UAH

# COVID-19 EN REDES SOCIALES: ANÁLISIS DE SENTIMIENTO EN SUDAMÉRICA, 2020

**Máster en Business Intelligence y Data Science**


**Presentado por:**

**Dª JILLIE VANESSA CHANG KCOMT**

**Dª PILAR DENISSE VILLENA GUZMÁN**

**Dirigido por:**

**Dr. LINO GONZÁLEZ GARCÍA**


**Alcalá de Henares, a 13 de Diciembre de 2020**

## Índice

# 1. Introducción

## *1.1 Contextualización del trabajo*

La actual crisis generada por la enfermedad por coronavirus (COVID-19)[1] reportada en diciembre de 2019[2] ha obligado a los gobiernos de todos los países del mundo a tomar medidas para controlar su propagación. Diversas acciones, como la restricción en la movilidad de la población, han generado un impacto no solo en la economía, sino también en el estado ánimo de las personas. En los últimos meses, a partir de información disponible en internet, se han desarrollado algunos estudios que utilizan la técnica del análisis de sentimiento (AS) para observar las sensaciones generadas en la población respecto al COVID-19. A continuación, se realiza una breve descripción de la literatura encontrada.

En Dubey, A. D. (2020) se aplica la técnica de AS con el objetivo de observar cómo los ciudadanos de 12 países europeos lidian con la situación provocada por el COVID-19 a partir de los *tweets* emitidos entre el 11 y 31 de marzo de 2020. A través de la librería Syuzhet en R basada en la utilización de un diccionario, se clasifican los *tweets* en sentimientos básicos (negativos y positivos) y en 8 emociones (miedo, alegría, previsión, ira, asco, tristeza, sorpresa y confianza). Se concluye que la mayoría de la población tenía sentimientos positivos y esperanzadores, aunque con instantes de tristeza y preocupación. Además, se señala que cuatro de esos países mostraban signos de desconfianza y molestia en mayor proporción que los otros. Dicho análisis brinda una aproximación de cuál era la percepción de la población en las primeras semanas del confinamiento.

Asimismo, Samuel, J. et ál. (2020) analiza los *tweets* emitidos por la población de Estados Unidos con respecto al COVID-19 para el periodo comprendido entre febrero y marzo de 2020. A través de modelos de clasificación de Naïve Bayes y regresión logística, se identifica un progresivo aumento de sentimientos negativos y miedo a medida que los niveles de contagio aumentaban. Además, de realizar el análisis de los sentimientos de la población, se analiza la efectividad de ambos modelos en función de la longitud de los *tweets*. Así, se demuestra que la precisión de clasificación mejora en los *tweets* cortos, siendo el de Naïve Bayes es superior al de regresión logística.

Similarmente, Barkur, G. et ál. (2020) utiliza la misma librería en R mencionada anteriormente para evaluar el impacto del anuncio del confinamiento en la salud mental de la población de India a partir de los tweets publicados entre el 25 y el 28 de marzo de 2020. Se concluye que la población en su mayoría estuvo de acuerdo con las medidas tomadas por el gobierno. Sin embargo, se encontraron sentimientos de molestia y preocupación ante la demora del establecimiento de la cuarentena. Este estudio recomienda realizar un análisis antes y después de la cuarentena de los *tweets* emitidos para analizar los cambios de la población al comienzo y al final del confinamiento.

---

[1] Para mayor información sobre la enfermedad por COVID-19, visitar el portal web de la Organización Mundial de la Salud (OMS) disponible en: https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019, actualizado al 12 de octubre de 2020.

[2] Según la OMS el brote de enfermedad por COVID-19 fue notificado por primera vez en Wuhan (China) el 31 de diciembre de 2019. Información disponible en: https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019?gclid=Cj0KCQjw6PD3BRDPARIsAN8pHuH6P-lEKcXe8N8J04OzIre6NpQblg9biDtlBvWOwmUmF6Uz_SYz64QaAoxIEALw_wcB, actualizado al 09 de julio de 2020.

Por último, Sharma K et ál. (2020) realiza un análisis de la malinformación en Twitter en 20 países y todos los estados de Estados Unidos del 1 de marzo al 5 de junio de 2020. En particular, se destaca la información incorrecta sobre el acceso y control de precios de las vacunas investigadas contra el COVID-19, así como afirmaciones falsas sobre figuras políticas que intentaban maliciosamente empeorar la crisis. Por otro lado, a través del modelado de topics se encuentra que al principio de marzo la discusión estuvo centrada en los brotes del COVID-19, restricciones de viaje, medidas de prevención, síntomas y el presupuesto y respuesta de los gobiernos. Además, se encuentra que las impresiones de las personas sobre las medidas de relacionadas con el teletrabajo y distanciamiento social fueron positivas.

Todas las investigaciones mencionadas previamente analizaron el estado de ánimo de la población respecto al COVID-19 utilizando información de Twitter en países de Europa, Estados Unidos y la India. Sin embargo, hasta la fecha no se encontró ningún estudio que investigue cuál es la situación en Latinoamérica. Siendo esta una crisis provocada a nivel mundial no vista en los últimos 100 años, se cree necesario saber cuál es el sentir de la población ante el COVID19 y sobre las diversas medidas que toman las autoridades para contenerla en Latinoamérica. **Este trabajo de investigación busca realizar dicho análisis a través de un enfoque de data science.**

*1.2 Fundamentos teóricos*

**Definición Análisis de Sentimiento**

El Análisis de Sentimiento, también conocido como minería de opinión, se refiere al uso del Procesamiento de Lenguaje Natural (PLN) para determinar automáticamente el sentimiento que el autor está expresando en un extracto de texto. Este sentimiento puede ser clasificado de manera binaria (positivo/negativo), terciaria (positivo, negativo, neutro), o con múltiples categorías (neutro, feliz, triste, cólera, odio) (Zhang, M., Ng, J.).

**Aplicaciones**

El análisis de sentimiento en el ámbito de las redes sociales, es una técnica que presenta muchos beneficios ya que nos permite conocer las opiniones sobre determinados temas y poder así tomar mejores decisiones. Entre sus múltiples aplicaciones se encuentran las siguientes:

- Análisis del contenido en internet sobre un tema social/político específico.
- Mejorar la atención al cliente a través de mejorar las recomendaciones.
- Investigación de mercado.
- Permite medir el impacto de las acciones en redes sociales.
- Manejo de la reputación de marcas, etc.

**Procesamiento del texto**

Para poder elaborar un modelo que permita estimar el sentimiento de un texto, se debe primero realizar un conjunto de procedimientos que

- Preprocesamiento de los datos.- antes de a aplicar cualquier algoritmo de clasificación, es necesario primero realizar algunas transformaciones a los datos para mejorar el resultado final. Así, el procesamiento cubre las siguientes acciones:

- Normalización: se refiere al conjunto de acciones que transforman un texto a una forma más conveniente para su tratamiento posterior. Estas acciones generalmente incluyen: convertir todas las letras a minúsculas, eliminar hashtags, signos de puntuación, links de páginas webs, etc.
- Stopwords: las "stopwords" son palabras que no ofrecen información adicional al procesamiento automático (por ejemplo: de, por, tu, etc.). En algunas ocasiones se considera su eliminación como una buena práctica. Sin embargo, se debe tener en cuenta que remover alguna stopword importante, podría quitar el verdadero sentido del texto.
- Lematización: permite agrupar diferentes formas de una misma palabra en una sola que es identificada por su lema. Este proceso toma en cuenta el análisis morfológico de las palabras. Para realizar esto es necesario tener diccionarios que permitan asociar las palabras a su "lema".
- Stemming: es el proceso mediante el cual se reduce una palabra a su raíz morfológica ("Stem").

- Tokenización.- proceso que consiste en dividir un texto en unidades lingüísticas que constituyen un dato en sí misma denominadas tokens. Generalmente estos tokens están formados por una palabra, sin embargo, también pueden estar compuestos por hasta N palabras (N-grams)

- Modelado de espacio de vectores (o codificación del texto del documento).- permite obtener una representación numérica de cada documento para que pueda ser procesada por los algoritmos de aprendizaje automático.

  - Bolsa de palabras.- este método consiste en contar, para un documento, el número de veces que aparece cada token. De esta manera se obtiene una representación matricial del documento. Se debe tener en cuenta que para este caso
  - Term Frequecy- Inverse Document Frequecy (TF-IDF).- este método calcula la importancia de una palabra, la cual se incrementa proporcionalmente al número de veces que esa palabra aparezca en un documento.

**Clasificación del sentimiento del texto**

Para la realización del Análisis de Sentimientos, existen diferentes métodos que pueden ser agrupados en:

- Clasificación de sentimientos usando aprendizaje supervisado.- se refiere al uso de técnicas de aprendizaje automático (machine learning) para clasificar un sentimiento (generalmente en positivo/negativo). Se necesita contar con una base de entrenamiento previamente etiquetada con el sentimiento. En algunas ocasiones se cuenta con alguna variable que pueda definir el sentimiento asociado al texto, como las reseñas en internet de películas o libros que tienen puntuaciones asignadas por los mismos usuarios que emiten los comentarios. Así, por ejemplo, si estas puntuaciones se dan en forma de cantidad de estrellas, se puede determinar que los comentarios que son valorados con 4 o 5 estrellas son positivos; mientras que los que cuentan con 1 o 2, son negativos.

  Para aplicar las técnicas de análisis de sentimiento a los datos de Twitter, se ha optado en algunos casos por el etiquetado manual de los tweets. Esto sin embargo demanda gran cantidad de tiempo por lo que muchos de los trabajos aplicados a Twitter,

etiquetan los sentimientos usando emoticones (denominado método ruidoso o supervisión a distancia).

Con la base de entrenamiento debidamente etiquetada, se procede a emplear cualquiera de las técnicas de clasificación dentro del aprendizaje automático, como Naive Bayes, Máxima Entropía, Máquinas de vectores de soporte, etc., las cuáles se describirán brevemente a continuación:

- Regresión logística.- es un algoritmo de clasificación considerado como una extensión del modelo de regresión lineal. Es ampliamente usado en el sector bancario a fin de crear modelos scoring para otorgar préstamos

- Naive Bayes: Es un algoritmo de clasificación basado en el teorema de Bayes. Se asume que las variables explicativas son independientes entre sí

- K vecinos más cercanos.- algoritmo de clasificación basado en la determinación de K grupos dentro de un conjunto de datos.

- Árboles de decisiones.- son modelos que se basan en la partición recursiva de un conjunto de datos, resultando grupos cada vez más homogéneos. Este modelo puede usar como variables predictoras de tipo categóricas y continuas, y tiene como principal ventaja que es una técnica robusta antes ante outliers.

- Vector soporte clasificación lineal.- el objetivo de este algoritmo es ajustar los datos, y presentar el mejor hiperplano que divida o categorice estos datos. Se puede predecir los datos de acuerdo al lado del hiperplano en que se encuentren.

- Análisis de sentimiento español.- es una librería de Python que utiliza redes neuronales para predecir el sentimiento de un texto en español. Este modelo tuvo como base de entrenamiento 800 000 opiniones en línea de los usuarios de páginas web como Decathlon, TripAdvisor, filmaffinity, eltenedor y ebay. Para etiquetar la reseña se usó el puntaje que le asignó cada usuario. Este modelo obtuvo como medida de precisión un valor de 88% (en la base test) y retorna un número entre 0 y 1, que indica la probabilidad de que el texto sea “positivo”. Esto quiere decir que valores cercanos a 1 indican que es positivo y valores cercanos a cero señalan que es negativo. Los valores intermedios se señalan como neutros.

- Clasificación de sentimientos usando aprendizaje no supervisado.- se refiere a clasificar el sentimiento de un texto a partir de la orientación semántica de las palabras o frases que lo conforman. Entre los principales enfoques de este tipo de aprendizaje se encuentran:

  - Enfoque basado en diccionarios.- el término “diccionarios” se refiere a un listado de palabras o conjunto de palabras que se asocian a un determinado sentimiento o emoción (positivo-negativo, ira, etc.). Consiste en detectar en el texto coincidencias con las palabras en los diccionarios y predecir el sentimiento en base al número de coincidencias encontradas. El problema en este enfoque es que no tiene en cuenta el contexto en el que se encuentra la palabra.
  - Enfoque basado en corpus.- a diferencia del enfoque basado en diccionario, este método se enfoca en las relaciones lingüísticas. Se usa un corpus subyacente como un inventario de datos lingüísticos, de la cual se extrae material para el

conocimiento intuitivo que permita cuantificar fenómenos lingüísticos (Storjohann, P. 2005).

**Evaluación de resultados de los modelos de clasificación supervisada.**

Existen diversas medidas que nos permiten evaluar el rendimiento de un modelo, entre los cuales tenemos:

- Matriz de clasificación.- es una tabla que nos muestra los valores reales y los pronosticados y presenta los siguientes datos:

  Verdaderos Positivos (VP).- Cantidad de registros positivos clasificados como positivos.
  Falsos Positivos (FP).- Cantidad de registros negativos clasificados como positivo.
  Falsos Negativos (FN).- Cantidad de registros positivos clasificados como negativos.
  Verdaderos Negativos (VN).- Cantidad de registros negativos clasificados como negativos.

  De estos valores se pueden obtener diferentes medidas que nos ayudan a evaluar un modelo.

- Exactitud (accuracy).- mide el porcentaje de casos en los que el modelo ha acertado. Si bien es un buen indicador general del modelo, debe evaluarse siempre junto con otras medidas como la precisión

$$Exactitud = \frac{VP\ + VN}{VP\ + VN + FP + FN}$$

- Exhaustividad (Recall)- porcentaje de positivos clasificados correctamente como positivos

$$Exhaustividad = \frac{VP}{VP + FN}$$

- Precisión (precisión).- porcentaje de los clasificados como positivos, que realmente son positivos

$$Precision = \frac{VP}{VP + FP}$$

- Valor F (F1-score).- este valor combina las dos medidas anteriores, exhaustividad y precisión.

$$Valor\ F = \ 2\frac{Precision * Exhaustividad}{Precision + Exhaustividad}$$

## 2. Objetivos

Junto con la crisis sanitaria que ha producido el COVID19, también se ha manifestado una crisis social de miedo masivo y fenómenos de pánico que han afectado a la población, a veces por información incompleta o inexacta. Esto genera que sea importante medir el sentimiento de la población de modo que los Gobiernos puedan transmitir mensajes

apropiados y oportunos a sus ciudadanos. Asimismo, es importante que el Gobierno escuche las necesidades y opiniones sobre las políticas implementadas a modo de extraer retroalimentación que busque la mejora continua.
En tal sentido, este trabajo de investigación tiene los siguientes objetivos:

i. Medir el sentimiento de los ciudadanos de algunos países de Sudamérica con respecto a las medidas tomadas por sus Gobiernos con relación al COVID19. Con esta información se realizará un análisis comparativo de los países.

ii. Identificar cuáles han sido los temas de interés durante las medidas adoptadas ante el COVID19.

Este trabajo es una de las primeras investigaciones en Latinoamérica que estudia el efecto del COVID19 sobre las emociones u opiniones de la población a partir del análisis de sentimientos. A la fecha, la literatura sobre el análisis de sentimientos en los países latinoamericanos no ha sido muy extensa y los estudios sobre los efectos del COVID19 son relativamente nuevos. En ese contexto, este trabajo busca ampliar la literatura, brindándoles a los gobiernos latinoamericanos una herramienta para que tengan una mejor comprensión de cómo se sienten sus ciudadanos y puedan recibir una retroalimentación de sus medidas.

## 3. Plan de Trabajo/Material y Métodos:

Para cumplir con los objetivos planteados, se utilizará la siguiente metodología:

i. Selección de red social: Se elige la red social que se utilizará como fuente de información. Se decide utilizar Twitter en lugar de Facebook, pues cuenta con una API de donde se puede extraer la información, lo cual puede ser más accesible. Además, en Twitter, se obtienen directamente las publicaciones de los usuarios con relación a distintos temas o hashtags, lo cual puede ser más transparente para recolectar la información deseada. En contraste, extraer la información de Facebook presenta más restricciones y la información no es necesariamente por usuarios (hay fanpages por ejemplo).

ii. Selección de países de Latinoamérica: se identifican los países que tengan mayor penetración en Twitter para los cuales se realizará el análisis. Se tendrá en cuenta la cantidad total de usuarios totales. No se usará información de Brasil, Surinam, Guyana ni Trinidad y Tobago por cuestiones de idioma.

En el Cuadro 1, se puede apreciar la lista de los países de Sudamérica y la cantidad de usuarios de Twitter y otras redes sociales. Teniendo en cuenta este porcentaje y la cantidad de usuarios, se seleccionan los siguientes países: Argentina, Chile, Colombia, Perú, Uruguay y Ecuador

| País | Población | Cant. Usuarios Twitter | | Cant. Usuarios Social Media | |
|---|---|---|---|---|---|
| | Cant. (mill) | Cant. (mill) | Part. | Cant. (mill) | Part. |
| Argentina | 44.9 | 4.98 | 11% | 34.0 | 76% |
| Chile | 19.0 | 2.47 | 13% | 15.0 | 79% |
| Colombia | 50.3 | 3.20 | 6% | 35.0 | 70% |
| Perú | 32.5 | 1.24 | 4% | 24.0 | 74% |
| Uruguay | 3.5 | 0.82 | 24% | 2.7 | 78% |
| Ecuador | 17.4 | 1.11 | 6% | 12.0 | 69% |
| Paraguay | 7.0 | 0.48 | 7% | 4.0 | 57% |
| Venezuela | 28.5 | 1.29 | 5% | 12.0 | 42% |
| Brasil | 211.0 | 12.15 | 6% | 144.0 | 68% |
| Bolivia | 11.5 | 0.38 | 3% | 7.5 | 65% |
| Trinidad y Tobago | 1.4 | Sin inf | Sin inf | Sin inf | Sin inf |
| Surinam | 0.6 | Sin inf | Sin inf | Sin inf | Sin inf |
| Guyana | 0.8 | Sin inf | Sin inf | Sin inf | Sin inf |
| **Total** | | | | | |

***Cuadro 1.*** Países de Sudamérica y su participación en redes sociales

iii. Identificación de medidas implementadas en cada país: para cada país se identificarán las medidas (se revisa si se utilizaron hashtags) y las fechas en las que se implementaron. Se clasificará a los países según el grado de restricción a la movilidad que cada Gobierno impuso sobre la población: más restrictivo (toque de queda, multas o posibilidad de encarcelamiento por no respetar el confinamiento obligatorio) a menos restrictivo (libre movilidad).

iv. Extracción y preparación de datos de Twitter de países seleccionados. Se extraen los tweets que contengan palabras claves: covid, cuarentena, pandemia y coronavirus. Se utiliza la librería GetOldTweets3 de Python para extraer la información de los países seleccionados para las fechas desde 1 enero hasta 31 agosto 2020. Se aplican procesos de normalización, stemming, tokenización, lematización, entre otros que se consideren necesarios. Se realiza un análisis de palabras y de frases (top unigramas, bigramas, etc.). Asimismo, se estudia el resto de las variables asociadas a los mensajes, como la ubicación, dispositivo utilizado, etc. Se considerarán los "emoticones" para categorizar los tweets y desarrollar un modelo de predicción.

v. Modelamiento a partir del Sentiment Analysis. Se analiza la polaridad de los tweets a partir de dos enfoques: enfoque basado en Machine learning y enfoque basado en léxico (uso de diccionario). De esta forma, se comparan los resultados obtenidos. Una alternativa para clasificar los Tweets es a través del paquete Sentiment-Spanish de Python. El primer paquete asigna un puntaje de 0 a 1, lo que permite clasificar a los tweets en positivo, neutro y negativo. Además, se explorará si es posible aplicar el modelo de representación de lenguaje BERT para clasificar la polaridad de los mensajes.

## 4. Desarrollo del Trabajo/Resultados y discusión[3]:

### *4.1 Datos*

#### 4.1.1 Extracción de datos

La información empleada en este trabajo se obtuvo a partir de la red social Twitter[4] para el periodo comprendido desde el 1 de enero hasta el 31 de agosto de 2020. A través de las palabras claves "covid", "pandemia", "coronavirus" y "cuarentena", se extrajeron 402 229 tweets en español relacionados con la pandemia en las capitales de los 6 países seleccionados. Como se muestra en el Gráfico 1, la cantidad de tweets se empieza a incrementar en cada ciudad entre finales de febrero e inicios de marzo cuando se identifican los primeros casos de coronavirus en Sudamérica.

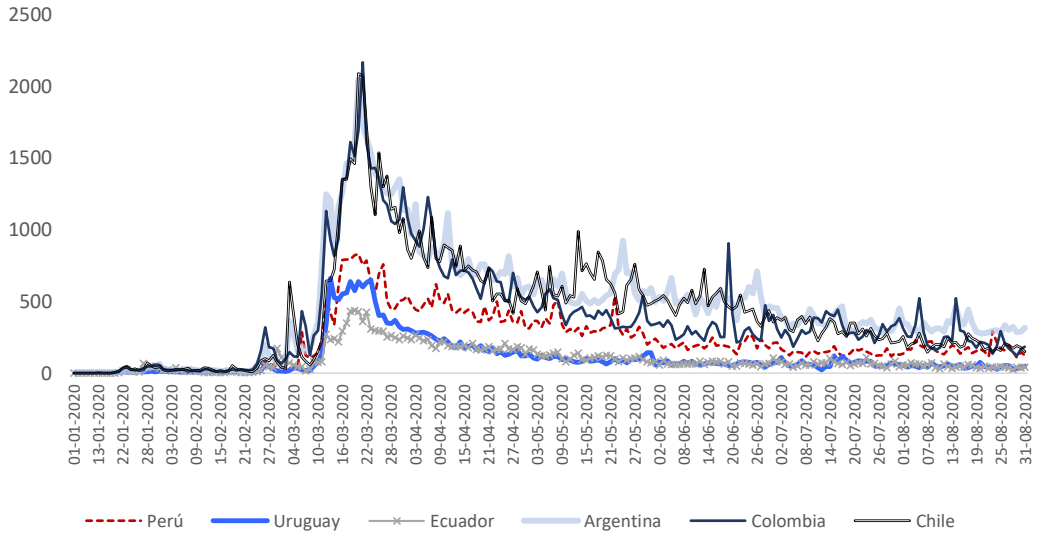


***Gráfico 1***. Evolución de la cantidad de tweets por país, 01-01-2020 al 31-08-2020.

Para extraer dicha información, se utilizaron dos librerías en Python 3: GetOldTweets3[5] y Tweepy[6]. Por un lado, se utilizó GetOldTweets3 para obtener la información histórica de los Tweets con las palabras claves mencionadas en las ciudades seleccionadas. A través de los parámetros "geocode" y "distances" de dicha librería, se ubicaron las coordenadas de una zona y se descargargaron los tweets contenidos en un radio determinado[7]. Por otro lado, se utilizó la librería Tweepy para complementar la información descargada, pues GetOldTweet3 solo permite descargar la información histórica de manera limitada. En el Cuadro 2, se muestran todas las variables descargadas a partir de ambas librerías.[8]

---

[3] La base de datos y el código utilizado para el desarrollo de este trabajo se encuentra disponible en: https://github.com/TFMChangVillena/AnalysisSentimentCovidSudamerica

[4] Disponible en: https://twitter.com, actualizado al 01 de octubre de 2020.

[5] Disponible en: https://pypi.org/project/GetOldTweets3/, Actualizado al 10 de octubre de 2020.

[6] Disponible en: https://pypi.org/project/Tweepy/, Actualizado al 10 de octubre de 2020

[7] En el Anexo 1 se muestra en detalle el proceso de descarga de la información.

[8] Inicialmente se trató de descargar la información a través de la versión estándar del API de Twitter; sin embargo, esta tiene algunas limitaciones y solo permite descargar los tweets de forma gratuita los últimos 7 días (con un máximo de 5000 tweets). Para acceder a los tweets de mayor antigüedad a través de la API de Twitter, se tendrían que utilizar los servicios Premium o Enterprise, que permiten acceder a tweets con 30 o más días de antigüedad, pero se tendría que pagar por el servicio. En este trabajo, se decidió utilizar la librería GetOldTweets3 de Python 3, que es una alternativa a la API de Twitter y permite descargar el historial de tweets de forma gratuita especificando el nombre de usuario, popularidad, fecha, hashtags, locación a través de un radio en una zona geográfica, entre otros. Aunque, esta librería no llega a descargar todos los campos que pueden existir en un tweet, como por ejemplo la fuente desde donde se emitió el tweet, sí logra descargar los campos suficientes para poder elaborar este trabajo de investigación.

| Variable | Descripción | Librería |
|---|---|---|
| Tweet_Id | Identificador del Tweet | |
| Tweet_User_Id | Identificador del Usuario | |
| Tweet_User | Nombre del usuario | |
| Text | Texto | |
| Retweets | Cantidad de retweets | GetOldTweets3 |
| Favorites | Cantidad de favoritos | |
| Replies | Cantidad de respuestas | |
| Datetime | Fecha del tweet | |
| hashtags | Etiquetas | |
| Pais | Capital del país identificado a partir del "geocode" y "distances" | |
| Tweet_Source | Fuente de origen de tweet | Tweepy |
| lang | Idioma | |

***Cuadro 2***. Variables que conforman la base de datos del estudio según la librería utilizada para su descarga

Es importante mencionar que la información presentada tiene en cuenta las siguientes consideraciones: i) se filtraron los tweets en idiomas distintos al español, ii) no se incluyeron a los tweets cuyos usuarios registraron un perfil privado, iii) se eliminaron los tweets duplicados, iv) se eliminaron los tweets de usuarios que publicaban información repetida frecuentemente en distintos días (por ejemplo usuarios que publicitaban anuncios de productos en distintos días) y v) no se consideraron retweets dentro de la base de datos.

### 4.1.2 Análisis descriptivo de datos

| País | Cant. Tweets | | Cant. Users | | Tweets promedio | Principales Hashtags |
|---|---|---|---|---|---|---|
| | Cant. | Part. | Cant. | Part. | | |
| Argentina | 109 845 | 27% | 15 528 | 25% | 7,1 | coronavirus, cuarentena, quedateencasa, yomequedoencasa, covid19, argentina, pandemia,coronavirusargentina,buenos aires |
| Chile | 99 558 | 25% | 15 166 | 24% | 6,6 | coronavirus, covid_19, cuarentena, quedateencasa, chile,pandemia, cuarentenatotal, covid19Chile, |
| Colombia | 92 210 | 23% | 16 342 | 26% | 5,6 | coronavirus, covid_19, cuarentena, colombia, quedateencasa, yomequedoencasa, bogota, pandemia, coronavirusencolombia |
| Perú | 53 100 | 13% | 8 002 | 13% | 6,6 | cuarentena, coronavirus, yomequedoencasa, covid_19, quedateencasa, peru, lima, pandemia, coronavirusperu |
| Uruguay | 25 144 | 6% | 4 348 | 7% | 5,8 | coronavirus, covid_19, cuarentena, quedateencasa, uruguay, coronavirusenuruguay, yomequedoencasa, |
| Ecuador | 22 372 | 6% | 3 654 | 6% | 6,1 | covid_19, coronavirus, ecuador, cuarentena, quedateencasa, quito, covid_19ec, urgente, yomequedoencasa |
| **Total** | **402 229** | **100%** | **63 040** | **100%** | **6,4** | |

***Cuadro 3.*** Cantidad de tweets, cantidad de usuarios y principales hashtags por país

En el Cuadro 3 se muestra el análisis de las variables descargadas por país, en donde se encontró que Argentina fue el país que registró mayor cantidad de tweets (27%), seguido de Chile (25%), Colombia (23%), Perú (13%) , Uruguay (6%) y Ecuador (6,%). En todos los países, cada usuario publicó entre 6 y 7 tweets en promedio durante el periodo analizado. Asimismo, los hashtags más utilizados estuvieron relacionados con las palabras claves que se utilizaron para descargar los datos. Sin embargo, también aparecieron otros hashtags como #quedateencasa y los relacionados a la ubicación: #colombia, #argentina, #lima, entre otros.

Con relación a las fuentes desde donde se emitieron los tweets, los Android fueron la principal fuente en todos los países, seguido del iPhone, Instgram e iPad. Como se registra en el Cuadro 4, en todos los países entre el 48% y 68% de tweets provino del Android.

| Fuente | Perú | | Uruguay | | Ecuador | | Argentina | | Colombia | | Chile | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Cant. | Part. | Cant. | Part. | Cant. | Part. | Cant. | Part. | Cant. | Part. | Cant. | Part. | Cant. | Part. |
| Android | 35 980 | 68% | 16 336 | 65% | 10 807 | 48% | 72 646 | 66% | 58 449 | 63% | 60 426 | 61% | 254 644 | 63% |
| iPhone | 12 840 | 24% | 8 030 | 32% | 9 877 | 44% | 27 107 | 25% | 28 821 | 31% | 30 206 | 30% | 116 881 | 29% |
| Instagram | 3 865 | 7% | 633 | 3% | 1 336 | 6% | 9 424 | 9% | 4 450 | 5% | 8 198 | 8% | 27 906 | 7% |
| iPad | 330 | 1% | 99 | 0% | 172 | 1% | 297 | 0% | 266 | 0% | 374 | 0% | 1 538 | 0% |
| Otros | 85 | 0% | 46 | 0% | 180 | 1% | 371 | 0% | 224 | 0% | 354 | 0% | 1 260 | 0% |
| **Total** | **53 100** | **100%** | **25 144** | **100%** | **22 372** | **100%** | **109 845** | **100%** | **92 210** | **100%** | **99 558** | **100%** | **402 229** | **100%** |

***Cuadro 4.*** Tweets según principales fuentes de emisión por país (cantidad y participación en %)

### 4.1.3 Analítica de textos


***Gráfico 2.*** Wordclouds por país

En el Gráfico 2 se identificaron las palabras que aparecieron con mayor frecuencia en los tweets a través de un gráfico de nube de palabras (“wordclouds”). En todos los países destacan las palabras “cuarentena”, “covid”, “pandemia” y “coronavirus”, que fueron las palabras claves con las que se descargó la información. Sin embargo, también se observan otras palabras relacionadas al gobierno (presidente, el apellido del presidente peruano Vizcarra, entre otras), la zona geográfica de los usuarios (la ciudad o país registrados por los usuarios), entre otras.

Además de realizar un gráfico de nube de palabras, se realizó el análisis de las cadenas de texto en pares (bigramas) y en grupos de tres palabras (trigramas). En el Gráfico 3, se observa que en primer lugar la palabra “cuarentena” es la más utilizada. Luego, están las palabras “covid”, “coronavirus” y “pandemia”, que casi tienen la misma cantidad de apariciones. Tras esas palabras, están “gente”, “ahora”, “días”, “salud”, entre otras. Con relación a los bigramas, los pares más populares fueron “día, cuarentena” y “cuarentena, total”. Probablemente pueda hacer mención a la cantidad de días de cuarentena transcurridos que las personas afrontaron tras las medidas obligatorias de los gobiernos. También destaca “buenos, aires”, que hace referencia a la ciudad de Buenos Aires. Lo mismo ocurre en los trigramas, en donde varios de los grupos están relacionados con dicha ciudad. Estos grupos de palabras deberían ser considerados como un solo token en el análisis posterior.

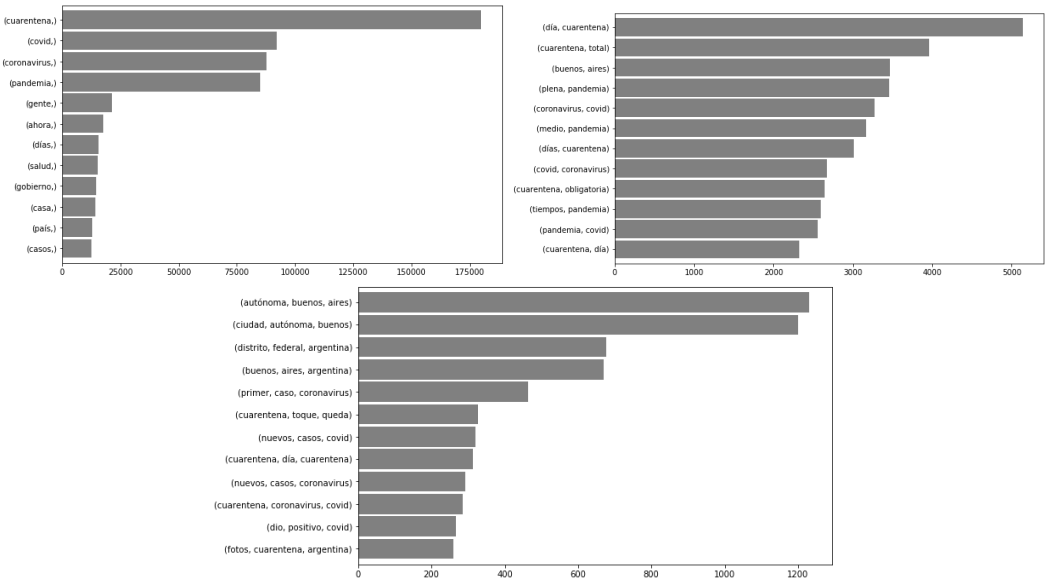


***Gráfico 3***. N-gramas

### 4.1.4 Procesamiento de los datos

Se creó un corpus de textos a través del agrupamiento del conjunto de 402 229 tweets, compuesto por 65 692 728 palabras. A dicho corpus se le aplicaron los siguientes procesos para el tratamiento posterior de los datos y análisis mediante métodos del Proceso del Lenguaje Natural: normalización, eliminación de stopwords, stemming y tokenización. En el Cuadro 5 se detallan las acciones realizadas para cada proceso.

| Proceso | Acciones |
|---|---|
| Normalización | Se aplicó el siguiente conjunto de tareas:<br>- Cambiar las mayúsculas por minúsculas<br>- Reemplazar enlaces we que contengan "http" o "www" por "link"<br>- Reemplazar los nombres de usuarios por "usuario"<br>- Reemplazar las palabras que empiecen con # por "usuario"<br>- Eliminar los emoticones<br>- Elimina desde el "at" hasta el final del tweet (hace referencia al lugar)<br>- Normaliza las jergas como "d" por "de" y "fav" por "favorito"<br>- Cambia las lisuras por la palabra "insultos"<br>- Se eliminan las tildes<br>- Normaliza risas, como jajajaja, jaaa, jejeje, jiji, jaja por "risas"<br>- Eliminar signos de puntuación<br>- Eliminar números<br>- Borrar espacios excesivos y espacios al final |
| Stemming | Se utiliza SnowballStemmer de la librería NLTK |
| Eliminación de stopwords | - Se consideraron las stopwords de la libería NLTK en español |
| Tokenización | Se utilizó TweetTokenizer de la libreía NLTK |

***Cuadro 5.*** Procesos y acciones realizadas para procesamiento de textos

Se debe precisar que se probará si los procesos de stemming y eliminación de stopwords mejoran los resultados. Por lo tanto, se realizarán las estimaciones excluyendo e incorporando dichos procesos.

### 4.1.5 Extracción y ponderación de características

Luego de limpiar y tokenizar los tweets, se procedió a extraer y ponderar las características de los mensajes para representarlos de forma matricial y luego ser capaz de procesaros con algoritmos de aprendizaje automático. En particular, se trabajó con las palabras a nivel de unigramas y se utilizaron dos modelos de la librería de Scitik-learn basados en espacios de vectores: modelo de bolsa de palabras (BoW) y modelo TF-IDF. En el caso del modelo BoW, se empleó la ponderación binaria: dada una lista con todas las características de los tweets, se coloca 1 a las características que forma parte de mismo y 0 en otro caso. Por otro lado, en el caso del modelo TF-IDF, se recogió la frecuencia de los términos en cada tweet, así como la importancia del término en el conjunto de tweets (el valor del TF-IDF se incrementa proporcionalmente a la cantidad de veces que una palabra aparece en el tweet, pero es compensado por la frecuencia de la palabra en la colección de tweets).

## *4.2 Modelos de clasificación de sentimientos*

Para seleccionar un modelo que permita predecir el sentimiento de un tweet, se dividió la muestra en tres partes: muestra de entrenamiento (se utilizaron los datos para obtener los parámetros), muestra de validación (datos utilizados para la selección del mejor modelo) y muestra de prueba (datos utilizados para calcular el error de predicción del modelo seleccionado). A partir de estas muestras, se utilizaron métodos supervisados y no supervisados para clasificar la polaridad de los tweets en positivos o negativos.

En el caso de los métodos supervisados (los tweets de entrada deben ser etiquetados previamente), se estimaron los siguientes modelos: Regresión logística, Vector de soporte

de clasificación lineal, Naive Bayes, k-vecinos más cercanos y árboles de decisiones[9]. Para ello, primero se clasificaron los tweets que presentaban emoticones: los tweets que contenían símbolos como 😍 😊, 😘, entre otros fueron clasificados como mensajes "positivos"; mientras que los que contenían símbolos como 😠,🤬,🤮, entre otros fueron clasificados como "negativos". Luego, a partir de estos tweets previamente etiquetados, se estimaron los modelos con la muestra de entrenamiento y se realizó una validación cruzada (cross-validation) de los mismos.

Adicionalmente, se empleó la librería de Python Sentiment-spanish[10], con la cual se predice la probabilidad de que un texto sea "positivo" a través del uso de redes neuronales sobre la información de las opiniones de usuarios brindada por eltenedor, Decathlon, TripAdvisor, filmaffinity y ebay. Los valores cercanos a 0 significan que existen bajas probabilidades de que el texto sea negativo y valores cercanos a 1 significan que existe altas probabilidades de que el texto sea positivo. Se considera que los valores cercanos a 0.5 corresponden a sentimientos neutrales.

Por otro lado, en el caso de los métodos no supervisados (los tweets de entrada no necesitan ser etiquetados previamente, se utilizó el método basado en redes neuronales propuesto por Google: BERT (Bidirectional Encoder Representations Transformer), el cual es un sistema no supervisado bidireccional profundo para el preentrenamiento de textos sin etiquetar. Utilizando los textos de Wikipedia y BookCorpus y a través del uso de transformadores (un tipo de redes de codificación-decodificación), se preentrena un modelo y se consigue realizar representaciones contextualizadas de las palabras y oraciones que permite entender el lenguaje de forma general (BERT enmascara el 15% de las palabras y realiza cambios aleatorios en ciertas palabras, prediciéndolas a partir del contexto). Luego que se obtiene el modelo preentrenado, se agrega una capa de salida adicional para realizar tareas específicas.

En este trabajo se utilizó la versión del modelo BERT "BERT-Base Multilingual Cased", que contiene 104 lenguas, 12 capas de entrada, 768 capas escondidas, 12 capas de salida y 110 parámetros. A este modelo se le agregó una capa adicional a partir de los tweets previamente etiquetados con los emoticones y se ajustó el modelo para realizar la tarea de clasificación de polaridad de tweets. Se pudo aplicar el BERT directamente, pero se prefirió afinarlo con esta información, convirtiéndolo en un modelo semi-supervisado[11].

En el Cuadro 6, se muestran los resultados de las estimaciones de los modelos previamente señalados. Para todos los modelos primero se normalizaron los tweets y luego se consideraron distintos escenarios, como la inclusión o no de los procesos de Stemming y eliminación de stopwords. Además, se compararon los resultados utilizando las técnicas de extracción de características de BoW e IT-IDF. Como se observa en el cuadro, los mejores resultados fueron obtenidos a partir del modelo de Regresión logística con los mensajes normalizados, sin eliminar stopwords y aplicando el proceso de stemming. Este modelo obtuvo un AUC de 0.79 y un F1W de 0.72, superiores a los estimados por los otros modelos supervisados y los enfoques basados en BERT y uso de diccionarios.

[9] Para la estimación de estos modelos se utilizó un código adaptado de Sobrino, J (2018). ANÁLISIS DE SENTIMIENTOS EN TWITTER.
[10] Disponible en: https://pypi.org/project/sentiment-analysis-spanish/, actualizado al 10 de octubre de 2020
[11] Para la estimación del modelo BERT se utilizó un código adaptado de Akoksal, A. (2020), BERT Sentiment Analysis Turkish.

| Stemming | Elimina Stopwords | Normalización | Extracción de caracterítica s | Diccionario | | Bert | | Regresión Logística | | Lineal SVC | | Multinomial Nbayes | | Kneighbors Classifier | | Árboles de decisión | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | AUC | F1W | AUC | F1W | AUC | F1W | AUC | F1W | AUC | F1W | AUC | F1W | AUC | F1W |
| No | Sí | Sí | BOW | - | - | - | - | 0.76 | 0.70 | 0.72 | 0.67 | 0.78 | 0.71 | 0.65 | 0.57 | 0.64 | 0.63 |
| No | No | Sí | BOW | 0.61 | 0.46 | 0.69 | 0.67 | 0.77 | 0.70 | 0.73 | 0.67 | 0.78 | 0.71 | 0.66 | 0.50 | 0.63 | 0.63 |
| Sí | Sí | Sí | BOW | - | - | - | - | 0.77 | 0.70 | 0.73 | 0.67 | 0.78 | 0.71 | 0.66 | 0.56 | 0.63 | 0.63 |
| Sí | No | Sí | BOW | 0.58 | 0.19 | 0.67 | 0.65 | 0.77 | 0.70 | 0.73 | 0.68 | 0.78 | 0.71 | 0.67 | 0.51 | 0.62 | 0.62 |
| No | Sí | Sí | TF-IDF | - | - | - | - | 0.78 | 0.71 | 0.76 | 0.69 | 0.78 | 0.71 | 0.72 | 0.67 | 0.62 | 0.62 |
| No | No | Sí | TF-IDF | - | - | - | - | 0.78 | 0.71 | 0.76 | 0.69 | 0.79 | 0.71 | 0.73 | 0.68 | 0.61 | 0.61 |
| Sí | Sí | Sí | TF-IDF | - | - | - | - | 0.78 | 0.71 | 0.76 | 0.69 | 0.78 | 0.71 | 0.73 | 0.68 | 0.62 | 0.62 |
| **Sí** | **No** | **Sí** | **TF-IDF** | - | - | - | - | **0.79** | **0.72** | **0.77** | **0.70** | **0.79** | **0.71** | **0.73** | **0.68** | **0.61** | **0.61** |

***Cuadro 6.*** Selección del mejor modelo (capacidad predictiva)

Sobre la base del modelo de regresión logística seleccionado, se elaboró una matriz de confusión, a partir de la cual se estimaron indicadores de capacidad predictiva con la muestra de prueba. Como se aprecia en los cuadros 7 y 8, la sensibilidad (casos de tweets positivos estimados correctamente con relación a los tweets la cantidad total de tweets positivos observados) asciende a 68.0%; mientras que la sensitividad (casos de tweets negativos estimados correctamente con relación a la cantidad total de tweets negativos observados) es relativamente más alta y asciende a 75.9%. Además, la exactitud (tweets positivos y negativos clasificados correctamente con relación a la cantidad total de tweets) es 72.0%, mientras que la precisión (tasa de aciertos de tweets positivos con relación a los tweets estimados por el modelo como positivos) es 74.0%. Por último, el indicador F1 weighted (una combinación de precisión y sensibilidad) es 70.9% y el valor de AUC es 0.79.

| Matriz de confusión | | Polaridad estimada | | |
|---|---|---|---|---|
| | | Tweets postivos | Tweets negativos | Total |
| Polaridad observada | Tweets postivos | 3260 | 1531 | 4791 |
| | Tweets negativos | 1144 | 3611 | 4755 |
| | Total | 4404 | 5142 | 9546 |

***Cuadro 7.*** Selección del mejor modelo (capacidad predictiva)

| Indicador | Porcentaje |
|---|---|
| Sensibilidad/Recall (tasa positiva real) | 68.0% |
| Especificidad (tasa negativa real) | 75.9% |
| Precisión | 74.0% |
| Exactitud (Accuracy) | 72.0% |
| F1-ponderado (F1W) | 70.9% |

***Cuadro 8.*** Indicadores de capacidad predictiva

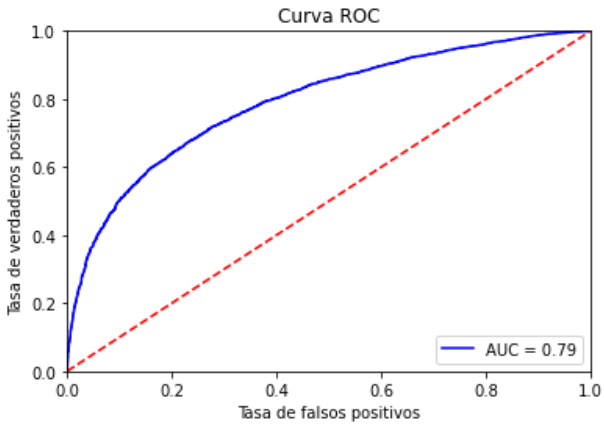


***Gráfico 4.*** Curva ROC

*4.3 Resultados: Análisis de sentimiento*

Como se registra en el Gráfico 5, todos los países presentaron valores menores a 40% en el periodo febrero hasta agosto de 2020; es decir, los textos relacionados al Coronavirus estuvieron asociados a sentimientos negativos. En todos los países, al principio en el mes de febrero hasta la tercera semana de marzo aproximadamente, la positividad fue relativamente baja (menor a 23%) y luego en la medida que se presentó el primer caso de coronavirus, la primera defunción y los gobiernos fueron implementando medidas de reactivación económica y de lucha contra la difusión del Covid19 (cierre de fronteras, cuarentenas obligatorias, entre otras), la positividad se incrementa significativamente.

Como señala Zhang, M y Ng, J. (2020), esto podría explicarse a que al principio cuando el Covid aún no afectaba directamente al país, la población que comenta en Twitter generalmente se ocupa en el desarrollo del Covid de otros países y cargan más sentimientos negativos. Sin embargo, cuando la enfermedad la afecta directamente, surgen mensajes de esperanza y más personas empiezan a apoyar ideas positivas como la de la campaña "quédate en casa" que siguieron estos países (#yomequedoencasa, #mequedoencasa, entre otros). A continuación, se muestra un ejemplo de mensajes del 26 de febrero de Colombia y Perú, que reflejan comentarios sobre el coronavirus en Estados Unidos; mientras que otros tweets del 31 de marzo indican mensajes de esperanza frente a la pandemia.

Analizando la evolución de la positividad en cada país, en el Gráfico 5, se observa que Argentina fue el país con una mayor cantidad de postergaciones del levantamiento de la cuarentena. En particular, se registra que en la cuarta postergación la positividad cae de 33% a 25%. Igualmente, en la octava postergación vuelve a observarse una caída de 28% a 24%. Esto puede reflejar una caída en el ánimo de la población tras las múltiples postergaciones.

**Tweets negativos**

Soy el único que piensa que el coronavirus, es una manipulación de EEUU para desestabilizar la economía y el orden social mundial. #Covid_19 #coronaviruscolombia #Coronavirus

(puntaje de 0.084 según el modelo)
Tweet del 26 febrero, Colombia

La OMS dijo: Ningún país se librará del virus. Así que dejen de andar diciendo que es histeria colectiva o que es una estrategia por ser potencia mundial. El coronavirus existe y está pasando ! #coronavirus

(puntaje de 0.02 según el modelo)
Tweet del 26 febrero, Colombia

**Tweets positivos**

Vibras positivas siempre aunque las olas sean altas #vamosquevamos #cuarentena #positivevibes #positivo #energia #quarantine

(puntaje de 0.99 según el modelo)
Tweet del 31 de marzo, Perú

¡Se apareció un arco iris al comenzar el toque de queda! Ojalá ese arco iris sea un signo que las cosas mejorarán... #quedateencasa #coronavirus #covid19 #vamosperu #cuarentena en San Juan de Miraflores

(puntaje de 0.773 según el modelo)
Tweet del 31 de marzo, Perú

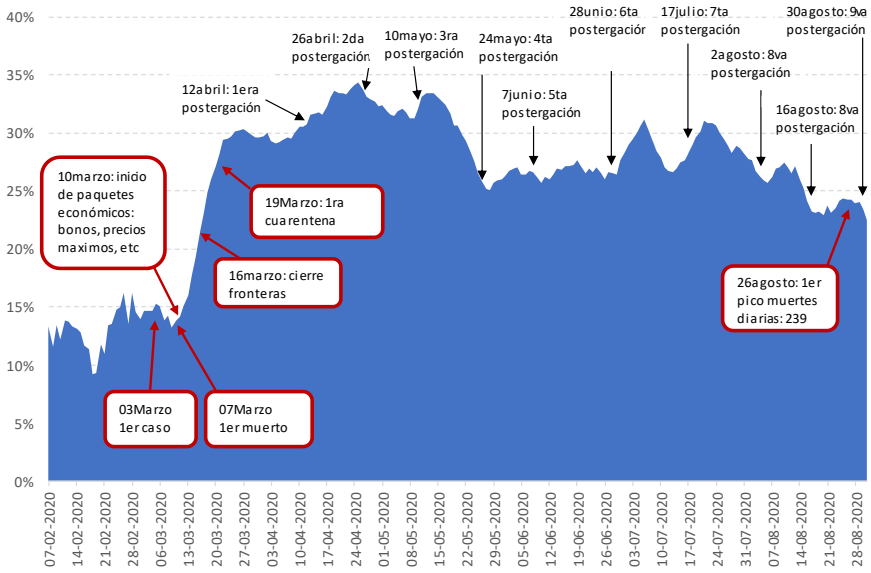


***Gráfico 5***. Ratio de positividad de Tweets en Argentina (media móvil de los últimos 7 días)[12]

En el caso de Perú, la positividad alcanza su valor máximo el 26 de marzo, cuando la prensa internacional y local opina favorablemente sobre las medidas tempranas implementadas contra el coronavirus. Luego, se muestra una tendencia negativa en positividad tras 3 postergaciones del levantamiento de la cuarentena, enfrentar el segundo pico de muertes en el país y pasar a convertirse en el país con mayor mortalidad por Covid, como se muestra en el Gráfico 6. Del mismo modo, en Colombia, la positividad cae tras 3 postergaciones y luego, a diferencia de Perú, se mantiene constante a partir de julio, como se registra en el Gráfico 7.

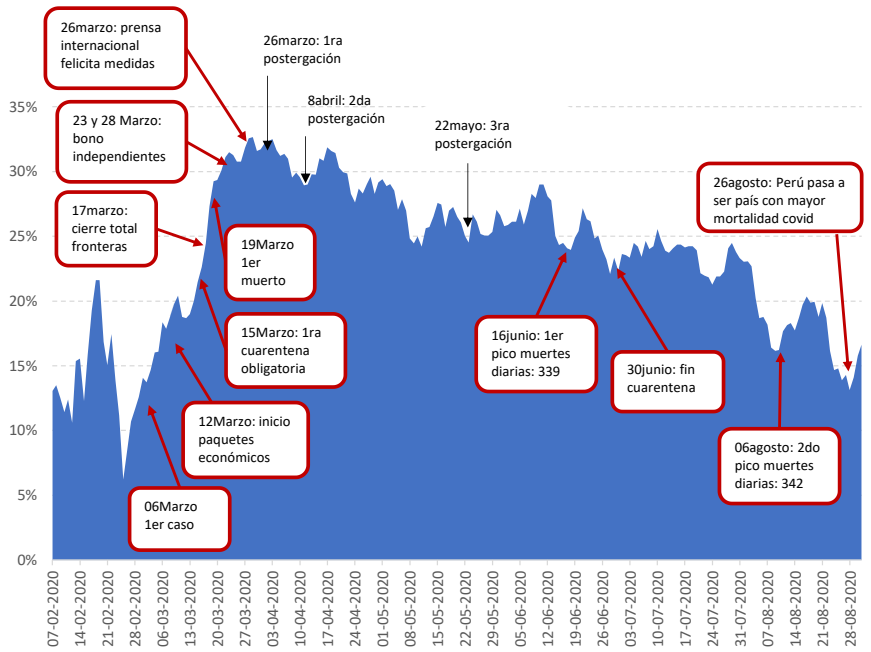


***Gráfico 6.*** Ratio de positividad de Tweets en Perú (media móvil de los últimos 7 días) [12]

[12] El ratio puede tomar los valores de 0% a 100% y corresponde al porcentaje de tweets positivos sobre la cantidad de tweets totales registrados en el día. Un tweet es positivo si el puntaje con el modelo es mayor a 0.6.

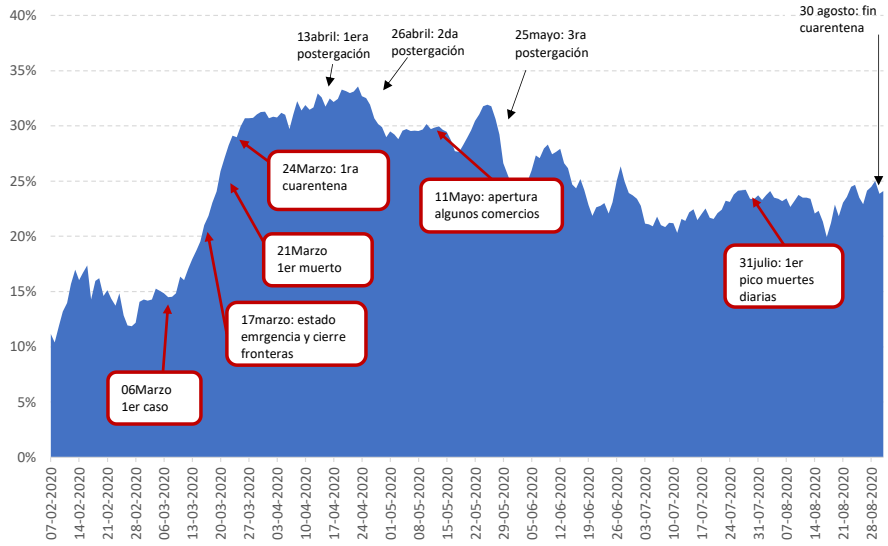


***Gráfico 7.*** Ratio de positividad de Tweets en Colombia (media móvil de los últimos 7 días) [12]

En Ecuador, la positividad se mantiene alrededor de 35% desde finales de marzo hasta el primero de junio. Este país fue el que tuvo el pico de número de muertes más tempranamente y (434 casos) también fue el primero en finalizar la cuarentena. El 7 de julio se observa una caída en la positividad, luego de que la cantidad de muertos se volvió a incrementar, lo cual pudo atemorizar a la población nuevamente. Este pico no fue tan alto como el primero y no volvieron las medidas restrictivas, por el contrario, se flexibilizaron más medidas, como la apertura de las playas, y la positividad volvió a mostrar un incremento.

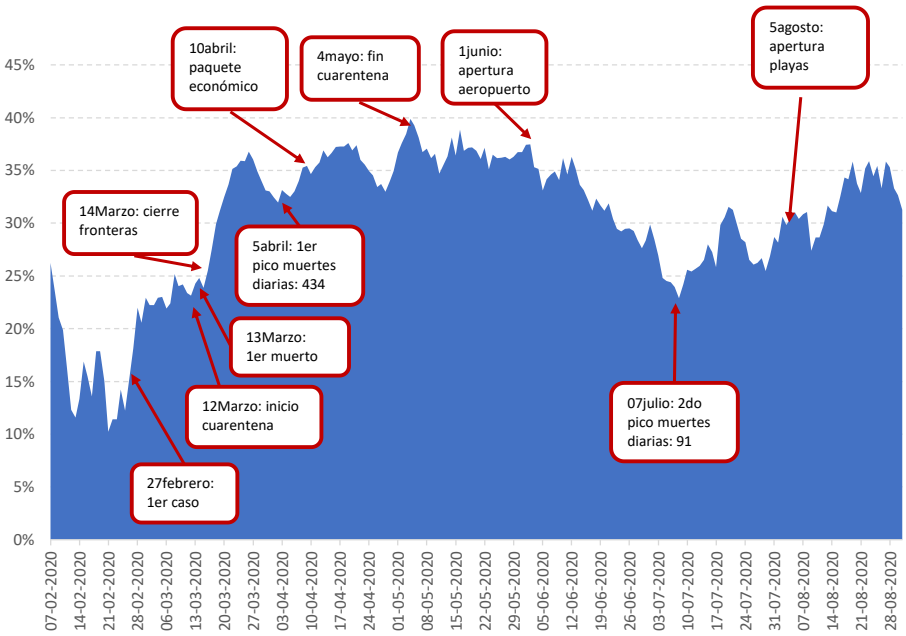


***Gráfico 8***. Ratio de positividad de Tweets en Ecuador (media móvil de los últimos 7 días) [12]

En Chile, el punto más alto de positividad se presentó a principios de abril, luego tuvo una ligera caída y se mantuvo prácticamente constante hasta finales de agosto, como se muestra en el Gráfico 10.

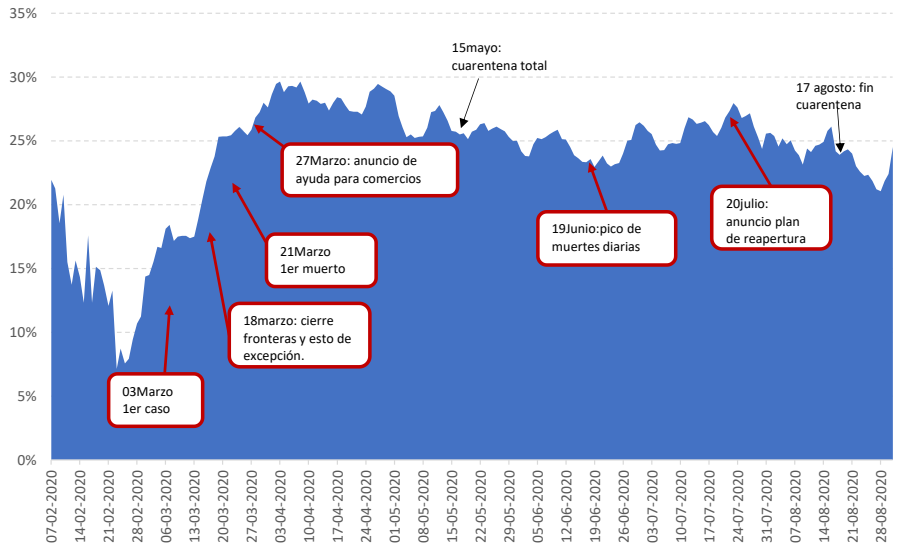


***Gráfico 10.*** Ratio de positividad de Tweets en Chile (media móvil de los últimos 7 días) [12]

Por último, en el caso de Uruguay, al igual que en el resto de países, la positividad se incrementó tras las medidas implementadas de lucha ante el Covid y la primera defunción. Aunque Uruguay no implementó la cuarentena de forma obligatoria, el 90% de la población adoptó la medida de forma voluntaria y afrontaron el cierre de fronteras y suspensión de espectáculos y clases presenciales13. Además, como se observa en el gráfico 11, se registró una reducción de la positividad el 29 de mayo, 17 de julio y 17 de agosto debido a que las personas criticaron la movilización de paros o huelgas ante la situación de pandemia que atravesaba el país.

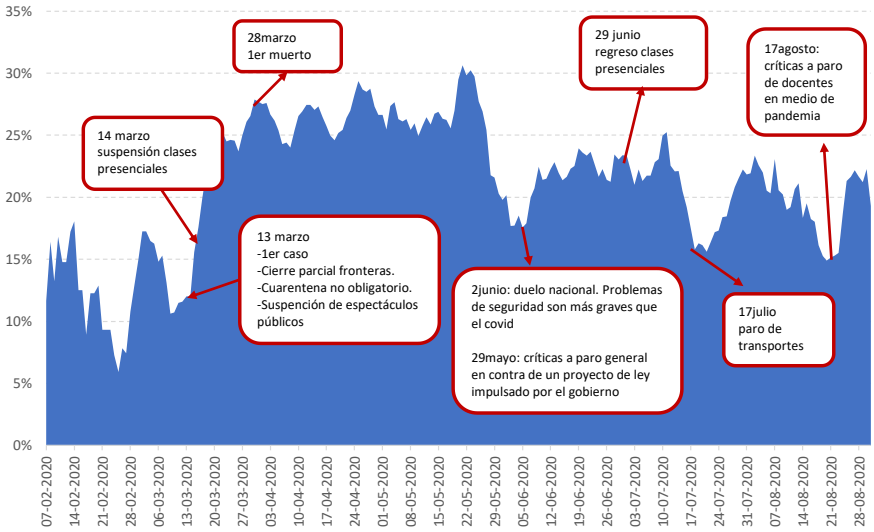


***Gráfico 11.*** Ratio de positividad de Tweets en Uruguay (media móvil de los últimos 7 días) [12]

Es importante mencionar que, en promedio, Ecuador es el país analizado que presenta mayores niveles de positividad a lo largo de los meses de marzo a agosto. Posiblemente sea

[13] Información disponible en: https://www.bbc.com/mundo/noticias-america-latina-52837193, actualizado al 12 de setiembre de 2020

debido a que fue el país que relajó las medidas de confinamiento meses antes que los otros países. Por otro lado, Uruguay y Perú mostraron los niveles de positividad más bajos.

**Palabras clave:**

Se realizó un análisis de los tweets que contienen ciertas palabras que han cobrado interés en esta época de pandemia. En el Gráfico 12 se puede observar el porcentaje de positividad de los tweets asociados a las palabras "Quédateencasa", "Gobierno", "China", "Salud". Notamos que los tweets que presentan el hashtag Quédateencasa tienen un porcentaje de positividad mucho mayor que el de aquellos con la palabra "Cuarentena" aunque ambos términos signifiquen lo mismo. Según Zhang, M., Ng, J., esto podría significar una oportunidad para que los gobiernos enfoquen mejor sus maneras de comunicación de las medidas ante posibles nuevos rebrotes que vuelvan a obligar a la población a permanecer en sus casas. Por otro lado, los tweets con la palabra "Gobierno" y "China" presentan ratios de positividad muy bajos, con un valor promedio de 5% y 14% respectivamente a lo largo de todos los meses analizados. Finalmente, los tweets asociados a "salud" muestran radios de positividad similares a los de la base total analizada, que se incrementan en quincena de marzo al comenzar la cuarentena.

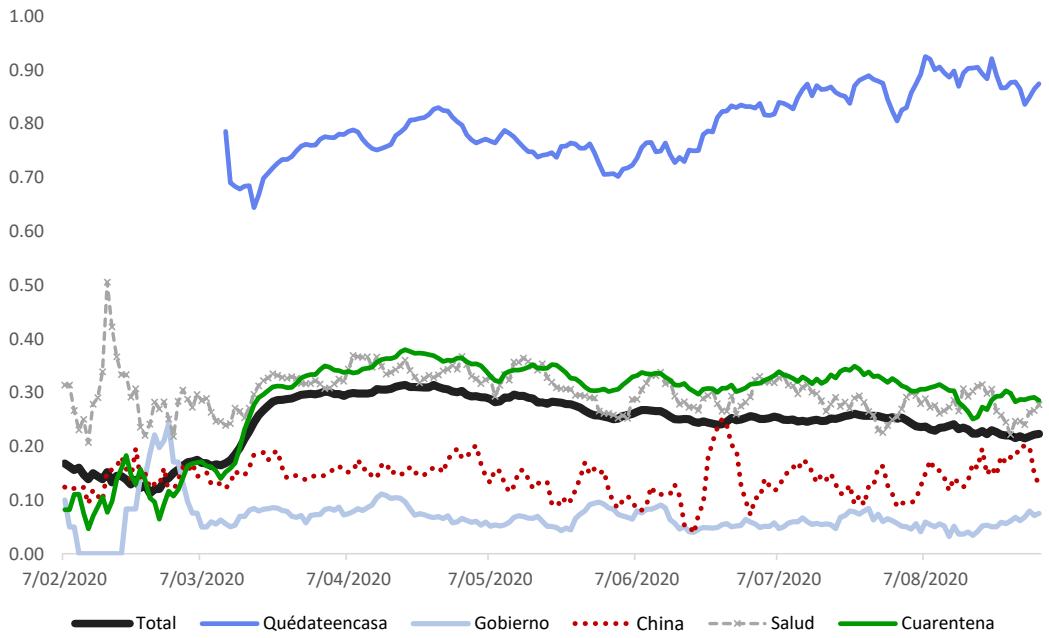


***Gráfico 12.*** Ratio de positividad de Tweets que contienen las palabras: "Quédateencasa", "Gobierno", "China" y "Salud" y de la base total (media móvil de los últimos 7 días) [12].

## 5. Conclusiones:

En el presente trabajo se buscó medir el sentimiento de algunos países de Latinoamérica respecto al coronavirus y a las medidas adoptadas por sus gobiernos. Se usó la técnica de Análisis de Sentimiento analizando los datos de Twitter emitidos en las capitales de dichos países desde el primero de enero hasta el 31 de agosto, obteniendo las siguientes conclusiones:

- Se evaluaron distintos modelos de clasificación a fin de encontrar el que permitiera predecir un sentimiento de manera más acertada. El modelo seleccionado fue el de

Regresión logística, usando como método de ponderación de características el modelo TF-IDF, realizando un proceso de normalización y stemming sin remover las stopwords. El valor de AUC de este modelo fue de 0.79 en los datos de evaluación. Este modelo representa una considerable mejora con respecto a la librería de análisis de sentimiento en español de Python.

- Con este modelo se estimó el sentimiento asociado a los tweets de cada país, analizándose la evolución de la positividad a través del tiempo y vinculándose con las medidas adoptadas por los gobiernos de estos países:

  - Se encontró que todos los países mostraron valores menores a 40% en el periodo analizado; es decir los tweets relativos al coronavirus se asociaron a sentimientos negativos como era de esperarse.

  - Al principio en el mes de febrero hasta la tercera semana de marzo aproximadamente, la positividad fue relativamente baja y luego en la medida que se presentó el primer caso de coronavirus, la primera defunción y los gobiernos fueron implementando medidas de reactivación económica y de lucha contra la difusión del Covid19, la positividad se incrementa significativamente. Esto podría explicarse a que al principio, cuando el Covid aún no afectaba directamente al país, la población que comenta en Twitter generalmente se ocupa en el desarrollo del Covid de otros países y cargan más sentimientos negativos. Sin embargo, cuando la enfermedad la afecta directamente, surgen mensajes de esperanza y más personas empiezan a apoyar ideas positivas como la de la campaña "quédate en casa" que siguieron estos países (#yomequedoencasa, #mequedoencasa, entre otros).

  - En promedio, Ecuador y Argentina son los países analizados que presentan mayores niveles de positividad a lo largo de los meses de marzo a agosto. Por otro lado, Perú y Uruguay mostraron los niveles de positividad más bajos, manteniendo niveles casi similares a lo largo de todo el periodo evaluado.

  - Respecto a los temas más comentados durante la pandemia se encontró que las palabras de mayor interés fueron "gente", "gobierno", "salud" y "casa". Esto refleja que una de las preocupaciones de la población estuvo relacionada con la salud y con la actuación del gobierno frente a la pandemia. Asimismo, la palabra casa está relacionada con la política de confinamiento de quedarse en casa durante la pandemia.

  - Los tweets con el hashtag Quédateencasa presentan un sentimiento de positividad mucho mayor que el de aquellos con la palabra "Cuarentena" lo que podría significar una oportunidad para que los gobiernos enfoquen mejor la manera de comunicación de las medidas ante posibles nuevos rebrotes que vuelvan a obligar a la población a permanecer en sus casas.

## 6. Bibliografía:


Barkur, G., Vibha, y Kamath, G. (2020). Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. *Asian Journal of Psychiatry. (51).* https://doi.org/10.1016/j.ajp.2020.102089

Dubey, A. D. (2020). *Twitter Sentiment Analysis during COVID-19 Outbreak.* http://dx.doi.org/10.2139/ssrn.3572023

Köksal A. (2020). *BERT Sentiment Analysis Turkish* https://github.com/akoksal/BERT-Sentiment-Analysis-Turkish

Korkut, U., Foley, J. y Ozduzen, O. (2020). The Digital Publics of #Schengen and #Eurozone During the Coronavirus Crisis. *Respond. (3).* https://drive.google.com/file/d/1f8uokB9rptS9GwNpQe-beyDS0JpQ1wRg/view

Liu, B., (2012). *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

Samuel, J., Nawaz, G., Rahman, M., Esawi y E., Samuel, Y. (2020). *COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification.* https://doi.org/10.3390/info11060314

Sharma, K., Seo, S., Meng, C., Rambhatla, S. y Liu, Y. (2020). *COVID-19 on Social Media: Analyzing Misinformation in Twitter Conversations.* https://arxiv.org/abs/2003.12309

Sobrino, J.C. (2018). *Análisis de Sentimientos en Twitter* [Tesis de maestría, Universidad Oberta de Catalunya]. http://openaccess.uoc.edu/webapps/o2/bitstream/10609/81435/6/jsobrinosTFM0618memoria.pdf

Storjohann, P. (2005), *Corpus-driven vs. corpus-based approach to the study of relational patterns* https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/5006/file/Storjohann_Corpus_driven_vs_corpus_based_approach_to_the_study_of_relational_patterns_2005.pdf

Zhang, M., Ng, J., (2020), *Twitter Sentiment Analysis: What does Social Media tell us about coronavirus concerns in the UK?* https://www.actuaries.org.uk/system/files/field/document/Twitter%20Sentiment%20Analysis.pdf

## 7. Anexos:

Anexo 1

Para obtener los tweets de una determinada ciudad, se usó la librería GetOldTweets3 que permite filtrar lugares específicos a través de los parámetros “geocode” y “distances”. Esto es posible dado que cada tweet contiene información de las geocoordenadas en la API pública que ayuda a identificar el lugar de donde fueron emitidos.

A través de la página web mapdevelopers.com se obtuvieron las diferentes coordenadas y radios para extraer la información de la manera más aproximada posible. Así por ejemplo, como se muestra en el Gráfico 11, para la ciudad de Montevideo se usaron los datos de seis zonas que conforman dicho territorio (Los tweets repetidos de las zonas que se superponen serán eliminados).

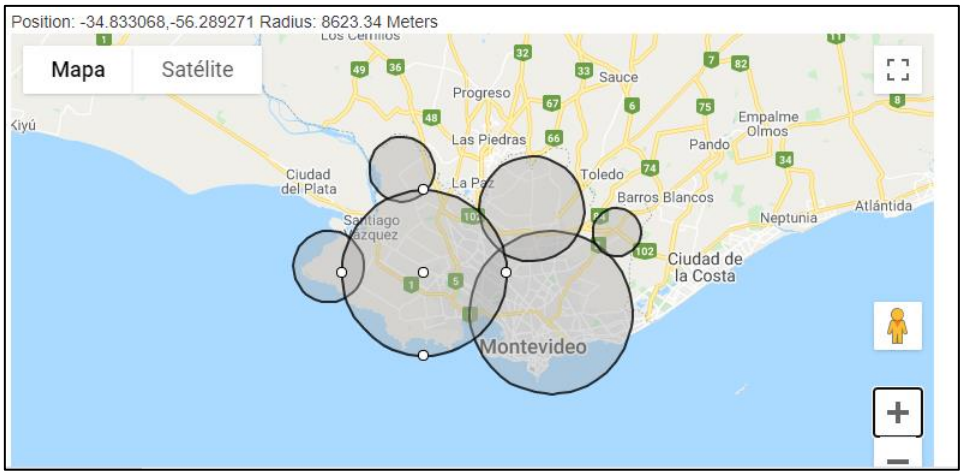


Gráfico 11: Determinación de las coordenadas y radios para la ciudad de Montevideo.

Con las coordenadas y radios obtenidos se procedió a elaborar los códigos en Python que permitieran obtener los tweets de las zonas señaladas como se muestra a continuación.

```
geocode = "-34.826967,-56.398748"
km = "3.71705km"
text_query = 'covid OR coronavirus OR pandemia OR cuarentena'
since_date = "2020-1-01"
until_date = "2020-9-01"
count = 2500000
tweetCriteria = got.manager.TweetCriteria().setQuerySearch(text_query)\
.setNear(geocode).setWithin(km) \
.setSince(since_date) \
.setUntil(until_date) \
.setMaxTweets(count) \
.setEmoji("unicode")
tweets = got.manager.TweetManager.getTweets(tweetCriteria)
tweets_list = [[tweet.id, tweet.author_id, tweet.username, tweet.text, tweet.retweets, tweet.favorites,
tweet.replies, tweet.date, tweet.hashtags] for tweet in tweets]
tweets_uruguay1 = pd.DataFrame(tweets_list, columns = ['Tweet_Id', 'Tweet_User_Id', 'Tweet_User',
'Text','Retweets', 'Favorites', 'Replies', 'Datetime', 'hashtags'])
```