Maestría en Explotación de Datos
y Descubrimiento del Conocimiento
Universidad de Buenos Aires

FACULTAD DE CIENCIAS EXACTAS Y NATURALES

FACULTAD DE INGENIERÍA

Data Mining

Trabajo Práctico N° 2

GRUPO N° 14

Carrasco, Lisandro

Opazo, F. Ayelén

Sotelo, Santiago

EQUIPO DOCENTE

Banchero, Santiago

Fernandez, Juan Manuel

Piccoli, Eloisa

Julio, 2021

Emocionalidad y positividad de Spotify en el año de la pandemia

Carrasco, Lisandro; Opazo, F. Ayelén; Sotelo, Santiago

Julio, 2021

RESUMEN

El presente estudio se motivó a partir de resultados obtenidos anteriormente en cuanto a diferencias en features de canciones en situación de pandemia por Covid-19 comparado al 2019. El objetivo principal fue buscar patrones ocultos en datos de Spotify, relacionados a la emocionalidad en dicho contexto y mediante reglas de asociación. Se halló que pese a la disminución de *valence* entre el 2019 y 2020 el vocabulario de las canciones del top 5 en pandemia se asoció en general a emociones positivas, y por el contrario, el vocabulario del 2019 se asoció a términos más negativos; sin embargo, en presencia de *valence* alta en pandemia ocurren letras con emociones negativas. No se encontraron reglas de asociación robustas que indiquen que los temas de *valence* bajo hayan sido muy escuchados durante la pandemia y, por el contrario, hay más confianza para pensar que los temas de 2020 eran más positivos en términos acústicos. Se observó relación entre la cantidad de palabras utilizadas en la canción y la sensación de felicidad, energía y demás features que proyectan sensaciones positivas.
Palabras clave: pandemia, Spotify, emociones, valence

INTRODUCCIÓN

Dominio de aplicación y objetivo general

El presente trabajo se aborda como una continuidad de datos encontrados en el *TP N°1*, más específicamente en los resultados de la tercera pregunta: *¿existen variaciones entre las características presentes en las canciones del top 1 durante el año de la pandemia con respecto al año previo, 2019?* Allí se encontraron algunas diferencias de features en promedio entre ambos años que podrían hacernos sospechar sobre factores emocionales, dado el contexto. El objetivo general de este estudio consiste en buscar patrones ocultos en datos de Spotify, relacionados a la emocionalidad a partir de la pandemia por Covid-19. Se describen, a continuación, los tipos de variables que componen el dataset seleccionado y las metodologías utilizadas para su preprocesamiento. A su vez, se muestran distintas coocurrencias en las características presentes en los datos, determinadas mediante reglas de asociación. Se presentan, además, su robustez y correspondiente análisis. Finalmente, se discuten los hallazgos más interesantes.

Objetivos específicos

Entre las diferencias halladas en el *TP N°1* se destacan las siguientes: entre el año 2019 y 2020, se vio una caída de la positividad (*valence*) de los temas que ocuparon la primera posición del ranking de Spotify. También se encontró un 26% menos de palabras habladas (*speechiness*) en las canciones que ocuparon el top 1 con respecto al 2019. . En función de esta información se abre el rango al top 5 y se definen las siguientes preguntas que guiarán el resto del informe:

1. ¿Pueden asociarse las letras del top 5 de 2019 a un vocabulario emocionalmente más positivo y los temas de 2020 a un vocabulario más negativo?
2. ¿Qué reglas se pueden establecer entre la positividad (*valence*) y su posición en el top 200 de Spotify? ¿Las reglas confirman lo observado en el *TP N°1*?
3. ¿Con qué tipo de canciones se asocian los temas con menor cantidad de palabras en el top5?

DATOS

Selección del dataset - Integración

Para el desarrollo del trabajo se utilizó como base el dataset generado en el trabajo práctico anterior, compuesto por las colecciones de Spotify *charts* y *artist_audio_features_solo_art,* con sus respectivos NA previamente imputados por ***cold-deck*** a través del paquete **'*spotifyr*'**. Ese preprocesamiento se encuentra detallado en aquel informe. A este dataset se le adiciona la colección *lyrics* de más de 6000 documentos mediante un merge similar al realizado entre las otras colecciones. Siguiendo los objetivos planteados con anterioridad se filtran los charts menores a posición 6, es decir, se toma el top 5, y se realiza una agregación por la media de los *features* que se deciden utilizar, para generar una única fila por cada letra de canción, para no analizarla más de una vez. Los datos se agrupan, entonces, por *artist, track_name, album_name, track_id* y *lyrics,* y se obtiene un dataset con las variables reflejadas en la *Tabla 1* (ver ANEXO). El dataset generado contiene 65 instancias con 13 variables. Notar que se trae también *week_star*t (se elige indistintamente la semana, ya que sólo importa el año a los fines de los objetivos). Para ampliar el espectro de lo observado en el top 5, para la segunda pregunta de este trabajo se han utilizado todas las canciones presentes en el top 200, agregadas de la misma manera.

Transformaciones

A partir de las variables definidas se realizan una serie de transformaciones necesarias para responder cada pregunta.

Lyrics. En primer lugar, se genera un corpus de *lyrics* removiendo palabras vacías en inglés (ya que el top 5 contiene mayoritariamente canciones en inglés), espacios en blanco, puntuaciones, y todo lo que no sea alfanumérico. Luego de esto se genera la matriz término-documento binaria, indicando con 1 que el término está presente y con 0 que no está. La matriz obtenida tiene una dimensión 65x3258.

Emociones. Para profundizar en el tipo de vocabulario que contienen las letras del top 5 y vincularlos a emociones se utiliza la librería ***syuzhet*** y su función *get_nrc_sentiment* del léxico Emoción NRC de Mohammad. Según Mohammad, "*el léxico de emociones de NRC es una lista de palabras y sus asociaciones con ocho emociones (ira, miedo, anticipación, confianza, sorpresa, tristeza, alegría y disgusto) y dos sentimientos (negativos y positivos)*"[1]. Para poder utilizar esta función primero se convierte en vector a *lyrics* y se le aplica el idioma a utilizar -inglés-. Se genera una matriz 65x10

[1] *http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm*

donde por cada fila se obtienen puntajes de las distintas emociones y sentimientos según palabras reconocidas en las letras. Luego se transforman estos puntajes a 0 y 1, dejando en 0 aquellas emociones no encontradas en cada letra y 1 a las emociones que se reconocieron al menos 1 vez.

Features - discretización. Para generar las reglas de asociación se precisa trabajar con variables dicotómicas o categóricas, por lo cual los features numéricos se discretizan en labels según la documentación de Spotify[2], otras fuentes externas para interpretar los atributos musicales[3] o en función de su distribución univariada.

- *Danceability* se categorizó en “Baja”, “Media” y “Alta” con valores entre 0, 0.25, 0.5 y 1.
- *Speechiness* entre “Baja”, “Media” y “Alta” con valores 0, 0.33, 0.66 y 1.
- *Acousticness* y *energy* en “Baja” y “Alta”, entre 0, 0.5 y 1.
- *Liveness* se ha dividido por encima y por debajo de 0.8, siendo los temas de valores superiores altamente probables de ser en vivo. Ningún tema ha quedado por encima de 0.8.
- *Duration_ms* se ha dividido en función de los cuantiles de orden 0.33 y 0.66.
- *Tempo* se ha discretizado como “Lento”, “Moderado”, “Alegre” y “Rápido”, con quiebres en los valores 90, 110 y 135.
  En el caso de *valence* se discretiza en “Baja” y “Alta”, entre 0 a 0,5 y 0,6 a 1 respectivamente.
- También se ha incluido una variable *cat_position,* que categoriza la mejor posición alcanzada por cada track siendo ‘Muy alta’ la categoría para las posiciones entre 1 y 50, ‘Media’ entre 51 y 125 y ‘Baja’ entre 126 y 200.
- Se crearon dos variables categóricas cat_cantidad_palabras y cat_longitud a partir de las variables artificiales cantidad_palabras y longitud_palabras respectivamente. Se utilizó el segundo cuartil para definir el umbral entre la categoría “Baja” y “Alta” que se situó en 64 para cantidad_palabras y 5.032 para longitud_palabras.
- Por último, se incluye la variable *cat_año* con 2 valores: “pandemia”, “no pandemia” de acuerdo al criterio de si *week_start* es menor o mayor que la fecha 11/03/2020, fecha en que se declaró la pandemia oficialmente[4].

Finalmente se unen las matrices de *término-documento* y *emociones* y las variables discretizadas con el dataset de partida (ver *Tabla 1*) y se genera un único set de 65x3291, del cual se utilizan los datos de la columna 14 en adelante, dejando de lado los *features* numéricos del dataset original. Luego se genera un nuevo data frame con una columna *TID* por instancia (65 TID en total) y una columna *item* compuesta por los términos de las letras, las emociones y las variables discretizadas, todos con sus respectivos valores y un prefijo según corresponda (“TERM_”, “EMOCION_” y “cat_”). Se muestra un ejemplo en la *Tabla 2* (ver ANEXO). A este dataframe se lo guarda como un archivo *.txt* para poder leerlo posteriormente como un objeto ***transactions*** y generar las reglas de asociación.

RESULTADOS

1. Tipo de vocabulario presente en canciones del año de la pandemia

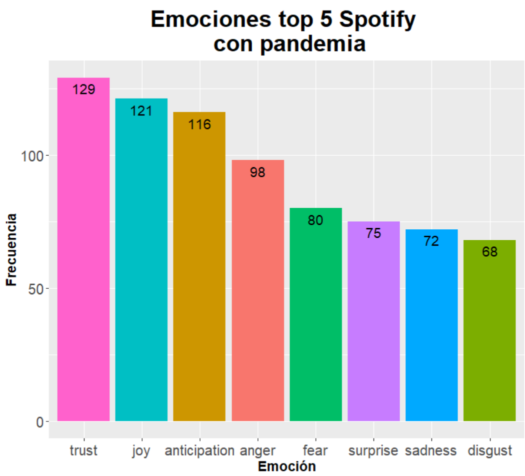
Previo a buscar asociaciones entre vocabulario de canciones del top 5 de Spotify y emociones positivas o negativas, se exploraron los términos más frecuentes de las letras, así como las emociones reconocidas con mayor frecuencia; comparando entre contexto de pandemia y no pandemia. En el wordcloud de la izquierda se observan las palabras más frecuentes con pandemia *(yeah* 115 ocurrencias*, like* 111*, know* 88 *y rain 66*), mientras que en el de la derecha están las

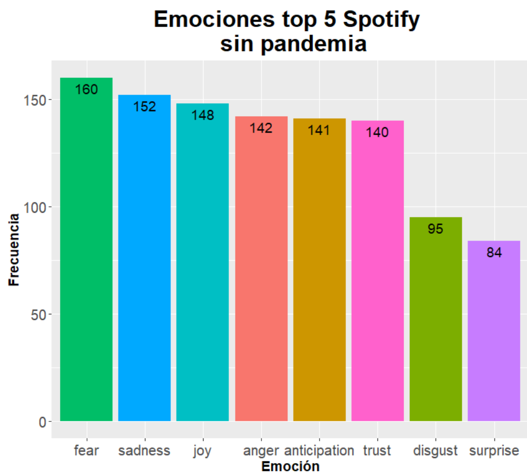
[2] *https://developer.spotify.com/documentation/web-api/reference/#object-audiofeaturesobject*
[3] *https://es.wikipedia.org/wiki/Tempo*
[4] *https://www.who.int/es/news/item/27-04-2020-who-timeline---covid-19*

palabras frecuentes antes de la pandemia (*yeah* 200 veces, *love* 112, *like* 108). A simple vista, se observa bastante similitud entre ambos contextos.

Con pandemia | **Sin pandemia**

Por otra parte, se buscaron las emociones más frecuentes en las letras de canciones escuchadas, dentro del top 5, en los dos contextos. En los gráficos de barras se observan emociones positivas con mayor frecuencia en contexto de pandemia, como *trust, joy* y *anticipation*; mientras que sin pandemia las más frecuentes son *fear, sadness* y *joy*; sin embargo, esta segunda distribución es más uniforme que la primera. Según NCR, además, en ambos años se detectaron mayores casos de sentimientos positivos, siendo 262 términos positivos y 170 negativos con pandemia y 339 positivos y 271 negativos sin pandemia (ver ANEXO).





Se realizó un filtro de ítems de emociones, términos y año o no de pandemia y se probaron distintos parámetros para la elección de las reglas de asociación, definiendo un soporte mínimo de 0.2 y confianza de 0.6. Los resultados indican mayor co-ocurrencia de vocabulario reconocido como positivo en pandemia y de emociones negativas dado el año de no pandemia con mayor soporte, **contrariamente a lo esperado**. Sin embargo, tomando como antecedente el *valence* de los temas se observó que en contexto previo a la pandemia, sus valores altos se asocian a palabras relacionadas a emociones positivas, mientras que en contexto de pandemia un valence alto se asocia a emociones negativas. Las emociones más frecuentes en pandemia (*trust* y *joy*), positivas, acompañadas de de las palabras *know* y *yeah* co-ocurren robustamente con el término *cause*, mientras que en no pandemia y acompañadas de la emoción anticipation co-ocurren con *way*. Las palabras más frecuentes en ambos años (*yeah* y *like*) se asocian a la emoción trust en pandemia, pero no se hallaron asociaciones entre estas palabras y *trust* en no pandemia.

LHS	RHS	S	C	L
{cat_año=pandemia} =>	{EMOCION_positive}	0.2	1	1.41
{cat_año=no pandemia} =>	{EMOCION_negative}	0.57	1	01.03
{cat_año=pandemia,cat_valence=Alta} =>	{EMOCION_negative}	0.23	1	1.3
{cat_año=no pandemia,cat_valence=Alta}=>	{EMOCION_positive}	0.28	1	1.3
{cat_año=pandemia, TERM_like, TERM_yeah} =>	{EMOCION_trust}	0.26	1	1.5
{cat_año=pandemia, EMOCION_joy, TERM_know, EMOCION_trust, TERM_yeah} =>	{TERM_cause}	0.2	0.93	1.72
{cat_año=no pandemia, EMOCION_anticipation, EMOCION_joy, EMOCION_trust} =>	{TERM_way}	0.2	0.76	1.91

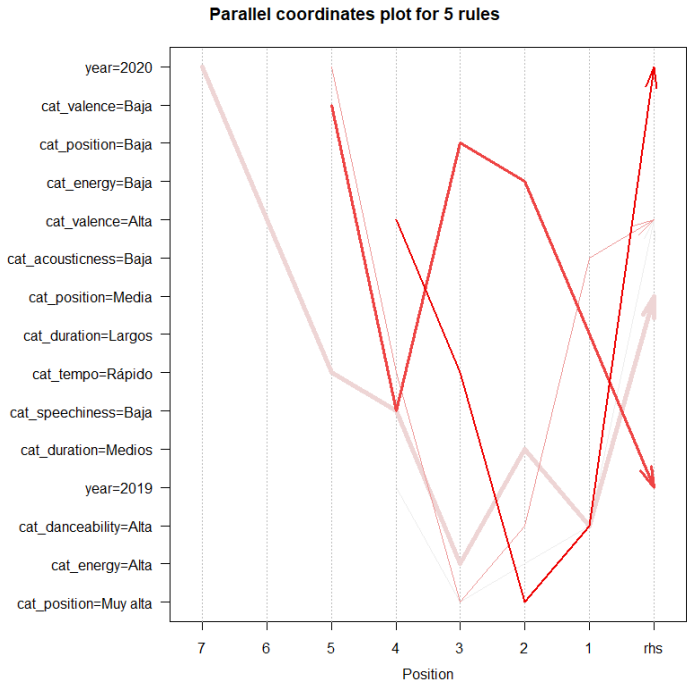
2. Asociación entre *valence* y posición

Para evaluar la relación entre el atributo que Spotify define como positividad y sus vínculos con otros atributos musicales, pero principalmente con la posición que ocuparon en los rankings en los diferentes momentos evaluados, se amplió el espectro de investigación a todos los temas que hayan integrado cualquier posición del top 200. En este caso, no se han utilizado las letras de las canciones, para complementar el enfoque de texto de la pregunta anterior.

Bajo el algoritmo a priori, con un soporte mínimo de 1% y una confianza de 50%, se generaron 91.458 reglas. Estas reglas fueron inspeccionadas filtrando en detalle por la aparición de las diferentes categorías de *valence* tanto en el antecedente como en el consecuente, incluyendo también otros atributos en el subseteo, tales como *posición*, *año* y *bailabilidad*.
Las reglas seleccionadas para esta pregunta se presentan en la siguiente tabla:

LHS	RHS	S	C	L
{cat_acousticness=Baja, cat_danceability=Alta, cat_position=Muy alta, cat_tempo=Rápido, cat_año=pandemia}	{cat_valence=Alta}	0.010	0.77	1.61
{cat_danceability=Alta, cat_energy=Alta, cat_position=Muy alta,cat_año=no pandemia}	{cat_valence=Alta}	0.010	0.56	1.18
{cat_danceability=Alta, cat_duration=Medios, cat_energy=Alta, cat_speechiness=Baja, cat_tempo=Rápido, cat_valence=Alta,cat_año=pandemia}	{cat_position=Media}	0.013	0.64	1.3
{cat_energy=Alta, cat_speechiness=Baja, cat_tempo=Rápido, cat_valence=Baja, explicit,cat_año=pandemia}	{cat_position=Baja}	0.012	0.5	1.24
{cat_danceability=Alta, cat_duration=Breves, cat_energy=Alta, cat_speechiness=Baja, cat_valence=Baja, explicit, cat_año=no pandemia}	{cat_position=Baja}	0.012	0.59	1.45

Evaluando los temas que han estado en posiciones muy altas del ranking, comparando al año de la pandemia con el año anterior, se puede decir con mayor confianza para 2020 que para 2019 que los temas que **ocuparon una posición entre las 50 primeras tienen una positividad alta**. Hay una diferencia notable en las confianzas generadas por estas reglas en función de los años.



Más allá de esta observación, otra regla señala con un 0.64 de confianza que si un tema de 2020 tuvo positividad y bailabilidad alta, entre otros atributos, es más probable que ocupe una posición media en el top 200 que una posición muy alta.

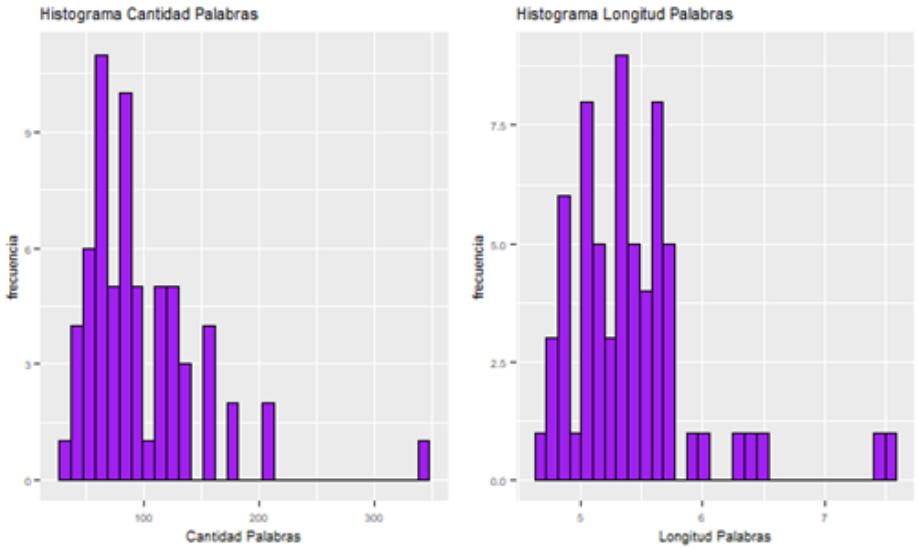
En otro aspecto, las reglas señalan que, a pesar de haber observado una baja en la media de los temas más escuchados en 2020 en relación a los de 2019, no se puede decir con confianza que durante 2020 se hayan escuchado temas necesariamente “pocos positivos”. De hecho, las reglas señalan que canciones con baja positividad y lenguaje explícito son más propensas a quedarse en posiciones bajas del ranking, sea en el año que sea. De igual manera, estas reglas son menos robustas que las anteriores, dado que su confianza es muy cercana al umbral mínimo establecido de 0.5

3. Características de las canciones menos habladas del Top 5

En virtud de lo observado respecto de la merma de la cantidad de palabras en las canciones del top 1 se procedió a analizar las características de las canciones que tuvieron menos palabras ampliando el scope del trabajo al top 5. Con el propósito de agregar valor la información del dataset, se creó un campo con la cantidad de palabras de cada canción y otro para analizar la longitud promedio de cada palabra que componen los temas, de forma de poder captar la complejidad del vocabulario utilizado. Una vez calculados se calcularon sus medidas descriptivas. Se observa una leve asimetría positiva en ambas variables.

Indicadores	cantidad_palabras	Longitud Palabra
Promedio	97.63	5.41
Minimo	35.00	4.66
1er Cuartil	64.00	5.03
Mediana	82.00	5.35
3er Cuartil	121.00	5.61
Maximo	347.00	7.52
Desvio	51.52	0.54
Coeficiente de Variacion	0.53	0.10

En ambos casos se tomó el primer cuartil como el corte para caracterizar la categoría de cantidad baja de palabras y longitud promedio baja. Se crearon variables categóricas para poder sumarlas al análisis de las reglas. Luego se analizó la diferencia entre los features de las canciones menos habladas con las demás.



variable	danceability	acousticness	energy	speechiness	tempo	valence	streams	Longitud
Cantidad_altas	0.7420	0.1510	0.6240	0.1410	120.0260	0.5495	30,891,923	5.4292
Cantidad_bajas	0.6310	0.3710	0.4880	0.0598	101.9930	0.3760	28,788,883	5.0317
Variacion	-15%	146%	-22%	-58%	-15%	-32%	-7%	-7%

Dentro de las variaciones más importantes, se observó que en el año de la pandemia las canciones menos habladas fueron un 15% menos bailables, 146% más acústicas, 22% con menos enérgicas, con un tempo 15% más lento y 32% de menor *valence*. En menor medida se observa que las canciones tienen palabras con una longitud promedio y cantidad de streams más bajas.

Es decir, hay **un principio de evidencia que dentro de las canciones menos habladas hay una leve merma en la complejidad del vocabulario utilizado y del interés de la audiencia**. A su vez, estos temas son más acústicos, razonablemente menos bailables y proyectan una menor energía. Como consecuencia de ser más acústicos, tienen un tempo menor, es decir el ritmo de las canciones es más despacio. Y naturalmente, al ser temas acústicos, son más íntimos y menos enérgicos, las canciones proyectan menos felicidad o euforia, hecho que se ve reflejado con un menor *valence*.Las reglas seleccionadas para esta pregunta se presentan en la siguiente tabla:

LHS	RHS	S	C	L
{cat_cantidad_palabras=baja,cat_energy=Baja}	{cat_acousticness=Alta}	0.10	0.67	2.89
{cat_cantidad_palabras=baja}	{cat_energy=Baja}	0.14	0.53	1.5
{cat_cantidad_palabras=baja}	{cat_valence=Baja}}	0.17	0.65	1.36
{cat_cantidad_palabras=baja}	{cat_danceability=Alta}	0.23	0.63	1.14
{cat_cantidad_palabras=baja,cat_energy=Baja,TERM_love}}	{cat_longitud=Baja}	0.11	0.7	2.84
{cat_cantidad_palabras=baja}	{cat_streams=Baja}	0.15	0.59	1.16
{cat_cantidad_palabras=baja}	{cat_tempo=Baja}	0.17	0.65	1.31

DISCUSIÓN

Aunque el análisis exploratorio del trabajo anterior presentó una disminución notable de algunos aspectos que pueden considerarse positivos en términos de las emociones que transmite la música, esas conclusiones no se vieron reforzadas por los métodos de text mining y de reglas de asociación a lo largo de esta investigación.

De hecho, algunas reglas de asociación y el análisis de emociones en el texto pueden sugerir que hay condiciones para decir lo opuesto a nuestro argumento inicial. Sin embargo, creemos que no se puede descartar tajantemente la conclusión del trabajo anterior y asegurar algo diferente.

Si bien esta investigación arrojó resultados interesantes en términos de exploración de las emociones, consideramos que la robustez de este método en este caso puede no ser suficiente para hacer aseveraciones sobre reflejos de un estado de ánimo colectivo.

Por el contrario, consideramos que hay aportes valiosos a tener en cuenta tanto desde el análisis de las features musicales como desde el enfoque de text mining y que ambos métodos son valiosos para tener un panorama más amplio y extender el conocimiento sobre el conjunto de datos en cuestión. Principalmente, haber complementado ambos enfoques nos llevó a encontrar que muchas canciones asociadas a un *valence* alto (positividad en el sonido musical definido por Spotify), pueden incluir letras de emociones negativas. Es decir, lo que el modelo de Spotify evalúa sobre las cuestiones acústicas de un tema, no necesariamente van de la mano con lo que sus letras expresan. El hecho de que durante la pandemia haya habido temas posicionados en el top 5 con una medida de *valence* alta, pero con emociones positivas en sus letras, puede ser una señal que refuerce el pesimismo expresado en la hipótesis original. Por último, es posible resaltar que la cantidad de palabras que componen la letra de una canción está relacionada con los features y a menor valor de ellas una percepción más íntima y acústica y menos feliz ha sido evidenciada en el presente estudio.

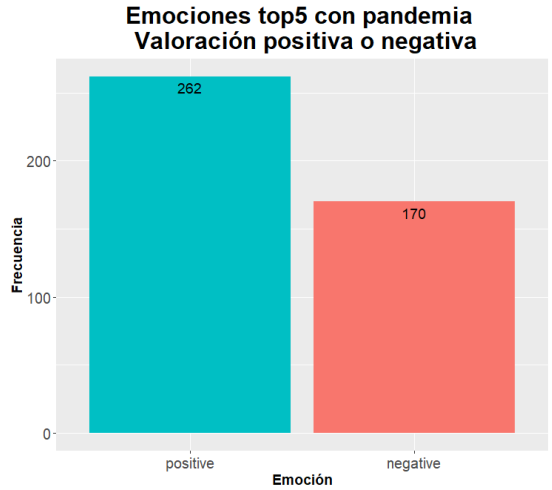
ANEXO

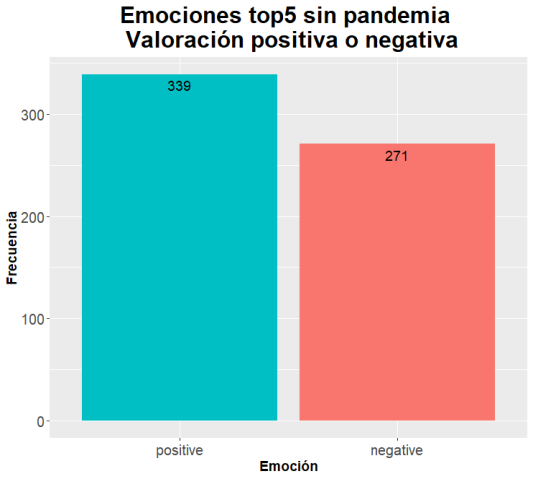
artist	track_name	album_name	lyrics	danceability	energy	tempo	liveness	speechiness	valence	streams	position	week_start
24kGoldn	pod (feat. iann di	Mood (feat. iann dior	Mood Lyrics	0.700	0.722	66234	0.2720	0.369	0.7560	33935925	1	2020/08/12
6ix9ine	GOOBA	GOOBA	GOOBA Lyrics	0.611	0.688	78126	0.2510	0.3410	0.3930	30772787	4	2020/05/23
Ariana Grande	7 rings	thank u, next	7 rings Lyrics	0.778	0.317	83288	0.881	0.3340	0.3270	40159840	1	2019/09/04

Tabla 1. 3- head dataset agrupado por medias de features

TID	item
44	cat_año=pandemia
23	TERM_baby
5	EMOCION_joy

Tabla 2. DF con los item presentes en cada TID





REFERENCIAS

Jockers, Matthew L. *Syuzhet: "Extract Sentiment and Plot Arcs from Text"*, 2015. https://github.com/mjockers/syuzhet

Jockers, Matthew L. *"Introduction to the Syuzhet Package"*, CRAN R Project, 2017.