¿Quiénes son los que generan el contenido?

Bertolini, Lucas; Di Paola, Matías; Vilca, Stella

Resumen

Twitter es una red social que genera muchos datos. Allí se hace eco el humor social. Cualquier situación que este ocurriendo en el mundo offline tiene su complemento en el mundo online. El trabajo práctico se da en un contexto de pandemia global y distanciamiento social provocados por el COVID-19, y toma tweets relacionados a ello.

Encontramos que se distribuye más información de la que se genera. Se generaron distintas categorías para clasificar a los usuarios y entender cómo es que circula la información. Buscamos entender quiénes crean el contenido que tiene más difusión.

Creemos que esta es una primera aproximación, simple pero potente, con baja necesidad de recursos, que puede ser el punto de partida para análisis más completos y complejos, sobre muestras de datos de mayor tamaño.

Analizando quiénes son los generadores de contenido

El presente trabajo práctico tiene como finalidad aplicar los conceptos que fuimos viendo en la materia Data Mining, parte de la maestría en UBA. Así mismo entrar en la dinámica de KDD, Knowledge Discovery in Databases, que fue el marco de trabajo estudiado durante el cursado de la materia. Es muy interesante poder llevar a un contexto más real, lo visto como contenido teórico.

Partimos de dos archivos que contenían tweets y usuarios. Los tweets fueron tomados de la api rest de la red social y giraban en torno a la situación del COVID-19. Es el tema recurrente en el último tiempo, desde principio de año cuando comenzó como un rumor que venía desde oriente, hasta el día de hoy, y muy probablemente por lo que quede de 2020.

El primer paso fue entender una generalidad de los datos que teníamos. Encontramos 90 columnas en el archivo de tweet distribuidas de una forma heterogénea. Si bien seguía siendo una cantidad de atributos manejables, se parecía más a una tabla rasa, con muchos datos faltantes y vacíos. Nos tocó entender como Twitter daba la información y qué significado tenía cada uno de los campos.

Una segunda situación que tuvimos que enfrentar fue la profundidad de la información. En cada línea de la colección había información que pertenecía al tweet final, al usuario y al tweet original (en caso de haber sido un retweet o una cita). Lo primero ahí fue ver ya una duplicidad de información. Supimos que había bastantes decisiones de transformación que deberíamos tomar para generar perspectivas de esos datos que nos ayudarán a comprenderlo mejor.

El proceso, como muchos de los trabajos en data science, se dió desde la combinación de miradas heterogéneas, provenimos de tres áreas relativamente diferentes, e intentamos usar eso para potenciar el trabajo. Hubo un primer trabajo, en el que cada uno de los integrantes del grupo fue aproximándose al dataset, sacando algunos primeras preguntas, conclusiones de los datos y posibles caminos a seguir. Posterior a eso, hubo una puesta en común, donde coordinamos un sentido general para el trabajo. Ahí formulamos algunas primeras preguntas, ya más generales y orientandonos en un cierto sentido. La última parte correspondió a integrar distintas ideas entorno al eje que habíamos pensado, quienes generan el contenido en la red, además repasamos este análisis alrededor de un eje de contenido correspondiente a los distintos países del continente.

Análisis exploratorio

Hubo un primer momento de exploración sobre las distintas columnas del dataset. No es relevante exponer aquí cada conclusión, cada columna, el porcentaje que correspondía a nulos. Pero si dar una idea de las primeras decisiones que tomamos:

- Todos los tweets son del idioma español y no están protegidos, esa nula variabilidad hizo que las descartemos.
- Los datos relacionados a las reacciones de los tweets son de poca información. Esto se debe a la forma en que fueron capturados, y representa un primer impedimento para medir el alcance de un tweet, o la popularidad de un usuario y su tweet. Pero aquí también pudimos diseñar estrategias para avanzar sobre un aspecto más que importante dentro de twitter y el análisis de cualquier dataset.
- Otro aspecto que tuvimos que descartar fue la temporalidad de los tweets. Inicialmente

vimos un pico en el día dos de mayo y nos llamó poderosamente la atención. Era más raro aún que no coincidía con ninguna noticia en particular, ni el día del trabajador, ni una prolongación de la cuarentena[1]. Para poder descartar este tema, analizamos con mayor granularidad el campo fecha en los distintos tweets. Fuimos disminuyendo la ventana temporal hasta llegar a la cantidad de tweets por minuto. Adjuntamos a continuación el gráfico que muestra esa distribución.

Gráfico 1 - Distribución de la variable temporal en días/minutos

El gráfico muestra, distribuído por día, y ordenado por minuto, la cantidad de tweets. Podemos ver que efectivamente hay una mayor cantidad en el día dos de mayo, pero también que el tiempo que estuvo capturando tweets fue mayor y sostenido por varios minutos. No hay un patrón de captura, ni una regularidad apreciable. Parece, más bien, fechas y horas aleatorias. Esto nos lleva a tener que descartar esta característica del dataset.

Viendo estos primeros datos, percibimos la necesidad de ser selectivos en nuestros esfuerzos. No por evitar hacer trabajo o transformaciones, sino entendiendo que había mucha información y muy desestructurada. Con dos posibles consecuencias, que inclusive se presentan como contrapuestas, realizar muchísimo trabajo de saneamiento de datos que no íbamos a usar; o de enfrentarnos a una tarea enorme y que, siendo muy desafiante, nos invite a extender más allá de las 15 páginas el trabajo práctico.

Caracterización de los tweets

Conociendo una estructura general de los datos tocaba:

- Seleccionar las variables relevantes
- Crear las preguntas que nos iban a guiar
- Reducir dimensiones y condensar información
- Tratar los nulos

El proceso fue una combinación de estas ideas. Es difícil enmarcar cada paso, decisión o criterio en uno de estos temas. Como un primer paso fuimos generando categorías que nos

[1] Los anuncios de prolongación del aislamiento social y obligatorio, en Argentina, fueron en las fechas: 19/3 - 10/4 - 25/4 - 8/5 - 20/5 - 5/6

ayudarán a resumir el dataset y caracterizar a los dos componentes principales: Usuarios y Tweets. Y a partir de complementar los datos que ya teníamos, con estas nuevas características, pudimos responder (o aproximar una respuesta) a lo que nos planteamos en el trabajo práctico.

En algunas categorías la ausencia de valor representaba información en sí mismo. Por ejemplo, is_retweet podía tener tres valores: Verdadero, Falso y Nulo, que significaba que era solamente una cita. Esto se debe a que en la información del tweet el dato no existe, porque no corresponde.

Eso nos llevó a uno de los primeros análisis. No todos los registros eran del mismo tipo. Identificamos 4 variedades.

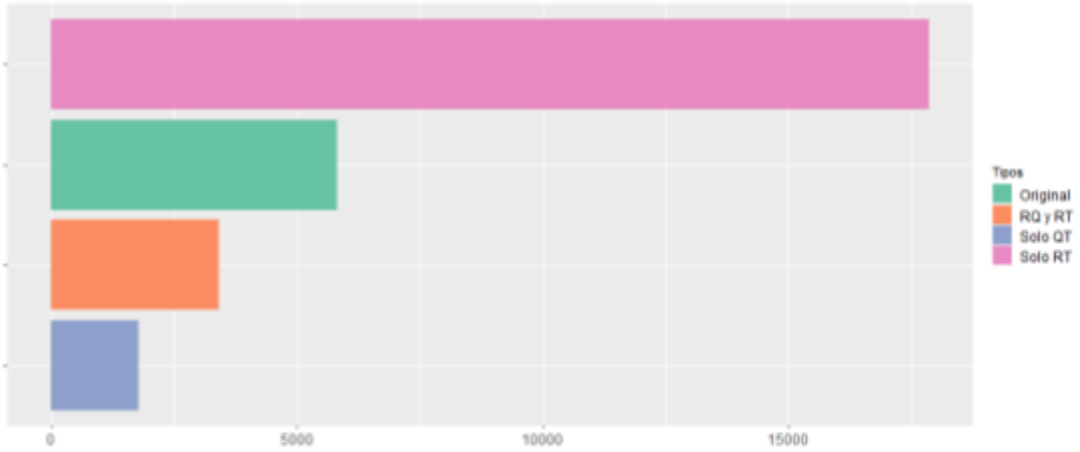

Gráfico 2 - Distribución de los tipos de tweets

En la imágen se puede ver la proporción de cada uno. No es menor pensar que el contenido original es una minoría del tráfico de la red. Twitter ya se presentaba en esta primera aproximación como una red, donde la mayor parte del flujo amplifica opiniones que alguien dio. Podemos decir que es una forma de opinar, difundir el tweet de otra persona, es reforzar mi creencia en que eso es verdadero o válido.

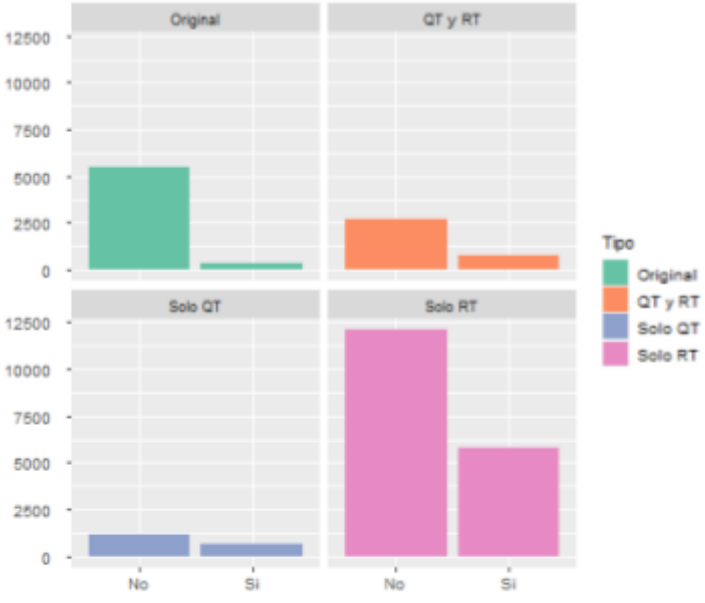

Gráfico 3 - Distribución de verificados por tipo de tweet

Algo que se figuraba como importante, era el campo verificado (Verified, retweet_verified, quoted_verified). Estos campos nos dicen si quien creo el tweet fue una cuenta verificada. Estas tiene un aval de twitter, una validación de autenticidad. Esto representa una mayor confiabilidad a lo que se escribe. Más pensando tantas noticias de bots[2] y situaciones que llaman la atención.

[2] https://www.infobae.com/tecno/2019/08/09/satisface-a-mauricio-caricia-significativa-y-otras-frases-insolitas-viralizadas-en-twitter-abrieron-un-debate-sobre-los-bots-en-campana/

Se aprecia a simple vista como la proporción de usuarios no verificados es mayor. Lo cual era esperable. Pocos usuarios eligen pasar por un proceso de verificación. Suele estar reservado a personas públicas, negocios o perfiles de usuarios que buscan tener un mayor cuidado de su imágen en internet.

Es interesante percibir que en el contenido difundido hay una mayor presencia de cuentas verificadas. Ya podemos empezar a ver que esto realmente se toma como un signo de confianza. En solo retweets la proporción ya es más alta, pero principalmente en solo citas (Quotes), donde representa más de la mitad de los valores.

Teniendo en cuenta una primer recuento de la cantidad de Tweets y retweet por usuario vimos un desbalance entre creadores y difusores. En este punto era vago el concepto pero nos servía para entender esa primera idea que luego profundizaríamos.

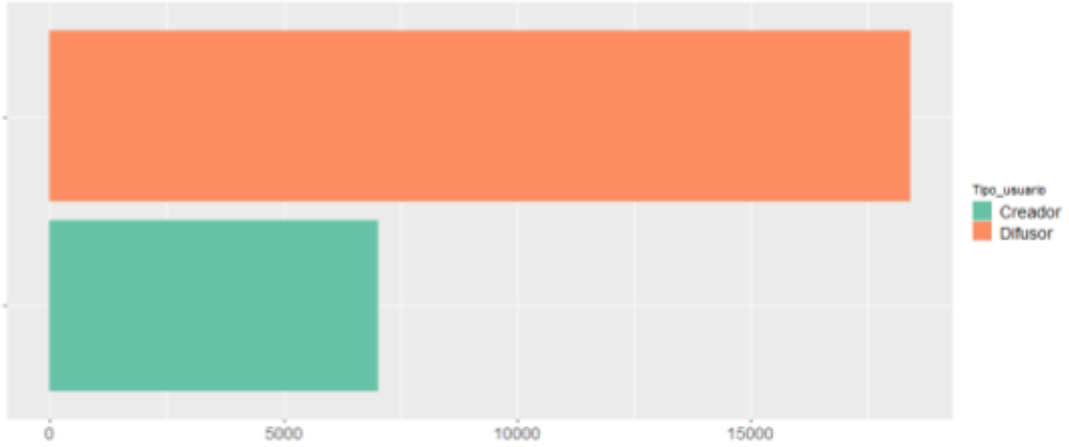

Gráfico 4 - Distribución de creadores y difusores

Caracterización de los “Generadores de contenido”

De todos los atributos antes considerados, aquellos que consideramos relevantes del usuario para continuar con el análisis son: **screen_name** (nombre o alias con que los otros usuarios lo ven y arroban)**, followers_count** (usuarios que lo “siguen”)**, friends_count** (usuarios “seguidos”)**, statuses_count** (cantidad de “posteos”) y **account_created_at** (fecha de creación de la cuenta).

Creación de nuevas variables. Teniendo en cuenta que nuestro objetivo es identificar a los creadores de contenido ,pensamos que dos atributos que podrían definirlos son “popularidad” y “actividad”. El perfil buscado debería ser altamente popular y tener un nivel de actividad alto.

- **Popularidad** definida como: { **followers_count / friends_count + e** }

La idea subyacente es que “popularidad” para este tipo de perfil no es solo tener una gran cantidad de seguidores, sino también tener una proporción de amigos mucho menor. De alguna manera “no le interesa” tener amigos y seguir gente, sino más bien que lo sigan para que consuman lo que comparte o dice. Al cálculo se le suma un número pequeño **e** para evitar dividir por cero (aquellos perfiles que no siguen a nadie).

- **Actividad** definida como: { **statuses_count** / 2021 - **account_created_at[year]** }

Al solo tener unas "fotos" de unos días de tweets, la manera de medir la actividad histórica de un determinado usuario fue obtener la relación entre la cantidad de posteos totales realizados y la cantidad de años transcurridos desde la creación de la cuenta, o más fácil, posteos realizados por año.

Una vez obtenidas las nuevas variables de interés, nos quedamos con aquellas numéricas y se pasó a analizar su distribución, ruido, presencia de outliers y redundancia y luego aplicar las transformaciones requeridas para cada caso.

Re-escalado.Lo primero que se nota es que las variables tienen una distribución de larga cola[3], muy similar a lo que nos sucedió en las variables del dataset ruidoso que vimos en clase. Para subsanar el sesgo se aplicó una transformación Logarítmica, en base 10, ya que esto permite asemejar la distribución de los datos a una distribución normal. Podemos ver el resultado de la transformación sobre "Posteos por año" en el gráfico a continuación, en este caso resultó con una leve asimetría negativa.

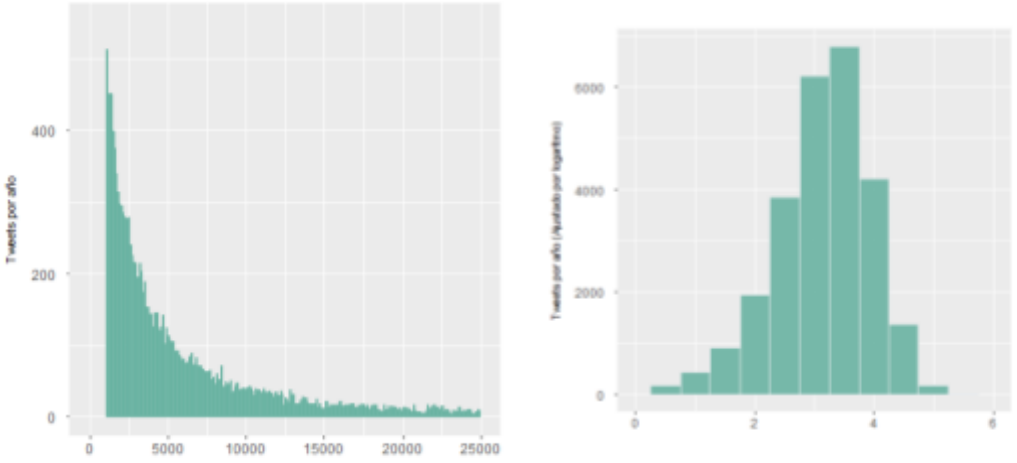

Gráfico 5 - Distribución de la variable "Posteos por año" antes (izq) y luego de aplicar log10 (der)

Ruido, outliers y Binning. Al analizar la variabilidad aplicando un boxplot se puede apreciar que muchas de las observaciones se encuentran por fuera de los "bigotes" (IQR*1.5), habiendo mayor cantidad por debajo del bigote inferior, por lo explicado más arriba. Se decidió aplicar una discretización por binning con la intención de reducir el ruido, los outliers y al mismo tiempo categorizar las variables ya que será más fácil aplicar los filtros deseados cuando se necesite.

Se eligió el binning de igual ancho ("equal width") frente a igual frecuencia ("equal freq") ya que nuestro interés es quedarnos con los usuarios que se encuentren en los bins superiores (más populares y con más actividad) y no que en cada bin tengamos igual cantidad de frecuencia de observaciones. También se probó aplicar el método floor(), pero lo descartamos ya que preferíamos tener una "mayor resolución" de bins y tener más libertad cuando aplicamos los filtros.

[3] https://es.wikipedia.org/wiki/Larga_cola

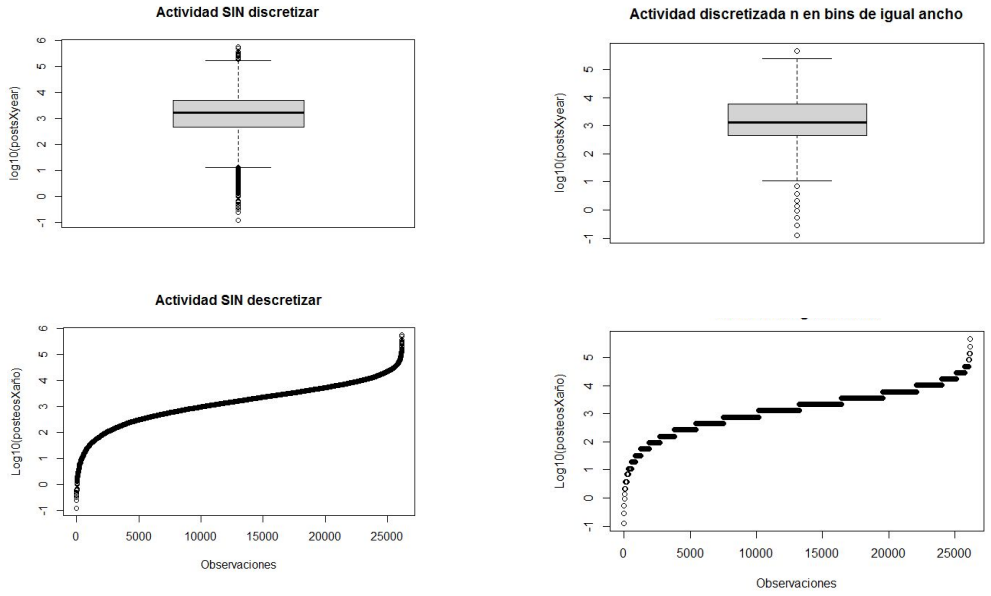

Gráfico 6 - *Boxplot de la variable "Actividad" (posteosXaño) antes (izq) y luego de discretizar por binning de igual ancho (der).*

Correlación y reducción de dimensionalidad. una vez que teníamos las variables transformadas decidimos analizar la correlación de las mismas. La pregunta que nos hicimos fue si la creación de las nuevas variables, actividad y popularidad, correlacionaban con statuses_count y followers_count, y hasta qué punto eran utilizar una frente a otras. Es decir, necesitamos validar la elección de las nuevas variables. Como se esperaba, la mayor correlación se produjo entre **statuses_count** y **actividad** (0.93) y entre **followers_count** y **popularidad** (0.6). Utilizamos la libreria caret y generamos una matriz de correlación con un cutoff de 0.75 y nos arrojó que los atributos más correlacionados, y por lo tanto a ser descartados, eran **statuses_count** y **followers_count.**

Hipótesis. En este punto del análisis contamos una comprensión más cabal de las variables, y estamos en condiciones de definir nuestras hipótesis para caracterizar el perfil de los usuarios que producen contenido, informan y/o generan opinión:

1. **Son altamente Populares y presentan gran Actividad**

Filtramos las observaciones que se encuentren simultáneamente en los bins más altos de estas variables discretizadas, elegimo para ello bins > 3, para cada variable. Si bien el número es arbitrario, lo elegimos lo suficientemente alto pero al mismo tiempo nos arroje una cantidad de usuarios para poder seguir filtrando. En este primer filtrado obtuvimos 168 usuarios.

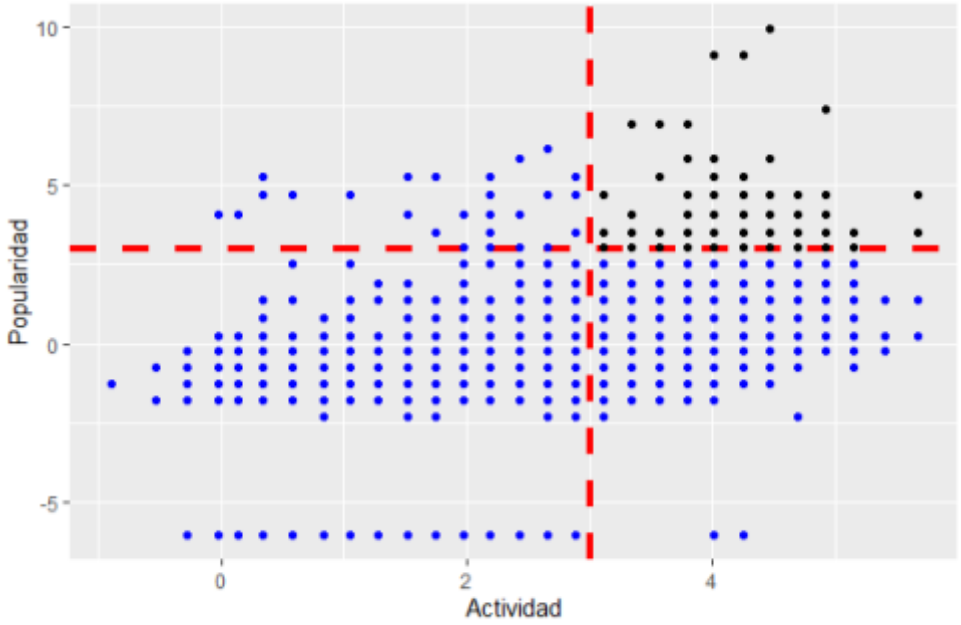

Gráfico 7 - Distribución de usuarios según actividad y popularidad

2. Tweetean, no Re-tweetean

Desde el lado del tipo de tweet, suponemos que mayoritariamente producen tweets propios más que retweetean ajenos. Se filtró entonces por la categoría "TW", quedaron **133 usuarios.**

3. C**omparten en su tweets, algo más que texto**

Cuando analizamos los tweets (categoría "TW"), pudimos observar que un alto porcentaje contenía urls. Por lo tanto creamos la variable **urls presentes por tweet**, y filtramos aquellos con 1 o más urls por tweet, quedaron **124 usuarios.**

Validación por retweet. La manera que encontramos de validar si efectivamente encontrábamos a los usuarios generadores de contenido, fue tomar los 100 usuarios más retweeteados de nuestro dataset, y comparar si estaban presentes (o no) respecto a los que habíamos obtenido a partir de nuestras hipótesis. Para ello necesitábamos rankear los valores obtenidos luego de aplicar los diferentes filtros, por lo que tomamos las dos variables numéricas, popularidad y actividad, las normalizamos por zscore, y luego las promediamos para cada una de las observaciones. Finalmente ordenamos de manera decreciente los primeros 100 resultados . Lo mismo hicimos para followers_count y statuses_count de manera de tener un control.

Filtros/Hipotesis	100 usuarios más retweetados encontrados
Popularidad + Actividad + TW	39
Popularidad + Actividad + TW + URL	39
control variables descartadas	
followers_count + statuses_count	43
followers_count + statuses_count + TW+ URL	42

Como se puede apreciar en la tabla presentada más arriba, si bien se pudo detectar casi la mitad de los usuarios más retweeteados parecería no haber diferencia entre las variables construidas y las originales del dataset. Tal como había predicho el método de detección de redundancia las variables se encuentran altamente correlacionadas y no se distingue el efecto en la detección del perfil deseado.

Gráfico 8 - Nube de palabras a partir de los 100 primero usuarios filtrados a partir de las hipótesis planteadas.

Dividir pequeños universos

Una idea que se fue tejiendo en paralelo a la caracterización del dataset, fue encontrar una historia en estos datos, y una forma de utilizar estas categorías. Aplicando los conceptos en pequeñas muestras, podríamos validar también las conclusiones y categorías a las que nos acercabamos en lo general.

La primera aproximación a cómo dividir el universo surgió explorando los hashtags. Desde un comienzo son muy famosos los trending topic[4]. Un hashtag que se torna trending topic escala de una forma inimaginable. Son muy interesantes dentro de la plataforma porque actúan como agregadores de contenidos y sensaciones. Si bien nuestro universo era pequeño y no podíamos tomar esto para generalizar, si podríamos encontrar palabras o temas con eco en el universo.

Lo más mencionado claramente tenía que ver con COVID-19, en múltiples formas de escribirlo, pero una vez normalizado el texto[5] y filtrada la temática coronavirus surgieron estas 20 palabras:

ultimahora	envideo	**dedicareestemayo**	confirmado	china
cuba	loultimo	urgente	**bartlett**	**imss**
atencion	envivo	afp[6]	**proteccionyaccion**	**snteunidoyfuerte**
venezuela	**mexico**	**eeuu**	ahora	pandemia

Marcamos en negrita estos que fueron algunos que nos llamaron la atención. Todos se relacionaban al continente, y surgían algunas conexiones.

- México, Bartlett (Senador mexicano), snteunidoyfuerte (SNTE es Sindicato Nacional de Trabajadores de la Educación), imss (Instituto Mexicano del Seguro Social)

[4] Tema del momento

[5] Convertir todos los hashtags a minúsculas, y eliminar caracteres que molestaran al procesamiento (Puntuación, espacios al inicio o al final, caracteres fuera del alfabeto latino)

[6] AFP parecía interesante, porque es el acrónimo de la Administradora de Fondos de Pensiones en Chile. Sin embargo, en un segundo análisis, encontramos que dentro del hashtag se agrupaban variedad de temas: Agence France-Presse y el Administradoras de Fondos de Pensiones y de Cesantía de Colombia.

- Venezuela, proteccionyaccion
- Cuba, dedicareestemayo

Esto nos ayudó a preguntarnos y pensar en cómo la situación se iba desarrollando en distintos países del continente.

El primer problema fue la cantidad de datos faltantes en temas de geolocalización, como puede apreciarse en la siguiente gráfica “Grafico xxx”.

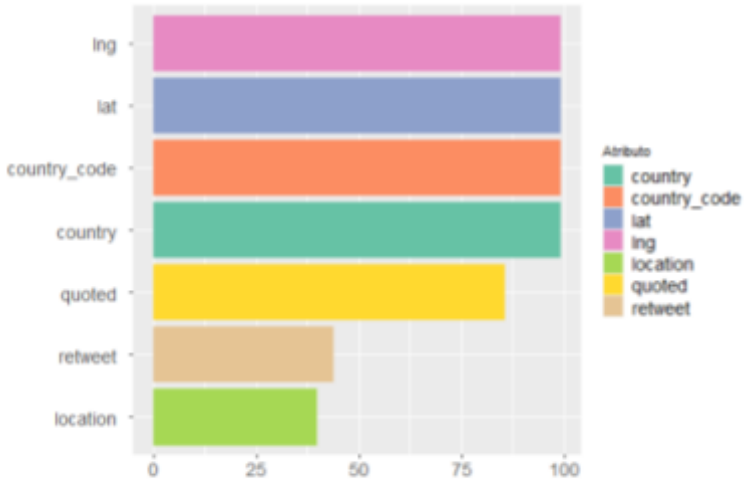

Gráfico 9 - Cantidad de nulos en variables de ubicación

De siete campos relacionado a la georreferencia del tweet. En los siete campos cinco tenían más del 75% de nulos. Location es un campo de dudosa confianza. El usuario tiene mayor libertad para completarlo y puede generar ubicaciones ficticias. Lo mismo pasa con retweet_location, donde además se suma la dificultad de vincular la ubicación del usuario que hizo el tweet y quién lo replicó.

Hubo un esfuerzo por extraer información desde el campo location para la disponibilidad de datos. Construimos dos archivos ‘.csv’, que contenían los países del mundo (En español y en inglés), junto con su continente y código ISO. Esto ayudó a disminuir la cantidad de nulos pero no fue suficiente.

Continente	Frecuencia
África	20
América	7623
Asia	73
Europa	167
Oceanía	4

Gráfico 10 - Tweets por continente

En el gráfico vemos la distribución de tweets según continente. Esto tenía bastante sentido porque el español es la lengua hablada por la mayoría de los países del continente. Pero nos generaba inquietud que decían esos tweets de África o Europa. Manualmente encontramos bastantes errores de matcheo, fuera de américa. Por ejemplo, “Georgia, USA” etiquetado como “Georgia”, o, “Miranda, Venezuela”, como “Iran”. Por otro lado, si nos encontramos algunos datos bien etiquetados, como la embajada de Cuba en Polonia, etiquetada como Polonia.

Esto nos llevó a cambiar de enfoque. *Lo que nos importaba eran entender sobre las noticias y retweets que se corresponden con un país*. La construcción de la realidad se da por la superposición de muchos factores, lo que dicen las propias personas del lugar, medios, políticos, pero también es relevante lo que se habla del país desde afuera.

Nuestro enfoque cambió para ser tweets que mencionan al país dentro de:

- El texto del tweet
- El texto del retweet
- El texto del tweet citado
- La ubicación (Location)
- El campo country

A partir de este criterio armamos vistas en MongoDB sobre: “Argentina”, “Chile”, “Cuba”, “Estados Unidos”, “México”, “Venezuela”. A partir de estos seis subconjuntos de datos repetimos algunas clasificaciones, verificando que se cumplía el mismo patrón.

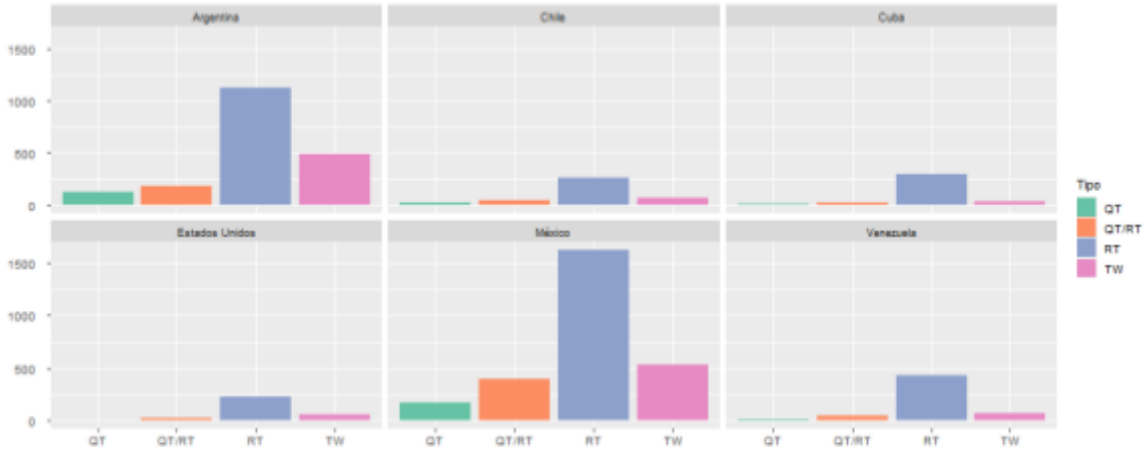

Gráfico 11 - Tipos de tweets por país

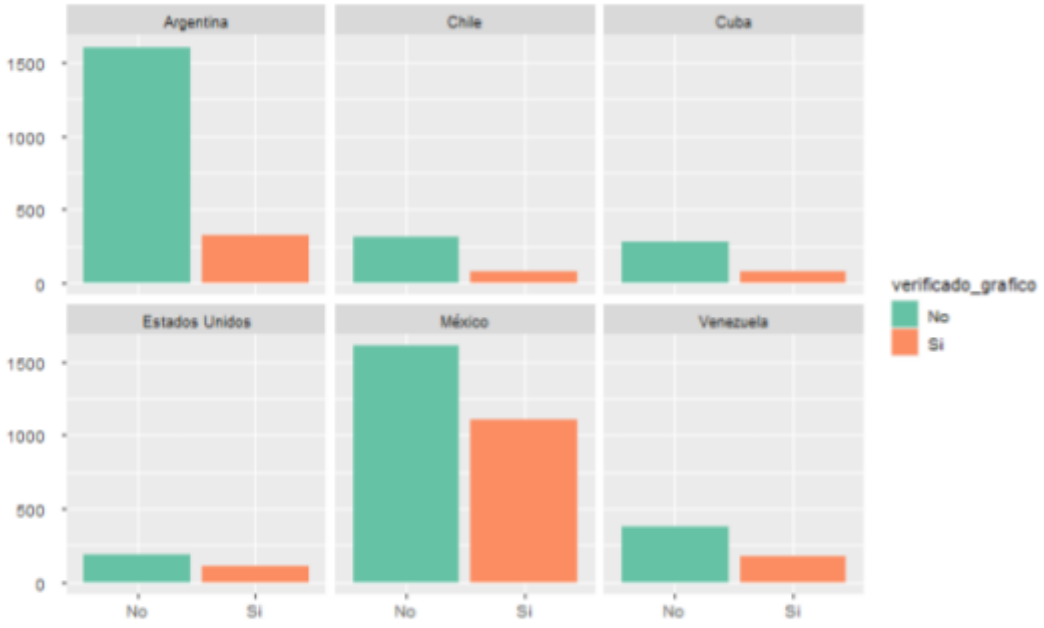

Gráfico 11 - Verificados por país

Podíamos ver la misma distribución, tanto de los tipos de tweets como de la proporción de verificados. Esto representaba una buena señal.

Hubo muchas iteraciones entre hacernos preguntas, seleccionar datos, preprocesarlos y generar algunas respuestas. Si bien tomamos los países como referencias y el filtrado fue por los campos textuales. La pregunta que nos guiaba en el trabajo práctico tenía que ver con quienes creaban tweets, y no con qué decían.

Una primera idea fue ver en una nube de palabras quiénes eran esos usuarios más retweeteados. Por una cuestión de familiaridad mostramos el caso de Argentina, ya que es más familiar para reconocer medios y políticos. A primera vista se ven muchos medios (Clarín, La nación, Todo noticias -TN-, el país), políticos (Alberto Fernandez -Alferdez-, Wolff Waldo, jlespert) y usuarios normales, personas.

Como mención, el usuario “LNV22” es un joven de Buenos Aires que twitteó “Yo aceptando propuestas para salir cuando termine la cuarentena” y obtuvó una gran aceptación en la red social. De igual tenor, es el usuario "AnguitaEXE" que twitteó "el coronavirus arruinó lo que podría haber sido de onda mi mejor año". Esto es algo que tiene cierta frecuencia en la red, porque es habitual que personas muy activas, eventualmente tengan un tweet que se masifique. De la misma forma, ese tweet pasa y continúan siendo usuarios normales.

Gráfico 13 - Nube de palabras de usuarios más retweeteados en Argentina

En el último cuadro, cerrando este apartado, podemos ver el top 10 de usuarios más retweeteados para cada universo. Se encuentran coloreados en rojo medios, y en azul, políticos.

Nro	México	Argentina	Venezuela	Estados Unidos	Chile	Cuba
1	HLGatell	elpais_america	elpais_america	Covid_19Time	elpais_america	GuerreroCuba
2	CiroGomezL	LNV22	GuerreroCuba	GuerreroCuba	XHespanol	DiazCanelB
3	CarlosLoret	XHespanol	ConElMazoDando	CMonteroOficial	alejandramatus	BrunoRguezP
4	elpais_america	todonoticias	NicolasMaduro	la_patilla	gabrielboric	PresidenciaCuba
5	MXvsCORRUPCION	AnguitaExe	WolffWaldo	UniNoticias	joseantoniokast	enplanxd
6	FelipeCalderon	clarincom	Mippcivzla	AliciaCastroAR	IPerezTuesta	luispazos1
7	ClaudioXGG	Lambofgala	la_patilla	BrunoRguezP	SosNinosename1	CubaMINREX
8	m_ebrard	LANACION	VTVcanal8	AlertaNews24	mirnaschindler	Yusnaby
9	XHespanol	alferdez	PartidoPSUV	CNNEE	ninebiker	DrDuranGarcia
10	Milenio	BrunoCHINOperez	AliciaCastroAR	Hispantv	andres20ad	InesMChapman

En los usuarios con más difusión se encuentran medios privados. En segundo muy parejo personas públicas y usuarios normales. Está clasificación se generó buscando palabras claves, dentro de las descripciones de los usuarios.

Algo que nos llama la atención fue la repetición de usuarios hablando de varios países. Lo cual, contextualizando al usuario, tiene sentido. GuerreroCuba, está entre los usuarios mas retweeteados hablando de Cuba, pero también habla de sobre Venezuela y EEUU. El diario “El país” está entre los más retweeteados hablando de paises latinoamericanos.

Venezuela y Cuba tienen muchos más tweets políticos que de los medios, y esto puede relacionarse con la fuerte presencia del estado.

Venezuela también tiene una connotación negativa en el mundo, y se usa como ejemplo de un proyecto de país no deseable, y eso se nota en que ,por ejemplo: “WolffWaldo”, diputado argentino, haya creado muchos tweets donde la menciona. Esto es un comportamiento que se vió en muchos lugares del mundo y favoreciendo cierta polarización de la opinión política. Por ejemplo en las elecciones de 2017 y 2019 en Argentina, cuando Venezuela era un tema muy mencionado.

Conclusión

Uno de los aprendizajes más importantes que obtuvimos al realizar el presente trabajo práctico fue que la naturaleza y las características del dataset determinan fuertemente las preguntas que se le pueden realizar. Gran parte del tiempo invertido fue en entender de qué se trataban las variables, cuales no se podían utilizar, y que nuevas variables se podían crear.

Consideramos que esta situación, estuvo asociada a la característica de multidimensionalidad de los tweets y a las limitaciones en la forma en que fueron capturados los datos.

Así nos percatamos, entre otras cosas, que la dimensión temporal no se podía utilizar, ya que era prácticamente aleatoria en la cantidad de tweets registrados por dia y horario, que localizar geográficamente a un usuario no era una tarea sencilla y requiere inferirse de otra manera ya que muchos de los campos de locación se encontraban vacíos, que algunas variables tenían valores nulos dado el modo de registro, y así fuimos iterando con las diferentes variables de las instancias "tweet" y las de usuario hasta acotar el universo de posibilidades e interiorizandonos con el dataset.

Luego de identificar los distintos tipos de tweets: tweets, retweets y quotes, uno de los principales disparadores del trabajo fue encontrar la gran diferencia que había entre la cantidad de tweets y la de retweets. Esto nos hizo pensar que en el universo twitter eran muchos más los usuarios que retweeteaban que los que tweeteaban, en otras palabras, eran pocos los usuarios que generaban contenido y eran muchos los que lo compartían. También, observamos que los tweets retweeteados correspondian a usuarios verificados.

Se planteó entonces una posible hipótesis que permitiera caracterizar e identificar a estos usuarios, que denominamos "Generadores de contenido", de una manera rápida y sencilla. En base a lo previamente analizado, creamos dos nuevas variables ("Popularidad" y "Actividad") e implementamos las diferentes técnicas de análisis e ingeniería de variables que vimos en clase. Si bien se pudo predecir a los 39 usuarios más "retweeteados" en un universo de 100, lo que nos pareció un buen número, también lo hicieron las otras dos variables utilizadas como control. Creemos que este es un primer paso en la búsqueda de patrones que permitan identificar al perfil deseado. Un primer análisis mostró que la mayoria de los usuarios encontrados corresponden a medios de comunicación, lo que se condice con el perfil de "generador de contenido".

Con respecto al análisis del contenido de los tweets, del estudio de los hashtags surgieron algunos temas de los que se hablaba en medio de la pandemia, tales como, casos de corrupción en México y problemas con los fondos de previsión de algunos países de la región, entre otros. Esto nos llevó a intentar profundizar el análisis en el dato de ubicación y decidimos indagar en los usuarios que hablaban sobre un país. Agrupamos y comparamos este tipo de publicaciones de tweet y entre los usuarios más retweeteados, clasificados en medios/políticos/usuarios normales, observamos que los medios de comunicación eran los generadores de contenido con más alcance. También surgieron unos usuarios clasificados como normales que encontraron un trending topic, y tuvieron un tweet que se viralizó.

Finalmente, el balance del trabajo práctico es muy positivo. Pudimos aplicar y entender de una forma práctica los conceptos de KDD y de las distintas técnicas que vimos en clases. Nos enfrentamos a la complejidad de un dataset real, tuvimos que tomar decisiones que orientarán el trabajo, dar atención a los principales problemas de un preprocesamiento, tratar con valores nulos, datos faltantes, categorizar, integrar los datos para poder procesarlos e intentar generar conocimiento.