
## iswash 1.069e+00 1.036e+00 1.032 0.3022   
## isracist -6.505e-01 4.344e+00 -0.150 0.8810   
## isasian 9.432e-01 3.484e+00 0.271 0.7866   
## ischinese -5.986e-01 7.883e-01 -0.759 0.4476   
## isinfectious 2.796e-01 2.541e+00 0.110 0.9124   
## ch.confirmed 5.806e-03 9.589e-03 0.605 0.5449   
## it.confirmed 4.441e-04 7.745e-04 0.573 0.5664   
## ot.confirmed -3.377e-04 4.403e-04 -0.767 0.4431   
## us.confirmed 5.392e-03 4.803e-03 1.123 0.2616   
## ch.deaths -5.664e-03 3.186e-02 -0.178 0.8589   
## it.deaths 7.023e-04 6.285e-03 0.112 0.9110   
## ot.deaths -5.302e-03 1.534e-02 -0.346 0.7295   
## us.deaths -1.520e-01 1.781e-01 -0.853 0.3935   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 25.06 on 41970 degrees of freedom  
## Multiple R-squared: 0.0005786, Adjusted R-squared: -0.000112   
## F-statistic: 0.8379 on 29 and 41970 DF, p-value: 0.7141

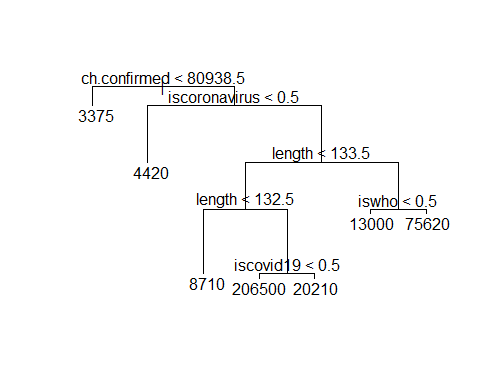
#Generalized Linear Models (GLM) - is it a retweet?  
isretweeted.glm <- glm(isretweeted ~ length + favorites + retweets + ishealth +  
 ispandemic + isvirus + isemergency + isdeaths + iswho +   
 iscdc + isnih + isdisease + isquarantine + isrecover +   
 isban + iscoronavirus + iscovid19 + iswash + isracist +   
 isasian + ischinese + isinfectious + ch.confirmed +  
 it.confirmed + ot.confirmed + us.confirmed + ch.deaths +  
 it.deaths + ot.deaths + us.deaths,   
 data = covid19current)  
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 +   
## ch.confirmed + it.confirmed + ot.confirmed + us.confirmed +   
## ch.deaths + it.deaths + ot.deaths + us.deaths, data = covid19current)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.00302 0.07424 0.10557 0.13617 3.01912   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.358e+01 8.966e+00 2.630 0.008539 \*\*   
## length 6.365e-03 1.022e-04 62.297 < 2e-16 \*\*\*  
## favorites -1.621e-03 6.412e-05 -25.286 < 2e-16 \*\*\*  
## retweets 1.546e-06 7.463e-08 20.713 < 2e-16 \*\*\*  
## ishealth 2.431e-02 7.707e-03 3.153 0.001615 \*\*   
## ispandemic 4.498e-03 1.405e-02 0.320 0.748821   
## isvirus -3.704e-02 5.722e-03 -6.473 9.7e-11 \*\*\*  
## isemergency -1.644e-02 2.000e-02 -0.822 0.411131   
## isdeaths -1.878e-02 1.333e-02 -1.409 0.158974   
## iswho -1.064e-02 8.072e-03 -1.318 0.187369   
## iscdc 2.672e-02 1.550e-02 1.724 0.084721 .   
## isnih 2.132e-02 3.735e-02 0.571 0.568142   
## isdisease 1.317e-02 1.788e-02 0.736 0.461458   
## isquarantine 4.225e-02 1.291e-02 3.271 0.001072 \*\*   
## isrecover -7.396e-02 2.812e-02 -2.630 0.008533 \*\*   
## isban -1.712e-02 1.148e-02 -1.492 0.135769   
## iscoronavirus 2.325e-02 6.391e-03 3.639 0.000274 \*\*\*  
## iscovid19 3.829e-02 3.203e-03 11.955 < 2e-16 \*\*\*  
## iswash -4.347e-02 1.361e-02 -3.194 0.001404 \*\*   
## isracist 2.193e-02 5.707e-02 0.384 0.700770   
## isasian 3.102e-02 4.577e-02 0.678 0.497860   
## ischinese 3.071e-02 1.035e-02 2.966 0.003023 \*\*   
## isinfectious 3.196e-02 3.338e-02 0.957 0.338356   
## ch.confirmed -3.421e-04 1.260e-04 -2.716 0.006613 \*\*   
## it.confirmed -3.616e-05 1.017e-05 -3.555 0.000379 \*\*\*  
## ot.confirmed 1.485e-05 5.784e-06 2.567 0.010269 \*   
## us.confirmed -1.151e-04 6.308e-05 -1.825 0.068064 .   
## ch.deaths 1.281e-03 4.185e-04 3.061 0.002209 \*\*   
## it.deaths -8.517e-05 8.255e-05 -1.032 0.302220   
## ot.deaths 5.558e-04 2.014e-04 2.759 0.005801 \*\*   
## us.deaths -3.019e-03 2.340e-03 -1.290 0.196952   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.1083265)  
##   
## Null deviance: 5125.0 on 41999 degrees of freedom  
## Residual deviance: 4546.4 on 41969 degrees of freedom  
## AIC: 25874  
##   
## 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 +   
 ch.confirmed + it.confirmed + ot.confirmed + us.confirmed +  
 ch.deaths + it.deaths + ot.deaths + us.deaths,   
 data = covid19current)  
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 +   
## ch.confirmed + it.confirmed + ot.confirmed + us.confirmed +   
## ch.deaths + it.deaths + ot.deaths + us.deaths, data = covid19current)  
## Variables actually used in tree construction:  
## [1] "ch.confirmed" "iscoronavirus" "length" "iscovid19"   
## [5] "iswho"   
## Number of terminal nodes: 7   
## Residual mean deviance: 101200000 = 4.251e+12 / 41990   
## Distribution of residuals:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -206500 -3370 -3218 0 -1808 74820

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 + ch.confirmed +   
 it.confirmed + ot.confirmed + us.confirmed + ch.deaths +  
 it.deaths + ot.deaths + us.deaths,   
 data = covid19current)  
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 + ch.confirmed + it.confirmed +   
## ot.confirmed + us.confirmed + ch.deaths + it.deaths + ot.deaths +   
## us.deaths, data = covid19current)  
## Variables actually used in tree construction:  
## [1] "retweets" "favorites"  
## Number of terminal nodes: 5   
## Residual mean deviance: 0.01419 = 595.8 / 42000   
## Distribution of residuals:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -0.98890 0.01108 0.01108 0.00000 0.01108 0.99300

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

