Capítulo III

El procesamiento de lenguaje natural para actividades deportivas durante la pandemia por Covid-19

DOI:https://doi.org/10.58299/utp.268.c929

María Beatriz Bernábe Loranca
Facultad de Ciencias de la Computación
Benemérita Universidad Autónoma de Puebla
beatriz.bernabe@gmail.com
https://orcid.org/0000-0003-3014-4139

Melissa Isaaly Mendoza Bernábe
Universidad Iberoamericana Puebla
melissa.mendoza@iberopuebla.mx
https://orcid.org/0009-0006-9097-2939

Marleny Reyes Monreal
Escuela de Artes Plásticas y Audiovisuales
Benemérita Universidad Autónoma de Puebla
marleny.reyes@correo.buap.mx
https://orcid.org/0000-0003-0493-4786

Resumen
Las actividades físicas son uno de los principales factores que impactan positivamente en la salud. El presente trabajo describe la aplicación del procesamiento de lenguaje natural PLN para identificar l a percepción sobre cuatro actividades deportivas en tiempos de la pandemia por COVID 19 durante 2020-2021, con el fin de reconocer el significado y relevancia de la actividad física. La extracción de información y análisis del tema de deportes se centra en los comentarios de los usuarios de Twitter. El procesamiento se basa en tres pasos principales: descarga de información, creación de diccionarios y análisis de sentimientos. Los resultados indican que las personas tuvieron una importante reflexión para ejercitarse aún más durante la pandemia.

Introducción

México se encuentra entre los países con mayor índice de obesidad del mundo, según la Encuesta Nacional de Salud y Nutrición (ENSANUT) más del 75% de la población mayor a 20 años tiene problemas de sobrepeso y obesidad (Instituto Nacional de Estadística y Geografía, 2018). Entre las causas determinantes del sobrepeso se encuentran la vida sedentaria, una alimentación desequilibrada y la poca actividad física. El 59.7% de la población mexicana mayor a 18 años no realiza actividades físicas o deportivas (Kánter-Coronel, 2021) y la Organización Mundial de la Salud (OMS) afirma que en el mundo más del 80% de adolescentes tienen actividad física insuficiente (Organización Mundial de la Salud, 2022).

La relación del sobrepeso con problemas de salud está ampliamente documentada, entre los efectos se encuentra el desarrollo de enfermedades crónicas como la diabetes tipo 2, enfermedades cardiovasculares y osteoarticulares. La OMS asegura que “se podrían evitar hasta 5 millones de fallecimientos al año con un mayor nivel de actividad física de la población mundial” (Organización de las Naciones Unidas, 2021, párr. 2).

Durante el confinamiento producido por la pandemia de COVID-19 la actividad física de las personas se redujo ampliamente. Como consecuencia, distintas campañas de salud se enfocaron en alertar a la población a través de recomendaciones basadas en una buena alimentación o la realización de distintas actividades físicas. Sin embargo, en contadas ocasiones dichas campañas toman en cuenta lo que el público expresa desde su percepción con respecto a la actividad física (Polero *et al.*, 2021).

El trabajo que se expone busca interpretar las opiniones de usuarios de Twitter para establecer algunas condiciones que puedan incidir en los distintos promotores de la salud. Contar con información acerca de las opiniones, intereses y tendencias de los usuarios en la web sobre el deporte, se asume que mejorará el entendimiento del problema y las posibilidades de generar estrategias. Con PLN y utilizando la red social Twitter, se persigue encontrar la motivación y sentir de las personas que tomaron la decisión de realizar actividades deportivas, de ese modo, es posible comprender los motivos que los usuarios de Twitter tuvieron sobre el

ejercicio en tiempos de pandemia y la manera en que el análisis de sentimientos puede usarse en el estudio de tweets sobre el deporte y la actividad física durante la pandemia.

La estructura del trabajo es la siguiente: en la sección uno se presenta la introducción para posteriormente hablar sobre los antecedentes en la sección 2, preliminares y estado del arte. A partir de este punto, se inicia el desarrollo del trabajo que comienza con la extracción de información, posteriormente en la creación de diccionarios, se especifica la forma en que dicha información ha sido conjuntada y el inicio de su análisis descriptivo. Posteriormente, se menciona el análisis de sentimientos implementado en los tweets descargados y el trabajo realizado en cuanto a *clústering* en la sección 7. Finalmente, se revelan los resultados obtenidos tras el procesamiento de toda la información recolectada y se establecen las conclusiones del estudio.

Antecedentes y preliminares

PLN ha sido de mucha utilidad como herramienta para interpretar información proveniente de Twitter debido a que la red social es una fuente rica en interacciones en tiempo real de un gran número de usuarios, en 2022 Twitter contaba con alrededor de 217 millones de usuarios activos (TwitterIR, 2022). Debido a la multiplicidad de temas de discusión en Twitter y a la gran cantidad de información, se hace necesario crear herramientas usando PLN para extraer información de las redes sociales. Esta tendencia ha convertido el texto en un componente clave para la comunicación en la sociedad, consolidándolo como adecuado para el intercambio de información y ayuda para entender la opinión, valoraciones y perspectivas del público sobre un tema determinado. El análisis de sentimientos provee información clave para comprender las opiniones de los usuarios.

Se han planteado algoritmos y modelos computacionales con PLN para identificar y analizar patrones o tendencias sociales tanto en centros de investigación como de desarrollo tecnológico. Dichos modelos se basan en métodos de aprendizaje a nivel de análisis morfológico, sintáctico y léxico (Díaz-Mendivelso & Suarez-Baron, 2019).

El cuidado de la información procesada tanto en cantidad como en contenido es relevante dado que la cantidad de datos que circula en las redes sociales ha alcanzado la masa crítica, de tal manera que es inevitable el uso de una computación sensata para el procesamiento de datos que responde a la cantidad de información que aumenta exponencialmente. Se entiende así que la democratización de la creación de contenidos en línea ha dado lugar a un aumento de acceso a la Web generando retazos, lo que afecta inevitablemente, incluso de manera negativa a la recuperación y extracción de información (Cambria & White, 2014).

La información que circula en Twitter es múltiple y no se encuentra estructurada, entonces, procesar las opiniones de los usuarios para un determinado problema es importante y complejo. En tal situación, el análisis de textos tiene un rol importante para facilitar la obtención y el procesamiento de datos, en particular, el estilo breve, informal y ruidoso de Twitter presenta serios desafíos, por ejemplo, la información no se encuentra en un solo idioma, y consecuentemente, distintos trabajos se han ocupado de tratar al menos dos retos para examinar los textos de Twitter: clasificación de la polaridad (análisis de sentimientos) y reconocimiento de entidades (extracción de información que tiene por objeto localizar y clasificar entidades nombradas en texto).

Para el problema que se presenta en este documento, la herramienta conveniente para analizarlo es PLN (Becerra-Pozas, 2018). Aunque PLN ha sido muy popular en los últimos años, sus inicios se estiman en 1930 y su objetivo, entre muchos otros, ha sido principalmente relacionar palabras. Para poder aplicar herramientas de PLN, es útil acudir a tecnologías de software como Python para recolectar la información necesaria que se ajuste al problema de las actividades físico-deportivas. Se justifica el uso de Python por su *tipado* dinámico que le permite ser una herramienta consistente para el desarrollo de *scripts*. De igual manera, Python es excelente debido a su acoplamiento con la API de *Tweepy*, la cual es una biblioteca que permite tener acceso a la información requerida de Twitter, además, Python concede utilizar muchas otras bibliotecas para facilitar el desarrollo de este trabajo.

Estado del arte

Dentro de los temas del estudio de tweets durante la pandemia de COVID-19 mediante PNL destacan las investigaciones sobre la percepción de la enfermedad misma y la salud mental del público, la polarización política y el uso de vacunas. Cabe destacar el estudio sobre salud mental: *Using Natural Language Processing to Explore Mental Health Insights From UK Tweets During the COVID-19 Pandemic: Infodemiology Study* (Marshall *et al.*, 2022) cuyo objetivo principal es exponer el valor del estudio por medio de PNL para respaldar investigaciones en salud pública. Se examinan tweets relacionados con salud mental durante la pandemia COVID-19 en Reino Unido, lo cual se relaciona con los objetivos de este trabajo.

Son pocas las investigaciones cuyo tema central es el análisis de la percepción de los usuarios sobre la actividad física y deportiva durante la pandemia de COVID-19 por lo que se presenta la revisión bibliográfica en dos vertientes 1) estudios del tema desde otras disciplinas, 2) el estudio de temas relacionados mediante el uso de PNL.

La UNESCO publicó en 2020 el libro titulado El deporte en tiempos de pandemia (Cáceres-Andrés, 2020) que analiza el impacto del COVID-19 para comprender las fortalezas y debilidades de los sistemas deportivos. Destaca la importancia de la actividad física y su relación con la salud. Se encontró que las actividades físicas en Iberoamérica se caracterizan por el informalismo, por lo que es indispensable la gobernanza del deporte y fortalecer el valor de las actividades físicas desde las instituciones.

Por otra parte, el tema de la percepción de la actividad física en redes sociales ha sido trabajado desde las humanidades. Por ejemplo, en texto Redes sociales en tiempos de COVID-19 el caso de la actividad física describe problemáticas del deporte con el uso de tecnología digital, redes como Twitter y la influencia de personas a seguidores jóvenes. Destaca propuestas como el día de la educación física en casa #DEFC2020 que se implementó en España (Piedra, 2020).

Otro texto sobre el tema es la tesis *Sports Analytics with Natural Language Processing: Using Crowd Sentiment to Help Pick Winners in Fantasy Football* (Hendricks, 2022); en tal documento se desarrolla un método para maximizar el potencial en el problema de selección de ganadores de análisis deportivo por medio del análisis de sentimientos a gran escala basado en BERT de tweets sobre jugadores. Los resultados sugieren un nuevo enfoque para el modelado de predicción deportiva que se basa en el procesamiento del lenguaje natural y modelos de lenguaje de última generación para incorporar la sabiduría de la multitud.

En el artículo *Research on Keyword Extraction of Word2vec Model in Chinese Corpus* (Zhang *et al*., 2018) muestra los resultados de la aplicación de Word2vec a las noticias deportivas y se utilizó para explorar el corpus chino y la extracción de palabras clave. Los resultados experimentales muestran que Word2vec puede hacer coincidir palabras relacionadas en orden descendente de similitud para que los lectores puedan comprender mejor las noticias.

Finalmente, el estudio Aplicación Web para la gestión de competiciones de deporte electrónico usando el *Framework Ruby on Rails* (Rozalén-Barberán, 2020) muestra el proceso de desarrollo de una aplicación Web que unifique las competiciones de deporte electrónico con la comunidad, permitiendo que jugadores no profesionales puedan vivir la experiencia de competir en torneos de deporte electrónico.

Recolección de Información

La información se ha extraído de Twitter por ser una red que cuenta con rutas y capacidad para acceder a datos mediante las distintas bibliotecas de Python.

Definición del problema

Para iniciar el desarrollo del problema planteado se identificaron las palabras clave que tuvieran una asociación más cercana al tema y destacó la palabra *running*, vocablo muy difundido y popular en el ámbito cotidiano-deportivo. Debido a la facilidad con que las personas pueden participar en una actividad deportiva que implica bajo costo económico, flexibilidad de tiempo, entrenamiento al aire libre, entre otras ventajas, ha colocado a *running* como la palabra inicial de la exploración

de opiniones en Twitter a través de descargas. Adicionalmente, en una encuesta a 30 entrenadores de 4 distintos clubs deportivos, se encontró que el 90% asegura que correr, trotar o caminar es la actividad deportiva más difundida dada su accesibilidad económica, no necesita experiencia y puede ser realizada por cualquier persona, por tanto, se justifica a *running* como la palabra indicada para iniciar las descargas de tweets.

Vicinitas

El tratamiento de scripts en Python es viable con el uso de distintas bibliotecas y con ello, poseer control sobre los datos recolectados. Para este trabajo, Vicinitas ha sido la herramienta elegida debido a diversas cualidades donde destaca la manera en que se da el acceso a la web, es decir, se aloja en un espacio virtual que funciona como una interfaz a un código que utiliza la API de Tweepy. En este sitio, se introduce la palabra base para la búsqueda de tweets. En el desarrollo de este trabajo, el inicio consistió en introducir en Vicinitas la palabra "*running*", consecuentemente, los tweets se descargaron diariamente durante un mes, alcanzando de esta manera una gran cantidad de información correspondiente al problema.

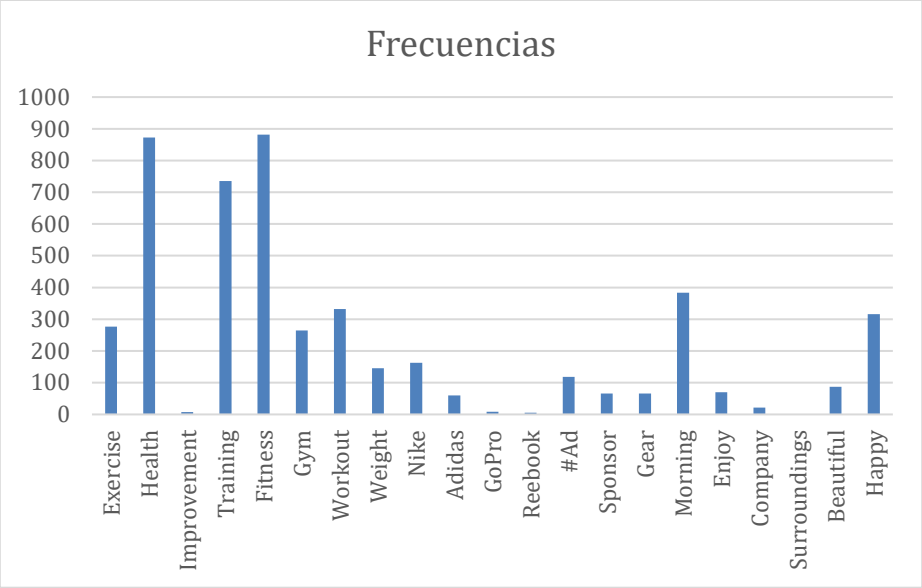
Tras analizar la información, se confirmó que la palabra clave *running* es la más popular en el proceso de descargas y representa al problema en el contexto de PLN como se había estimado. Para organizar las descargas, diseñar consultas, contabilizar y ordenar las palabras que formarán los diccionarios, se utilizó una base de datos relacional para ejecutar consultas en conexión con las palabras más repetidas en los tweets obtenidos.

Por otro lado, los adjetivos relacionados con la periodicidad de las palabras derivadas de *running* son importantes no solo por su influencia sobre los resultados finales, también por la evidente presencia que induce a construir los respectivos diccionarios. La lista de palabras con sus frecuencias influyó en los términos que determinarían los conjuntos de palabras directamente afines con tres grupos de enfoque del deporte que se reconocieron al principio de las descargas con *running* como el principal *hashtag*: 1) salud, 2) marketing y 3) ocio. Estos grupos se

enumeraron con las palabras repetidas de cada sub-grupo, entonces fue claro apreciar la relevancia de aspectos de salud en los tweets alcanzados hasta el momento y la relación con los conceptos de belleza y felicidad, ver Figura 1.

Figura 1

Gráfica de barras correspondiente a las palabras de los tres grupos de enfoque para la actividad running



Fuente: Elaboración propia.

Al momento de estos resultados, el estudio se encuentra limitado considerando que los léxicos encontrados son insuficientes para un nivel de evaluación PLN, por tanto, se reestructuraron las descargas de acuerdo con las palabras que se observan en la Figura 1.

Ampliación de descargas

Una vez que el primer proceso de descargas nos acercó al problema, se procedió a cambiar la manera de obtener tweets dando paso a bibliotecas, principalmente las relacionadas a la obtención de tweets (Tweepy) y la creación de documentos (CSV). Este código de script descarga únicamente la fecha de creación, el autor, el texto, el id, la cantidad de reacciones *like* y re-tweets, además, el *script* genera la cantidad de tweets deseados con la restricción de descargar de 2000 tweets en 15 minutos.

Las frecuencias confirmaron que la palabra *running* representa al problema, pero sin protagonismo, es decir, se obtuvo una descripción extendida del problema más allá de una única actividad, de esta manera y considerando la presencia de los vocablos *hiking*, *jogging* y *cycling*, se incorporaron estas palabras al estudio para su análisis. Dichas actividades físicas son importantes tanto para el enfoque de salud como para el aspecto de marketing y económico. En este punto del procesamiento, nos concentramos en el tema de la salud, en un estudio posterior, atenderemos la situación de marketing.

Sumadas *hiking*, *jogging* y *cycling* a la palabra inicial *running*, una nueva fase de extracción de información en Twitter es necesaria. La conexión de estas cuatro palabras se justifica por la semántica compartida de cardio y a la facilidad de práctica.

El nuevo proceso de descargas fue durante 3 meses diariamente durante el periodo de la pandemia por COVID-19 en el punto crítico entre noviembre 2020-enero 2021, con ello se han acumulado más de 4 meses para las descargas de tweets.

Creación de diccionarios

La segunda etapa del presente trabajo es la creación de diccionarios. Estos diccionarios reúnen todos los términos correspondientes a los tweets obtenidos mediante el script de Python con las frecuencias adjuntas. El llenado de los diccionarios se alimenta con las descargas diarias, libres de caracteres innecesarios y que además se encuentran en contacto con las cuatro actividades *running*, *hiking*, *jogging* y *cycling*. Seleccionados los archivos para la generación del diccionario, el script recorre cada uno de los archivos verificando por columna los términos de cada uno de los tweets. Cuando el script tropieza con un nuevo término, se actualiza la frecuencia respecto a todos los tweets de los distintos archivos. En caso de contar con la frecuencia tope, el término se agrega al diccionario, de lo contrario se continúa con el recorrido del archivo en busca del siguiente nuevo término. Todos los términos con una frecuencia de 50 o superior se incluyen en el archivo final.

Las especificaciones de filtrado se ajustaron en un script donde el filtrado principal consistió en agregar al diccionario palabras de cuatro letras o más. En tal restricción se satisface implícitamente detener muchos conectores que no aportan nada, sin embargo, existen palabras que teóricamente no significan nada, pero contrastándolos con los cambios presentados en distintos periodos de tiempo, se consigue entender el problema en magnitud y contenido, e incluso, esta situación en algunos casos precede al análisis de sentimientos.

En esta segunda etapa de descargas, no solo se ampliaron y definieron con precisión los diccionarios, se distinguieron también expresiones más consistentes del problema, por ejemplo, el uso de los hashtag *#ukrunchat* o *#people*, que se interpretaron como actividades para ser exploradas debido al impacto de ejecutarse en grupos o comunidades con intereses muy similares. Del mismo modo surgieron *#nature*, *#photography* y *#photo* las cuales se entienden como mensajes y fueron muy recurrentes donde se asumían importantes dado el placer de compartir evidencias visuales sobre los momentos deportivos; por último, se incorporó el hashtag *#fitness*.

Análisis de Sentimientos

El propósito del análisis de sentimientos para este trabajo se centra en observar los posibles cambios emotivos de los usuarios respecto a las actividades deportivas y a su vez, la incidencia en el público. La biblioteca CLiPS Pattern es útil para el análisis de los sentimientos de un texto y divide una oración con el fin de identificar sustantivos, adjetivos, verbos, adverbios, etc. Se tienen así una base de datos de palabras con una ponderación entre -1 y 1 siendo (negativo y positivo respectivamente), es decir, calcula la polaridad del texto.

El análisis de sentimientos se procesó sobre las descargas de los tweets de las actividades deportivas respecto a los vocablos *running, jogging, hiking* y *cycling*. Se procedió a agrupar las descargas por actividad específica y después por fecha para analizar el cambio de los sentimientos de los tweets descargados. La cantidad de descargas por día fue de alrededor de1000 tweets por actividad, dando un total

de 4000 tweets. Los resultados se agruparon en 3 categorías de sentimientos: Positivos > 0, Neutral = 0 y Negativos < 0.

Clustering

Para el problema que se trata en este artículo, se aplicó k-medias para agrupar tweets en distintas categorías.

Debido a que k-means recibe una matriz de valores numéricos, es necesario transformar los tweets a números. Para este proceso se deben de separar las distintas palabras del texto (tokenize) y continuar con el cálculo con las palabras con el fin de asignarles un valor numérico. Esta tarea se resolvió empleando TF-IDF (Term Frequency-Inverse Document Frequency), la cual es una estadística numérica que demuestra lo importante que es una palabra para un cuerpo (Díaz-Mendivelso & Suarez-Baron, 2019). La frecuencia de término señala la relación entre el número de palabras actuales y el número de todas las palabras del documento o la cadena, etc. Esta estadística se calcula con la formula $tf(t,d) = n_t / \sum_k n_k$ bajo el uso de la biblioteca de scikit-learn con Python, un recurso muy útil en el análisis de datos.

El texto que procesa el clustering incluye hashtags, etiquetas, abreviaciones, sinónimos de palabras, etc. Específicamente, el programa en Python que resuelve el clustering en este trabajo permite entre otras funciones, que el usuario manipule tanto etiquetas como hashtags para ir construyendo el modelo. Por otra parte, para reducir la redundancia en las palabras, dos técnicas conocidas como *Stemming* y *Lemmatization* fueron eficientes para suprimir palabras que significan lo mismo, por ejemplo, cuando una de ellas es singular y otra plural (Cambria & White, 2014). Otra situación común es cuando se presentan verbos en distintas conjugaciones como “corre y correr”, en este contexto, las palabras se reducen al singular en tercera persona (corre). Tanto *Stemming* como *Lemmatization* tienen la misma finalidad, pero el proceso lo realizan de formas muy distintas. *Stemming* modifica la palabra basado en su terminación, como lo es el caso de “fotos y foto”. Esta técnica es conveniente para algunas palabras, pero produce malos resultados en casos como “canciones y canción” (el resultado obtenido sería “cancione”). Para este caso, se

recurre a la *lemmatization* cuya función es transformar las palabras a su forma base mediante el significado de ellas, lo cual se resuelve revisando los diccionarios de palabras y relacionándolas para transformarlas a su forma original. El procedimiento se muestra en la Figura 2.

Figura 2

Resultados obtenidos del clustering de tweets en Python

```
❯ python3 tweets_clusteing.py
Nombre del archivo a analizar: #running 17-04-2020 10-51-12.csv
Tweets encontrados = 500
Cantidad de tweets a analizar: 400
Remover etiquetas de tweets (@usuario)? s/n: s
Remover hashtags de tweets (#running)? s/n: n
Reducir las formas de las palabras a una forma base? s/n: s
 -Methodo ha utilizar? (1.-Stemming 2.-Lemmatization): 2
Numero de clusters a generar: 4
Número de iteraciones maximas: 100
Valor de n_init (Número de veces que el algoritmo de k-means se ejecutará con diferentes semillas centroides.) : 5
Top terms per cluster:   Cluster 1:     Cluster 2:     Cluster 3:
Cluster 0:                socks          news           run
 runner                   win            start          fitness
 run                      running        fitspo         running
 runnersofinstagram       giveaway       crossfit       mask
 runhappy                 freebie        gymlife        exercise
 instarunners             competition    train          health
 runners                  retweet        gym            runners
 instarun                 enter          fitfam         day
 runningmotivation        click          workout        best
 marathontraining         comp           good           morning
 nikeplus
```

Fuente: Elaboración propia.

Es posible construir el tamaño de clusters que el usuario desee siempre que sea consistente con el problema, por tanto, la decisión de especificar k=4 en el clustering responde a las tres diferentes actividades derivadas de la única palabra base *running* que se identificó en los tweets de descargas en la primera etapa (*hiking, jogging* y *cycling*). Recordemos que estas tres actividades se escogieron debido a la fuerte relación con *running*. Es oportuno mencionar que existen distintos parámetros que podrían inducir a otro tipo de relaciones no triviales para *running* no obstante, esta tarea es complicada y el proceso no se terminó a pesar de que algunas pruebas fueron realizadas sin conseguir aunar otros vocablos a *running* que señalaran un agrupamiento con mayor significado que el que se expone en esta sección. Aun así, fue de mucha utilidad incluso cuando no se pudo explotar al máximo, pero se estima continuar con este proyecto en un futuro trabajo.

Ampliación del Tema y Cambio de Tecnología

Tras concluir con el primer acercamiento al PLN, se dio lugar a un cambio en la forma en que se obtienen los tweets. Este cambio se propuso dada la gran cantidad de información basura o de poca relevancia e implicó al desarrollo de un script usando el Python para aprovechar sus bibliotecas, principalmente las relacionadas a la obtención de tweets (Tweepy) y la creación de documentos (CSV). El código de dicho script descarga únicamente la fecha de creación, el autor, el texto, el id, la cantidad de *likes* y re-tweets de un tweet a partir de la palabra que el usuario ingrese también a la cantidad de Tweets a descargar (la restricción es de 2000 tweets en 15 minutos).

El siguiente paso es la incorporación de nuevas actividades físicas que extienda el contenido, principalmente para el enfoque de salud y en segundo lugar marketing-economía para conectar con marcas deportivas. A partir de este punto, se agregan al marco de estudio las actividades *hiking, jogging* y *cycling* (excursionismo, trotar y ciclismo por sus traducciones al español). Estas actividades se incorporaron por similitudes con *running* respecto a facilidad de entrenamiento, relevancia en temas de salud y su impacto en el área comercial.

Se hace necesario un nuevo ciclo de descargas en dos horarios de manera diaria. Este proceso se enfrenta a las limitaciones técnicas que presenta la API pero con la ventaja de obtener multiplicidad de términos en las descargas. Se gastaron alrededor de 2 meses tratando de realizar las descargas de manera diaria y en horarios similares cada día. Este periodo fue en el punto crítico de la pandemia por COVID-19.

Cabe destacar que en el primer análisis se encontraron como palabras cercanas a marcas deportivas como Adidas o Nike. Esto se puede explicar por una parte porque las marcas mismas utilizan estas actividades que han detectado lideres en sus ventas como parte de sus contenidos difundidos en Tweeter y por otra parte en la relación del público que realiza dichas actividades deportivas con el uso de estas marcas.

Resultados

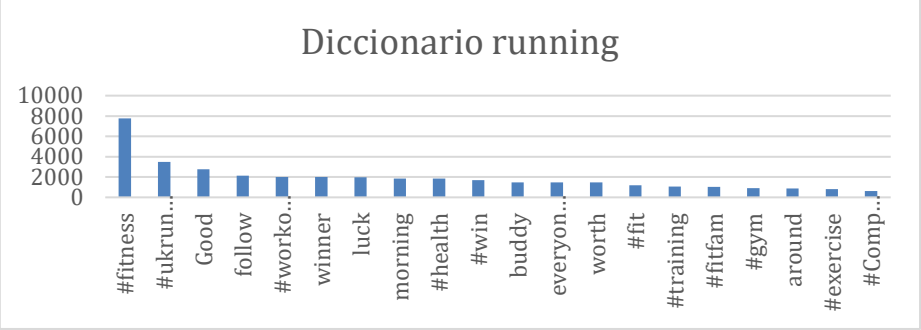
De las secciones, se obtuvieron los resultados que se ilustran desde la Figura 3 hasta la Figura 10. Se presentan primero los diccionarios para las cuatro actividades deportivas analizadas para dar lugar al análisis de sentimientos, igualmente para *running*, *jogging, hiking* y *cycling*.

Diccionarios

Considerando que el estudio se desarrolló en los meses difíciles de la pandemia por COVID-19, se asume que la percepción sobre actividad física fue afectada por la manera en que la gente se confinó y cambió la forma de ejercitarse.

Figura 3

Resultados de análisis de diccionarios de la actividad running – top 20 palabras

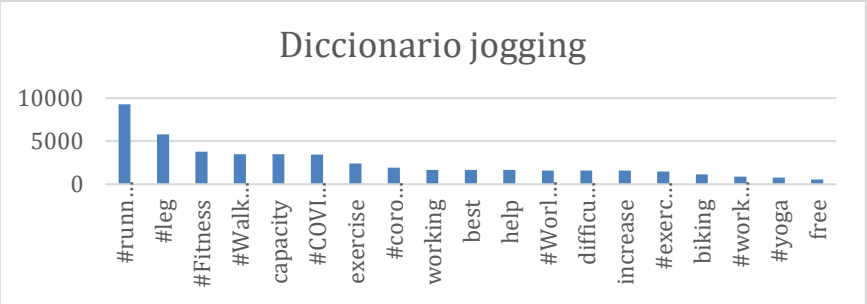


Fuente: Elaboración propia.

En la Figura 3 se observa poca variabilidad entre la palabra más frecuente *running* y los vocablos relacionados, lo cual es consistente por la asociación semántica de los tweets.

Figura 4

Gráfica de barras correspondiente a los resultados de análisis de diccionarios de la actividad running – top 20 palabras

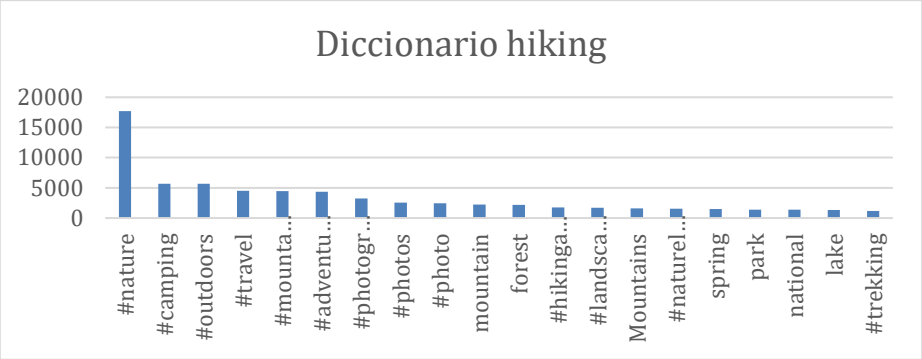


Fuente: Elaboración propia.

En la Figura 4, en el diccionario *jogging* la palabra más insistente es *running* dada su popularidad y por ser la primera expresión que inició este estudio. Por otra parte, un hashtag que destaca es *#WorldHealthDay* aunque apareció por un periodo corto, la gente cree que el ejercicio es necesario como parte de una vida saludable, asociado a un día conmemorativo a la salud.

Figura 5

Gráfica de barras correspondiente a los resultados de análisis de diccionarios de la actividad hiking – top 20 palabras

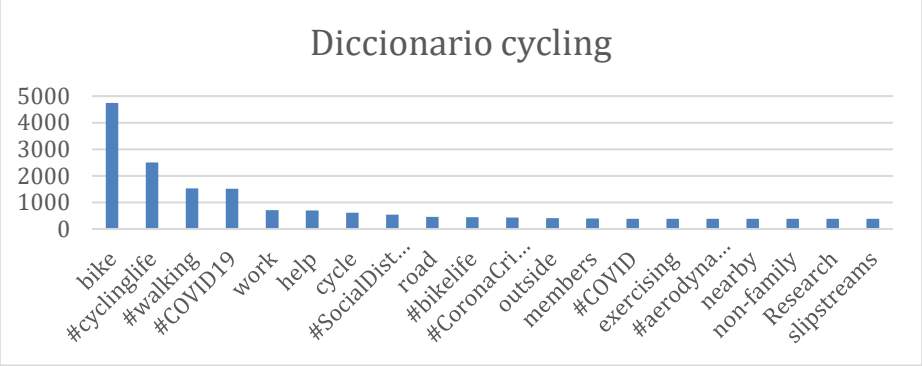


Fuente: Elaboración propia.

En la Figura 5, *hiking*, se muestra que el mayor interés no es precisamente la salud o el *fitness* como en otras actividades, también aparece como necesario el contacto con la naturaleza y su disfrute. Paralelamente, la fotografía y la aventura resaltan, se supone que es por el interés de capturar el momento.

Figura 6

Gráfica de barras correspondiente a los resultados de análisis de diccionarios de la actividad cycling – top 20 palabras



Fuente: Elaboración propia.

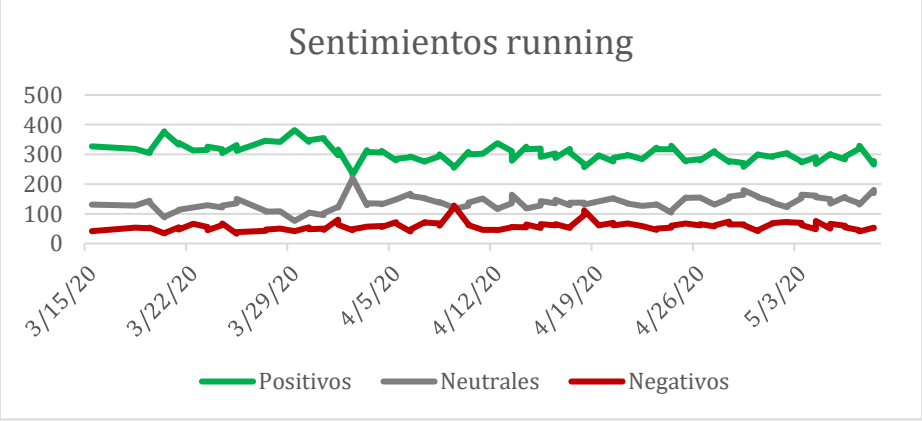
Al analizar el diccionario *cycling* de la Figura 6, se dedujo que esta actividad influenciada por el COVID-19 de forma más determinante, debido principalmente al confinamiento. Su presencia se distingue en los 4 lugares en el top 20 de las palabras con mayor frecuencia. De igual forma se observa que es una actividad atractiva por el instrumento para realizarla y ocupa el primer lugar.

Sentimientos

En las gráficas de las siguientes figuras, no se observan cambios sustanciales respecto a los sentimientos de los tweets, excepto por algunas fechas (por ejemplo, el día mundial de la salud: 7 de abril y el día de la tierra).

Figura 7

Gráfica de emociones de los tweets de *running*

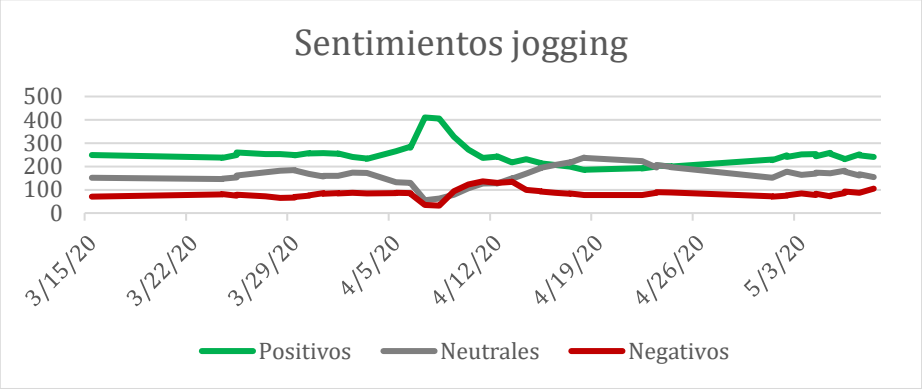


Fuente: Elaboración propia.

El *running* fue una actividad que como podemos observar en la Figura 7 no sufrió muchos sesgos importantes en cuanto a las emociones en los tweets. En los casos donde se diferencian cambios obedece principalmente a que se promovieron eventos de la actividad y este tipo de tweets se compartieron mucho durante estas fechas.

Figura 8

Gráfica de emociones de los tweets de jogging

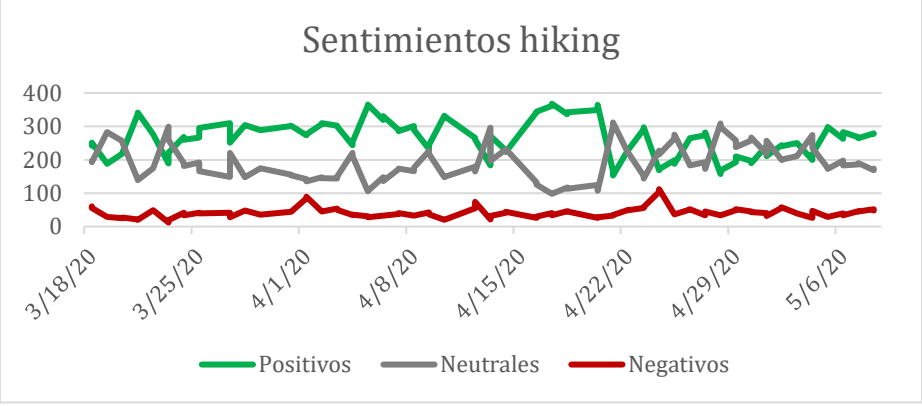


Fuente: Elaboración propia.

Al analizar la gráfica generada de los sentimientos en *jogging* de la Figura 8, se aprecia un aumento de tweets positivos entre el 5 de abril y el 12 de abril, lo cual probablemente se debió a la promoción de eventos como fue el día mundial de la salud.

Figura 9

Gráfica de emociones de los tweets de hiking

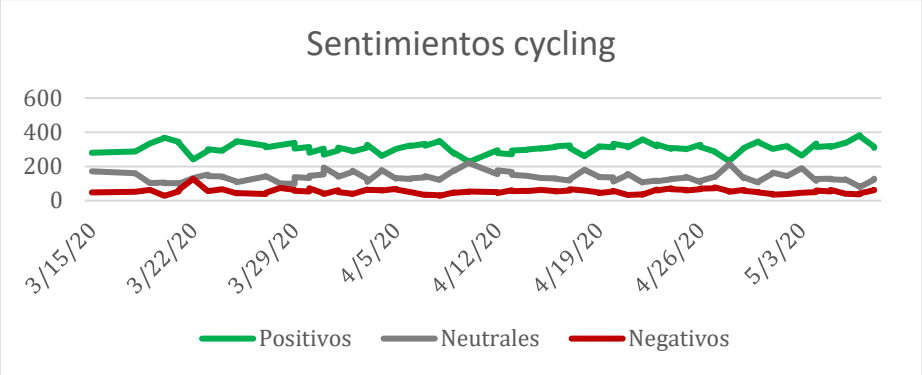


Fuente: Elaboración propia.

El análisis de sentimientos de tweets sobre *hiking* es consistente en proporción con los cambios entre sentimientos positivos y neutrales. Esta actividad no se distorsionó por festividades ni eventos públicos, pero si fue menos solicitada debido a las cuarentenas o aislamientos sociales.

Figura 10

Gráfica de emociones de los tweets de cycling



Fuente: Elaboración propia.

Los resultados obtenidos para *cycling* expresados en la Figura 10 no se vieron muy afectados en la polaridad positiva.

En general hubo una presencia poco importante de sentimientos negativos en los tweets, la mayor parte de sentimientos encontrados fueron positivos y después los neutrales. La pandemia del COVID-19 tuvo impacto en los resultados de las emociones, sin embargo, se ve que el impacto negativo en los sentimientos de las personas que practican estas actividades no fue tan considerable como se asumió al principio.

Conclusión

Los resultados obtenidos son consecuencia del proceso basado en PLN. Los hallazgos revelan que el mayor interés de las personas en Twitter con respecto a las cuatro actividades del estudio se centra en el impulso de compartir a la comunidad de la red social Twitter la realización del deporte. Las actividades deportivas más citadas fueron determinantes para la construcción de los diccionarios, las cuales se identificaron al principio del estudio y nutrieron los diccionarios al contabilizar los términos relacionados a grupos de personas que

practican alguna de las cuatro actividades que destacaron en la primera parte del estudio.

Poseer resultados de esta magnitud permite elevar conciencia sobre el enfoque primordial de los usuarios de la red social y así, generar nuevas formas de promover la práctica del deporte y de otras nuevas que puedan relacionarse adecuadamente. Es evidente que los eventos acompañados de una campaña de hashtag en Twitter impactaron en la diseminación de los deportes, así como de los sentimientos positivos.

Los supuestos de este trabajo indican que la existencia e insistencia de palabras en Twitter que hablen sobre el ejercicio, representan preocupación e interés en cuidar la salud a través de una actividad física, sobre todo en tiempos de pandemia. En este punto, se cree que a pesar de que esta idea ha estado presente durante décadas, en la pandemia se tuvo mayor especulación sobre la urgencia de ejercitarse debido a que los casos graves de COVID estaban relacionados a la obesidad, hipertensión y otras enfermedades crónicas relacionadas con la falta de actividad física.

Por otra parte, esta información es relevante para empresas y diseñadores de productos deportivos, ya que es posible construir otros diccionarios para marketing al encontrar un nuevo enfoque para sus campañas publicitarias. Adicionalmente, con el fin de identificar causas y personas que recurren a productos milagro, se hace necesario un estudio independiente sobre estos productos para bajar de peso o de pastillas con supuestos nutrientes etc.

De igual manera, el impacto reflexivo que la pandemia tuvo con respecto a las actividades fue importante, incluso cuando la gente se encontraba en confinamiento, estaba preocupada por ejercitarse. Es claro que el COVID-19 generó implicaciones negativas en todo ámbito de la vida cotidiana, pero se observó que la gente cambió hábitos de sedentarismo con propósitos de mejorar su estilo de vida.

A partir de la investigación realizada nos quedan más entusiasmo para aplicar PLN a los resultados generados y a problemas relacionados con la salud, entonces, es posible promover no solo las actividades deportivas, sino invitar a seguir buenas conductas de alimentación y evitar ser seducidos por campañas

sucias que prometen falsamente solucionar la obesidad y distintas enfermedades por medio de charlatanería, independientemente de la época.

Referencias

Becerra-Pozas, J. L. (2018). *¿Qué es y cómo funciona el procesamiento del Lenguaje Natural en Inteligencia Artificial?* CIO México. *https://cio.com.mx/funciona-procesamiento-del-lenguaje-natural-en-inteligencia-artificial/*

Cáceres-Andrés, F. (2020) *El deporte en tiempos de pandemia: Una mirada desde Iberoamérica.* Organización de las Naciones Unidas para la Ciencia y la Cultura, Consejo Iberoamericano del Deporte. https://bit.ly/3ONBUqN

Cambria, E. & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Computational Intelligence Magazine, 9*(2), 48-57. https://doi.org/10.1109/MCI.2014.2307227

Díaz-Mendivelso, J. D. & Suarez-Baron, M. J. (2019). Análisis social aplicando técnicas de lenguaje natural a información extraída de Twitter. *Scientia Et Technica*, 24(3), 496-503. https://doi.org/10.22517/23447214.21731

Hendricks, B. (2022). *Sports Analytics with Natural Language Processing: Using Crowd Sentiment to Help Pick Winners in Fantasy Football* [Tesis de maestría, Harvard University Division of Continuing Education]. Digital Access to Scholarship at Harvard. https://bit.ly/3sufdk1

Instituto Nacional de Estadística y Geografía. (2018). *Encuesta Nacional de Salud y Nutrición: Presentación de resultados* [diapositivas de PowerPoint]. Secretaría de Salud, Instituto Nacional de Salud Pública https://bit.ly/3PdaH20

Kánter-Coronel, I. J. (2021). Magnitud del sobrepeso y obesidad en México: Un cambio de estrategia para su erradicación. *Mirada Legislativa*, (197), 1-24. http://bibliodigitalibd.senado.gob.mx/handle/123456789/5127

Marshall, C., Lanyi, K., Green, R., Wilkins, G.C., Pearson, F., Craig, D. (2022) Using Natural Language Processing to Explore Mental Health Insights From UK Tweets During the COVID-19 Pandemic: Infodemiology Study. *JMIR Infodemiology*, 2(1), e32449. https://doi.org/10.2196/32449

Organización Mundial de la Salud. (2022, 5 de octubre). *Actividad física* https://www.who.int/es/news-room/fact-sheets/detail/physical-activity

Organización de las Naciones Unidas. (2021, 14 de octubre). *Hacer ejercicio puede evitar hasta cinco millones de muertes al año.* https://news.un.org/es/story/2021/10/1498412#:~:text=La%20actividad%20f%C3%ADsica%20mejora%20la,poblaci%C3%B3n%20mundial%20fuera%20m%C3%A1s%20activa.

Piedra, J. (2020). Redes sociales en tiempos del COVID-19: El caso de la actividad física. *Sociología del deporte*, *1*(1), 41-43. https://doi.org/10.46661/socioldeporte.4998

Polero, P., Rebollo-Seco, C., Adsuar, J.C., Pérez-Gómez, J., Rojo-Ramos, J., Manzano-Redondo, F., Garcia-Gordillo, M. A. & Carlos-Vivas, J. (2021). Physical Activity Recommendations during COVID-19: *Narrative Review. International Journal of Environmental Research and Public Health, 18*(65), 1-24. http://dx.doi.org/10.3390/ijerph18010065

Rozalén-Barberán, S. (2020). *Aplicación Web para la gestión de competiciones de deporte electrónico usando el Framework Ruby on Rails* [tesis de licenciatura, Escola Tècnica Superior d'Enginyeria Informàtica, Universitat Politècnica de València]. Repositorio Institucional UPV. https://bit.ly/3KXWSSq

TwitterIR. (2022). *Quarterly results*. https://bit.ly/45JUnv9

Zhang, C., Wang, X., Yu, S., & Wang, Y. (2018). Research on Keyword Extraction of Word2vec Model in Chinese Corpus. En W. Xiong, S. Xu, H. Lee & W. Shang (Eds.), *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)* (pp. 339-343). https://doi.org/10.1109/ICIS.2018.8466534