



**UNIVERSIDAD DEL AZUAY**

**FACULTAD DE CIENCIA Y TECNOLOGÍA**

**ESCUELA DE INGENIERÍA ELECTRÓNICA**

**ANÁLISIS DE SENTIMIENTOS SOBRE LA VACUNA PARA COVID-19 EN LA RED  
SOCIAL TWITTER EN EL CONTEXTO ECUATORIANO**

**Trabajo de graduación previo a la obtención del título de:  
INGENIERO ELECTRÓNICO**

**Autor:  
Pedro Xavier Vallejo Cabrera**

**Director:  
MSc. Francisco Salgado**

**CUENCA-ECUADOR**

**2022**

## RESUMEN

### ANÁLISIS DE SENTIMIENTOS SOBRE LA VACUNA PARA COVID-19 EN LA RED SOCIAL TWITTER EN EL CONTEXTO ECUATORIANO

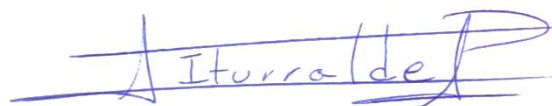
Se realizó un análisis de sentimientos sobre tweets levantados referentes a la vacunación contra el “SARS-COV-2”, durante la primera fase del plan nacional de vacunación, entre los meses de enero y julio de 2021 en el contexto ecuatoriano. Debido a la escasa relación cognitiva entre los contenidos de la técnicos carrera y sus formas o estructuras, el procesamiento de los datos y los resultados potenciales presentan limitaciones que se resuelven -contextualmente- al final del estudio. Se exponen parámetros definitivos sobre el impacto que la tecnología puede llegar a tener sobre el comportamiento social, obtenidos a través de algoritmos programáticos. Los resultados han sido analizados de manera general, para identificar percepciones o aproximarse a ello y se han realizado pruebas para determinar diferencias significativas entre los mismos.

**Palabras clave:** Twitter, COVID, Vacuna, Análisis de Sentimientos, Diferencias Significativas, Filosofía.



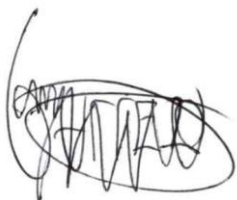
MSc. Francisco Salgado

Director del trabajo de titulación.



Ing. Daniel Iturralde Ph.D.

Coordinador de la Escuela de Ingeniería  
Electrónica.



Pedro Vallejo

Autor.

## ABSTRACT

### SENTIMENT ANALYSIS ABOUT COVID-19 VACCINE ON TWITTER IN THE ECUADORIAN CONTEXT

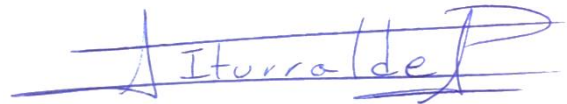
This study showed a sentiment analysis on several tweets referring to the vaccination program against the “SARS-COV-2” virus, during the first phase of the national plan, between January 2021 and July 2021 in the Ecuadorian context. Due to the poor cognitive relationship between the technical contents of the career and its forms or structures, data processing and possible results present limitations that get resolved -contextually- by the end of this report. Definitive parameters have been shown about the impact of technology on social behavior obtained by programmatic algorithms. Results have been analyzed to identify perceptions or approximate perceptions and tested to determine significant differences between the data.

**Keywords:** Twitter, COVID, Vaccine, Sentiment Analysis, Differences, Significance, Philosophy.



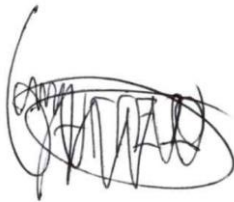
MSc. Francisco Salgado

Director del trabajo de titulación.



Ing. Daniel Iturralde Ph.D.

Coordinador de la Escuela de Ingeniería  
Electrónica.



Pedro Vallejo

Autor



# Análisis de sentimientos sobre la vacuna para COVID-19 en la red social *Twitter* en el contexto ecuatoriano.

Pedro Xavier Vallejo Cabrera  
Universidad del Azuay  
Cuenca, Ecuador  
pexvacab@es.uazuay.edu.ec

**Abstract—** Se realizó un análisis de sentimientos sobre tweets levantados referentes a la vacunación contra el “SARS-COV-2”, durante la primera fase del plan nacional de vacunación, entre los meses de enero y julio de 2021 en el contexto ecuatoriano. Se exponen parámetros definitivos sobre el impacto que la tecnología puede llegar a tener sobre el comportamiento social, obtenidos a través de algoritmos programáticos. Los resultados han sido analizados de manera general, para identificar percepciones o aproximarse a ello y se han realizado pruebas para determinar diferencias significativas entre los mismos. Se estima la percepción sobre el tema en la red social y se cuantifica en un espectro de “positivo-neutro-negativo”.

## I. INTRODUCCION

Las *sociedades* están en constante cambio, *evolución* o crecimiento, y esta característica implica una dinámica de las *estructuras* sociales, así como de los valores -que bien podrían denominarse “cultura”- en esas sociedades. [1]

Esa constante dinámica de *valores* provoca una *deconstrucción* de las estructuras que constituyen una sociedad. [1] En ese sentido, la interacción social *moderna* ha incluido elementos tecnológicos, cuya relevancia en momentos conlleva cierto impacto en la *naturaleza* misma de las relaciones humanas, el uso del internet y la masificación de las *redes sociales* virtuales han facilitado que aquel impacto no solo aumente en cantidad sino también en profundidad.

Se realiza un estudio sobre el uso de las redes sociales virtuales respecto de un tema específico -la vacuna para la COVID-19- en un contexto determinado -el territorio ecuatoriano- y, de manera breve, en un sentido conceptual.

A través de una lectura de *tweets* públicos, emitidos por usuarios dentro del territorio ecuatoriano, tanto en tópicos de discusión como en parámetros geográficos, se pueden realizar análisis lingüísticos para obtener cualidades *sentimentales* acerca de esas mismas expresiones públicas.

Usar, por un lado, herramientas del lenguaje para determinar percepciones y, por otro, recursos (herramientas) tecnológicos para optimizar y automatizar análisis de sentimientos relativos, podría traducirse como el objetivo principal de este estudio.

Para construir la herramienta que realice la comparación directa de términos y sus significados e intencionalidad, es preciso referirse al uso pragmático del lenguaje, definirlo y cuantificarlo. Con este objetivo -de manera interna, en las librerías- existen bases de datos de diccionarios que incluyen la información necesaria para dicha comparación, es decir, nubes de palabras con etiquetas para positivo y negativo. [2]

Para propósitos “técnicos” (ontológicos), el reporte se centrará en los resultados tanto estadísticos del levantamiento de datos como de las pruebas de significancia.

En otros trabajos realizados sobre análisis de sentimientos en *twitter* se abordan -de manera muy superficial- revisiones sobre el uso de técnicas de *machine learning* para la clasificación de textos e incluso sus “sentimientos”. [3] Los clasificadores bayesianos, por ejemplo, funcionan de una manera eficaz en cuanto a la identificación de los textos, las palabras más comunes y su función en las oraciones o frases. [3]

El análisis de sentimientos o “minado de opiniones” suelen pensarse como el mismo concepto, sin embargo, hay pequeñas pero consistentes diferencias entre ambos. Esclarecer aquellos conceptos, y por ende sus consecuencias, es una exigencia que -a pesar de ser ignorada sistemáticamente- radicaliza el valor de cualquier análisis. Suele pensarse que el minado de opiniones procura identificar la subjetividad de una frase mientras que la clasificación o análisis de sentimientos estudia la polaridad de una frase. [4] Pero, sin una conceptualización (definición) consistente y convencional, cualquier resultado y su estudio posterior termina siendo puramente interpretativo. Para ejemplificar mejor esta idea:

Considerando la frase:

“La calidad de imagen de mi nueva cámara Nikon V3 es genial”

Se nota que el sujeto de la opinión es “Nikon V3”, “calidad de imagen” es el aspecto del sujeto, el sentimiento es “positivo”, quién emite la “opinión” es el usuario y la fecha es parte de la *metadata* del *tweet*. [4] La clasificación del sentimiento, a pesar de parecer bastante directa, tiene implicaciones lingüísticas, conceptuales y contextuales que no han sido tomadas en cuenta para este análisis, de forma que, cualquier resultado estadístico sobre dicha clasificación, a pesar de tener información con cierto valor, no es sino una interpretación pura que se concreta al utilizar aquella estadística en cualquier actividad pragmática. Es decir, un grupo de *tweets* que se consideren “positivos o negativos” sobre un tópico puntual -antes de tener definido concretamente que es lo que significa emitir un comentario “positivo” alrededor de ese tópico- son una sencilla interpretación de lo que un comentario intenta mostrar.

Con eso en mente, se procede.

## II. METODOLOGIA

### A. Esquema del estudio

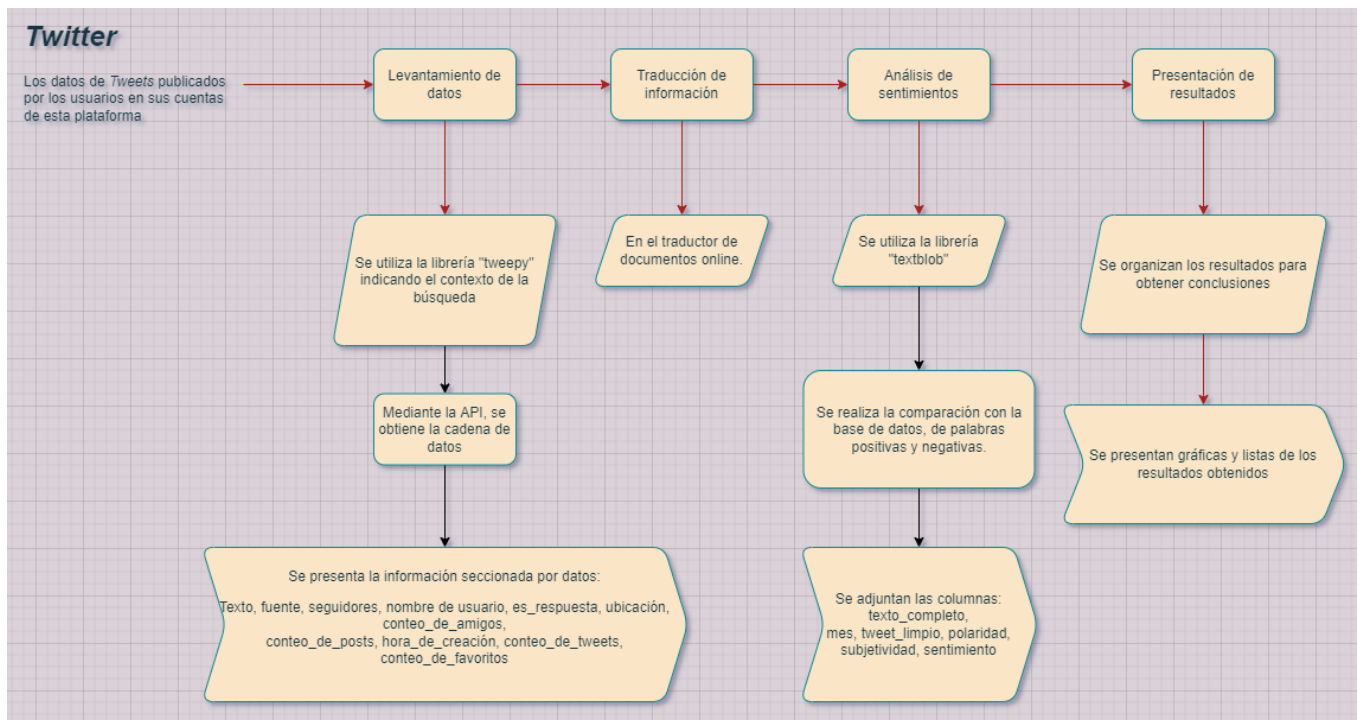
En primera instancia, es necesario describir una secuencia operativa para el levantamiento y tratamiento de

los datos, de forma que el trabajo pueda ser dividido en partes y realizado de manera óptima.

Se inició el levantamiento de los datos mediante el uso de librerías disponibles públicamente para manejar el API (*Application Programming Interface* – una aplicación que permite la comunicación entre dos o más programas [5]) de *Twitter* y poder acceder a los datos de la plataforma.

Luego, se procedió a realizar una traducción de todos los *tweets* obtenidos, pues, la librería que realiza el análisis de sentimientos está construida -con la totalidad de funciones- en idioma inglés. Esto nos obliga a perder un poco de semántica en el estudio, pero nos permite obtener una cuantificación más amplia de los datos ingresados.

Para realizar el análisis de sentimientos, se utiliza otra librería, denominada “textblob” de acceso libre, y desarrollada para el lenguaje *Python* que obtiene los datos. Se presenta un diagrama de flujo del estudio:



Diag. 1. Diagrama de flujo del estudio.

### B. Levantamiento de datos

Para obtener los datos de *tweets* publicados se utilizó la librería *tweepy*, de acceso libre y codificada para *Python*. Dicha librería puede comunicarse con el API de la plataforma y así realizar tareas de manera automática, por ejemplo, es capaz de realizar búsquedas con parámetros específicos, mismos que se encuentran disponibles en la documentación de la librería.

De manera puntual se ha realizado una búsqueda por tópicos, tiempo y por ubicación geográfica. Se ha focalizado la búsqueda en *tweets* provenientes de cuentas cuya etiqueta de ubicación es “Ecuador” y hayan sido publicados entre los meses de enero y julio de 2021 (la fase 1 de vacunación en el país), las palabras clave a buscar fueron: “*Vacuna COVID 19*”. [6]

La información obtenida de cada *tweet* se detalla así:

de la traducción y devuelve una cuantificación paramétrica de sentimiento e imparcialidad de cada oración, frase o -en esta oportunidad- *tweet*. Es decir, retorna un valor numérico por cada parámetro de “sentimiento” (positivo, negativo, neutro, objetivo, subjetivo) en cada *tweet*.

De manera adicional, se realizarán pruebas para determinar si las diferencias establecidas en los datos originados en tres ciudades son significativas entre sí.

**Texto, fuente, seguidores, nombre de usuario, es\_respuesta, ubicación, conteo\_de\_amigos, conteo\_de\_posts, hora\_de\_creación, conteo\_de\_tweets, conteo\_de\_favoritos, texto\_completo, mes, tweet\_limpio, polaridad, subjetividad, sentimiento.**

### C. Traducción de los datos

Los datos almacenados, con su orden secuencial, necesitan ser traducidos para poder operarlos con la librería “textblob”.

Para realizar la traducción masiva de datos se utilizó un servicio web que recibe archivos completos y los traduce usando el API de *Google*.

Dicho servicio está ubicado en una página web de acceso público.

#### D. Análisis de sentimientos

Para realizar el proceso de análisis, se utilizó la librería *textblob*, que es un *add-on* del lenguaje *Python*.

Esta librería recibe frases, las analiza utilizando algoritmos de *estructuración* lingüística, reconociendo características constitutivas de las expresiones para poder identificar técnicamente la información a procesar. [7] Es decir, identifica las partes de la frase, como el verbo, el sustantivo, etc.

El proceso de análisis recurre a una *base de datos* de palabras y frases (en inglés), -cada una con su identificación contextual y lingüística- en la que puede reconocerse, según parámetros *estructurales*, la *intencionalidad* de las expresiones dentro de un espectro de *positivo-negativo* y *objetividad-subjetividad* cuantificando cada dato.

La cuantificación *positivo-negativo* está en el rango de -1 a 1, siendo -1 muy *negativo* y 1 muy *positivo*.

Por otro lado, la cuantificación *objetividad-subjetividad* está en el rango de 0 a 1, siendo 0 muy *objetivo* y 1 muy *subjetivo*. [7]

La librería de análisis de sentimientos, de manera interna, realiza una comparación elemento a elemento -palabra a palabra- de cada frase o *tweet* con una nube de datos de palabras etiquetadas entre positivas y negativas, de forma que obtiene, según la frecuencia de aparición, una cuantificación que puede retornar como resultado.

Seguidamente, la comparación se realiza con cierta intención contextual, comparando no solo la palabra analizada, sino también, las palabras que la acompañan según la configuración de los parámetros del procedimiento.

Las técnicas que la librería del análisis utiliza son varias, podría generalizarse la aplicación de métodos de modelo lingüístico y memoria a corto y largo plazo, ambos catalogados como *machine learning* o *deep learning*. [8]

Sobre la aplicación de redes neuronales, para el modelamiento del lenguaje, puede afirmarse que es necesario para obtener información contextual, reconocer la forma de las palabras leídas. Entonces, el uso de técnicas de inteligencia artificial es, y quizá de manera implícita, imperativo en este estudio.

#### E. Presentación de resultados

Los resultados fueron filtrados para el análisis. Se presentan en forma gráfica y acorde a comparaciones geográficas, con la intención de interpretar de manera contextual las ocurrencias vislumbradas durante el proceso.

Se usan gráficas estadísticas sobre la cantidad y recurrencia de cada sentimiento en los *tweets* analizados. Se expondrá el desarrollo cronológico de las publicaciones y se analizarán posibles relaciones entre los datos.

Mediante un análisis exploratorio, que destaca la información levantada en las tres ciudades principales del Ecuador: Quito Guayaquil y Cuenca, se establece una serie de clasificaciones que describen de forma conjunta la interpretación de los *tweets* levantados.

#### F. Pruebas de Hipótesis

Con los resultados que el análisis devuelve, se ha propuesto una interpretación focalizada en los datos que

puedan aportar información relacionada con diferencias geográficas, presentado por ciudades, considerándose las tres más grandes y pobladas.

Se han realizado tres tests para determinar la significancia de las diferencias entre los datos de esas tres ciudades. En primera instancia el test de Fisher, luego el chi-cuadrado de Pearson, y por último el test de McNemar.

Todos los tests procuran estimar una cuantificación de la significancia de las diferencias entre dos o más instancias. [9]

Debido a que los datos se distribuyen entre variables independientes, es decir, cada ciudad y cada sentimiento tiene un proceso de variación individual y no se relacionan entre sí y a que los datos son cuadrados, es decir son tres ciudades versus tres estados de sentimientos (positivo, negativo, neutral) se han elegido las pruebas a realizar. [10]

Se han ajustado los datos para ampliar la información obtenida, es decir, dado que Twitter permite ingresar manualmente la localización o ubicación geográfica de una cuenta, existen datos que no contienen nombre correcto de la ciudad de origen, por ejemplo, hay casos en los que los usuarios han digitado las letras "uio" para indicar la ciudad de Quito, "gye" para Guayaquil o "cue" para Cuenca, de igual manera, se han tomado las ubicaciones que tienen especificada la provincia como parte de los datos originados en esas ciudades, entonces, un *tweet* etiquetado con la provincia del Azuay, se han considerado dentro de los datos obtenido en la ciudad de Cuenca, Guayas aparece como Guayaquil y Pichincha como Quito.

### III. DESARROLLO

#### A. Levantamiento de datos

En el lenguaje de programación *Python* se codificó una secuencia de comandos que interactúan con *Twitter* a través de su *API* y la librería *Tweepy* para obtener los datos de *tweets* que contengan palabras específicas y en un rango espacial y temporal también específico.

La tarea se ejecuta a través de dos líneas de código, una en la que se prepara la cadena de texto con palabras clave y los parámetros de búsqueda, y otra en donde se efectúa la búsqueda y levantamiento de datos.

Se encuentra, en los anexos, el código para el levantamiento de datos.

Los datos, que se levantan bajo demanda, deben ser almacenados y organizados para su próximo análisis. De modo que las líneas de código siguientes sirven para solicitar, ordenar y almacenar los datos desprendidos de la búsqueda.

Cada *tweet* contiene información "*metadata*" que incluyen datos del usuario y del *tweet* publicado, la especificación de los datos obtenidos dependerá de -y puede ser encontrada en- la *API* de la plataforma, de manera que aquí se detallarán únicamente los datos referentes al estudio actual.

Entre los parámetros obtenidos están las fechas de publicación, los usuarios y sus características (número de seguidores, procedencia, etc), la cantidad de *RT* (*re-tweets*) y las características del *tweet* (original, *retweet*, respuesta, etc).

En los anexos -de manera gráfica-, se detallan los puntos de datos específicos almacenados. Para finalizar se almacenan los datos en una hoja de *Excel*.

### B. Traducción de los datos

Hasta la fecha de realización de este trabajo, en la página web: *onlinedoctranslator*<sup>1</sup>, se puede encontrar un servicio de traducción de documentos gratuito, que utiliza el motor de traducción de *Google* para operar el servicio.

Para el objetivo de esta investigación, se precisa una traducción al idioma inglés, debido a que la librería de análisis de sentimientos funciona íntegramente en idioma inglés. [11]

Como se mencionará más a detalle en las conclusiones, la traducción presenta problemas y limitaciones por sí misma, pues, la terminología utilizada transforma la interpretabilidad de cada frase, y esto representa una desventaja, por lo menos, cuando se precisa el contexto de los datos analizados.

### C. Análisis de sentimientos

La librería encargada de realizar el análisis se denomina *Textblob*, tiene una estructura de algoritmos *inteligentes* que le permite realizar comparaciones con *bases de datos* de palabras y sus intencionalidades.

La librería se encuentra disponible de manera gratuita para el lenguaje *Python*, tiene características *sintácticas* que simplifican su uso y puede obtener información sobre el texto.

El índice *TF* indica la relación entre la cantidad de veces que un término se repite en un *tweet* y la cantidad de palabras en él, mientras que el *IDF* es una relación entre la cantidad de veces que una palabra se repite en un *tweet* y el resto de *tweets* levantados.

De forma matemática, se podría expresar:

$$idf(t, D) = \log \frac{N}{\text{cuanta}(d \in D: t \in d)}$$

En donde *N* es el número de documentos (en este caso *tweets*), *D* es el cuerpo (en este caso el *tweet* analizado) y *t* el término o palabra repetido. [12]

Índice	Descripción
<b>TF</b>	Indica la frecuencia de uso de las palabras.
<b>IDF</b>	Indica la repetitividad de palabras en un texto.

Tabla 1. Índices para realizar el análisis.

Adjunto en los anexos, se muestra el código para la obtención de los índices de análisis lingüístico.

La primera tarea que se debe realizar, previo al análisis, es la limpieza del texto levantado, es decir, eliminar elementos gráficos, enlaces web, palabras comunes -*stop words*-, símbolos especiales o lingüísticos -tildes, números, etc- y preparar el texto para enviarlo a estudiar.

De igual manera, se encuentran anexados los códigos para eliminar palabras y símbolos y limpiar los textos para proceder con el análisis, respectivamente. También se anexa

el código para obtener la cuantificación del análisis de sentimientos.

### D. Pruebas de Hipotesis

Las hipótesis consideradas en las pruebas han sido las siguientes:

**$H_0$ : Los sentimientos son independientes, por tanto, no varían respecto de la ciudad en donde se emitió el tweet.**

**$H_a$ : Los sentimientos son dependientes, por tanto varían según la ciudad en donde se emitió el tweet.**

Cabe señalar que para la realización de los *tests* que indican diferencias significativas, se han agrupado los datos según la ciudad en la que fueron publicados. Los datos se han seleccionado de todas las etiquetas que, directa o remotamente, se han identificado en una de las ciudades, también fueron limpiados los datos, eliminando información inexacta o que no guarde relación con la premisa del análisis en los *tests*. Específicamente se han filtrado todos los *tweets* cuya ubicación no esté identificada como dentro de una de las tres ciudades consideradas para las pruebas (Quito, Cuenca, Guayaquil).

Luego de aplicado dicho filtro, los datos contabilizan un 43,4% de la cantidad original de información obtenida.

Las premisas son excluyentes entre sí, es decir, si una prueba afirma la hipótesis  $H_0$  entonces, anula simultáneamente, la hipótesis  $H_a$ .

## IV. RESULTADOS

Los resultados que devuelven los códigos adjuntados en los anexos, han sido expuestos en forma gráfica para simplificar su interpretación. Los parámetros elegidos para realizar el análisis han devenido de la forma de funcionamiento de las librerías utilizadas, la búsqueda realizada y los objetivos planteados durante el diseño de la investigación.

El análisis retorna una lista de *tweets* que está dividida por meses, y tiene columnas agregadas que contienen la información del análisis. Se anexa una tabla que muestra los detalles de los datos.

### A. Sentimiento

La primera gráfica muestra los datos en general, sintetizados a partir de la misma tabla mencionada anteriormente.

<sup>1</sup> El enlace del sitio traductor es el siguiente:

<https://www.onlinedoctranslator.com/en/translationform>

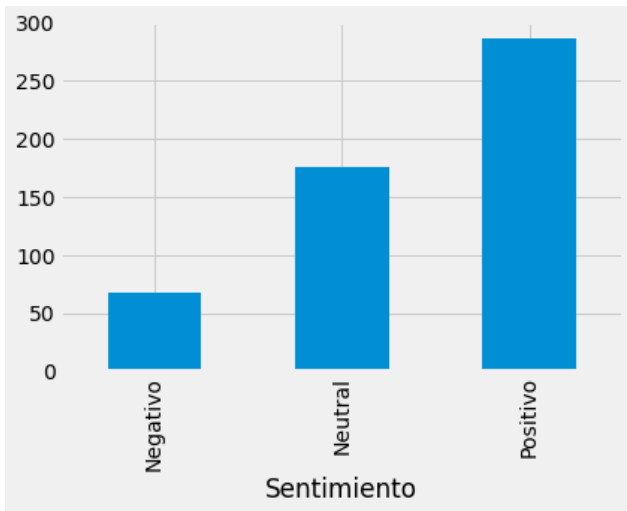


Fig. 1. Distribución general de los datos obtenidos.

El eje vertical representa la cantidad de tweets etiquetados con cada sentimiento y el eje horizontal muestra cada uno de los sentimientos que la librería define.

### B. Sentimiento por mes

La siguiente gráfica presentada muestra la primera clasificación de los datos considerada en este estudio, una especificación cronológica por mes, con la intención de exponer una relación o desarrollo de la cantidad y forma que toman los tweets en el tiempo.

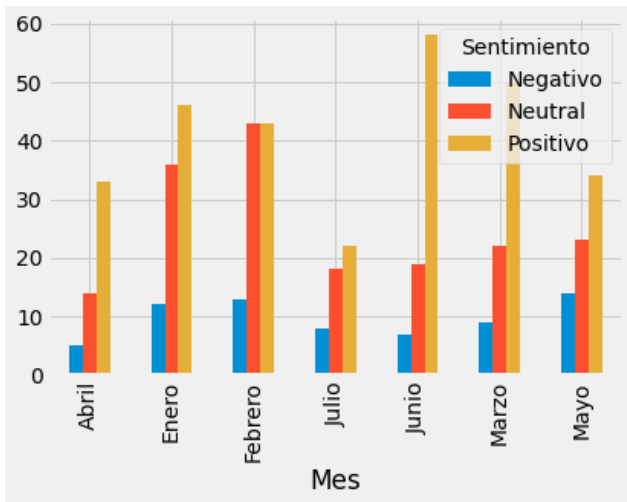


Fig. 2. Presentación de los datos de manera cronológica.

Otra vez, la gráfica utiliza el eje vertical para representar la cantidad de tweets mientras que la horizontal representa el mes en cuestión.

Aquí se expone una propiedad de los datos con respecto al plan de vacunación, pues, se puede notar que la “percepción” de los usuarios de twitter sobre la vacuna se vuelve mayormente “positiva” mientras avanza la ejecución del plan. A pesar de que el número de tweets no es constante, la cantidad de expresiones positivas aumenta en cada mes, hasta llegar a un punto medianamente estable, en donde empieza a decrecer lentamente.

Queda para investigaciones futuras continuar con el monitoreo cronológico del resto del plan de vacunación.

### C. Nube de palabras

Otra de las clasificaciones realizadas con los resultados se produce al contar la frecuencia de aparición de las palabras, es decir, contar las palabras usadas en los tweets de manera directa, para construir una “nube de palabras” que representa gráficamente dicho conteo.

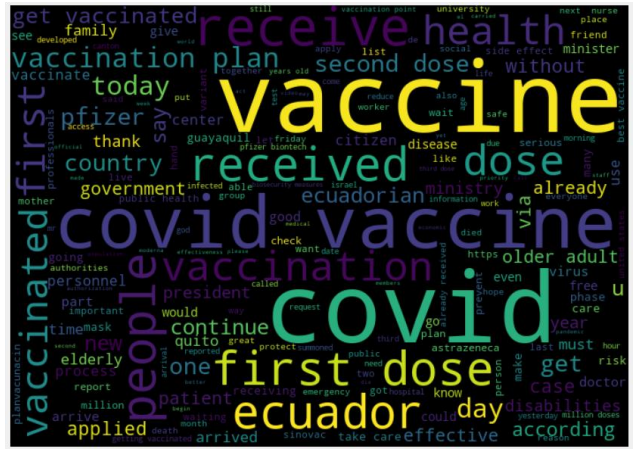


Fig. 3. Nube de palabras.

Las palabras se indican según su tamaño, la relación es directamente proporcional, las palabras más grandes y visibles son las más frecuentes.

Puede observarse, sin mayor sorpresa, que las palabras de la búsqueda y las que describen el contexto son las más frecuentes, sin embargo, se visibilizan también algunas marcas y dosis de vacunas que han sido mencionadas.

En este punto, no se considera el sentimiento del texto pues, el cálculo es puramente cuantitativo, es decir, no se considera la semántica de las palabras sino únicamente su repetitividad y frecuencia.

Cabe mencionar que en este punto los tweets han sido traducidos y debido a eso, aparecen en el idioma del análisis.

### D. Palabras más utilizadas

Luego de este conteo simple, se precisa realizar un análisis contextual, es decir, considerando que muchas de las palabras que aparecen tienen que ver con la búsqueda y aparecen por coincidencias literales más no aportan información específicamente analítica.

Entonces, la librería aplica técnicas lingüísticas para poder obtener información con mayor valor semántico en el procesamiento. La primera de estas técnicas tiene que ver con la frecuencia de aparición de las palabras, no de manera general, sino por tweets, entonces, se cuenta las palabras que más aparecen y los tweets que más aparecen para clasificar la frecuencia de las mismas en una forma un poco más contextual.



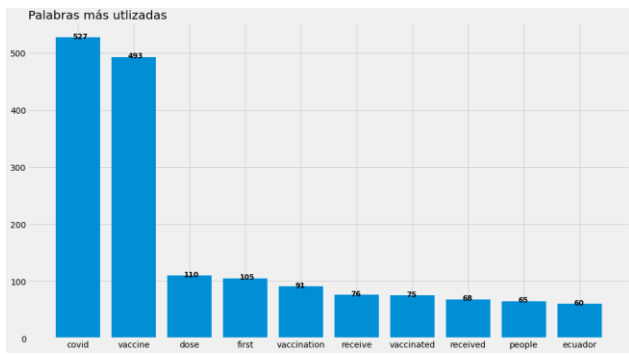


Fig. 4. Palabras más utilizadas.

En esta gráfica, los datos se representan por su cantidad en el eje vertical y por palabra en el eje horizontal. Se especifica la cantidad de veces que aparece cada palabra.

### E. Índices

La siguiente técnica que se utiliza implica la obtención de los índices descritos en la sección C de la parte “desarrollo” de este documento.

Los dos índices consideran, también, las palabras aledañas o que acompañan tanto previamente como después de cada una de las palabras en un *tweet*. Es decir, la repetitividad de las palabras que representan los índices, indica también la repetitividad de las palabras alrededor de la analizada. Esta técnica es también de un análisis contextual.

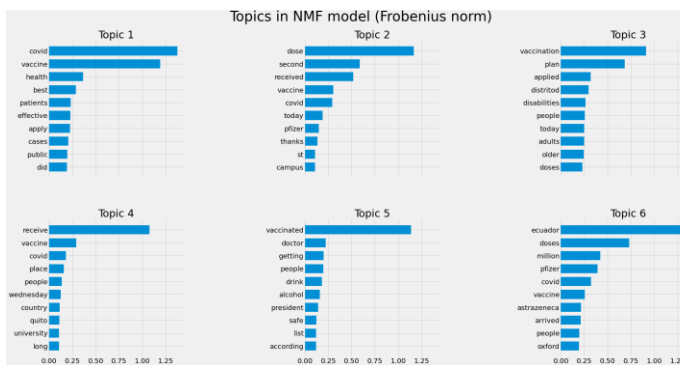


Fig. 5. Repetitividad de palabras.

De igual manera, y para consideraciones igualmente contextuales (análisis semánticos sobre tópicos específicos), se presenta una gráfica de palabras con relativa poca frecuencia de aparición cuando están acompañadas entre sí.

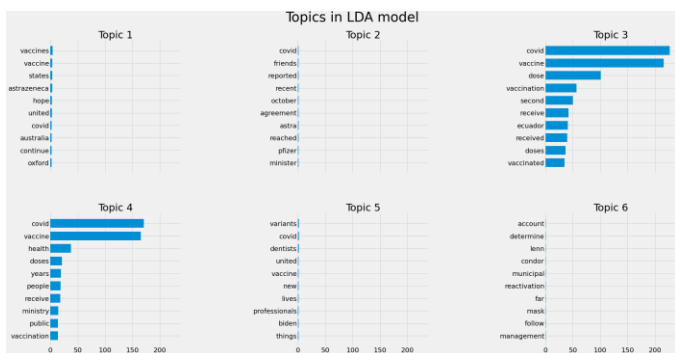


Fig. 6. Palabras de menor aparición.

Las gráficas, puntualmente, muestran las seis (6) primeras palabras de mayor repetición junto a los términos

que con mayor frecuencia las acompañan, se muestra también una cuantificación de la repetitividad de las mismas. En el caso de la primera gráfica se muestran las palabras más frecuentes y en la segunda gráfica se muestran las palabras de menor uso o frecuencia.

### F. Resultados de tests

Luego de filtrar y limpiar los datos, se han pasado a través de los algoritmos de las pruebas de hipótesis.

Los datos, ordenados por ciudad, no comparten la misma cantidad, es decir, cada ciudad tiene un número diferente de datos y por tanto muestran una cantidad independiente de “positivos” o “negativos”.

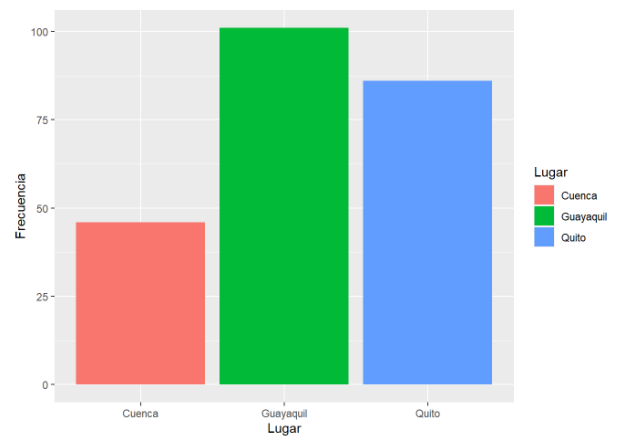


Fig. 7. Datos limpiados y ordenados por ciudad.

Para valorar los resultados, se entiende como diferencias significativas un resultado de “p” menor a 0.05. Por el contrario, un valor mayor a ese umbral refiere a una diferencia significativa entre el comportamiento de las variables.

Para probar la relación de dependencia entre los datos variantes por ciudad se ha realizado la prueba de independencia de Pearson. Este procedimiento es una prueba no paramétrica que requiere una distribución específica de los datos para poder aplicarse, exige que los datos sean cualitativos, de forma que su distribución es desconocida previa al análisis.

Los datos ingresados a la prueba se detallan en la tabla de contingencia:

Sentimiento	Lugar		
	Cuenca	Guayaquil	Quito
Negativo	8	10	17
Neutral	11	33	24
Positivo	27	58	45

Tabla 1. Tabla de contingencia.

La intención de este test es determinar si las proporciones de una variable (las variaciones) dependen en alguna forma de los valores que adquiere otra variable. El resultado nos ayudará a indicar si la percepción sobre la vacuna en una ciudad tiene alguna relación con la percepción de otra. [13]

El test de independencia de Pearson, por su parte, ha devuelto un resultado de:

$$p = 0.350$$

<i>Test</i>	<i>P</i>
<i>Pearson</i>	0.35

Tabla 2. Resultados de las pruebas de hipótesis.

Debido a la cantidad limitada de tweets, la separación de datos ha minimizado la clarificación de estas pruebas pues, para mejor fiabilidad de ellas es necesaria una base de datos más grande. Podrían extenderse las fechas de levantamiento de datos para obtener más información.

La aparente contradicción entre los resultados es completamente factible y viable, pues, tanto los datos son relativamente cortos, como las pruebas son metodológicamente distintas.

## V. DISCUSIONES

A lo largo del trabajo de investigación, han surgido cuestionamientos importantes sobre la metodología de levantamiento, análisis y filtrado de filtros, así como sobre la presentación e interpretación de resultados.

Primero, vale la pena calificar la calidad del levantamiento de datos, la sencillez de su aplicación y la fiabilidad del mismo. Debido al uso de la *API* de *Twitter* se puede simplificar bastante el proceso de búsqueda, levantamiento y análisis de los datos. La *API* retorna datos originales, por lo que los se puede confiar en el origen y almacenamiento de la información. Puesto que existen varias librerías programadas para interactuar directamente con esta aplicación nativa, basta con un par de líneas de código para acceder a los *tweets*.

Es probable que la mayoría de librerías, o por lo menos las más eficientes, usen un algoritmo similar, con interacción directa con la *API*, sin embargo, queda planteada la posibilidad de mejorar los datos cambiando la metodología de levantamiento de datos.

Seguidamente, al analizar la metodología de análisis de sentimientos cabe cuestionarse si el uso de herramientas lingüísticas considera el cambio de idioma para realizar análisis en idiomas diferentes al inglés nativo. Cabe considerar, así mismo, si las palabras tienen la intención que el algoritmo determina, es decir, figuras lingüísticas como el sarcasmo, la ironía o inclusive la propia ortografía pueden arrojar resultados erróneos, debido a la dificultad de reconocer la semántica en las expresiones.

Dentro del mismo ámbito, se debe tener en cuenta el hecho de que los textos son imperativamente cortos, por lo que términos como modismos o anglicismos toman significado de manera contextual, acompañados incluso de *emoticones* que terminan por cerrar el sentido de la frase.

Por otra parte, la interpretación de los resultados propone por sí misma una metodología que no exime responsabilidad de la intención investigativa o del autor. Entonces, siempre que los resultados sean interpretados con una intención específica, podrán encontrarse relaciones, dependencias u otras interacciones entre los datos que son capaces de aportar distintas índoles de información.

Para finalizar, de manera puntualmente técnica, cabe cuestionarse si las metodologías usadas en cada etapa de la investigación merecen ediciones programáticas que aporten

o efectivicen los análisis realizados. El uso de inteligencia artificial permite evolucionar los algoritmos para afinarlos cada vez mejor a las intenciones propuestas en los resultados a obtener.

## VI. CONCLUSIONES

Para poder iniciar una redacción acerca de las posibles conclusiones que este análisis puede devolver se vuelve menester definir algunos conceptos. La misma plataforma y sus conceptos, deben ser considerados previo a dicha redacción.

De esta forma, luego de revisar aquellas conceptualizaciones, se procura establecer -de manera *textual*- un conjunto de conclusiones que abordan, o intentan abordar, la mayoría de campos vislumbrados durante este ejercicio. Entonces, son los propios resultados del estudio los que, antes de clasificarse de manera axiomática, exigen un repaso conceptual para poder comprender la dimensión de la información develada.

### A. Conclusiones Técnicas

Para catalogar los aspectos técnicos abordados -a manera de conclusiones-, se enumeran varios puntos encontrados a lo largo de la investigación.

1) Las librerías utilizan herramientas lingüísticas avanzadas que superan el contenido de la carrera, lo que permite, o en algunos casos, obliga a trabajar con diversos temas y tópicos que pueden ser aportados por otros profesionales, especializados en las herramientas necesarias. Probablemente la inclusión de profesionales en ciencias sociales, dentro de un equipo multidisciplinario, amplie los campos de investigación y conclusión.

2) El uso de estas herramientas es simple, pues, estas utilizan algoritmos avanzados para análisis de todo tipo. Podrían incluirse dentro de los conocimientos impartidos de manera formal en los programas de estudio de la carrera.

3) Los modelos o estrategias para realizar este tipo de análisis son tan dinámicos como las formas de generación de contenido textual, es decir, las diferentes formas que adquieren los datos implican, cada una, su propia conceptualización, contextualización y -por ende- estudio. Las bases de datos de diccionarios de palabras constituyen una sola de esas formas, podrían realizarse análisis semánticos o sintácticos para identificar la intencionalidad de un texto.

4) La complejidad y rigurosidad de una investigación de este tipo, estará siempre ligada -inseparablemente- a las capacidades interpretativas, no solo de los sistemas activos, sino del/la propio/a autor/a del mismo. La subjetividad que implica el uso del lenguaje -complementariamente, de sus herramientas técnicas- es un vasto parámetro que influye en el procedimiento.

5) De acuerdo a la exposición de los resultados, pueden obtenerse algunas conclusiones. Por un lado, puede notarse la gran aceptación de la vacuna en la opinión pública, que, por otro lado, guarda interesantes similitudes cronológicas con la aplicación de la vacuna. Puntualizar -peor aún, cuantificar- la relación entre esos dos indicadores es una tarea que exige un análisis todavía más amplio, incluyendo consideraciones contextuales y el propio plan de

vacunación, por tanto, bastará con mencionar que, numéricamente, el avance del proceso de vacunación se aproxima al crecimiento de la aceptación de la misma en el universo de ciudadanos -usuarios- analizado.

6) El análisis de frecuencias y uso de palabras ha arrojado un conjunto de palabras específicas que llaman la atención respecto de la vacuna. La cantidad de palabras positivas es, relativamente, mucho mayor a la cantidad de palabras negativas y cada una de ellas está relacionada directamente con una nación o marca fabricante de una vacuna específica, pudiendo así notarse una inclinación contextual hacia una marca o dosis específica.

7) Las gráficas presentadas agrupan un alto número de datos en cada una de las categorías consideradas. Podría realizarse otro tipo de categorización de los datos para encontrar todavía más detalles sobre el impacto, en todo nivel, del uso de la red social *Twitter*.

8) La traducción presenta una desventaja importante a la hora de concretar el análisis, por dos cosas puntuales, primero el hecho de perder semántica contextual por el cambio de idioma, como se mencionó previamente, y luego porque es una parte del proceso independiente, es decir, separa el levantamiento de datos del análisis de sentimientos, entonces, obliga a romper la secuencia programática de los algoritmos, haciendo imposible que se realice todo en un solo código. A pesar de que se puede escribir en un solo script todo el código, este precisa almacenar los datos después de cada instancia para poder proceder con la secuencia.

9) La prueba de Pearson ha dejado notar que no existe evidencia suficiente para determinar una relación de dependencia entre las percepciones de las ciudades analizadas.

## B. Reflexiones y Recomendaciones

a) Para obtener datos de distinta naturaleza, pero en el mismo ámbito o tópico, podrían considerarse otras variables, incluidas en los datos, pero que no han sido protagonistas de este proyecto, por ejemplo, datos sobre demografía, etnia, clase, frecuencia de uso de la plataforma, popularidad entre otros usuarios, naturaleza de la información expuesta, información sobre los usuarios o las cuentas que son fuentes de datos, etc.

b) Podría ser muy conveniente formalizar las ramas científicas que, a pesar de ser ajenas a la materia *técnica* de la carrera, tienen importancia vital sobre los fundamentos teóricos y filosóficos que pretenden sentar base para los conocimientos estrictamente técnicos. Probablemente, se simplifiquen muchos conceptos -de alguna forma *abstractos*- que se utilizan de manera accesoria en los desarrollos tecnológicos.

c) Plantear cuestionamientos académicos podría desarrollar la potencialidad de la misma academia. Analizar los modelos educativos, los procedimientos escolásticos -sobre todo universitarios- y las maneras de aplicarlos resulta extremadamente importante para expandir las capacidades de la profesionalidad ofrecida. Analizar, con herramientas técnicas, información de índole -conceptualmente opuesta- social permite verticalizar la extensión profesional.

d) La traducción como tal, según Derrida, es un proceso deconstructivo, una estrategia que reúne varias técnicas lingüísticas para trasladar un mensaje, elemento a elemento (palabra a palabra) -y solo luego de su desmantelamiento del texto- desde un espacio hasta otro, transformando la superficialidad de un mensaje sin alterar su esencia. Sin embargo, un análisis como el realizado precisa con vital importancia considerar la construcción literal de una idea expresada, lo que -lamentablemente- es imposible sin mantenerse en el mismo idioma. Probablemente, la amplitud, exactitud y precisión de los resultados aumente cuantiosamente si las herramientas analíticas conservasen las palabras del idioma original.

e) Las opciones de investigaciones futuras que este desarrollo deja abierto son amplias, extensas, numerosas y relativas a cada intención investigativa. Los datos aportan mucha información valiosa sobre distintos tipos de abordajes, de hecho, las discusiones aquí presentadas demuestran aquella característica de los datos en este tipo de estudios, pues se analizan los resultados algorítmicos desde un punto de vista completamente disruptivo. Incluso, podría decirse que según esta visión particular del análisis es preciso definir conceptos tan complejos que, concluir de manera técnica sobre el mismo se vuelve -según el autor- prácticamente imposible. Por lo mismo, se plantea realizar más y mayores investigaciones a los tópicos aquí expuestos con características, metodologías y objetivos propios.

f) Se plantea realizar el mismo análisis levantando datos de fechas pertenecientes a otras fases del plan de vacunación y compararlos con los aquí presentados, con la intención de determinar una relación evolutiva en las cuestiones sentadas a lo largo de esta investigación.

## VII. REFERENCIAS

- [1] C. Castoriadis, *Transformación Social y Creación Cultural*, París: Sociologie et sociétés, 1979.
- [2] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, "Sentiment Analysis of Twitter Data," in *Proceedings of the Workshop on Language in Social Media*, Oregon, Association for Computational Linguistics, 2011, pp. 30-38.
- [3] A. Go, L. Huang and R. Bhayani, "Twitter Sentiment Analysis," Stanford University, Palo Alto, 2009.
- [4] A. Gianchanou and F. Crestani, "Like It or Not: A Survey of Twitter Sentiment Analysis Methods," *ACM Comput.*, 2016.
- [5] J. Ofoeda, R. Boateng and J. Effah, "Application Programming Interface (API) Research: A Review of the Past to Inform the Future," *International Journal of Enterprise Information Systems*, vol. 15, no. 3, pp. 76-95, 2019.
- [6] Tweepy, "tweepy.org," [Online]. Available: <https://docs.tweepy.org/en/stable/>. [Accessed 11 2022].

- [7] Textblob, "textblob.io," [Online]. Available: <https://textblob.readthedocs.io/en/dev/>. [Accessed 11 2022].
- [8] "A Novel Machine Learning Approach for Sentiment Analysis," *Molecular Diversity Preservation International (MDPI) - ASI (Applied System Innovation)*, vol. 5, no. 13, 2022.
- [9] W. G. Cochran, "Approximate Significance Levels of the Behrens-Fisher Test," *International Biometric Society, Biometrics*, vol. 20, no. No. 1, pp. 191-195, 1964.
- [10] R. Joaquín Amat, "Test estadísticos para variables cualitativas: test exacto de Fisher, chi-cuadrado de Pearson, McNemar y Q-Cochran," *cienciadedatos.net*, Enero 2016. [Online]. Available: [https://www.cienciadedatos.net/documentos/22.2\\_test\\_exacto\\_de\\_fisher\\_chi-cuadrado\\_de\\_pearson\\_mcnemar\\_qcochran.html](https://www.cienciadedatos.net/documentos/22.2_test_exacto_de_fisher_chi-cuadrado_de_pearson_mcnemar_qcochran.html). [Accessed 30 03 2022].
- [11] onlinedoctranslator, "onlinedoctranslator.com," [Online]. Available: <https://www.onlinedoctranslator.com/en/translationform>. [Accessed 11 2022].
- [12] Z. Yun-Tao, G. Ling and Y.-c. Wang, "An improved TF-IDF approach for text classification," *Journal of Zhejiang University SCIENCE*, vol. 6, no. I, pp. 49-55, 2004.
- [13] J. Amat Rodrigo, "cienciadedatos.net," 01 2016. [Online]. Available: [https://www.cienciadedatos.net/documentos/22.2\\_test\\_exacto\\_de\\_fisher\\_chi-cuadrado\\_de\\_pearson\\_mcnemar\\_qcochran#Introducci%C3%B3n](https://www.cienciadedatos.net/documentos/22.2_test_exacto_de_fisher_chi-cuadrado_de_pearson_mcnemar_qcochran#Introducci%C3%B3n). [Accessed 15 04 2022].

## VIII. ANEXOS

### Anexo 1:

```

1. keyword = 'Vacuna COVID 19 -is:retweet place_country:"EC"'
2. noOfTweet = 100
3.
4.
5. tweets = tweepy.Cursor(api.search_full_archive, environment_name='ScientificResearch',
6.                        query=keyword,
                           fromDate=202107300000, toDate=202108300000).items(noOfTweet)

```

Anexo 1. Código para el levantamiento de datos. (tweets)

### Anexo 2:

```

1. for tweet in tweets:
2.     #print(tweet.extended_tweet.full_text)
3.     status_id.append(tweet.id_str)
4.     status = api.get_status(tweet.id_str, tweet_mode="extended")
5.     try:
6.         rt_status.append(status.retweeted_status.full_text)
7.     except AttributeError:
8.         rt_status.append(status.full_text)
9.
10.    tweet_list.append(tweet.text)
11.    created_at.append(tweet.created_at)
12.    source.append(tweet.source)
13.    in_reply_to_status_id.append(tweet.in_reply_to_status_id_str)
14.    in_reply_to_screen_name.append(tweet.in_reply_to_screen_name)
15.    retweet_count.append(tweet.retweet_count)
16.    favorite_count.append(tweet.favorite_count)
17.    location.append(tweet.user.location)
18.    follower.append(tweet.user.followers_count)
19.    screen_name.append(tweet.user.screen_name)
20.    friends_count.append(tweet.user.friends_count)

```

```
21. statuses_count.append(tweet.user.statuses_count)
22. created_at_us.append(tweet.user.created_at)
```

Anexo 2. Código para el levantamiento de datos. (características de tweets)

Anexo 3:

```
1. print("Extracting tf-idf features for NMF...")
2. tfidf_vectorizer = TfidfVectorizer(max_df=0.95, min_df=2,
3.                                 max_features=n_features,
4.                                 stop_words='english')
5.
6. mattfidf = tfidf_vectorizer.fit(frame["cleaned_tweet"])
7. tfidf = mattfidf.transform(frame["cleaned_tweet"])
8. print("Finish...")
9.
10. pd.DataFrame(tfidf.toarray(), columns=tfidf_vectorizer.get_feature_names_out())
11.
12. # Use tf (raw term count) features for LDA.
13. print("Extracting tf features for LDA...")
14. tf_vectorizer = CountVectorizer(max_df=0.95, min_df=2,
15.                                max_features=n_features,
16.                                stop_words='english')
17.
18. tf = tf_vectorizer.fit_transform(frame["cleaned_tweet"])
19. print("Done ... ")
20. print()
21.
22. pd.DataFrame(tf.toarray(), columns=tf_vectorizer.get_feature_names_out())
23.
24. # Fit the NMF model
25. print("Fitting the NMF model (Frobenius norm) with tf-idf features, "
26.       "n_samples=%d and n_features=%d..."
27.       % (n_samples, n_features))
28.
29. nmf = NMF(n_components=n_components, random_state=1, l1_ratio=.5).fit(tfidf)
30.
31. tfidf_feature_names = tfidf_vectorizer.get_feature_names_out()
32. plot_top_words(nmf, tfidf_feature_names, n_top_words, 'Topics in NMF model (Frobenius norm)',3,2)
33.
34. print('\n' * 2, "Fitting LDA models with tf features, "
35.       "n_samples=%d and n_features=%d..."
36.       % (n_samples, n_features))
37. lda = LatentDirichletAllocation(n_components=n_components, max_iter=5,
38.                                 learning_method='online',
39.                                 learning_offset=50.,
40.                                 random_state=0)
41. lda.fit(tf)
42. print("done...")
43.
44. tf_feature_names = tf_vectorizer.get_feature_names_out()
45. plot_top_words(lda, tf_feature_names, n_top_words, 'Topics in LDA model',3,2)
```

Anexo 3. Código para obtención de índices.

Anexo 4:

```
1. def word_tokenize_clean(sentence, to_lower=True, remove_special_chars=True,
2.                          remove_numbers=True, remove_stopwords=True):
3.     token_clean = word_tokenize(sentence)
4.     if to_lower:
5.         token_clean = [i.lower() for i in token_clean]
6.     if remove_special_chars:
7.         token_clean = [re.sub(r'\W+', '', i) for i in token_clean]
8.     if remove_numbers:
9.         token_clean = [re.sub(r'[0-9]+', '', i) for i in token_clean]
10.    if remove_stopwords:
11.        token_clean = [i for i in token_clean if i.lower() not in stopwords.words('spanish')]
12.    token_clean = [i for i in token_clean if i!='']
```

```
13. return token_clean
```

Anexo 4. Código para eliminar palabras y símbolos. (tweets)

Anexo 5:

```
1. def clean_text(text):
2.     pat1 = r'@[^ ]+'
3.     pat2 = r'https?://[A-Za-z0-9./]+'
4.     pat3 = r'\s'
5.     pat4 = r'\#\w+'
6.     pat5 = r'& '
7.     pat6 = r'^A-Za-z\s]'
8.     combined_pat = r'|'.join((pat1, pat2, pat3, pat4, pat5, pat6))
9.     text = re.sub(combined_pat, "", text).lower()
10.    return text.strip()
11.
12. frame["cleaned_tweet"] = frame["texto"].apply(clean_text)
```

Anexo 5. Código para la limpieza de los datos. (tweets)

Anexo 6:

```
1. print("Procesando sentimiento ..... ")
2. for row in frame.itertuples():
3.     tweet = frame.at[row[0], 'cleaned_tweet']
4.
5.     #run sentiment using TextBlob
6.     analysis = TextBlob(tweet)
7.
8.     #set value to dataframe
9.     frame.at[row[0], 'polarity'] = analysis.sentiment[0]
10.    frame.at[row[0], 'subjectivity'] = analysis.sentiment[1]
11.
12.    #Create Positive / negative column depending on polarity
13.    if analysis.sentiment[0]>0:
14.        frame.at[row[0], 'Sentimiento'] = "Positivo"
15.    elif analysis.sentiment[0]<0:
16.        frame.at[row[0], 'Sentimiento'] = "Negativo"
17.    else:
18.        frame.at[row[0], 'Sentimiento'] = "Neutral"
19.
20. print("Listo el procesamiento")
21.
22. frame.to_excel("Resultados\Análisis.xlsx")
```

Anexo 6. Código para el análisis de sentimientos.

Anexo 7:

	texto	Mes	cleaned_tweet	polarity	subjectivity	Sentimiento
0	Johnson & amp; Johnson requests authorization for his vaccine against Covid-19 in the US <a href="https://t.co/KSU2iRvLth">https://t.co/KSU2iRvLth</a>	Enero	johnson amp johnson requests authorization for his vaccine against covid in the us	0	0	Neutral
1	A new batch of vaccines is scheduled to arrive in Ecuador on February 15 to complete the second phase of the vaccination dose against Covid-19. #vaccine #coronavirus #trendspulse <a href="https://t.co/RmGSAqDLTV">https://t.co/RmGSAqDLTV</a>	Enero	a new batch of vaccines is scheduled to arrive in ecuador on february to complete the second phase of the vaccination dose against covid	0.078788	0.284848	Positivo

2	AstraZeneca ensures that its vaccine against covid-19 protects 100% from serious cases and deaths <a href="https://t.co/2S4iuExwrY">https://t.co/2S4iuExwrY</a>	Enero	astrazeneca ensures that its vaccine against covid protects from serious cases and deaths	-0.33333	0.666667	Negativo
3	And so I received my first dose of the COVID 19 vaccine. To wait 21 days for the 2nd dose. The wait has been worth it <a href="https://t.co/7XCS2AZJIG">https://t.co/7XCS2AZJIG</a>	Enero	and so i received my first dose of the covid vaccine to wait days for the nd dose the wait has been worth it	0.275	0.216667	Positivo
4	@ALBERTOACOSTAB Finally, the vaccine is classified as a global public good, so it will be necessary to see how Mexico does, which is the first nation to approve that the private sector import and sell vaccines but not it will be easy. <a href="https://t.co/szmucn0ZNi">https://t.co/szmucn0ZNi</a>	Enero	finally the vaccine is classified as a global public good so it will be necessary to see how mexico does which is the first nation to approve that the private sector import and sell vaccines but not it will be easy	0.172917	0.526042	Positivo
5	@ALBERTOACOSTAB There is also the WHO COVAX initiative that brings together 172 countries that will receive the vaccine after commitments with manufacturers. <a href="https://t.co/shSM2ZMzmN">https://t.co/shSM2ZMzmN</a>	Enero	there is also the who covax initiative that brings together countries that will receive the vaccine after commitments with manufacturers	0	0	Neutral
...	...	...	...	...	...	...

Anexo. 7. Tabla de resultados del análisis de datos.

