Análisis de Sentimientos en Publicaciones de Twitter Durante la Pandemia

Roberto Osciel Romero Obispo

8 de mayo de 2024

Resumen

Durante la última semana de enero de 2021, un periodo aún marcado por intensos desarrollos en la pandemia de COVID-19, las redes sociales ya eran un vehículo crucial para que las personas expresaran sus emociones y sentimientos. Este reporte técnico aborda un estudio sobre estas emociones y sentimientos manifestados en publicaciones de Twitter (ahora X). Se analiza cómo, a través de diferentes herramientas tecnológicas como R, se pueden obtener variados resultados en el análisis de sentimientos. Además, se destaca la personalización de diccionarios léxicos, asignando significancia a cada palabra para determinar si es positiva o negativa, lo cual permite adaptar el análisis a necesidades específicas. Este enfoque no solo revela las emociones predominantes entre los usuarios, sino que también refleja la diversidad de percepciones y reacciones ante la pandemia, evidenciando el impacto significativo de la personalización en la interpretación de datos de redes sociales.

Palabras clave: *análisis de sentimientos, análisis de emociones, procesamiento de lenguaje natural, Twitter, Covid-19*

1. Introducción

1.1. Motivación

Este reporte técnico explora un estudio exhaustivo sobre las emociones y sentimientos expresados en publicaciones de Twitter, específicamente durante la última semana de enero de 2021, un periodo marcado por intensos desarrollos en la pandemia de COVID-19. Se recopilaron tweets entre los días 21 y 28 de enero, capturando así un espectro representativo de la opinión pública en un momento crítico de la crisis sanitaria global.

La metodología de este estudio se centró inicialmente en un riguroso proceso de limpieza y preparación de datos para asegurar la calidad y relevancia del conjunto de datos analizado. Posteriormente, se llevó a cabo un análisis de sentimientos empleando varios enfoques para comparar resultados. Se utilizó un diccionario léxico personalizado junto con el Léxico ML-SentiCon, y se exploraron también los léxicos incluidos en paquetes de R como SentimentAnalysis y syuzhet. Esta estrategia multidimensional permitió obtener una variedad de perspectivas sobre los sentimientos expresados, enriqueciendo así la comprensión de las emociones manifestadas en las publicaciones durante el periodo estudiado.

Este análisis no solo subraya la utilidad de las redes sociales como herramientas de mapeo emocional en tiempos de crisis, sino que también destaca la importancia de entender las dinámicas sociales y emocionales que influencian la respuesta pública ante eventos globales significativos.

1.2. Objetivo

El objetivo principal del estudio fue identificar y cuantificar las principales emociones reflejadas por los usuarios de Twitter (X), proporcionando así una visión única de la psique colectiva durante un periodo de incertidumbre y cambio constante.

2. Selección y extracción de texto

En el análisis, se exploran diversos estilos expresivos, desde textos de opinión hasta críticas. Cada estilo posee una carga emocional inherente, que varía en intensidad. La metodología para analizar estos textos depende de si están clasificados según las emociones que expresan. Dicha clasificación puede ser básica, donde se asigna un valor de 1 para sentimientos positivos, 0 para neutros y -1 para negativos. Alternativamente, puede ser más detallada, identificando emociones específicas como tristeza, enfado o felicidad, y asignando a cada una una probabilidad de estar presente en el texto.

2.1. Filtrado y acondicionamiento

El preprocesamiento de lemguaje natural implica la limpieza y preparación de los datos obtenidos para adaptarlos a las necesidades específicas de un experimento. Este proceso utiliza diversos mecanismos que ayudan a optimizar el conjunto de datos para su análisis posterior. Algunos de estos mecanismos utilizados fueron:

- **Tokenización:** Consiste en dividir el texto en entidades llamadas tokens, con las que trabajaremos posteriormente. Además, se puede llevar a cabo la eliminación de signos de puntuación innecesarios.
- **Eliminación de Stop words:** Este proceso se refiere a la remoción de términos como 'de', 'el' e 'y', los cuales son frecuentemente utilizados pero no contribuyen significativamente al significado semántico del texto.
- **Conversión de Mayúsculas en Minúsculas:** Esta técnica se utiliza para conseguir una mayor simplicidad en el texto.

3. Implementación

Este capítulo tiene como objetivo presentar la metodológica al análisis de emociones y sentimientos. Para ello, describiremos en detalle los pasos conceptuales que hemos adoptado, las herramientas empleadas en el proceso (códigos de R), y los criterios que influyeron en la elección o rechazo de determinados métodos analíticos.

3.1. Base de datos

La base de datos utilizada para el análisis de sentimientos (*twitter.csv*) contiene varios registros de tweets recopilados entre el 21 y el 28 de enero de 2021. Estos datos fueron proporcionados por el Departamento de Modelación Matemática de Sistemas Sociales del Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas.

3.2. Preprocesamiento de los datos

Previo al proceso de limpieza, los datos se presentan de la siguiente manera:

1.35455E+18	VTVcanal8	2021-01-27-09:58:42	#EnVideo ☎ | Dip. @RicardoSanchezX informó que la Comis...
#VenezuelaSolidariaYHumanista https://t.co/1VmtLvBTii	es	null	null
1.35411E+18	InesArrimadas	2021-01-26-05:00:04	Un antiviral español reduce casi al 100% la carga viral del #...
https://t.co/omg6hUnML3	es	null	null
1.3543E+18	SinEmbargoMX	2021-01-27-05:15:00	La ciencia estalla en júbilo: dan con medicamento que reduc...
1.35471E+18	fmcuber	2021-01-28-08:51:16	La ivermectina es barata
1.35471E+18	luisherreramesa	2021-01-28-08:50:27	Un antiviral fabricado en España demuestra su eficacia contr...
1.35471E+18	Freebir92458311	2021-01-28-08:49:17	@FactaNverba1 @janogarcia_ https://t.co/HvBCRLqXJa
Responde mejor de lo esperado	es	null	null
1.35471E+18	Artemisa_Kc	2021-01-28-08:47:14	Avanza una nueva vacuna contra el COVID-19 con nanopartí...
1.35471E+18	lideresSC	2021-01-28-08:46:23	Presidente Duque: "La medicina y la ciencia son las que sanan
1.35471E+18	ElNacionalWeb	2021-01-28-08:45:00	Covid-19: estos son los posibles nuevos síntomas de la enfe...
1.35471E+18	JuandeBarte	2021-01-28-08:44:44	@felishuco1975 @Dani_Pellicer @mascaramarilla @emulen...
Aquí @LluisMontoliu lo explica muy bien.	es	null	null

Figura 1: Datos sin limpiar

Hemos observado que la base de datos contiene numerosos espacios en blanco y varias celdas llenas con el término "null". Para mejorar la organización de los datos, renombraremos la primera columna como ÏD", la segunda como ÜSERNAMEz la tercera como "DATE". Además, combinaremos varias columnas subsiguientes para consolidar la información principal de cada publicación. A este conjunto de columnas combinadas lo denominaremos "TWEET".

ID	USERNAME	DATE	TWEET
1.35455E+18	VTVcanal8	2021-01-27 09:58:42	#EnVideo ☎ | Dip. @RicardoSanchezX informó que la Comis...
1.35411E+18	InesArrimadas	2021-01-26 05:00:04	Un antiviral español reduce casi al 100% la carga viral del #...
1.3543E+18	SinEmbargoMX	2021-01-27 05:15:00	La ciencia estalla en júbilo: dan con medicamento que reduc...
1.35471E+18	fmcuber	2021-01-28 08:51:16	La ivermectina es barata segura y eficaz: recetarla contra el ...
1.35471E+18	luisherreramesa	2021-01-28 08:50:27	Un antiviral fabricado en España demuestra su eficacia contr...
1.35471E+18	Freebir92458311	2021-01-28 08:49:17	@FactaNverba1 @janogarcia_ https://t.co/HvBCRLqXJa
1.35471E+18	Artemisa_Kc	2021-01-28 08:47:14	Avanza una nueva vacuna contra el COVID-19 con nanopartí...
1.35471E+18	liderесSC	2021-01-28 08:46:23	Presidente Duque: "La medicina y la ciencia son las que san...
1.35471E+18	ElNacionalWeb	2021-01-28 08:45:00	Covid-19: estos son los posibles nuevos síntomas de la enfe...
1.35471E+18	JuandeBarte	2021-01-28 08:44:44	@felishuco1975 @Dani_Pellicer @mascaramarilla @emulen...
1.35471E+18	PacoMerlo8	2021-01-28 08:41:17	https://t.co/aqvk8agSCr Ciencia El covid-19 también afecta ...
1.35471E+18	fduran_ve	2021-01-28 08:40:19	fduran_ve: RT @IVIC_oficial: #62Aniversario || El lunes 01 de ...
1.35471E+18	BiofilosL	2021-01-28 08:38:41	La marihuana reduce riesgo de contagiarte de Covid-19 se...
1.35471E+18	elconfidencial	2021-01-28 08:38:24	Una duda que aún está por despejar es si la temperatura af...

Figura 2: Primera limpieza de datos

La columna correspondiente a la publicación (TWEET) incluye menciones, hashtags, palabras mal escritas y emojis. En esta etapa, es crucial emplear expresiones regulares, que nos facilitarán identificar y modificar los patrones que deseamos ajustar o eliminar. Utilizaremos el paquete "stringr"de R, aprovechando su función str_remove_all(), para llevar a cabo la limpieza de estos datos.

- **Eliminación de URL's**

```
str_remove_all(tweet, "https?://\\S+|www\\.\\S+")
```

  Los enlaces no sirven en nuestro análisis.

- **Eliminación de hashtags**

```
str_remove_all(tweet, "#\\S+")
```

  En las redes sociales, los hashtags consisten en una serie de caracteres precedidos por el símbolo de la almohadilla (#), sin incluir espacios. Se emplean para identificar y categorizar temas específicos.

- **Eliminación de menciones**

```
str_remove_all(tweet, "@\\w+")
```

  Las menciones se componen de una arroba (@) seguida del nombre de usuario. Al ser incluidas en un tweet, notifican al usuario mencionado. En este análisis, se considera que el nombre de usuario también funciona como una palabra vacía (stopword).

- **Eliminación de carácteres especiales y signos de puntuación**

```
str_remove_all(tweet, "[^\\w\\s]+")
iconv(tweet, to = 'ASCII//TRANSLIT')
```

  Es esencial llevar a cabo esta etapa después de eliminar URLS, menciones y hashtags, porque estos elementos incluyen signos de puntuación que son cruciales para identificarlos correctamente.

- **De mayúsculas a minúsculas**

```
tolower(tweet)
```

  Es recomendable tener todas las palabras en minúsculas (o mayúsculas), pues palabras cómo Çovidçon çovid"podrían ser diferentes.

- **Eliminación de duplicidad de letras**

```
str_replace_all(tweet, pattern = "([a-z])\\1{2,}", "\\1")
```

  Estandarizamos la duplicidad de las palabras, pues análizar palabras como "vaaaaamos"sería innecesario. En este ejemplo, solo buscamos analizar la palabra "vamos".

- **Eliminación de palabras similares a COVID**

```
str_remove_all(tweet, "\\s(covid|covid19|covid-19)\\s")
```

  Dado las fechas de origen de nuestros datos, es necesario ir quitando palabras relacionadas con çovid","sarscov2", çovid-19", etc.

Tras aplicar la función programada para la limpieza de la columna TWEET, la presentación de nuestros datos se muestra de la siguiente manera:

ID	USERNAME	DATE	TWEET
1.35455E+18	VTVcanal8	2021-01-27 09:58:42	dip informo que la comision de ciencia tecnologia e innovac...
1.35411E+18	InesArrimadas	2021-01-26 05:00:04	un antiviral espanol reduce casi al la carga viral del una gran...
1.3543E+18	SinEmbargoMX	2021-01-27 05:15:00	la ciencia estalla en jubilo dan con medicamento que reduc...
1.35471E+18	fmcuber	2021-01-28 08:51:16	la ivermectina es barata segura y eficaz recetarla contra eles...
1.35471E+18	luisherreramesa	2021-01-28 08:50:27	un antiviral fabricado en espana demuestra su eficacia contr...
1.35471E+18	Artemisa_Kc	2021-01-28 08:47:14	avanza una nueva vacuna contra elcon nanoparticulas que n...
1.35471E+18	lideresSC	2021-01-28 08:46:23	presidente duque la medicina y la ciencia son las que sanan ...
1.35471E+18	ElNacionalWeb	2021-01-28 08:45:00	covid estos son los posibles nuevos sintomas de la enferme...
1.35471E+18	PacoMerlo8	2021-01-28 08:41:17	ciencia eltambien afecta al reloj del juicio final estamos a se...
1.35471E+18	fduran_ve	2021-01-28 08:40:19	fduran_ve rt el lunes de febrero comenzamos nuestro ciclo ...
1.35471E+18	BiofilosL	2021-01-28 08:38:41	la marihuana reduce riesgo de contagiarte desegun estudio ...
1.35471E+18	elconfidencial	2021-01-28 08:38:24	una duda que aun esta por despejar es si la temperatura afe...
1.35471E+18	CaraotaDigital	2021-01-28 08:38:03	bill gates se sorprende de las locas y malvadas teorias cons...
1.35471E+18	mmMonterrubio	2021-01-28 08:38:00	conciencia social pruebas masivas tecnologia punta el secre...

Figura 3: Segunda limpieza de datos

3.2.1. Busqueda y eliminación de StopWords

Es fundamental recordar que las palabras vacías (stopwords) no contribuyen a nuestro análisis, por lo que es necesario eliminarlas de nuestra columna TWEET. Para ello, hemos utilizado un vector personalizado de stopwords junto con el vector de stopwords en español del paquete *tm* de R. Adicionalmente, hemos tokenizado la columna TWEET utilizando el paquete tidytext de R, lo que nos permite contabilizar cada palabra presente en cada registro. De esta manera, podemos identificar manualmente las stopwords que se repiten con mayor frecuencia. Esto nos permite obtener un tercer vector de stopwords que surgieron de nuestra base de datos.

```
> stop_words_es_3
 [1] "un"         "covid"      "19"         "coronavirus" "covid19"    "sars"       "100"
 [8] "cov"        "min"        "covid_19"   "topbqxf65a"  "dr"         "ee.uu"      "r422zkupwh"
[15] "2021"       "d"          "erik"       "dos"         "nelly"      "va"         "ola"
[22] "nqsxqasxdi" "uat3gkhiiu" "csic"       "21"          "4"          "2020"       "español"
[29] "españa"     "gabrielasjr" "vmp0zajvme" "oms"        "sino"       "cubana"     "paris"
[36] "2°"         "cpc"        "3yuggmwqvj" "Covid-19"    "sarscov"
```

Figura 4: Vector de stopwords utilizando las palabras que más se repiten

Al eliminar todas las stopwords obtenidas de nuestra columna TWEET, las palabras que más se repiten en nuestra base de datos se muestran en la Figura 5.

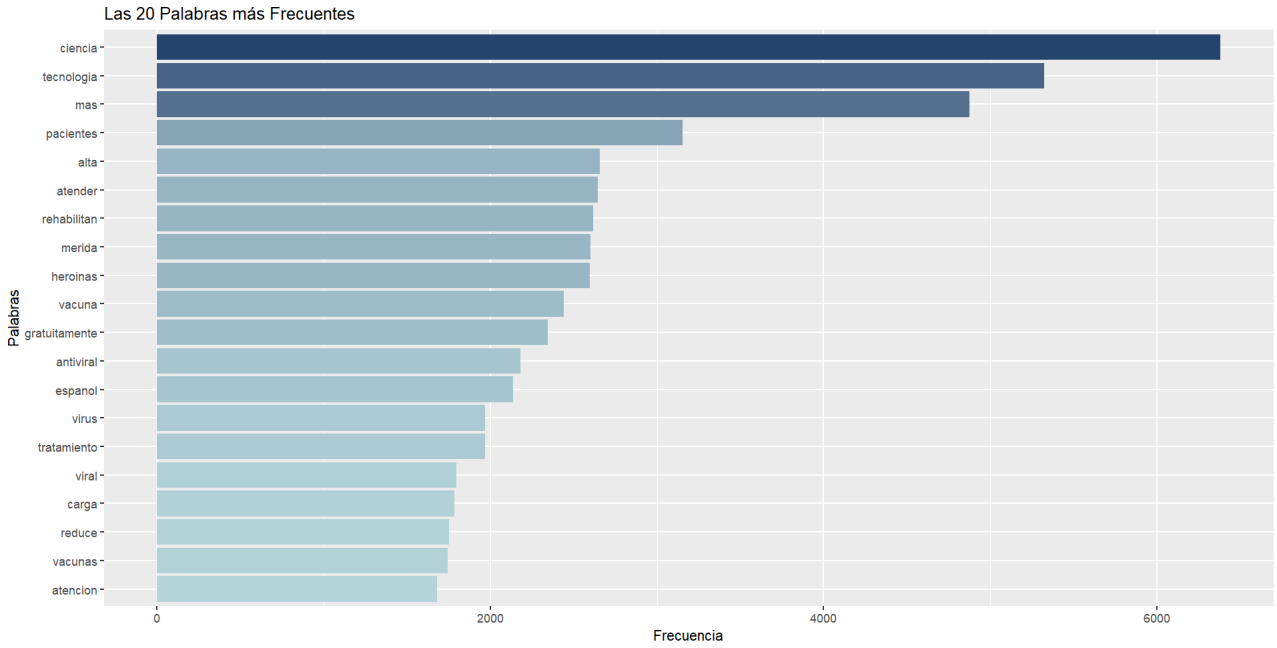

Figura 5: Palabras más frecuentes sin considerar stopwords

El proceso por el cual pasaron nuestros datos puede verse en el Cuadro 1.

Tipo	**Datos**
Tweet completo	Un antiviral fabricado en España demuestra su eficacia contra el coronavirus — Diario de Navarra https://t.co/wTnoO1cpB4
Limpieza 1	un antiviral fabricado en espana demuestra su eficacia contra el coronavirus diario de navarra
Tokens limpios	(un,antiviral,fabricado,en,espana,demuestra,su, eficacia,contra,el,coronavirus,diario,de, navarra)
Tweet sin stopwords	antiviral fabricado demuestra eficacia diario navarra

Cuadro 1: Ejemplo de Procesamiento de Tweet

3.3. Listados de léxicos

Hemos completado el preprocesamiento de los datos para mejorar su formato. El siguiente paso consiste en etiquetarlos según el sentimiento que expresan. Para lograr esto, desarrollaremos un sistema de etiquetado basado en el uso de léxicos. Utilizaremos una función matemática que combine la información de estos léxicos para asignar una etiqueta que refleje la polaridad asociada a cada sentimiento.

- **Léxico ML-SentiCon:** Disponemos de un conjunto de léxicos de polaridades semánticas para lemas en inglés, castellano, catalán, gallego y euskera. Estos léxicos se han desarrollado mediante un método mejorado basado en el utilizado para crear SentiWordNet. Este enfoque se estructura en capas: cada léxico consta de ocho capas, que están organizadas secuencialmente desde la primera hasta la octava. Las capas sucesivas incorporan todos los lemas de las capas anteriores y agregan nuevos lemas.

  En la Figura 6 se observa que, si las palabras en la columna TWEET de nuestra base de datos no coinciden con las del léxico, serán clasificadas como neutrales.

Sentimiento	Cantidad
Positivo	5568
Negativo	5974
Neutral	0

Figura 6: Distribución del léxico ML-SentiCon

Cálculo de polaridad

Dado que no tendremos ayuda de alguna función de un paqute en especifico para el cálculo de polaridad, lo que haremos será:

$$Polaridad\ \ texto = \sum Palabras\ \ positivas - \sum Palabras\ \ negativas$$

Si la polaridad del texto es positiva, podremos decir que, bajo el *léxico ML-SentiCon*,fue un tweet positivo. Del mismo modo, si la polaridad del texto fue negativa, podremos decir que, bajo el *léxico ML-SentiCon*, fue un tweet negativo. Por último, si la polaridad es igual a cero, se trata de un tweet neutral.

- **Diccionario_copalab**. Este léxico fue proporcionados por el Departamento de Modelación Matemática de Sistemas Sociales del Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas. Podemos observar que solo se consideran 10 palabras dentro de este léxico (6 palabras con polaridad positiva y 4 palabras con polaridad negativa)

Palabra	Puntuación	Word
avalen	2	endorse
efectividad	2	effectiveness
probar	2	prove
proteger	1	protect
éxito	2	success
ánimo	2	encouragement
decaiga	-2	decay
nuestras hermanas	-1	our sisters
desgraciadamente	-2	unfortunately
enfrentando	-2	facing up

Figura 7: Distribución del Diccionario_copalab

Cálculo de polaridad

Dado que no tendremos ayuda de alguna función de un paqute en especifico para el cálculo de polaridad, lo que haremos será:

$$Polaridad\ \ texto = \sum Palabras\ \ positivas - \sum Palabras\ \ negativas$$

Si la polaridad del texto es positiva, podremos decir que, bajo el *Diccionario_copalab*,fue un tweet positivo. Del mismo modo, si la polaridad del texto fue negativa, podremos decir que, bajo el *Diccionario_copalab*, fue un tweet negativo. Por último, si la polaridad es igual a cero, se trata de un tweet neutral.

- **Paquetería SentimentAnalysis**. La función *analyzeSentiment()* utiliza varios diccionarios léxicos para evaluar el sentimiento de los textos. Por defecto, esta función emplea el léxico Harvard IV-4 y el Lasswell dictionary, que están incluidos en el paquete. Además, también utiliza el léxico de General Inquirer, que es una combinación de varios diccionarios y clasificaciones de palabras según diversas categorías emocionales y semánticas.

  Estos léxicos contienen palabras clasificadas en categorías como positivas, negativas, fuertes, pasivas, entre otras, lo cual permite una evaluación detallada y contextual del sentimiento en los textos analizados.

- **Paquetería syuzhet.** La función *get_nrc_sentiment()* del paquete "syuzhet.[en] R utiliza el léxico NRC Emotion Lexicon, desarrollado por Saif Mohammad y Peter Turney. Este léxico, también conocido como NRC Word-Emotion Association Lexicon, clasifica las palabras en diez diferentes emociones y sentimientos: confianza, miedo, anticipación, sorpresa, tristeza, alegría, disgusto, enojo, positividad y negatividad.

3.4. Analisis de Sentimientos y emociones

Conforme a los léxicos y paquetes descritos en la subsección anterior, se procederá a analizar nuestra base de datos, específicamente la columna TWEET, para determinar cuál léxico o función ofrece resultados más coherentes o se ajusta mejor a nuestras necesidades.

- **ML_SentiCon.** Para utilizar un léxico propio, primero debemos tokenizar nuestros datos ya limpios. Esto implica descomponer el texto en palabras individuales, lo que nos permitirá comparar cada palabra con las que se encuentran en nuestro léxico. Para hacer esto, nos ayudaremos de paqueterias como *tidytext* y *dplyr*, donde el argumento *diccionario* es el léxico correspondiente:

```
tweets_unnested <- datos_limpios %>%
  unnest_tokens(word, TWEET_CLEAN) %>%
  inner_join(diccionario, by = c("word" = "Palabra"))
```

  Posteriormente, el calculo de la polaridad se realizará sumando el valor que obtuvo cada palabra del tweet, y se agrupará gracias al ID que tiene cada tweet.

```
tweets_polarity <- tweets_unnested %>%
  group_by(ID) %>%
  summarise(polarity = sum(Polaridad))
```

  Tanto en la figura 8 como en el Cuadro 2 podemos ver como se ajustaron los sentimientos a este léxico propuesto:

Positivos	Negativos	Neutral
13919	3435	14605

Cuadro 2: Resultados de análisis de sentimientos con ML_SentiCon.

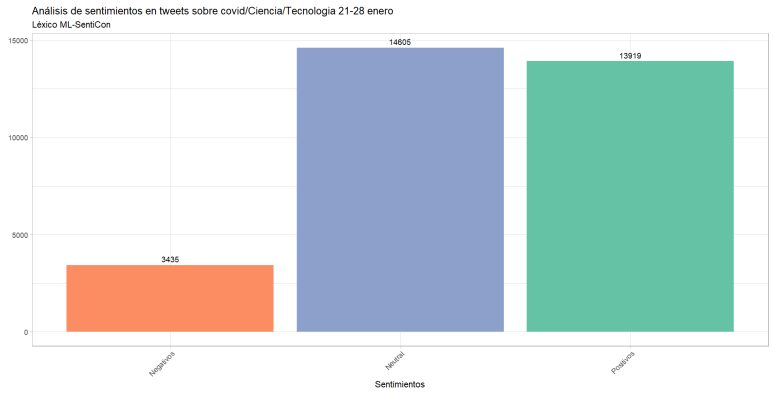

Figura 8: Conteo de sentimientos ML_SentiCon.

Se observa que una cantidad reducida de tweets fue clasificada como negativa utilizando este léxico, mientras que 13,919 tweets fueron identificados como positivos. Esto sugiere que nuestro léxico tiene una buena correlación con las palabras positivas presentes en los tweets, resultando en un número significativo de registros clasificados como positivos. Sin embargo, la escasez de registros negativos podría indicar una correlación limitada entre las palabras negativas del léxico y las palabras negativas potenciales presentes en nuestros tweets.

- **Diccionario_copalab.** Haciendo un procedimiento similar que se hizo con el léxico anterior obtendremos las siguientes estádisticas.

Estadístico	Valor
Mínimo	-2.00000
1er Cuartil	0.00000
Mediana	0.00000
Media	0.00478
3er Cuartil	0.00000
Máximo	2.00000

Cuadro 3: Resumen estadístico de los sentimientos con Diccionario_copalab.

Con estas estadísticas, es evidente que el diccionario no proporciona un análisis adecuado, recordando la puntuación asignada a cada palabra en la Figura 7. Podemos inferir que, en cada tweet, como máximo solo una palabra positiva o una palabra negativa coincidió con el diccionario. Al realizar el conteo basado en la polaridad asignada a cada tweet, obtendremos los siguientes resultados:

Positivos	Negativos	Neutral
186	52	31721

Cuadro 4: Resultados de análisis de sentimientos con Diccionario_copalab.

De acuerdo al Cuadro 4, podemos ver que muy pocos tweets tuvieron relación con el Diccionario_copalab.

- **Paquetería SentimentAnalysis.** Esta paquetería nos da la ventaja de limpiar un poco más nuestros datos creando un Corpus (conjunto de texto)

```
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, removeWords, stopwords("spanish"))
```

Con un Corpus lo suficientemente limpio, la función *analyzeSentiment()* nos hará el trabajo de calcular la valencia de cada tweet.

```
analyzeSentiment(corpus, language = "spanish")
```

Las estadísticas obtenidas se muestrán en el Cuadro 5.

Estadístico	Negativo	Positivo	Neutral
Mínimo	-1.00000	0.00000	0.00000
1er Cuartil	0.00000	0.00000	0.00000
Mediana	0.00000	0.00000	0.00000
Media	0.03679	0.2511	0.4936
3er Cuartil	0.10000	1.00000	1.00000
Máximo	1.00000	1.00000	1.00000
NA's	305	305	305

Cuadro 5: Resumen estadístico por utilizando la paquetería SentimentAnalysis

El problema al automatizar el análisis con la función *analyzeSentiment()* radica en que algunos registros fueron etiquetados como NA's. Esto indica que ciertos registros no pudieron ser clasificados como neutrales, negativos o positivos. Esto tal vez se atribuya al tipo de variable que se asigna a un Corpus.

El conteo de sentimientos que permitió este paquete se muestra en la Figura 9.

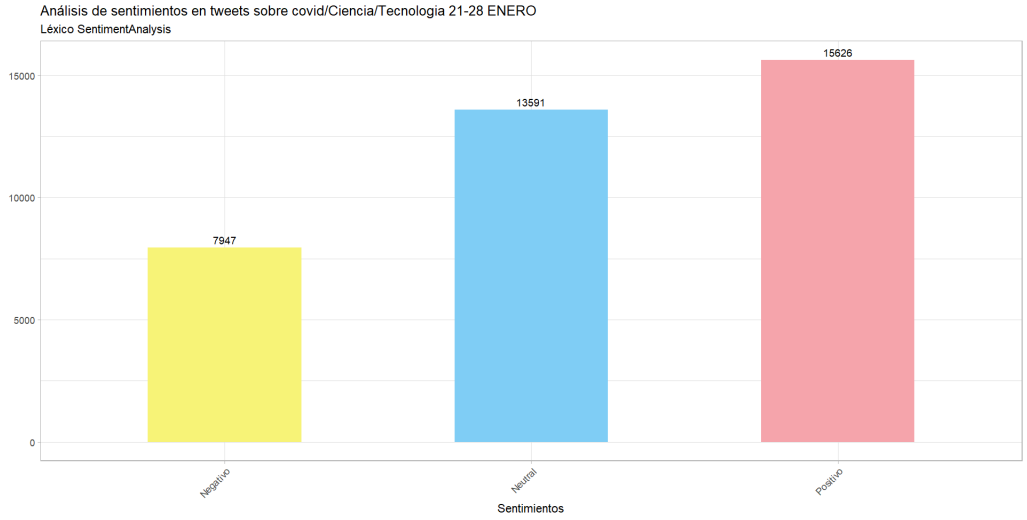

Figura 9: Conteo de sentimientos SentimentAnalysis.

- **Paquetería syuzhet.** asdas Al igual que la paquetería SentimentAnalysis, la función *get_nrc_sentiment()* nos facilita el trabajo de análisis de sentimientos. Simplemente necesitamos proporcionarle nuestra columna de tweets ya procesada.

```
get_nrc_sentiment(datos_limpios$TWEET_CLEAN, lang="spanish")
```

Las estadísticas obtenidas al utilizar esta paquetería son las correspondientes al Figura 10

```
     anger          anticipation      disgust          fear
Min.   :0.0000   Min.   : 0.0000   Min.   :0.0000   Min.   : 0.000
1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.000
Median :0.0000   Median : 0.0000   Median :0.0000   Median : 0.000
Mean   :0.1582   Mean   : 0.3252   Mean   :0.1309   Mean   : 0.342
3rd Qu.:0.0000   3rd Qu.: 1.0000   3rd Qu.:0.0000   3rd Qu.: 0.000
Max.   :9.0000   Max.   :12.0000   Max.   :5.0000   Max.   :12.000
      joy            sadness          surprise         trust
Min.   :0.0000   Min.   : 0.0000   Min.   :0.0000   Min.   : 0.0000
1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.0000
Median :0.0000   Median : 0.0000   Median :0.0000   Median : 0.0000
Mean   :0.1639   Mean   : 0.2166   Mean   :0.1147   Mean   : 0.5975
3rd Qu.:0.0000   3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.: 1.0000
Max.   :7.0000   Max.   :11.0000   Max.   :4.0000   Max.   :17.0000
    negative         positive
Min.   : 0.0000   Min.   : 0.000
1st Qu.: 0.0000   1st Qu.: 0.000
Median : 0.0000   Median : 1.000
Mean   : 0.5878   Mean   : 1.069
3rd Qu.: 1.0000   3rd Qu.: 1.000
Max.   :22.0000   Max.   :28.000
```

Figura 10: Estadisticas de emociones syuzhet.

Esta paquetería no solo nos permite identificar si cada tweet es positivo o negativo, sino que también proporciona información sobre las emociones presentes en los mismos. En la Figura 11, y basándonos en el léxico NRC, podemos observar que nuestros tweets exhiben las siguientes emociones:

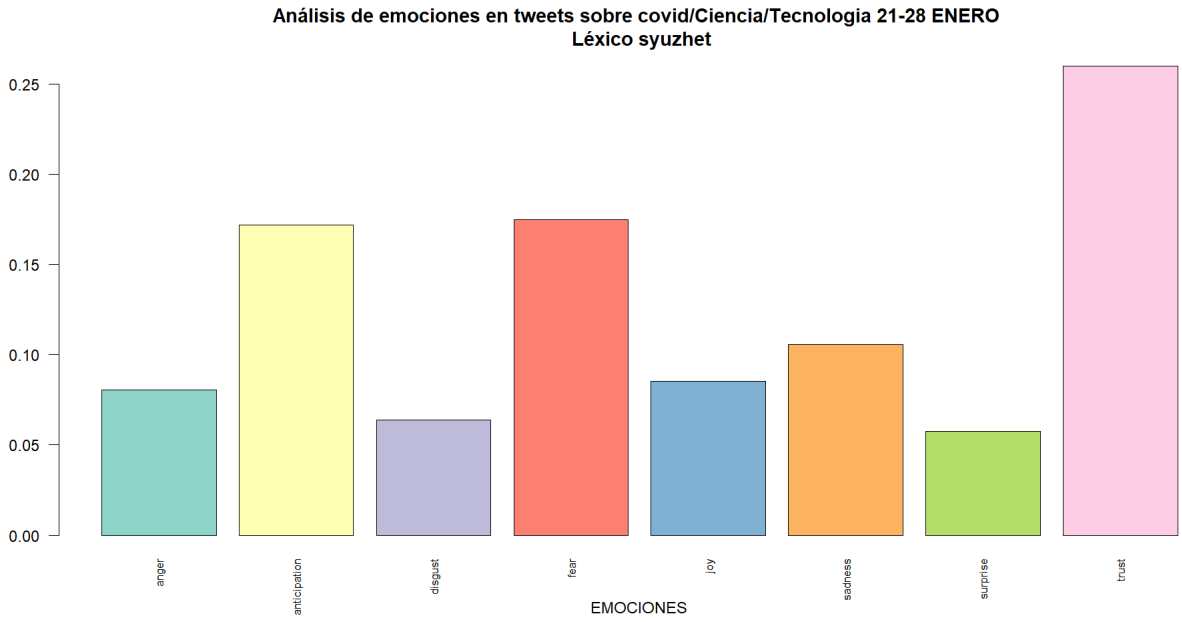

Figura 11: Emociones en los datos de acuerdo a syuzhet

Se puede observar que las emociones predominantes durante estos días fueron el miedo y la confianza.

Basandonos en noticias relevantes de esas fechas, tenemos registro de que:

- A finales de 2020, parecía que la pandemia estaba cerca de llegar a su fin. Varios países, incluyendo Europa, Israel, Estados Unidos y algunos países latinoamericanos, comenzaron la aplicación de vacunas contra el COVID-19 a sus poblaciones más vulnerables. Sin embargo, un fantasma de incertidumbre se cernió sobre este panorama alentador: el virus mutaba y daba origen a variantes más contagiosas.
- En febrero de 2021, una misión de 17 expertos internacionales fue enviada a Wuhan, China, por la OMS para investigar el origen del virus SARS-CoV-2, causante del COVID-19. Después de semanas de investigaciones en lugares clave de Wuhan, los investigadores no pudieron dar conclusiones firmes, pero indicaron que había cuatro posibles escenarios sobre el origen de la enfermedad.

Mientras tanto en México

- Enero de 2021 se convirtió en el mes con el mayor número de contagios y fallecimientos registrados desde que llegó la pandemia de COVID-19 a México. El país enfrentó desafíos significativos en la gestión de la crisis sanitaria

De acuerdo con los eventos ocurridos en esos días, es posible atribuir razones a las dos emociones predominantes observadas en las gráficas. Durante ese periodo, se llevaron a cabo investigaciones enfocadas en determinar el origen del Covid-19 y varios países habían logrado vacunar a una gran parte de su población vulnerable. Sin embargo, a pesar de estos avances positivos, el número de casos de Covid-19 continuaba en aumento, y se reportaba la aparición de nuevas variantes del SarsCov-2.

3.5. Otros resultados

Después de limpiar nuestros datos de varias maneras, podemos presentar algunos resultados gráficos generales. Por ejemplo, el día con más registros fue el 25 de enero de 2021. Sin embargo, no podemos sacar conclusiones definitivas sobre este hecho, ya que tanto el método de recolección de datos como el proceso de limpieza podrían haber influido en la actividad o frecuencia de las interacciones de las personas en ese día específico. Esta frecuencia se detalla en la Figura 12.

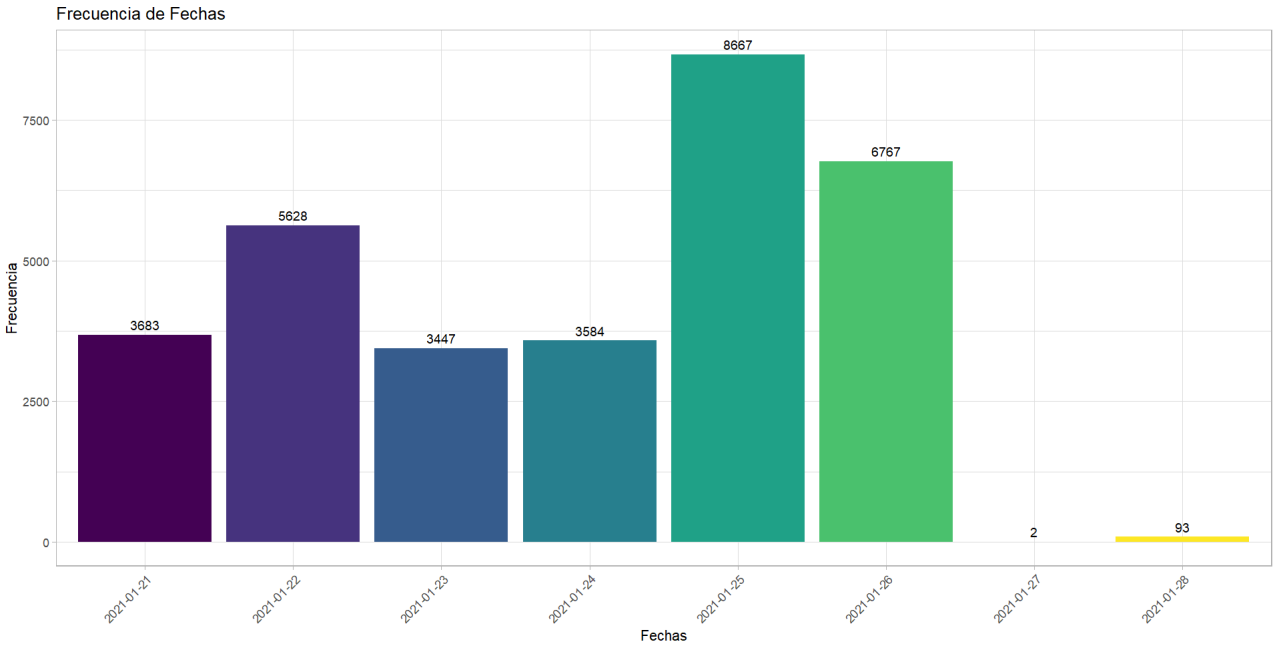

Figura 12: Días con mas actividad.

Otro análisis que podemos hacer, es el comportamiento de los sentimientos a través de estos días. Dicho comportamiento se visualiza en la Figura 13.

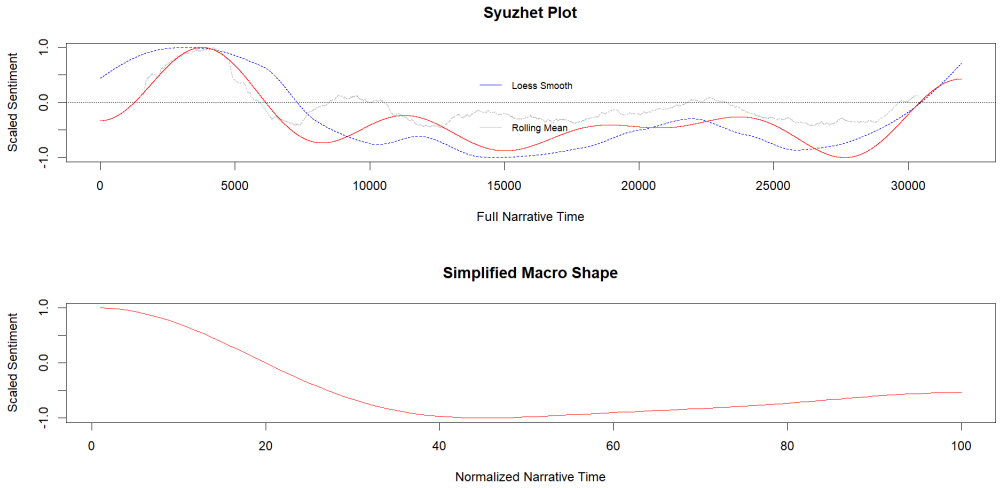

Figura 13: Días con mas actividad.

La gráfica superior ilustra con precisión cómo fluctuaron los sentimientos a lo largo del tiempo, mientras que la gráfica inferior nos muestra la tendencia general de estos sentimientos. Dado que la gráfica es cronológica y el día con más registros corresponde al 25 de enero de 2021, podemos afirmar con cierta

certeza que la tendencia general fue negativa durante el período analizado. A pesar de que el periodo comenzó con una cantidad significativa de tweets positivos, con el transcurso de los días, estos fueron dando paso a un predominio de sentimientos negativos.

4. Conclusiones

Hemos observado la variabilidad en los análisis según el léxico utilizado para llevarlos a cabo. Por otro lado, hemos aprendido que no siempre es aconsejable automatizar los análisis utilizando herramientas o paquetes proporcionados por alguna tecnología específica. No obstante, existen diversas opciones disponibles para identificar y cuantificar las emociones principales observadas durante este periodo. Paquetes como *syuzhet* ofrecen resultados que son relativamente fáciles de analizar.

Podemos concluir que el análisis más efectivo se logró utilizando el paquete *syuzhet* y el diccionario NRC asociado a este. Además, se demostró que un léxico limitado en número de palabras no es capaz de analizar adecuadamente los sentimientos o emociones en un conjunto extenso de textos.

5. Referencias

1. Sobrino Sande, J. C. (2018). Análisis de sentimientos en Twitter [Trabajo de fin de máster, Máster Universitario en Ingeniería Informática]. Universitat Oberta de Catalunya.
2. Calvo Madurga, A. (2020). Análisis de sentimientos y emociones en redes sociales usando ML [Trabajo de Fin de Grado, Grado en Ingeniería Informática]. Universidad de Valladolid.
3. Arsys. (2020). Análisis de sentimientos Python y Jupyter Notebooks. Blog de arsys.es; Arsys. https://www.arsys.es/blog/analisis-sentimientos-python-jupyter-notebooks
4. RPubs - Analisis de sentimientos. (s/f). Rpubs.com. Recuperado el 8 de mayo de 2024, de https://rpubs.com/elbuensato/1095352
5. Kuffo, L. (2020, mayo 11). Análisis de Sentimientos en textos — Te lo explico en 20 minutos. https://youtu.be/kRVJFhFDuYA?si=GOQoE4-5thIBXSPk
6. Jennifer Isasi, .ªnálisis de sentimientos en R con 'syuzhet'", Programming Historian en español 5 (2021), https://doi.org/10.46430/phes0051
7. Mendoza, G. (s/f). Análisis de sentimientos con R Léxico Afinn. Recuperado el 8 de mayo de 2024, de https://rpubs.com/jboscomendoza/analisis_sentimientos_lexico_afinn