Análisis de sentimientos de twitter sobre COVID-19 en paises hispanohablantes

Nicolas Garcés Rodríguez, Kevin Galarcio Torres, David Alejandro Bermúdez

***Abstract* – Este trabajo pretende utilizar analisis de texto y modelos de Machine Learning para explicar que elementos de la pandemia del COVID-19 y el manejo gubernamental influyen en el sentimiento de los Tweets de la poblacion Hispanohablante durante la crisis respecto al tema.**

***Palabras Clave* — Analisis de texto, Twitter, sentimientos, Arbor de regresion, Machine Learning.**

I. INTRODUCCIÓN

Con el paso de los años, Twitter ha logrado destacar y establecerse entre las demás redes sociales. Su formato de fácil acceso, enfocado en la interacción por medio de la difusión de texto plano, la ha convertido en un portal de flujo de información masiva, donde una buena parte la ocupa la opinión pública. La rapidez de las interacciones, las tendencias, los hashtags y los trinos son un reflejo de lo que está en boca de los actores sociales y conforme los usuarios asimilan más la plataforma, los trinos y tweets se convierten en la voz directa de estas personas y Twitter se transforma en un escenario real de discusión pública.

El gran valor que tiene el flujo de datos de Twitter ha llevado a la academia a tomarala como una fuente primaria de informacion, en especial de carácter critico. Rosenthal et al, en un paper del 2015, ahondan en el analisis de este tipo de informacion. El estudio desarrollo 6 herramientas de analisis : predicción de sentimientos, sentimiento expresado por una frase en el contexto, sentimiento general de un tweet, sentimiento hacia un tema en un solo tweet, sentimiento general hacia un tema en un conjunto de tweets y la polaridad previa de una frase. La metodología aplicada fue una clasificación de cada tweet, diferenciado entre palabras y frases y que se pudieran clasificar como positiva, neutral o negativa. La mayoria de las predicciones realizaron matrices de confusión y estadísticos F1-Score y Recall como bondad de ajuste. Los modelos fueron un éxito y su exactitud estuvo por encima del 90%(SemEval-2015: Análisis de sentimientos en Twitter; Rosenthal et al, 2015).

Un estudio mas aplicado en esta area, implementando machine Learning, con el fin de identificar sentimientos sobre las ofertas del BlackFriday en base a tweets, analizó la probabilidad de compra dividida en algunas características, como su estado civil, genero e ingresos (Identificando el sentimiento sobre las ofertas de #BlackFriday; José Saura et all; 2018). En esta investigación solo se usó el tema y los hashtags del tweet. Para realizar la clasificación se usó un algoritmo de análisis de sentimiento que se encuentra en la librería MonkeyLearn de Python. El estudio concluyó que Media Mark y Xiaomi fueron las que más tweets negativos recibieron, y Microsoft y Worten fueron la de Tweets mas positivos.

Las Metodologías implementadas en los ejemplos anteriores y en la mayoria de estudios consultados usan un algoritmo de clasificación sencillo como LPA (Label Propagation Algorithm).Otras realizan agrupación por Clustering para así identificar comunidades, y para evaluar el impacto de estos Tweets se usa SVP (Support Vector Machine) en algunas investigaciones.

Teniendo en cuenta estos estudios y la coyuntura internacional que se está presentando por cuenta de la pandemia "COVID-19", es de especial interés analizar las opiniones y sentimientos de las personas en cuanto a el desarrollo de virus en sus respectivos países, conjunto a las diferentes medidas que toman las autoridades para su contención. Twitter, como se ha mostrado, es el espacio propicio para encontrar esta información.

II. DATOS

Teniendo en cuenta la información perteneciente al paquete *COVID19*, se eligieron las siguientes variables explicativas, las cuales se dividen en dos grupos: datos reales y datos acerca de las medidas gubernamentales. Estas variables fueron elegidas porque son los principales datos que un individuo toma en consideración a la hora de opinar acerca de la coyuntura actual.

Datos reales

Tests: Hace referencia al número de pruebas de laboratorio realizadas en el país para la detección de este virus.

Confirmed: Número de personas en el país que tienen covid-19.

Recovered: Número de personas en el país que se tuvieron covid-19 y se han recuperado de este.

Deaths: Número de personas en el país que se tuvieron covid-19 y han fallecido a causa de este.

Medidas gubernamentales

Stay_home_restrictions: Hace referencia a las medidas de aislamiento impuestas. Donde 0 significa que no hay medidas, 1 el Estado recomienda no salir de Casa, 2 Restricción de salir de casa con ciertas excepciones, 3 Restricción de salir de casa con ciertas excepciones aún más rigurosas y severas.

Internal_movement_restrictions: Medidas en cuanto a la circulación dentro de la nación. Donde 0 significa que no hay medidas, 1 se recomienda no viajar entre regiones/ciudades, 2 restricciones internas de movimiento (dentro de la ciudad/municipio).

International_movement_restrictions: Medidas en cuanto a la circulación entre naciones. Donde 0 significa que no tiene restricciones, 1 se detectan las llegadas de internacionales, 2 si llega desde cierto país, debe permanecer en cuarentena preventiva, 3 prohibición de la entrada a personas de ciertos países, 4 prohibición de todos los países.

Workspace_closing: Medidas laborales. Donde 0 significa que no hay restricciones, 1 se recomienda el trabajo desde los hogares, 2 se requiere el cierre o trabajo desde casa de algunos sectores laborales, cierre para lugares de trabajo de carácter esencial (supermercados).

Gathering_restrictions: Medidas que restringen reuniones entre individuos. Donde 0 significa que no hay restricciones, 1 el límite es más de 1000 personas, 2 límite entre 101 y 1000 personas, 3 límite entre 11-100 personas, 4 limite es hasta 10 personas o menos.

Transport_closing: En cuanto a las medidas sobre el transporte público. Donde 0 significa que no hay restricciones, 1 reducción del volumen y rutas, 2prohibición a la mayoría de los ciudadanos el uso de estos medios.

Testing_policy: Sobre qué población se está realizando las pruebas del virus. Donde 0 significa que no hay políticas de prueba, 1 aquellos que tienen síntomas y cumplen ciertos criterios, 2 cualquier persona que tenga síntomas del virus, 3 pruebas públicas abiertas.

Contract_tracing: Seguimiento sobre personas con síntomas. Donde 0 significa que no hay seguimiento, 1 seguimiento para ciertos casos, 2 seguimiento integral para todos los casos identificados.

Debido a que las variables *Tests, Confirmed, Recovered y Deaths* tienen una varianza considerable, se estandarizaron mediante el logaritmo natural, a excepción de Variable *Tests* en cual, se usó el método de Min-Max debido a que Venezuela no ha realizado tests de acuerdo con los estándares internacionales y no son comparables con los demás.

En cuanto a nuestra variable de interés, el sentimiento de los tweets será un dato entre -5 y 5 diario para 11 países de habla hispana. La variable se denominará *Twitter sentiment.*

Bajo un análisis inicial de estadística descriptiva (Anexo2), podemos observar que, aunque el puntaje tiene un intervalo de –5 a 5, el sentimiento medio en cada día se encuentra en su mayoría sobre valores negativos y aún más relevante, estas puntuaciones están distribuidas de una manera muy compacta, evidenciado en que la varianza es de 0,034. Este factor puede incidir sobre los resultados y desempeños de los modelos.

De acuerdo los diagramas de dispersión (Anexo3), ex ante a proceder con el planteamiento de los diferentes modelos, se puede observar que no existe un patrón de correlación definido entre las variables predictivas y la variable de interés, por lo que se espera que los modelos no se desempeñen de la mejor manera.

III. METODOLOGÍA

La metodología del trabajo se dividirá en dos partes: 1) extracción y análisis de sentimiento de tweets y 2) planteamiento del modelo de predicción.

A. Extracción y análisis de sentimiento de tweets

Se realizó la extracción de tweets a lo largo de 20 días por medio del paquete "rtweet" en 11 países de habla hispana: Argentina, Bolivia, Colombia, Chile, Ecuador, España, México, Paraguay, Perú, Uruguay y Venezuela. Para filtrar la información en la función "search_tweets" se utilizó como palabra clave "COVID", no se incluyeron retweets, no se incluyeron cuentas verificadas (para dejar de lado cuentas de noticias y que no representaran la opinión de un individuo real) y se definió el lugar por medio de la longitud y latitud de las ciudades capitales de cada uno de estos países. Dada las limitaciones de twitter for developers, se descargaron 2500 tweets diarios por país, que después de pasar por los filtros anteriormente mencionados dejaba una base de alrededor de 1000 tweets diarios por país.

Para el análisis de sentimientos, se decidió aplicar un modelo de diccionario. Estos modelos utilizan colecciones de palabras ya categorizadas como positivas o negativas y al pasarlo por los tweets se queda solo con las palabras que coinciden tanto en el tweet como en el diccionario, para así poder dar una valoración promedio del sentimiento del tweet. En este caso se utilizó el diccionario "AFINN" que tiene palabras ya calificadas en el rango de -5 a 5. La versión utilizada es una traducción de esta misma base [6].

Para poder analizar los tweets con este diccionario se tuvo que hacer un preprocesamiento de la base. Se planteó una función para desglosar los tweets en palabras (tokenizar), que eliminaba espacios, puntos, links, entre otros caracteres no deseados. Una vez desglosados los tweets, se enlaza estas palabras con el diccionario y quedan las palabras principales de los tweets con su debida calificación. De esta base se elimina la palabra "no" por su ambigüedad y por su frecuencia, que no aporta mucho al análisis. Finalmente, después de todo el procesamiento, queda una lista de palabras utilizadas en todos los tweets con su respectiva valoración. De esta puntuación, se saca el promedio diario de la valoración de las palabras utilizadas y se obtiene la variable de interés.

Una vez creada la variable de interés (Twitter sentiment) se añadió esta información como una nueva columna por país de la base de datos Covid 19 Data Hub que tiene información diaria acerca la situación actual de cada país y las medidas que han tomado los diferentes gobiernos. De esta manera se conformó la base final utilizada para los modelos.

B. Modelo de predicción

Como la variable objetivo es numérica, se decidieron correr cinco modelos: *regresión lineal simple, regression tree, regresión lineal con cross-validatio, regression tree con cross- validation y un cubist model tree con cross-validation*, este último es solo una pequeña desviación del model tree clásico en el que las regresiones realizadas en cada nodo están suavizadas teniendo en cuenta las predicciones de nodos anteriores. Además, realiza regresiones entre nodos. Se incluyó este último modelo para tratar de mejorar la estimación del árbol de regresiones dada la poca varianza de la variable explicativa. En todos los modelos se utilizaron como variables explicativas todos los datos seleccionados de la base Covid 19 Data Hub con los ajustes ya mencionados y como variable explicada el *Twitter sentiment*. Para todos los modelos se utilizó como medida de ajuste la correlación del pronóstico con los datos del test y el error absoluto medio entre estas mismas dos variables. La Cross- validation utilizada en los últimos tres modelos fue un 10-fold cross-validation. Se decidió no realizar los modelos ridge y lasso debido a que como es mencionado anteriormente, la elección de las variables de predicción fue parte de un proceso de análisis previo y no era practico eliminar alguna de ellas.

IV. RESULTADOS

Para el modelo de regresión lineal las variables *Recovered* e *International movement restriction***s** fueron las únicas variables significativas al 5%, donde las dos tuvieron relaciones negativas respecto al sentimiento en Twitter, lo cual es algo contra intuitivo para *recovered* pero no para *international movement restriction*. El modelo explica un 25% de la varianza de los sentimientos de los tweets. tiene un R ajustado del 20%. El Error absoluto medio (MAE) es de 0.125. Si se tiene en cuenta que la varianza es de 0,03, el modelo no tiene un gran poder para pronosticar los datos. De todas formas, la distancia promedio entre lo pronosticado y el dato real está a menos de una desviación estándar, lo que significa que, como mencionado anteriormente, la poca varianza de la variable explicada limita la predicción del modelo.

Para el árbol de regresión las variables más importantes del son: *tests, deaths, confirmed, recoverd y testing policy*. De todas formas, ninguna de estas tiene una importancia significativa (ninguna pasa del 20%). El árbol hace splits en las mismas variables ya mencionadas, pues al ser las más explicativas también son las que permiten dividir los datos de forma más homogénea. El árbol tiene 8 nodos finales con distribuciones muy similares, lo que indica que el MAE puede ser grande dada la similitud entre las observaciones (Anexo 4). La correlación del modelo es de 53%. Es grande, pero eso no necesariamente es algo bueno. La poca varianza de los datos hace que esta medida no sea muy diciente de la capacidad predictiva del modelo. Si los datos están muy agrupados, la estimación será muy cercana a esos valores, pero no porque logre explicar cada variable, sino porque estadísticamente las predicciones tendrán más oportunidad de acertar si se encuentran en ese rango. El error absoluto medio (MAE) es 0.12909. Es muy similar al modelo anterior y presenta sus mismos problemas y limitaciones.

El modelo lineal con cross validation de 10-folds tiene un comportamiento similar al modelo lineal. Las variables significativas del modelo lineal normal siguen siendo las mismas con cross validation De todas formas "international movement restrictions" pasa de ser significativa del 5% al 10%. Hacer cross-validation para la regresión lineal no presenta ninguna mejora en el modelo. El MAE pasa a ser de 0.133, peor que en el modelo lineal normal.

El árbol de regresiones con cross-validation de 10-folds tampoco dista mucho del modelo original. la importancia de las variables aumenta un poco y el orden de importancia se mantiene constante. La correlación cae un poco, pero ya se sabe que esta medida no es muy diciente. El MAE empeora y pasa a ser 0.1367. Hacer cross-validation en este caso tampoco implica una mejora del modelo.

Por último, el cubist model tree es el modelo con mejor ajuste. El error promedio de los 10 árboles fue de 0.15. testing_policy es la variable explicativa con la que los splits crea grupos más homogéneos. Tiene correlación del 0.5437. El MAE de este modelo es 0.1223426 y por ende es el menor de todos. A pesar de ello, no dista mucho de los otros modelos y sigue teniendo los mismos problemas metodológicos ya mencionados.

V. LIMITACIONES

Como ya se ha mencionado, la poca varianza del análisis de sentimiento de Twitter limita mucho el estudio. La ventana de tiempo con la que se recolectaron los datos, al ser de corto plazo y al darse en un contexto de homogeneidad del desarrollo de pandemia entre los países hispanohablantes, no permite dar un contraste de eventos en los que las personas puedan reaccionar de forma diferente y por eso los sentimientos de los tweets mantienen un pesimismo constante.

Este mismo periodo de tiempo implico tener una baja recolección de datos, que, sumadas a los reportes erróneos en ciertos países, implico que el modelo se corriera con muy pocas observaciones (alrededor de 193) lo que les quita robustez a los modelos, especialmente a los modelos tipo árbol.

Por último, promediar la valoración de las palabras es una buena aproximación a lo que pueden estar sintiendo las personas, pero este tipo de análisis se encuentra muy limitada al diccionario utilizado, en el sentido de que pierde elementos clave como el contexto o la estructura de las frases que en ciertos casos pueden estar direccionados a un sentimiento distinto y puede proveer un análisis más preciso de los tweets.

VI. CONCLUSIONES

De todos los algoritmos utilizados, el cubist model tree fue aquel que tuvo mejor desempeño, las regresiones extra que hace entre nodos y el cross-validation con 10-folds permitió darle un mejor ajuste al pronostico

A pesar de las limitaciones, los modelos que se han utilizado apuntan constantemente a ciertas variables que pueden incidir en los sentimientos que quiera expresar una persona al escribir un tweet, como el número de muertes, el número de pruebas y las políticas para su uso. Estas herramientas pueden ser útiles para que autoridades estatales puedan tomar medidas incentivadas en el aumento de la confianza de la población.

El trabajo desarrollado en este Paper puede ser un punto de inicio para un estudio con una temporalidad más larga que pueda lograr opiniones más diversificadas, resultados significativos y un análisis más profundo acerca de la incidencia de las políticas gubernamentales sobre el sentimiento y opinión de la población hispanohablante en este o en próximos escenarios parecidos. De igual forma La expansión de este estudio también puede poner en tela de juicio la capacidad explicativa de los modelos de diccionario en el análisis de texto.

VII. REFERENCIAS

[1] BLAZQUEZ PARDO, B. P. (2019, septiembre). Análisis de redes sociales y minería de textos aplicadas a caracterizar comunidades de usuarios en Twitter. Recuperado de https://repositorio.uam.es/xmlui/bitstream/handle/10486/688795/blazquez_pardo_roberto_%28tfm%29.pdf?sequence=1&isAllowed=y

[2] ROALES GONZALEZ, R. G. (2014, junio). DETECCIÓN DE TENDENCIAS EN TWITTER UTILIZANDO MINERÍA DE DATOS ADAPTATIVA. Recuperado de https://repositorio.uam.es/bitstream/handle/10486/662510/roales_gonzalez_natalia_tfg.pdf?sequence=1

[3] SAURA RAMON, S. R., & REYES MELENDEZ , R. M. (2018). Un Análisis de Sentimiento en Twitter con Machine Learning: Identificando el sentimiento sobre las ofertas de #BlackFriday. Recuperado de http://www.revistaespacios.com/a18v39n42/a18v39n42p16.pdf

[4] Rosental Sara, R. S., Nakov Preslav, N. P., Mohammad Saif, M. S., Kiritchenko Svetlana , K. S., Ritter Alan , R. A., & Stoyanov Veselin , S. V. (2015). SemEval-2015 Task 10: Sentiment Analysis in Twitter. Recuperado de https://arxiv.org/pdf/1912.02387.pdf

[5] Singh Tajinder , S. T., & Kumari Madhu, K. M. (2016). Role of Text Pre-Processing in Twitter Sentiment Analysis. Recuperado de https://core.ac.uk/download/pdf/82655045.pdf

[6] Amat Rodrigo, J. (2017, diciembre). Text mining con R: ejemplo práctico en Twitter. Recuperado de https://rpubs.com/Joaquin_AR/334526

[7] Hale,T. Petherick ,A., Phillips,T(2020) Variation in government responses to COVID-19 recuperado de https://www.bsg.ox.ac.uk/sites/default/files/2020-04/BSG-WP-2020-032-v5.0.pdf

[8] Guidotti, E., Ardila ,A(2020) COVID-19 Data Hub. Recuperado de https://covid19datahub.io/

VIII. ANEXOS

ANEXO 1: CODIGO R

```
[1]  ##############################################
[2]  ######## DATOS COVID TWITTER ################
[3]  ##############################################
[4]
[5]  library(ROAuth)
[6]  library(streamR)
[7]  library(twitteR)
[8]  library(rtweet)
[9]  library(tidyverse)
[10] library(tidytext)
[11] # ACCESO A TWITTER ----------------------------------------------------
[12] app_name <- "BigDataCovid19"
[13] consumer_key <- "5XaBJzkUMS24xcT4xEUvsEvmx"
[14] consumer_secret <- "#####################################"
[15] access_token <- "########################################"
[16] access_secret <- "########################################"
[17] twitter_token <- create_token(app_name, consumer_key, consumer_secret, access_token, access_secret)
[18] # COLOMBIA -------------------------------------------------------------
[19] t_Col<- search_tweets(q = "covid" , type ="recent" ,n = 2500, lang="es",include_rts = FALSE, geocode = "4.624335,-74.063644,50mi" )
[20] t_Col<-as.data.frame(t_Col)
[21] t_col<- filter(t_Col,verified == F)
[22] Col<- select(t_col,screen_name,created_at,text,retweet_count)
[23]

[24] Col$created_at <- as.character(Col$created_at)
[25] col<-filter(Col, grepl('2020-05-23', created_at ))
[26] view(col)
[27]
[28] # ESPAÃ‘A ---------------------------------------------------------------
[29]
[30]
[31] t_Esp<- search_tweets(q = "covid" , type ="recent" ,n = 2500, lang="es",include_rts = FALSE, geocode = "40.416775,-3.703790,50mi" )
[32] t_Esp<-as.data.frame(t_Esp)
[33] t_Esp<- filter(t_Esp,verified == F)
[34] Esp<- select(t_Esp,screen_name,created_at,text,retweet_count)
[35]
[36]
[37] Esp$created_at <- as.character(Esp$created_at)
[38] Esp<-filter(Esp, grepl('2020-05-23', created_at ))
[39] view(Esp)
[40] # ECUADOR --------------------------------------------------------------
[41] t_Eq<- search_tweets(q = "covid" , type ="recent" ,n = 2500, lang="es",include_rts = FALSE, geocode = "-0.180653,-78.467834,50mi" )
[42] t_Eq<-as.data.frame(t_Eq)
[43] t_Eq<- filter(t_Eq,verified == F)
[44] Eq<- select(t_Eq,screen_name,created_at,text,retweet_count)
[45]
[46]
[47]
[48]
[49]
[50] Eq$created_at <- as.character(Eq$created_at)
[51]
[52] Eq<-filter(Eq, grepl('2020-05-23', created_at ))
[53] view(Eq)
[54] # MEXICO ---------------------------------------------------------------
[55] t_Mx<- search_tweets(q = "covid" , type ="recent" ,n = 2500, lang="es",include_rts = FALSE, geocode = "19.432608,-99.133209,50mi" )
[56] t_Mx<-as.data.frame(t_Mx)
[57] t_Mx<- filter(t_Mx,verified == F)
[58] Mx<- select(t_Mx,screen_name,created_at,text,retweet_count)
[59]
[60]
```

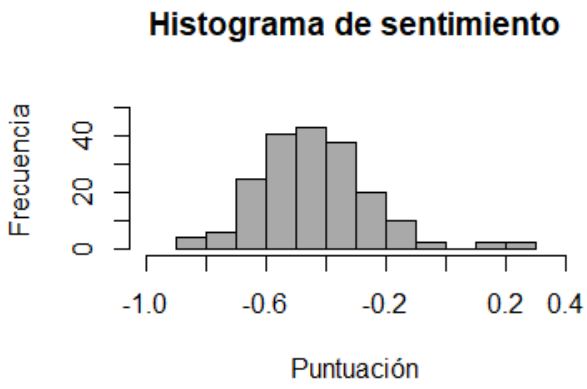
```
[61]
[62]
[63]
[64] Mx$created_at <- as.character(Mx$created_at)
[65] Mx<-filter(Mx, grepl('2020-05-23', created_at ))
[66] view(Mx)
[67]
[68] # PERU ----------------------------------------------------------------
[69]
[70]
[71] t_Pe<- search_tweets(q = "covid" , type ="recent" ,n = 2500, lang="es",include_rts = FALSE, geocode = "-12.046374,-77.042793,50mi" )
[72] t_Pe<-as.data.frame(t_Pe)
[73] t_Pe<- filter(t_Pe,verified == F)
[74] Pe<- select(t_Pe,screen_name,created_at,text,retweet_count)
[75]
[76]
[77] Pe$created_at <- as.character(Pe$created_at)
[78] Pe<-filter(Pe, grepl('2020-05-23', created_at ))
[79] view(Pe)
[80]
[81] # Paraguay ----------------------------------------------------------------
[82]
[83]
[84] t_par<- search_tweets(q = "covid 19" , type ="recent" ,n = 3000, lang="es",include_rts = FALSE, geocode = "-25.3006592,-57.63591,50mi" )
[85] t_par<-as.data.frame(t_par)
[86] t_par<- filter(t_par,verified == F)
[87] par<- select(t_par,screen_name,created_at,text,retweet_count)
[88] view(par)
[89]
[90] par$created_at <- as.character(par$created_at)
[91] par<-filter(par, grepl('2020-05-#DIA', created_at ))
[92] view(par)
[93]
[94] # Uruguay ----------------------------------------------------------------
[95]
[96]
[97] t_Uru<- search_tweets(q = "covid 19" , type ="recent" ,n = 3000, lang="es",include_rts = FALSE, geocode = "-34.9032784,-56.1881599,50mi" )
[98] t_Uru<-as.data.frame(t_Uru)
[99] t_Uru<- filter(t_Uru,verified == F)
[100]Uru<- select(t_Uru,screen_name,created_at,text,retweet_count)
[101]view(Uru)
[102]
[103]Uru$created_at <- as.character(Uru$created_at)
[104]Uru<-filter(Uru, grepl('2020-05-#DIA', created_at ))
[105]view(Uru)
[106]
[107]# Chile ----------------------------------------------------------------
[108]
[109]
[110]t_Chi<- search_tweets(q = "covid 19" , type ="recent" ,n = 3000, lang="es",include_rts = FALSE, geocode = " -33.4569400,-70.6482700,50mi" )
[111]t_Chi<-as.data.frame(t_Chi)
[112]t_Chi<- filter(t_Chi,verified == F)
[113]Chi<- select(t_Chi,screen_name,created_at,text,retweet_count)
[114]view(Eq)
[115]
[116]Chi$created_at <- as.character(Chi$created_at)
[117]Chi<-filter(Chi, grepl('2020-05-#DIA', created_at ))
[118]view(Chi)
[119]
[120]# Bolivia ----------------------------------------------------------------
[121]
[122]
[123]t_Bol<- search_tweets(q = "covid 19" , type ="recent" ,n = 3000, lang="es",include_rts = FALSE, geocode = " -16.5000000,-68.1500000,50mi" )
[124]t_Bol<-as.data.frame(t_Bol)
[125]t_Bol<- filter(t_Bol,verified == F)
[126]Bol<- select(t_Bol,screen_name,created_at,text,retweet_count)
[127]view(Bol)
[128]
[129]
[130]Bol$created_at <- as.character(Bol$created_at)
[131]Bol<-filter(Bol, grepl('2020-05-#DIA', created_at ))
[132]view(Bol)
[133]
[134]# Venezuela ----------------------------------------------------------------
[135]
[136]
[137]t_Ven<- search_tweets(q = "covid 19" , type ="recent" ,n = 3000, lang="es",include_rts = FALSE, geocode = "10.6222200,-66.5735300,50mi" )
[138]t_Ven<-as.data.frame(t_Ven)
[139]t_Ven<- filter(t_Ven,verified == F)
[140]Ven<- select(t_Ven,screen_name,created_at,text,retweet_count)
[141]view(Ven)
[142]
[143]
[144]Ven$created_at <- as.character(Ven$created_at)
[145]Ven<-filter(Ven, grepl('2020-05-#DIA', created_at ))
[146]view(Ven)
[147]
[148]# Argentina ----------------------------------------------------------------
[149]
[150]
[151]
[152]t_Arg<- search_tweets(q = "covid 19" , type ="recent" ,n = 3000, lang="es",include_rts = FALSE, geocode = "-34.6131516,-58.3772316,50mi" )
[153]t_Arg<-as.data.frame(t_Arg)
[154]t_Arg<- filter(t_Arg,verified == F)
[155]Arg<- select(t_Arg,screen_name,created_at,text,retweet_count)
[156]view(Arg)
[157]
[158]
[159]Arg$created_at <- as.character(Arg$created_at)
[160]Arg<-filter(Arg, grepl('2020-05-#DIA', created_at ))
[161]view(Arg)
[162]
[163]# Exportar a excel ----------------------------------------------------------
[164]###CAMBIAR FECHAS EN LOS NOMBRES DEL ARCHIVO
[165]
[166]write.table(par,file = "Par#DIAMAY.csv" ,row.names = FALSE,sep = ";")
[167]write.table(Uru,file = "Uru#DIAMAY.csv" ,row.names = FALSE,sep = ";")
[168]write.table(Chi,file = "Chi#DIAMAY.csv" ,row.names = FALSE,sep = ";")
[169]write.table(Bol,file = "Bol#DIAMAY.csv" ,row.names = FALSE,sep = ";")
[170]write.table(Ven,file = "Ven#DIAMAY.csv" ,row.names = FALSE,sep = ";")
[171]write.table(Arg,file = "Arg#DIAMAY.csv" ,row.names = FALSE,sep = ";")
[172]write.table(col,file = "Col23MAY.csv" ,row.names = FALSE,sep = ";")
[173]write.table(Eq,file = "Eq23MAY.csv" ,row.names = FALSE,sep = ";")
[174]write.table(Esp,file = "Esp23MAY.csv" ,row.names = FALSE,sep = ";")
[175]write.table(Mx,file = "Mx23MAY.csv" ,row.names = FALSE,sep = ";")
[176]write.table(Pe,file = "Pe23MAY.csv" ,row.names = FALSE,sep = ";")
[177]
[178]
[179]
[180]
[181]######### DATOS COVID ##########################
[182]
[183]
[184]library(COVID19)
[185]
[186]#Colombia
[187]Col_cov <-covid19(country=170, level= 1, end = Sys.Date(), vintage = FALSE, raw = FALSE, cache = TRUE)
[188]#Equador
[189]Eq_cov <- covid19(country= 218, level= 1, end = Sys.Date(), vintage = FALSE, raw = FALSE, cache = TRUE)
[190]#EspaÃ±a
[191]Esp_cov <- covid19(country = 724, level= 1, end = Sys.Date(), vintage = FALSE, raw = FALSE, cache = TRUE)
[192]#Mexico
[193]Mx_cov <- covid19(country = 484 , level= 1, end = Sys.Date(), vintage = FALSE, raw = FALSE, cache = TRUE)
```

```
#peru
Pe_cov <- covid19(country = 604 , level= 1, end = Sys.Date(), vintage = FALSE, raw = FALSE, cache = TRUE)
#ARGENTINA
arg_cov <-covid19(country=32, level= 1, end = Sys.Date(), vintage = FALSE, raw = FALSE, cache = TRUE)
#BOLIVIA
bol_cov <- covid19( country= 68, level= 1, end = Sys.Date(), vintage = FALSE, raw = FALSE, cache = TRUE)
#CHILE
chi_cov <- covid19(country = 152, level= 1, end = Sys.Date(), vintage = FALSE, raw = FALSE, cache = TRUE)
#PARAGUAY
par_cov <- covid19(country = 600, level= 1, end = Sys.Date(), vintage = FALSE, raw = FALSE, cache = TRUE)
#URUGUAY
uru_cov <- covid19(country = 858, level= 1, end = Sys.Date(), vintage = FALSE, raw = FALSE, cache = TRUE)
#VENEZUELA
ven_cov <- covid19(country = 862 , level= 1, end = Sys.Date(), vintage = FALSE, raw = FALSE, cache = TRUE)


write.table(Col_cov,file = "Col_covid.csv" ,row.names = FALSE,sep = ";")
write.table(Eq_cov ,file = "Eq_covid.csv" ,row.names = FALSE,sep = ";")
write.table(Esp_cov ,file = "Esp_covid.csv" ,row.names = FALSE,sep = ";")
write.table(Mx_cov ,file = "Mx_covid.csv" ,row.names = FALSE,sep = ";")
write.table(Pe_cov ,file = "Pe_covid.csv" ,row.names = FALSE,sep = ";")
write.table(arg_cov,file = "arg_covid.csv" ,row.names = FALSE,sep = ";")
write.table(bol_cov ,file = "bol_covid.csv" ,row.names = FALSE,sep = ";")
write.table(chi_cov ,file = "chi_covid.csv" ,row.names = FALSE,sep = ";")
write.table(par_cov ,file = "par_covid.csv" ,row.names = FALSE,sep = ";")
write.table(uru_cov ,file = "uru_covid.csv" ,row.names = FALSE,sep = ";")
write.table(ven_cov ,file = "ven_covid.csv" ,row.names = FALSE,sep = ";")
library(tidyverse)
library(readxl)
library(gmodels)
library(glmnet)
library(MASS)
library(stringr)
library(tokenizers)
library(stopwords)
library(tidyverse)
library(readxl)
library(quanteda)
library(tidytext)
library(tm)
library(zoo)
library(scales)
library(lubridate)
#############


################################################
######### ANALISIS DE SENTIMIENTOS ############
################################################


## DICCIONARIO ##

download.file("https://raw.githubusercontent.com/jboscomendoza/rpubs/master/sentimientos_afinn/lexico_afinn.en.es.csv",
              "lexico_afinn.en.es.csv")
afinn <- read.csv("lexico_afinn.en.es.csv", stringsAsFactors = F, fileEncoding = "latin1") %>%
  tbl_df()


## FUNCION LIMPIADORA ##

limpiar<- function(texto){
  nuevo_texto <- tolower(texto)
  nuevo_texto <- str_replace_all(nuevo_texto,"http\\S*", "")
  nuevo_texto <- str_replace_all(nuevo_texto,"[[:punct:]]", " ")
  nuevo_texto <- str_replace_all(nuevo_texto,"[[:digit:]]", " ")
  nuevo_texto <- str_replace_all(nuevo_texto,"[\\s]+", " ")
  nuevo_texto <- str_replace_all(nuevo_texto, ".+[:digit:].+", " ")
  nuevo_texto <- str_replace_all(nuevo_texto, "[:digit:].+", " ")
  nuevo_texto <- str_replace_all(nuevo_texto,".+[:digit:]", " ")
  nuevo_texto <- str_replace_all(nuevo_texto, "[:digit:]", " ")
  nuevo_texto <- str_replace_all(nuevo_texto,"@\\S*", " ")
}


## LIMPIANDO LOS DATOS ##

raw_tweets<- read.csv2("Pe23MAY.csv", stringsAsFactors = F, fileEncoding = "latin1")

## Los tweets tiene fecha de publicacion, nombre del usuario, contenido del tweet y numero de retweets

limpiar(raw_tweets$text)

## del contenido del tweet se eliminan los links, la mencion a otros usuarios, hashtags y otros
## elementos que no son importantes para el analisis.



tweets_clean<- raw_tweets%>%
  unnest_tokens(input = "text", output = "Palabra") %>%
  inner_join(afinn, ., by = "Palabra") %>%
  mutate(Tipo = ifelse(Puntuacion > 0, "Positiva", "Negativa"))

# se combina las palabras del diccionario con las palabras de los tweets, dejando solo las que coinciden
# en las dos bases. De esta forma, la nueva base tiene : las palabras, su clasificacion, su traduccion en ingles
# su puntaje y la misma informacion que tenia el tweet de donde proviene.


tweets_clean <-
  tweets_clean%>%
  group_by(created_at) %>% select_("screen_name","created_at","Palabra","Puntuacion","Tipo")
# se deja la informacion de interes


# Para eliminar la palabra no
tweets_clean <-
  tweets_clean%>%
  filter(Palabra != "no")

view(tweets_clean)



mean(tweets_clean$Puntuacion)
# Este es el valor que toma el analsis de sentimientos para ese grupo de tweets

####################
write.table(tweets_clean,file = "xxxxxxxx.csv" ,row.names = FALSE,sep = ";")
######################
################################################
######## MODELO COVID TWITTER ################
################################################



# APROXIMACION A LOS DATOS ----------------------------------------


base<-COVID_final
```

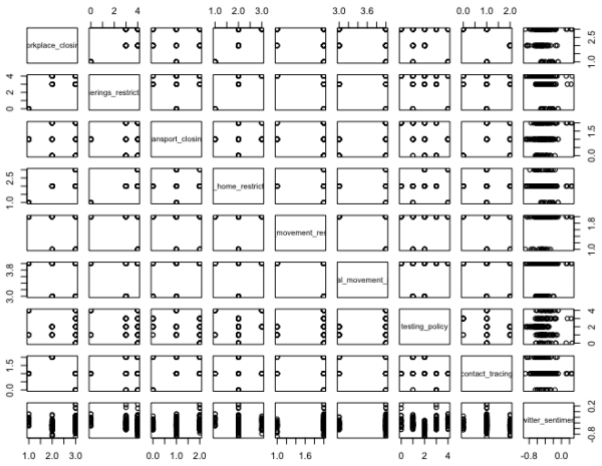
```
base<- base[,-c(1,2,13,16)] # se elimina la columna 13 porque no cambia para ninguna observacion y la 16 por exogeneidad.
base$twitter_sentiment <- as.numeric(base$twitter_sentiment)
view(base)
hist(base$twitter_sentiment)
summary(base$twitter_sentiment)
var(base$twitter_sentiment)#0.0344552
sd(base$twitter_sentiment)#0.1856211
# El analisis de sentimiento tiene una distrubucion normal con media -0.44.
# los datos varian entre -1 y 0, con 2 datos en la parte positiva
# Los datos tienen muy poca varianza, lo cual es problematico.

#tests,confirmed,recovered y deaths tienen mucha varianza, por ello normalizamos tests
# y aplicamos logaritmo a los demas. Se manipula de forma diferenciada el tests porque es
# la unica de stas variables con datos iguales a 0
normalize <- function(x)
{
  return((x-min(x))/ (max(x)-min(x)))
}

base$tests<-normalize(base$tests)
base$confirmed<- log(base$confirmed)
base$recovered<- log(base$recovered)
base$deaths<- log(base$deaths)

pairs(base[,c(2,3,4,13)])
# La variable objetivo no parece tener correlacion con las variables explicativas principales.
pairs(base[,c(5:13)])
# tampoco parece tener correlacion significativa con las variables de stringency-index

# CREACION DE GRUPOS TEST Y TRAIN ----------------------------------------

set.seed(2020)
train<- sample(193,155)
base_train<- base[train,]
base_test <- base[-train,]
covid_train<- base_train#[,c(2,3,4,8,9,10,13,14)]# EX ANTE SE DETERMINAN ESTAS VARIABLES
covid_test<- base_test#[,c(2,3,4,8,9,10,13,14)]
covid_lineal<-base#[,c(2,3,4,8,9,10,13,14)]
view(covid_train)


# MODELO LINEAL ----------------------------------------------------------

m_lineal<- lm(twitter_sentiment ~ . , data= covid_lineal)
summary(m_lineal)

# La variables "recovered" y "international_movement_restrictions", con un intervalo de confianza del 5%,
# son la unica variable con poder explicativo significativo sobre los tweets. A pesar de ello, "Recovered"
# se relaciona de forma negativa con el sentimiento de los tweets, lo que es algo conraintuitivo.Esto implica
#que entre mas recuperados, los tweets seran mas negativos. La variable "international_movement_restrictions" se
# relaciona de forma positiva (entre menos restricciones internacionales de mobilidad los tweets son mas positivos).
# El modelo explica un 25% de la varianza de los sentimientos de los tweets. tiene un R ajustado del 20%.

# Para evaluar el modelo se crea la funcion de Error Absoluto Medio
MAE <- function(actual, predicted){
  mean(abs(actual-predicted))
}
####

lin_pred<- predict(m_lineal,data=covid_test)
MAE(lin_pred,covid_lineal$twitter_sentiment)

# El Error absoluto medio es de 0.125. Si se tiene en cuenta que la varianza es de 0,03,
# el modelo no tiene un gran poder para pronosticar los datos. De todas formas, la distancia
# promedio entre lo pronosticado y el dato real esta a menos de una desviacion estandar, lo que significa
# que lo que limita el poder explicativo del modelo son los datos.


# REGRESION TREE ---------------------------------------------------------

library(rpart)
library(rpart.plot)

arbol<- rpart(twitter_sentiment ~ ., data=covid_train)
summary(arbol)
# Las variables mas importantes del arbol son: tests, deaths, confirmed, recoverd y testing policy.
# de todas formas, ninguna de estas tiene una importancia significativa (ninguna pasa del 20%)
rpart.plot(arbol)
# El arbol hace splits en tests, deaths, testing policy, confirmed y recovered
# El arbol tiene 8 nodos finales con distribuciones muy homogeneas , lo que indica que
# el MAE puede ser grande dada la similitud entre las observaciones.
arbol_pred <- predict(arbol,covid_test)

cor(arbol_pred,covid_test$twitter_sentiment)
#La correlacion del pronostico es grande, pero no necesariamente es algo bueno.
# La poca varianza de los datos hace que esta medida no sea muy diciente de la capacidad
# predictiva del modelo. Si los datos estan muy agrupados, la estimacion sera muy cercana
# a esos valores, pero no porque logre explicar cada variable, sino porque estadisticamente
# las predicciones tendran mas oportunidad de acertar si se encuentran en ese rango.


MAE(arbol_pred,covid_test$twitter_sentiment)
# El error absoluto medio es 0.12909
# presenta sus mismos problemas y limitaciones del MAE anterior.


# CROSS VALIDATION & REGRESION TREE ------------------------------------------

library(caret)
library(pls)
library(Cubist)
set.seed(2020)
ctrl <- trainControl(method = "cv", number=10)

## REGRESION LINEAL CON CROSS-VALIDATION ##

## MODELO LINEAL CON CROSS-VALIDATION##

lm_cv <- train(twitter_sentiment ~ ., data = covid_train,metric = "RMSE", method = 'lm', trControl = ctrl)
summary(lm_cv)
# Las variables significativas del modelo lineal normal siguen siendo las mismas con cross validation
# De todas formas "international_movement_restrictions" pasa de ser significativa del 5% al 10%
lm_cv_pred <- predict(lm_cv,covid_test)
MAE(lm_cv_pred,covid_test$twitter_sentiment)
#Hacer cross-validation para la regresion lineal no presenta ninguna mejora en el modelo. El MAE
# pasa a ser de 0.133, peor que en el modelo lineal normal.


## ARBOL DE REGRESION CON CROSS-VALIDATION##

Rt_cv <- train(twitter_sentiment ~ ., data = covid_train,metric = "RMSE", method = 'rpart', trControl = ctrl)
summary(Rt_cv)
```

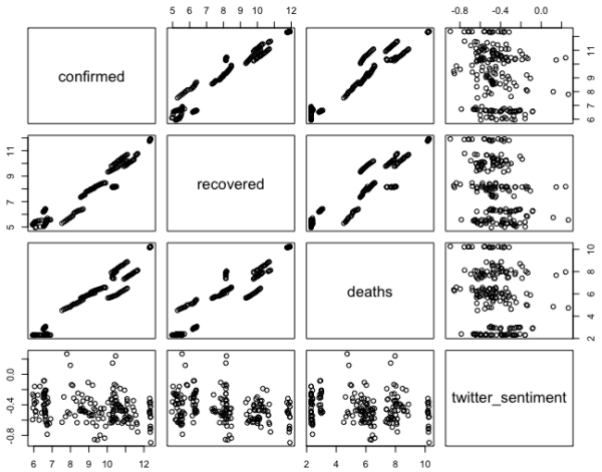
```
[448]# la importancia de las variables aumenta un poco y el orden de importancia se mantiene constante
[449]Rt_cv_pred <- predict(Rt_cv,covid_test)
[450]cor(Rt_cv_pred,covid_test$twitter_sentiment)
[451]# la correlacion cae un poco, pero ya se sabe que esta medida no es muy diciente.
[452]MAE(Rt_cv_pred,covid_test$twitter_sentiment)
[453]# Hacer cross-validation para el arbol de regresion tampoco presenta ninguna mejora. El MAE
[454]# empeora y pasa a ser 0.1367.
[455]
[456]
[457]
[458]
[459]
[460]## ARBOL DE REGRESION ALTERNATIVO CON CROSS-VALIDATION ##
[461]
[462]# se utiliza cubist para hacer el cubist model tree, la diferencia metodologica con el arbol de regresion es que
[463]#las regresiones realizadas en cada nodo están suavizadas teniendo en cuenta las predicciones de nodos anteriores.
[464]#Además, realiza regresiones entre nodos.Se incluyo este último modelo para tratar de mejorar la estimación
[465]#del árbol de regresiones dada la poca varianza de la variable explicativa.
[466]arbol_cv <- train(twitter_sentiment ~ ., data = covid_train,metric = "RMSE", method = 'cubist', trControl = ctrl)
[467]summary(arbol_cv)
[468]# el error promedio de los 10 arboles fue de 0.15 y el error relativo es 1.07
[469]# testing_policy es la variable explicativa con la que los splits crea grupos mas homogeneos.
[470]
[471]arbol_cv_pred <- predict(arbol_cv,covid_test)
[472]cor(arbol_cv_pred,covid_test$twitter_sentiment)
[473]# correlacion del 0.5437
[474]MAE(arbol_cv_pred,covid_test$twitter_sentiment)
[475]# El MAE de este modelo es 0.1223426, por lo cual, el cubist model tree es modelo
[476]# con el mejor pronostico de todas. a pesar de ello, no dista mucho de los otros modelos
[477]# y sigue teniendo los mismos problemas ya mencionados.
[478]
```

Anexo 2: Histograma de sentimiento



ANEXO 3: Diagrama de dispersión





ANEXO 4: Árbol de regresion