ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

UNIDAD DE TITULACIÓN

DESARROLLO DE UN APLICATIVO WEB PARA MONITOREAR EFECTOS PSICOLÓGICOS DE UNA PANDEMIA USANDO TÉCNICAS DE MINERÍA DE DATOS E INFORMACIÓN DE NOTICIAS Y TWEETS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

JULIO DAVID ROSERO GÓMEZ
julio.rosero@epn.edu.ec

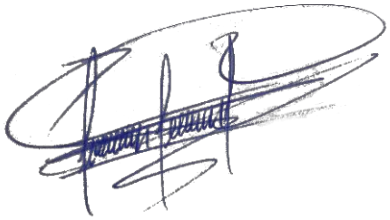
DIRECTORA:
DRA MARIA HALLO
maria.hallo@epn.edu.ec

Quito, Mayo 2022

DECLARACIÓN

Yo, Julio David Rosero Gómez, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

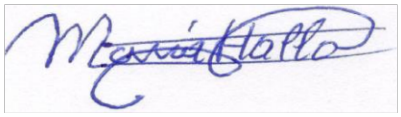
A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.



Julio David Rosero Gómez

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Julio David Rosero Gómez, bajo mi supervisión.



PhD. María Hallo.

DIRECTORA DEL PROYECTO

DEDICATORIA

A memoria de mi madre.

AGRADECIMIENTO

Agradezco a mi padre, Julio por inculcarme sus valores, principios y pasión por la ciencia. A mi novia Estefania por su amor, apoyo y motivación. A mi directora de proyecto, PhD. María Hallo por su ayuda y guía en la realización de este proyecto. Al psicólogo Damián Llumiquinga por su ayuda en este proyecto.

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

En el presente proyecto de titulación se construye un modelo de minería de datos para encontrar posibles efectos psicológicos al presentarse una pandemia. El análisis se lo realiza buscando las emociones y sentimientos expresados en la red social de Twitter. Las emociones en conjunto con un análisis de tópicos permiten conocer los posibles problemas psicológicos de una población. Para evaluar la evolución de una emoción o efecto psicológico se han obtenido datos en tres fechas consideradas de interés general.

El proyecto fue realizado usando la metodología CRISP-DM, metodología especializada para la minería de datos. La metodología cubre desde la recolección de datos hasta la presentación de un informe con los resultados del proceso de minería de datos. Esta metodología se caracteriza por su flexibilidad; dependiendo del proyecto sus fases tendrán mayor o menor importancia. En el caso de este proyecto la búsqueda patrones y la visualización de los datos son los puntos más importantes, por lo cual las fases de modelado y evaluación son las más relevantes. Adjunto a las fases de esta metodología se ha sumado una fase de desarrollo de un aplicativo web, para la visualización de los resultados.

Todo el proyecto se lo desarrolló con el lenguaje de programación Python desde la obtención de datos hasta la visualización de los resultados. La técnica de obtención de datos fue raspado web o web scraping usando la librería Selenium, para evitar las restricciones que presenta la API de Twitter. La obtención de sentimientos y emociones se lo realizó usando el diccionario de emociones National Research Council Canada (NRC) y su librería en Python. La técnica para obtener los tópicos fue etiquetado gramatical (part of speech) presente en la librería Stanza. Para la presentación de los cuadros de mando (dashboards) con los resultados se usó la librería Streamlit.

Los resultados finales permiten observar que problemas como la depresión, ansiedad y estrés son frecuentes en una pandemia. Independientemente de la población que se analice, las emociones predominantes son el miedo y la tristeza. Otras emociones como la ira y la aversión varían su intensidad dependiendo de la población que se estudie.

Palabras Clave: análisis de emociones, análisis de tópicos, CRISP-DM, minería de datos, pandemia, problemas psicológicos

ABSTRACT

In this degree Project, data mining mode is built to find posible psychological effects when a pandemic occurs. The analysis is carried out looking for the emotions and feeling expressed in Twitter. The emotion analysis added to the topics analysis allow knowing the posible psychologycal problems of a population. To evaluate the evolution of an emotion or psychological effect, data have been obtained on three dates considered to be of general interest.

The project was carried out using the CRISP-DM methodology, a specialized methodology for data mining. The methodology covers from data collection to the presentation of a report with the results of the data mining process. This methodology is characterized by its flexibility; depending on the project, its phases will have greater or lesser importance. In the case of this project, the search for patterns and the visualization of the data are the most important points, for which the modeling and evaluation phases are the most relevant. Attached to the phases of this methodology, a development phase of a web application has been added, for the visualization of the results.

The entire project was developed with the Python programming language from data collection to visualization of results. The data collection technique was web scraping using the Selenium package to avoid the restrictions presented by the Twitter API. Sentiments and emotions were obtained using the National Research Council Canada (NRC) dictionary and its Python package. The technique to obtain the topics was part of speech (PoS) present in the stanza package. For the presentation of the dashboards, Streamlit was used.

The final results allow us to observe that problems such as depression, anxiety and stress are frequent in a pandemic. Regardless of the population analyzed, the predominant emotions are fear and sadness. Other emotions such as anger and disgust vary their intensity depending on the population being studied.

Key words: emotion analysis, topic analysis, CRISP-DM, data mining, pandemic, psychological problems

1. INTRODUCCIÓN

1.1. Objetivos

1.1.1. Objetivo general

Desarrollar un aplicativo web que permita reflejar los resultados referentes a efectos psicológicos por pandemias obtenidos mediante minería de datos en sitios web y tweets.

1.1.2. Objetivos específicos

- Realizar un estudio previo de trabajos relacionados con la minería de datos referentes a problemas psicológicos que pueden surgir al desatarse una pandemia.
- Desarrollar un proyecto de minería de datos para el desarrollo de una aplicación web que permita la visualización de información referente a efectos psicológicos que surgen como consecuencia de una pandemia.
- Evaluar los resultados que se han obtenido mediante la minería de datos con el fin de llegar a conclusiones sólidas.

1.2. Minería de datos enfocado a la detección de problemas psicológicos en una pandemia

Las redes sociales ofrecen una gran fuente de datos para la investigación, incluida la salud mental (Conway, Hu, & Chapman, 2019). Un ejemplo de esto es el estudio realizado por Oladapo Oyebode (Oyebode, y otros, 2021) donde se analizan datos extraídos de Facebook, Twitter y YouTube con la finalidad de encontrar temas y agruparlos en categorías. De esta manera fue posible determinar los principales problemas psicológicos causados por el COVID-19. Según Oyebode la emoción que más sobresale es el miedo o el miedo por el COVID-19; como resultado de este miedo muchas personas recurrieron a compras de pánico, almacenando productos esenciales para limitar su exposición (Oyebode, y otros, 2021). Otras emociones a resaltar son la frustración, tristeza e ira; dichas emociones son producidas por la incertidumbre de situaciones futuras y al cambio del estilo de vida (Li, Wang, Xue, Zhao, & Zhu, 2020).

Los problemas de salud mental mencionados en el estudio (Oyebode, y otros, 2021) marcan la ansiedad, depresión estrés y trastorno obsesivo compulsivo como los principales problemas psicológicos. El estudio no menciona en concreto las causas de estos problemas, pero si las marca como consecuencias indirectas de las restricciones del COVID-19. Según Hao Yao (Yao, Chen, & Xu, 2020) los trastornos de la salud mental a causa de la preocupación por la

reciente pandemia están asociados a la ansiedad y depresión. El mismo estudio menciona que pacientes que padecen enfermedades mentales son más susceptibles a contagiarse.

Oyebode menciona que se extrajo 47 millones de tweets mediante la API de Twitter. Los tweets provienen de 26 hashtags relacionados al COVID-19 y comprendidos entre marzo y abril del 2020. Después de los procesos de limpieza y preparación de los datos, el total de tweets se redujo a 8 millones, de los cuales se trabajó únicamente con un total de 1 millón de tweets seleccionados al azar. La técnica usada para encontrar frases significativas y determinar emociones que presenten al grupo de tweets fue: etiquetado gramatical o part of speech (PoS) (Oyebode, y otros, 2021). La técnica de PoS consiste en separar un texto en palabras y asociar cada palabra con su clase, ya sea un verbo, adjetivo, sustantivo, etc.

El estudio realizado por Yinghui Huang (Huang, y otros, 2021) menciona que la población en China recurrió a plataformas de servicios de salud mental en línea. El comportamiento de los usuarios en dichas plataformas se categorizó en problemas psicológicos y factores influyentes. Los problemas psicológicos para destacar son: depresión, ansiedad, tendencias suicidas, fobia social, sentimientos de preocupación, miedo e ira. Por otro lado, los factores influyentes fueron: relaciones interpersonales, amor, familia, trabajo, psicoterapia y matrimonio. El estudio además menciona que la población que padeció problemas psicológicos fue por una fuerte presencia de emociones negativas, entre las que resaltan: preocupación, miedo e ira. Del mismo modo los trastornos mentales más grandes fueron: tendencias suicidas y depresión (PRC, 2021).

Los lugares de los cuales se obtuvieron los datos fueron plataformas de asesoramiento en línea. Dichas plataformas funcionan como foros de preguntas y respuestas de manera anónima. El total de registros obtenidos fue de alrededor de 40 mil comentarios provenientes de tres sitios web (xinli001.com, bazhuayu.com y WeChat). La técnica usada para la obtención de datos en los tres sitios web fue el raspado web o web scraping. La técnica de web scraping consiste en el uso de bots para extraer información de sitios web. Finalmente, para la clasificación de emociones y frases recurrentes se creó un diccionario enfocado al léxico de enfermedades mentales (Huang, y otros, 2021). El diccionario se usó en conjunto con el algoritmo de agrupamiento k-means. Este algoritmo tiene la finalidad de separar en grupos a un conjunto de datos.

La investigación realizada por Miguel Muñoz (Muñoz, Recéndez, & Ñánez, 2021) menciona a la depresión y ansiedad como resultado de la problemática de la pandemia por COVID-19. Menciona que estos trastornos surgen por la anticipación de algún peligro ya sea real o imaginario causado por el ámbito social. La emoción que resalta su estudio es el miedo proveniente a la incertidumbre surtida por la pandemia. Los autores también realizan un

análisis de tópicos para conocer la opinión pública, de estos los temas que sobresalen son: “ansiedad y depresión”, “salud mental”, “ataques de ansiedad” y “ansiedad social”. Estos tópicos brindan ayudan al conocer los problemas psicológicos más importantes de los que se está hablando, en este caso depresión y ansiedad.

La fuente de datos para el estudio de Miguel Muñoz fue Twitter utilizando técnicas de web scraping. El filtrado de datos se lo realizó mediante las palabras clave COVID y ansiedad. Los datos extraídos fueron obtenidos entre enero y mayo del 2021 en idioma español dando un resultado 1062 tweets. Para la determinar los principales problemas mentales los autores se basan en la obtención de frases frecuentes mediante la versión gratuita de la aplicación WordStat, creada para el análisis de textos. Las frases frecuentes son cotejadas con el DSM-5 (psiquiatría, 2014) de la asociación de psicólogos americanos. De esta manera crean una relación entre las frases frecuentes y manual DSM-5 para encontrar trastornos psicológicos y emociones predominantes.

1.3. Aspectos psicológicos asociados a una pandemia

La reciente pandemia por COVID-19 marca el más reciente confinamiento de una gran parte de la población, es de suponer que este tenga un impacto importante en el bienestar físico y psicológico de dicha población. El cierre de centros educativos, la paralización parcial o total de actividades económicas, el aumento del desempleo, la declaración de una cuarentena por varias semanas ha supuesto una situación que puede presentar múltiples estímulos que pueden llegar a generar estrés.

En el estudio realizado por Wang (Wang & Pan, 2020) se afirma que durante el confinamiento existen dos factores que son los más afectados, el bienestar físico y el bienestar psicológico, en dicho estudio se afirma que estas afecciones son causadas por la pérdida de rutinas, la interrupción de hábitos y el reemplazo de los mismos por otros poco saludables como, sedentarismo, mala alimentación y un mayor tiempo de exposición frente a una pantalla, podrían derivar en problemas físicos (Nekane Balluerka, 2020). Cabe esperar que aquellas personas que tengan un mayor grado de vulnerabilidad serán aquellas que presenten desventaja por edad, sexo, estructura familiar, nivel educativo, origen étnico, situación o condición física y/o mental (Nekane Balluerka, 2020).

La implementación de medidas como una cuarentena incrementa la posibilidad de problemas psicológicos (Huarcaya, 2020), esto principalmente por la ausencia de comunicación interpersonal, haciendo que los trastornos depresivos y ansiosos ocurran o empeoren. Los posibles problemas de salud mental que marca (Huarcaya, 2020) son: ansiedad, depresión y estrés.

1.3.1. Ansiedad relacionada con la salud

La ansiedad por la salud que una persona puede llegar a tener se caracteriza por una interpretación catastrófica de sensaciones y cambios corporales, creencias disfuncionales acerca de la salud y malos mecanismos adaptativos (Huarcaya, 2020). El ejemplo más claro de esto se puede encontrar en la pandemia de COVID-19, en donde las personas que presentaban niveles altos de ansiedad por la salud eran susceptibles a interpretar sensaciones corporales relativamente inofensivas como una sintomatología aludiendo que se encontraban infectados.

A pesar de que ciertas conductas son producto de las recomendaciones de la sociedad médica, las personas que presentan cuadros de ansiedad por la salud, llevan estas recomendaciones al extremo lo cual puede llevar a la compra en exceso de material de higiene personal, mascarillas, guantes, jabón, etc. En el otro extremo están aquellas personas que tienen bajo nivel de ansiedad por la salud, por lo cual tienden a creer que no pueden contagiarse por lo cual incumplen las recomendaciones de salud pública y distanciamiento social. (Huarcaya, 2020)

1.3.2. Depresión

Con la aparición del nuevo coronavirus (COVID-19) y el rápido aumento en el número de casos y muertes se han creado diferentes problemas psicológicos entre ellos depresión, la cual ha afectado tanto al personal médico como la población en general (Ozamiz Etxebarria, Dosil Santamaria, Picaza Gorrochategui, & Idoiaga Mondragon, 2020). Las primeras investigaciones relacionadas a trastornos por depresión vienen por estudios realizados en China (Ozamiz Etxebarria, Dosil Santamaria, Picaza Gorrochategui, & Idoiaga Mondragon, 2020) donde con 1210 personas se administró la “Escala de Depresión, Ansiedad y Estrés” (DASS-21) con el objetivo de conocer el nivel de impacto psicológico y de depresión en la etapa inicial del brote causado por el COVID-19, en este estudio se concluyó que el 16.5% de los participantes indicó síntomas depresivos que van desde moderados a graves.

1.3.3. Estrés psicológico

En un estudio realizado también en China con 52730 personas se concluyó que el 35% de los participantes presentó estrés psicológico principalmente en grupos del sexo femenino (Qiu, y otros, 2020). Algunos grupos como personas mayores de 60 años y personas entre 18 y 30 años presentaron niveles de estrés más elevados; esto puede deberse a que la mayor tasa de mortalidad por causa del COVID-19 está presente en adultos mayores, mientras que la causa para el grupo de personas entre 18 y 30 años puede estar en el uso de redes sociales,

pues se conoce que este grupo usa las redes sociales como principal medio de información, lo cual puede desencadenar el estrés más fácilmente (Qiu, y otros, 2020)

1.4. Fuente de datos y herramientas de desarrollo

La fuente de datos para el desarrollo de este proyecto fue Twitter y las herramientas usadas provienen de librerías de Python.

1.4.1. Twitter

La red social Twitter funciona en tiempo real permitiendo a sus usuarios crear y compartir ideas e información. Cuando se crea una cuenta en Twitter el usuario tiene la opción de agregar datos como: estados, su nombre, ubicación, y una breve descripción biografía (Twitter, 2021).

1.4.2. Python

Python es un lenguaje de programación que posee estructuras de datos de alto nivel con un enfoque orientado hacia la programación orientada a objetos. La principal característica de Python es su sintaxis sencilla de interpretar y por sus extensas librerías que se encuentran disponible gratuitamente (Python, 2021).

Para este trabajo se ha optado por Python como lenguaje de programación justamente por su gran cantidad de librerías y su inclinación hacia el análisis de datos. En este proyecto se han usado múltiples librerías, pero las de mayor importancia son:

- **Selenium:** Es un software que permite la automatización de navegadores web, permitiendo emular la interacción de un usuario frente a un navegador. (Fortner, 2021)

- **Pandas:** Es un proyecto de código libre de alto nivel usado para realizar análisis de datos de lenguaje natural en Python (pandas, 2018). Pandas proporciona estructuras de datos muy similares a los que están presentes en R.

- **Stopwords:** Es una librería de Python la cual trabaja con listas de palabras en más de 20 idiomas, dichas listas contienen palabras consideradas irrelevantes para un proceso, por lo general estas palabras no tienen un significado por sí mismas; comúnmente suelen ser artículos, pronombres y adverbios (Contributor, 2005).

- **NRCLex:** Es una librería que permite medir el efecto emocional de un texto. Usa el diccionario de *National Research Council Canada* (NRC) de emociones y sentimientos,

y los conjuntos de sinónimos de WordNet de la librería *Natural Language Toolkit* (NLTK) (Bailey, 2020)

- **Stanza:** Es un paquete de análisis de lenguaje natural, está construido en base a una red neuronal que permite convertir una cadena de texto en una lista de oraciones y palabras para obtener características morfológicas y reconocer temas mencionados. (Qi, Zhang, Zhang, & Bolton, 2021), usa técnicas como etiquetado gramatical y análisis de dependencias

- **Streamlit:** Es una librería de Python de código abierto que permite la creación de aplicaciones web encaminadas para el aprendizaje automático y la ciencia de datos. Streamlit permite convertir scripts de datos en aplicaciones web de tal forma que se pueda compartir y visualizar datos en tiempo relativos (Streamlit, 2020).

1.4.3. Técnica de obtención de información de un sitio web (web scraping)

La técnica de obtención de información de un sitio web es conocida como web scraping o raspado web. Existen diferentes tipos de raspado o scraper dependiendo de su funcionalidad, están el raspado web, raspado de pantalla, para este proyecto se utilizará el raspado web o web scraping. Zhao (Zhao, Web Scraping, 2017) lo define como una técnica para extraer datos de la World Wide Web para guardarlos en un archivo o una base de datos. El web scraping se centra en la transformación de datos no estructurados en datos que puedan ser almacenados en un fichero o una base de datos (Vargiu, 2013).

1.4.4. Diccionario National Research Council Canada (NRC)

Es un diccionario en inglés que consta de una lista de palabras asociadas con ocho emociones básicas: ira, miedo, expectación, confianza, sorpresa, tristeza, alegría y repulsión; además de dos sentimientos: positivo y negativo, la asignación de estas emociones se las realizó de manera manual mediante la asignación de tareas de clasificación a un grupo de personas, esta técnica es conocida en inglés como crowdsourcing (Mohammad & Turney, Crowdsourcing a Word-Emotion Association Lexicon, 2013). De esta manera a una lista de palabras se les asignó diferentes emociones y sentimientos

1.4.5. Diccionario de intensidad de emociones (Emolex)

Al igual que NRC, el diccionario Emolex, es un diccionario en inglés que intenta transmitir a una lista de palabras una o más emociones (ira, miedo, expectación, confianza, sorpresa, tristeza, alegría y repulsión) y en función de esto determinar la intensidad de cada emoción

presente en una palabra (Mohammad & Turney, Emotions Evoked by Common Words and Phrases Using Mechanical Turk to Create an Emotion Lexicon, 2010).

1.5. Análisis de emociones y sentimientos

La expansión continua de internet ha derivado en un crecimiento exponencial en información subjetiva, esta información ha despertado interés para el análisis de sentimientos, una tarea encargada del procesamiento de del lenguaje natural o por sus siglas en inglés (NPL), la cual busca identificar relaciones sobre un tema específico (Liu, 2011). Los datos continuamente son explotados por administraciones públicas, empresas y particulares.

1.5.1. Análisis de sentimiento en Twitter

En esencia el análisis de sentimientos en Twitter hace referencia a asignar a un mensaje publicado un valor relacionado con la carga emocional que intenta comunicar (Baviera, 2016), dicha carga emocional se puede dividir en varios tipos de variables.

- **Polaridad:** Es usado para indicar si el mensaje posee un sentimiento positivo, negativo o neutro.
- **Intensidad:** Da un valor numérico relacionado con la intensidad del sentimiento, al igual que la polaridad se puede distinguir entre intensidad positiva y una intensidad negativa
- **Emoción:** Al texto que es analizado se lo clasifica según diferentes tipos de emociones tales como tristeza, alegría, ira, miedo, etc.

1.5.2. Análisis de emociones

Las emociones tienen un papel importante en el comportamiento humano, por lo cual la integración de emociones con modelos computacionales puede mejorar los sistemas de interacción humano-computadora (Plaza & Ureña, 2019). Las emociones humanas pueden ser expresadas de manera verbal o escrita; en los últimos años las plataformas de mensajes, redes sociales y blogs son usados como medios para comunicarse con otros individuos, lo cual despierta el interés en el análisis de las redes sociales como medio para conocer las emociones de una población.

Los sistemas actuales que analizan el lenguaje natural aún se encuentran en desarrollo, sin embargo, la mayoría se basan en el diccionario de emociones Emolex. El analizar emociones en un texto puede ser un desafío pues, la interpretación puede depender del contexto de toda una oración. (Plaza & Ureña, 2019). Por este motivo la minería basada en texto es indispensable para conocer las emociones de un grupo de datos.

1.5.3. Etiquetado gramatical (part of speech *PoS*)

Es el proceso de etiquetar a cada palabra de un texto su correspondencia gramatical (Cutting, Kupiec, & Pedersen, 1992). Para esto es necesario entender el contexto de la oración pues una misma palabra puede tener varios significados. Existen dos maneras de realizar este etiquetado mediante aproximaciones lingüísticas y aproximaciones de aprendizaje (Klein & Simmons, 1963). En este proyecto se usó el método de aproximaciones de aprendizaje previamente ya desarrolladas que se encuentran presentes en la librería *Stanza* en el lenguaje de programación de Python.

1.5.4. Análisis de dependencias (dependency parcing)

Una oración puede ser entendida simplemente con sus palabras principales sin necesidad de artículos o pronombres; el análisis de dependencias o dependency parcing en inglés consiste en extraer las palabras principales de una oración y analizar sus dependencias con otras palabras. (Jurafsky & Martin, 2021). Estas dependencias se describen en relaciones binarias entre las palabras de una oración, es decir se asocian dos palabras principales las cuales pueden ser interpretadas como el tema central de la oración. En este proyecto la división de cada oración básicamente tratará de encontrar el núcleo de la oración, el objeto y los modificadores.

- **Núcleo:** Es la palabra de la cual depende directa o indirectamente todas las demás palabras.
- **Objeto:** Son las entidades que se encuentran participando en las acciones o las que complementan la semántica de otra palabra, por lo general suelen ser sustantivos.
- **Modificador:** Es la palabra que le da sentido a la oración

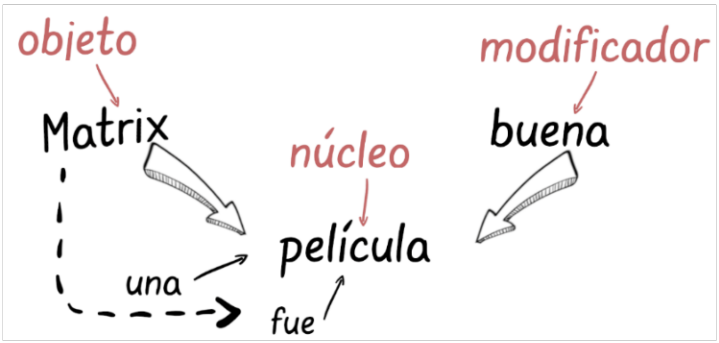


Figura 1. Relaciones de una oración

2. METODOLOGÍA DE DESARROLLO DE UN PROYECTO DE MINERÍA DE DATOS

La metodología abordada para esta solución fue Cross Industry Standard Process for Data Mining, o por sus siglas CRISP-DM la cual viene a ser la guía más usada en proyectos de minería de datos (Gallardo, 2010). CRISP-DM se encuentra dividido en 6 fases: compresión del problema, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. Estas fases se encuentran expresadas en la Figura 2 (Wirth & Hipp, 2000). Además de este modelo de ha añadido una fase de desarrollo de la aplicación web para la visualización de los resultados.

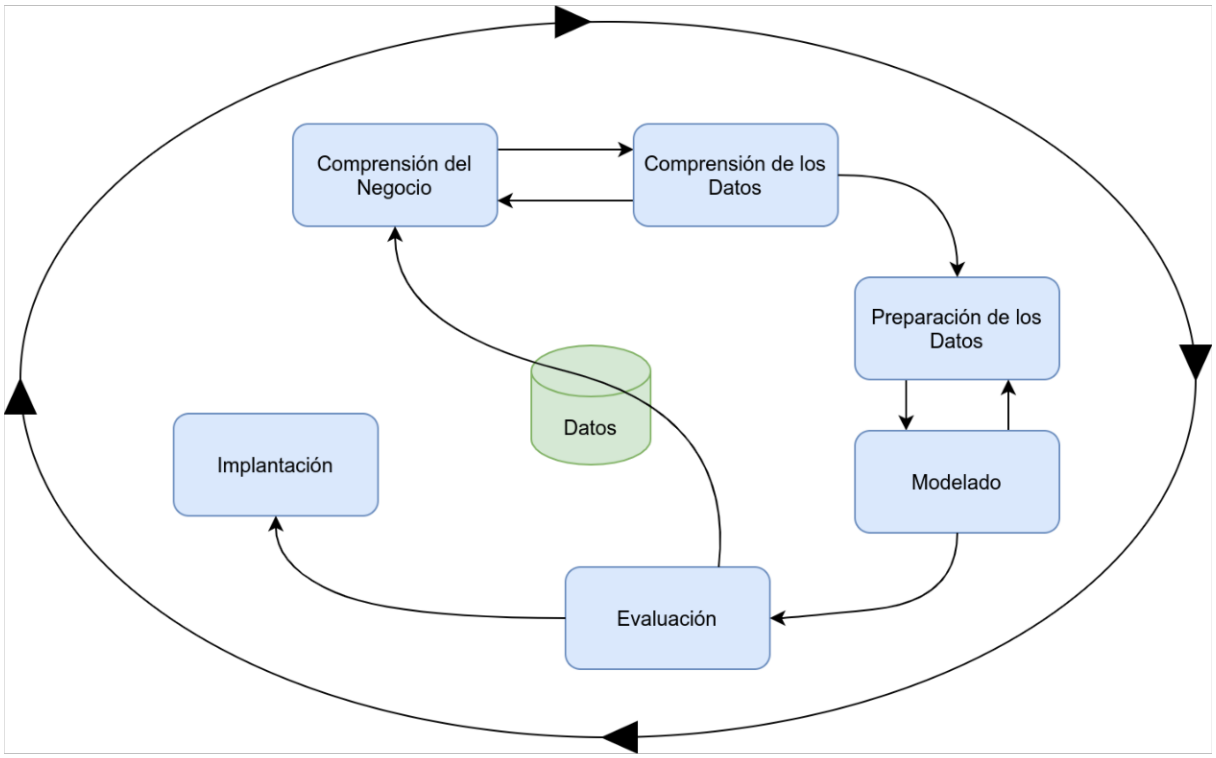


Figura 2. Modelo de procesos CRISPDM

2.1. Fase de Comprensión del problema

La primera fase de la metodología CRISP-DM es considerada la más importante pues, en esta fase se detallan los objetivos comerciales del proyecto para convertirlos en un problema de minería de datos (Gallardo, 2010). De esta forma es posible diseñar un plan estratégico que permita alcanzar los objetivos del negocio. Esta fase consta de 4 tareas representadas en la Figura 3.

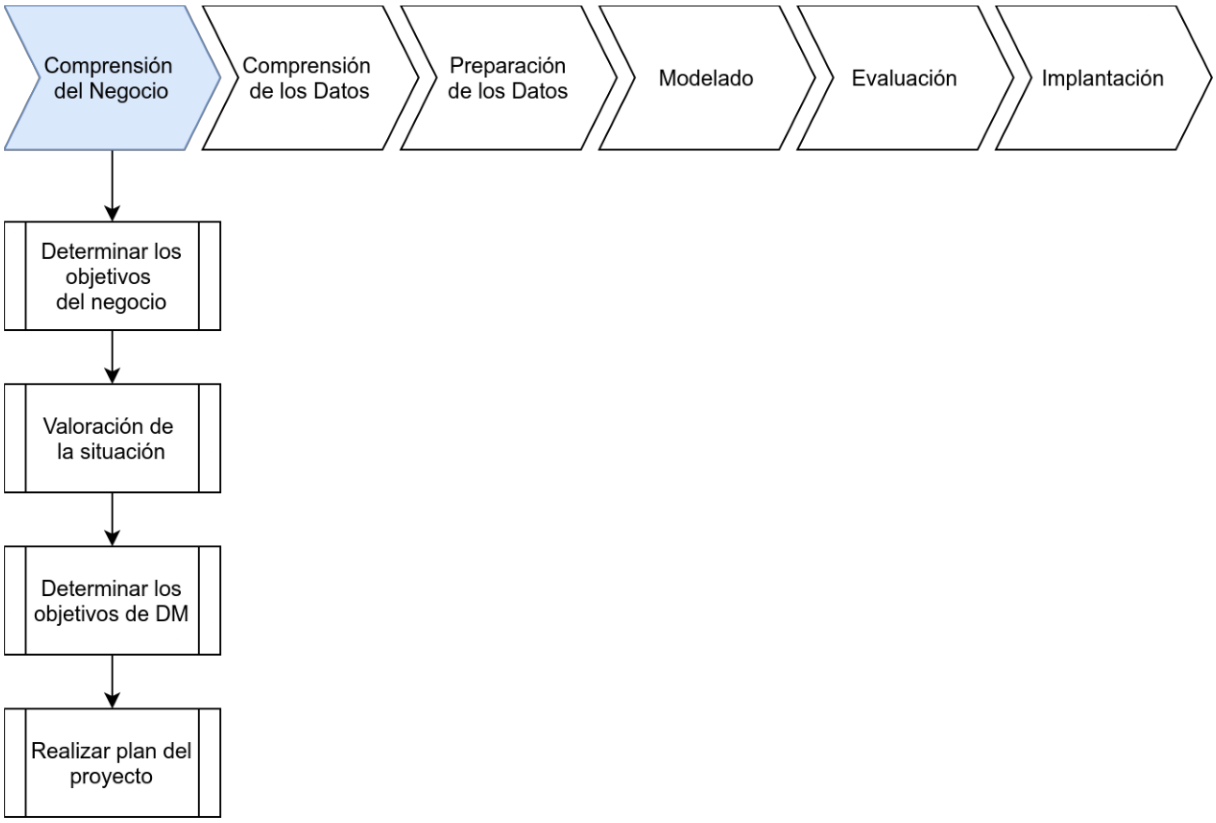


Figura 3. Fase de comprensión del negocio

2.1.1. Determinar los objetivos del problema

El objetivo principal de este estudio es conocer los efectos psicológicos que pueden llegar a tener determinados grupos de personas al momento en que se presenta una pandemia. Este proyecto pretende determinar los problemas psicológicos más comunes y las posibles emociones que los causan. La información con la que se pretende realizar el estudio son los comentarios presentes en la red social de Twitter.

2.1.2. Valoración de la situación

Es común ver en televisión o redes sociales el caos producido en una sociedad al presentarse un cambio brusco en su estilo de vida. Estos cambios pueden ser una catástrofe natural o la aparición de una pandemia, según el estudio de Chun & Geller (Xiang & James, 2013), realizado en 2013 durante un brote de SARS, el 23% de personas encuestadas admitió haberse comportado de manera irracional. Dentro de estas acciones se encuentran el realizar compras en exceso, buscar refugio y abastecerse con provisiones (Xiang & James, 2013). Así como este estudio, existen diferentes investigaciones referentes al comportamiento de una sociedad en la presencia de una pandemia.

En lo que refiere a la cantidad de datos que se puede llegar a requerir, Twitter es una de las mejores alternativas, pues ofrece una gran cantidad de información, la cual queda plasmada por sus usuarios al escribir un tweet. Según el portal Kinsta (Osman, 2021) en sus estadísticas de enero del 2021, al día se escriben aproximadamente 656 millones de tweets. Pese a la gran cantidad de tweets, Twitter solo permite una descarga limitada de información en su versión gratuita de desarrollo por medio de su API. Una alternativa para la recolección de tweets es la "hidratación de tweets", este método consiste en la descarga de tweets basándose en la obtención de un archivo con los "Id's" de tweets de una temática de interés. Estos archivos pueden encontrarse en sitios como Zenodo o Dataverse de la universidad de Harvard (León, 2020), dichos sitios brindan acceso abierto a datos, para su posterior estudio.

Otra posible alternativa es el raspado web o web scraping, una técnica que consiste en la extracción de datos presentes en un sitio web (Murillo & Saavedra, 2017). Al realizar web scraping es necesario considerar aspectos como la accesibilidad a Twitter, pues en muchos sitios web puede ser considerado como una práctica ilegal. En Twitter no existe una ley que marque esta práctica como ilegal "el web scraping es legal si los datos disponibles son públicos" (Grimes, 2021). Los datos presentes en Twitter pueden ser entendidos como de dominio público, lo cual puede desembocar en la obtención de datos de manera anónima y derivarse a posibles cuestionamientos éticos.

2.1.3. Determinar los objetivos de la minería de datos en relación con el problema en estudio

El objetivo general de este proyecto es encontrar patologías psicológicas que pueden aparecer como consecuencia de una pandemia, por lo cual la meta de la minería de datos es: obtener patrones (frases o palabras recurrentes) que permitan determinar qué tipo de trastorno psicológico pueden padecer cierto grupo de personas al encontrarse en situaciones de estrés, miedo, soledad, etc. En este aspecto también es necesario encontrar emociones primarias que marquen el estado de ánimo de una sociedad y puedan afectar su salud mental. Dentro de esto es posible obtener resultados distintos para grupos diferentes de individuos, pues es necesario tomar en consideración múltiples variables, como pueden ser el área geográfica, situación laboral, relación con la enfermedad en su entorno, situación personal, edad, etc. (Nekane Balluerka, 2020). Por último y para contrastar los resultados, se necesita contar con el criterio de un profesional en el área de ciencias psicológicas.

2.1.4. Realizar el plan del proyecto

El proyecto se lo realizará en varias etapas para facilitar su desarrollo y optimizar su tiempo de elaboración. El proyecto consta de las siguientes etapas:

- Etapa 1: Estudio de la mejor técnica de obtención de datos de Twitter.
- Etapa 2: Selección de fechas que encierren acontecimientos de interés general.
- Etapa 3: Obtención y preparación de los datos
- Etapa 4: Selección de técnicas de análisis y modelado de lenguaje natural
- Etapa 5: Análisis de los resultados conseguidos en las etapas anteriores
- Etapa 6: Revisión de los resultados con un especialista en problemas psicológicos
- Etapa 7: Presentación de los resultados finales.
- Etapa 8: Creación de aplicativo web

2.2. Fase de comprensión de los datos

El objetivo de esta fase es la recolección inicial de datos, con la finalidad de obtener una perspectiva general de la calidad, los problemas y las relaciones que pueden presentar los datos y así establecer las primeras hipótesis. (Gallardo, 2010).

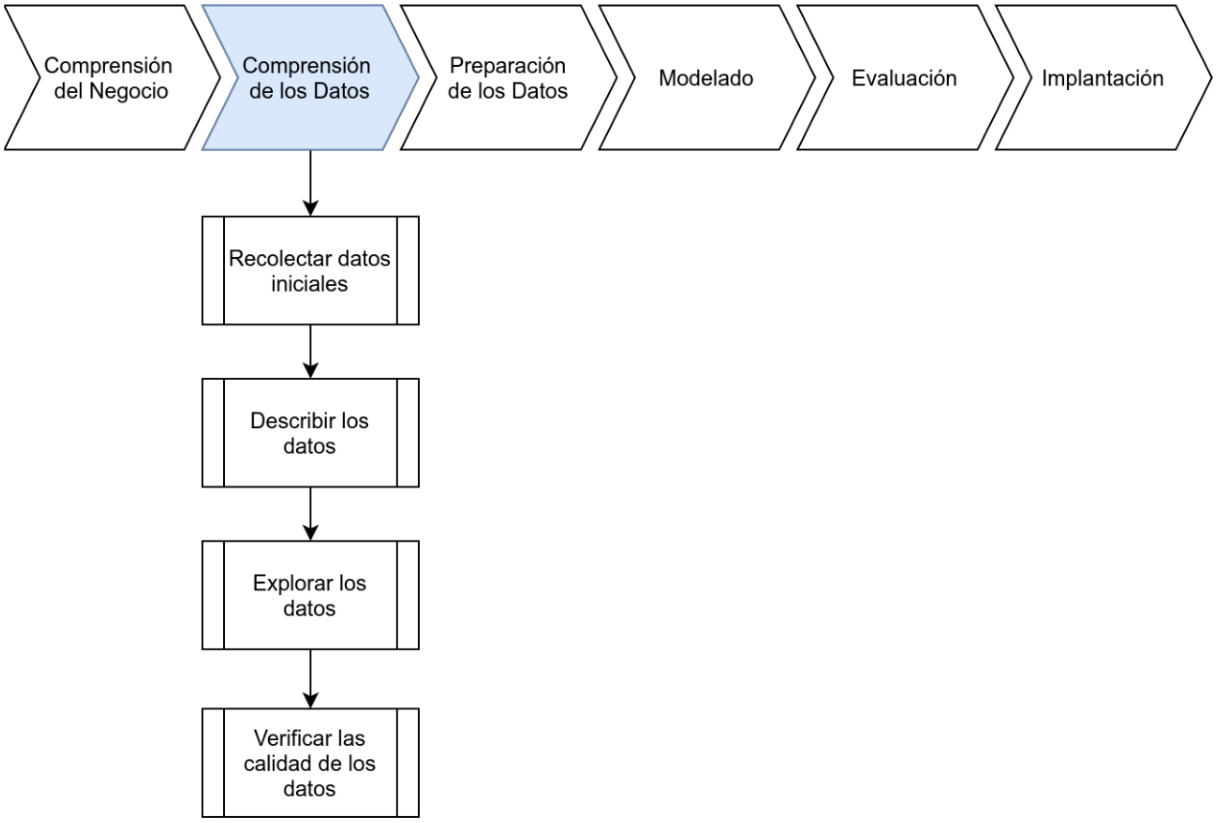


Figura 4. Fase de comprensión de los datos.

2.2.1. Recolectar datos iniciales

La técnica usada para la recolección de datos fue raspado web o web scraping, esta se realizó en la página de búsqueda de Twitter. Esta técnica consiste en extraer datos presentes en un sitio web, este procedimiento puede realizarse manualmente por un usuario o

automáticamente por un bot. (Zhao, Web Scraping, 2017). La extracción de datos se la realizó en tres diferentes momentos considerados de impacto social, y así evaluar las emociones presentadas en diferentes regiones. La obtención de los datos se ha limitado a Latinoamérica, España y Estados Unidos; es decir países donde existen un gran número de hispanohablantes. La técnica de web scraping aplicada en Twitter permite obtener tweets sin las limitantes de la API de Twitter y sin la necesidad de pagar una suscripción como desarrollador, como usualmente se usa para la minería de tweets. La desventaja de este es el tiempo de ejecución, este es mucho mayor que si se usará la API oficial de Twitter.

La herramienta utilizada para la automatización en el proceso de web scraping fue “Selenium” disponible para varios lenguajes de programación incluyendo Python. La búsqueda y filtración de tweets se lo realizó con la ayuda del formulario de “búsqueda avanzada” de Twitter. Este formulario ayuda en el filtrado de tweets mediante palabras clave, idioma y un rango de fecha. Las palabras clave usadas para la búsqueda de tweets fueron ‘covid’ y ‘ansiedad’, el idioma seleccionado fue español y las fechas utilizadas fueron momentos de impacto social como en abril del 2020 donde titulares como: “Cientos de cadáveres sobre el asfalto de las calles de Guayaquil…” (Lozano, 2020) o “Parte de Ecuador vuelve al confinamiento por la covid, ahora en fin de semana” (Brik, 2021) fueron noticia Internacional; además se buscó tweets referentes a fechas de gran interacción social como la navidad y año viejo de 2020.

Pese a que el idioma seleccionado fue español, a primera instancia no es posible conocer el país del autor de un tweet, por lo que fue necesario realizar otro raspado de información. Este raspado se lo realizó en el perfil de cada usuario para obtener su país de origen. La manera en que se llenó el formulario está representada en la Figura 5 y en la Figura 6; este formulario a pesar de ser de ayuda no cumple con todos los requerimientos necesarios para la obtención de los datos requeridos, por lo que fue necesario aplicar un segundo método de búsqueda.

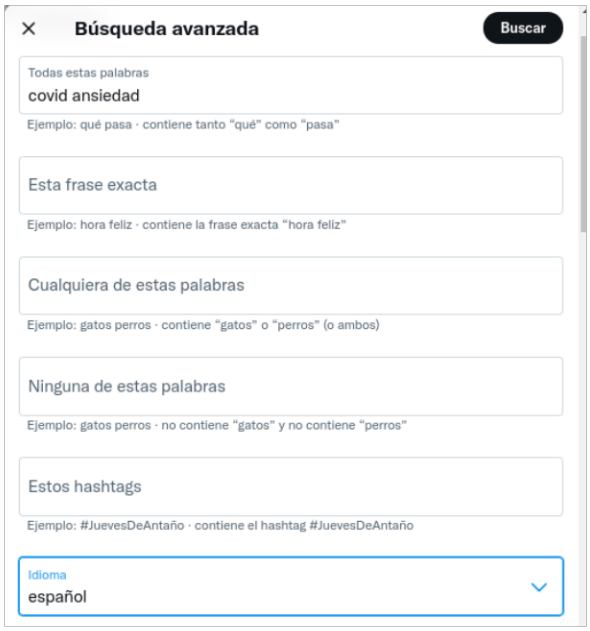


Figura 5. Formulario de búsqueda en Twitter por palabras clave

Figura 6. Formulario de búsqueda en Twitter por fecha

El propósito del segundo método de búsqueda es generar una página donde se visualicen todos los tweets que cumplen con la información ingresada en el formulario. En esta página se crea una instrucción que contiene los campos necesarios tal como lo muestra la Figura 7, dicha instrucción se usa en conjunto con la librería Selenium para automatizar y extraer los tweets. De esta manera la información recolectada de cada tweet se fue adjuntando en una estructura de datos o dataframe para su posterior escritura en un archivo CSV. Este proceso se repitió para cada noticia dando un total de tres búsquedas generales y tres archivos CSV, mismos que fueron unidos creando un solo documento. Los registros obtenidos pueden contener campos con errores o nulos, esto dependerá de la información que contenga Twitter referente a sus usuarios y tweets.

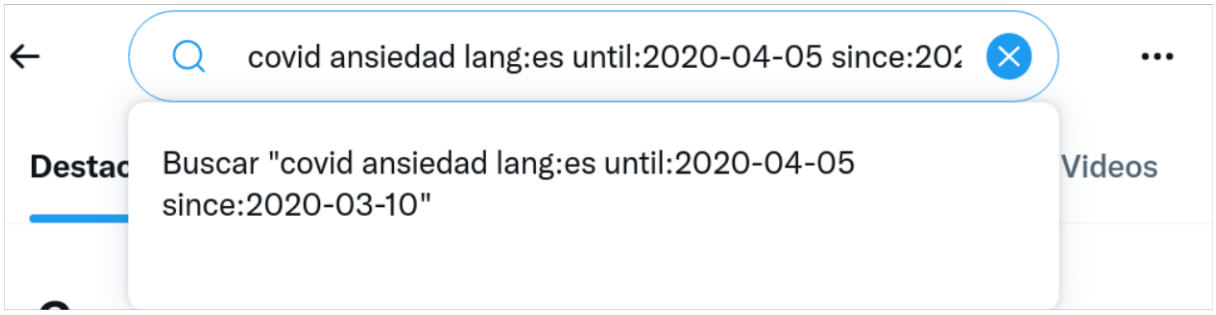


Figura 7. Frases con las instrucciones de búsqueda

2.2.2. Describir los datos

Los datos obtenidos en la fase anterior se han unido en un solo archivo, dando un total de 5456 registros y ochos campos. La cantidad de registros de cada noticia depende del total de

tweets referentes a dicha noticia, por este motivo se tiene diferente cantidad de registros para cada noticia. Como consecuencia la noticia referente a la crisis sanitara en Guayaquil entre las fechas 2020-04-05 hasta 2020-05-10, se obtuvieron 3088 registros, para los tweets referentes a navidad y año nuevo que va entre las fechas 2020-12-03 hasta 2021-01-03 se obtuvieron un total de 1927 registros, por último, los tweets referentes al nuevo confinamiento de únicamente fines de semana realizado en Ecuador estando entre las fechas 2021-04-23 y 2021-05-16 obteniéndose un total de 441 registros. Los campos de cada registro obtenido se les nombró como: user, handle, fecha, tweet, emoji, comentarios, likes y país. Los datos resultantes se los guardó en un archivo de texto plano en formato CSV, a lo largo de todo el proyecto se mantuvo ese formato para manejar los datos.

Los registros obtenidos pueden estar repetidos, tener campos nulos o vacíos, esto es en base a la información que se tenga de cada tweet. El tipo de dato de cada campo depende de su utilidad, en un principio todos los campos son de tipo “String” incluyendo número de likes, comentarios y fecha esto para su manipulación y posterior limpieza. La descripción de cada campo es la siguiente:

- **user:** representa el nombre de visualización que un usuario en Twitter posee.
- **handle**: en este campo se almacenan los nombres de usuario de Twitter, este siempre está precedido del símbolo “@”.
- **fecha**: representa la fecha y hora del tweet
- **tweet:** en este campo se almacena todo el texto del tweet, excluyendo emojis
- **emoji:** este campo acoge a todos los emojis presentes en un tweet, de no existir alguno, presentará un campo vacío.
- **comentarios**: este campo almacena la cantidad de comentarios que posee un tweet.
- **likes:** este campo almacena la cantidad de likes que posee un tweet.
- **locacion**: este campo indica la ubicación o el lugar de origen del usuario, este puede ser una ciudad o un país.

2.2.3. Explorar los datos

El análisis estadístico previo a la obtención de los registros no presenta una visión clara de los datos, a pesar de ofrecer una vista general de los campos numéricos. Esto en parte porque la mayoría de los campos son cadenas de texto. No obstante, hay que recordar que los datos obtenidos están pensados para realizar un procesamiento de lenguaje natural (PNL), aun así y gracias a las funciones estadísticas que ofrece la librería Pandas, se puede analizar los campos de comentarios y likes, tal como se aprecia en la Tabla 1.

	comentarios	likes
Cantidad	1524	1495
Media	2.75	10.77
Desviación estándar	19.19	54.78
Mínimo	1	1
Máximo	721	988

Tabla 1. Estadísticas de comentarios y likes

Los atributos que contiene texto no pueden ser analizados de la misma manera que los atributos numéricos; pero en el caso de 'user', 'emoji' y 'locacion', es posible analizar la cantidad de celdas en blanco (NaN) que poseen. Los atributos como: 'handle', 'fecha' y 'tweet', no tienen celdas en blanco, pues estos son los campos necesarios que implica la existencia de un tweet. La Tabla 2 muestra las celdas vacías para cada atributo.

	Celdas vacías	Porcentaje
user	57	1%
handle	0	0 %
fecha	0	0 %
tweet	0	0 %
emoji	3731	68.38 %
comentarios	3932	72.06 %
likes	3961	72,60 %
locacion	1539	28,20 %

Tabla 2. Celdas vacías en los campos de estudio

La Tabla 2 indica que los campos como 'comentarios', 'likes' y 'emoji' presentan un alto rango de celdas vacías, lo que implica que si se llegasen a usar reducirían notablemente el número de registros, por lo cual estos atributos vienen a ser candidatos a ser descartados.

2.2.4. Verificar la calidad de los datos

Los atributos como 'tweet' o 'locacion' pueden indicar cierta cantidad de repeticiones, provenientes no necesariamente del mismo usuario. En cuanto a 'tweet', al obtener la función de densidad de probabilidad representada en la Figura 8, indica que no hay una gran cantidad de tweets repetidos. La figura marca una máxima cantidad de repeticiones cercano a 50 y una media de tweets repetidos que viene a ser poco más de 20 repeticiones con una probabilidad menor al 1%.

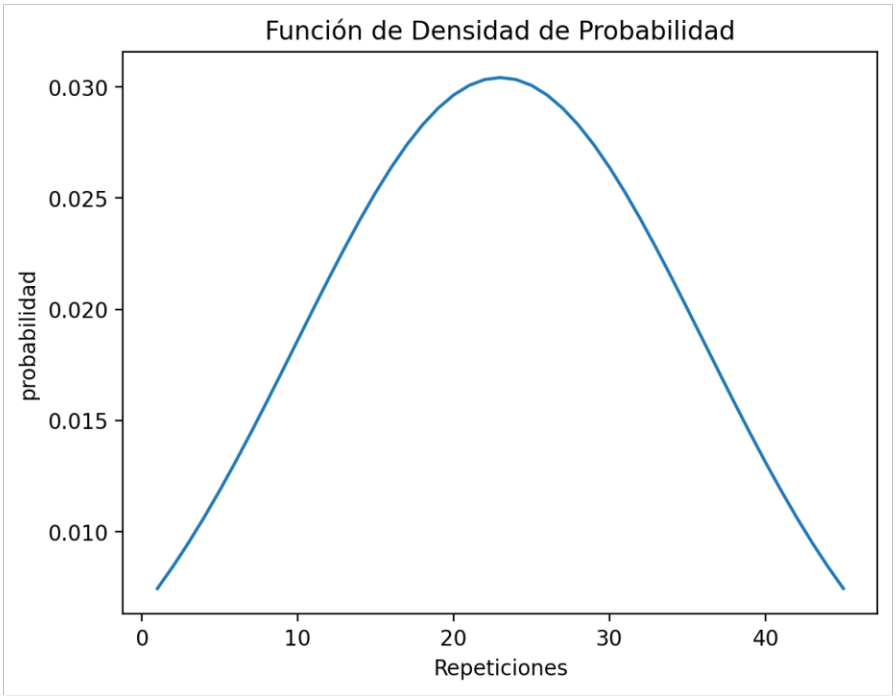


Figura 8. Función de densidad de probabilidad de tweets repetidos

El segundo campo, 'locacion' si debiese tener una cantidad representativa de repeticiones, pues este campo indica el país del usuario; sin embargo, el campo 'locacion' presenta datos dispersos, esto es en medida por la manera en que los usuarios han escrito su ubicación, es decir no existe un ingreso de datos estandarizado. En este sentido diferentes usuarios pueden escribir de formas diferentes su ubicación haciendo referencia al mismo país. El ejemplo más claro de esto se puede ver representado en la Tabla 3, donde, por ejemplo, para referirse a México, se pueden encontrar registros escritos como "México, D.F." o simplemente "México". De esta misma manera se aplica para cada país, por lo que viene a ser necesario una adecuación y estandarización del atributo de 'locación'.

locacion	**Frecuencia**
México	189
Caracas, Venezuela	118
Venezuela	77
Mexico, D.F.	68
Chile	53
CDMX	51
Mexico	44
España	41
Colombia	40
Bogotá, D.C., Colombia	38
Ciudad de México	37

Tabla 3. Diferentes formas de referirse a un mismo país

2.3. Fase de preparación de los datos

La preparación de datos consiste en la adecuación de estos para que sea posible aplicar técnicas de minería de datos, esto incluye técnicas como selección, limpieza, visualización, etc. (Gallardo, 2010). Toda la preparación de los datos se lo realizó utilizando el formato CSV esto con la finalidad de usar la librería Pandas, especializada en el manejo de estructuras de datos o dataframes, además de ser de código abierto.

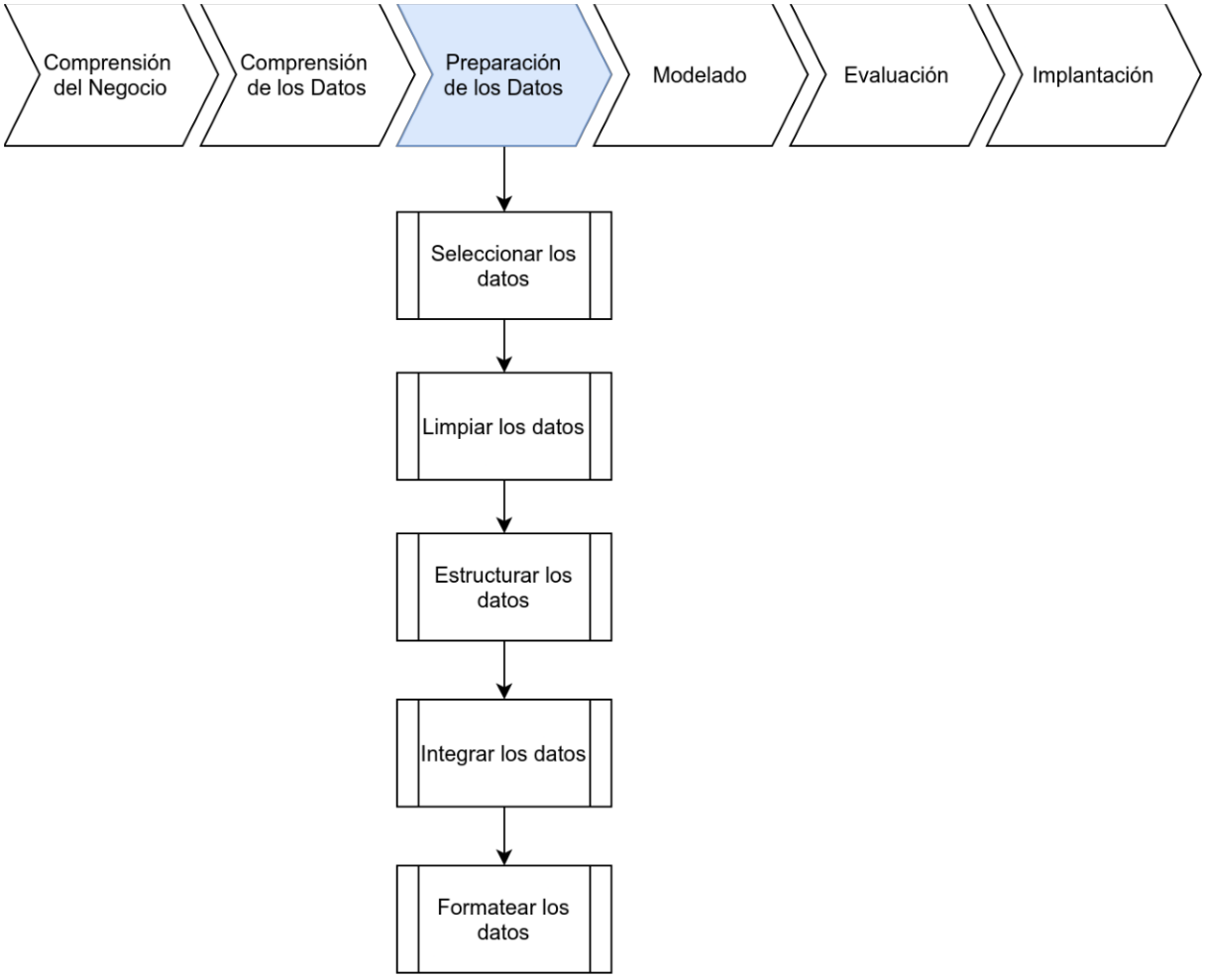


Figura 9. Fase de preparación de los datos

2.3.1. Seleccionar los datos

La selección de los campos se realizó en base al porcentaje de celdas vacías que posee cada atributo. En este sentido se han descartado los atributos de 'emoji', comentarios' y 'likes', pues poseen alrededor de 70% de celdas vacías. Frente a esto se optó por seleccionar los atributos: 'user', 'handle', 'fecha', 'tweet' y 'locacion'. A pesar de que el atributo 'locacion' posee un 28% de celdas vacías, este no es posible descartarlo, en parte porque este porcentaje no es tan elevado como los otros atributos y, además este campo es necesario para el análisis de la información. Como consecuencia el análisis de los datos va a estar enfocado en el tweet, la fecha y el país del usuario, y en menor medida el nombre y usuario del tweet.

2.3.2. Limpiar los datos

La limpieza de la información se la realizó en primera instancia, en los registros repetidos en el campo ‘tweet’, pasando a las celdas en blanco del campo ‘locacion’. La eliminación de registros repetidos se usó únicamente en el campo ‘tweet’ dado que campos como ‘locacion’, ‘fecha’ o ‘user’ es entendible que tengan campos repetidos. La eliminación de registros con celdas en blanco se aplicó en el atributo ‘locación’, los demás campos seleccionados no poseen celdas en blanco, a excepción del campo ‘user’, pero ese no es relevante para el análisis consecuente. Previo a esta limpieza es necesario realizar una adecuación de datos con el objetivo de estandarizar este campo. La limpieza de datos se retomó en los nuevos campos creados, con la finalidad de optimizar el tiempo de procesamiento.

Los registros repetidos en el campo ‘tweet’, fueron eliminados con ayuda de la función *drop_duplicates* presente en la librería Pandas. La eliminación de dichos registros se aplicó a todos estos exceptuando el tweet original. De este modo en la primera limpieza se tuvo una reducción de registros pasó de 5456 a 5096. En cuanto a este procedimiento de eliminación de registros repetidos se aplicó una sola vez y solo en el campo ‘tweet’. Para el análisis de emociones en este proyecto, la repetición de un tweet indica un análisis de lenguaje natural repetido, incrementando el tiempo de procesamiento y obteniendo los mismos resultados, es decir creando ruido.

El campo ‘locacion’ pasó por dos fases para su adecuación, en una primera instancia se procedió a estandarizar sus datos y consecuentemente pasó a su limpieza. La estandarización de este campo consistió en buscar palabras que representen un país o ciudades pertenecientes a algún país de habla hispana. De esta forma se consiguió estandarizar este campo por países hispanohablantes; todos los demás datos que no cumplen esta condición se procedieron a cambiar su valor a vacío. Finalmente, al haber cambiado todos los datos innecesarios del campo ‘locacion’ a vacío, se procedió a su eliminación. La eliminación de registros vacíos se realizó con la función *dropna* presente en la librería Pandas. En conclusión, para este campo fue necesario realizar primero una transformación sintáctica de los datos y posterior se pasó a la limpieza, esto para no repetir la misma tarea de eliminación de registros vacíos y optimizar el tiempo de procesamiento.

2.3.3. Estructurar los datos

Los campos nuevos que se han generado vienen a ser de dos tipos, el primero es una simple derivación de la información, mientras que el segundo tipo son aquellos campos creados en base a un análisis. El ejemplo más claro de un campo creado con una simple derivación es el campo ‘noticia’, el cual se deriva del campo fecha. Los campos creados en base a un análisis se detallan mejor en la integración de datos, teniendo como un ejemplo de estos la polaridad

del tweet. Es esencial entender que a medida que se repitió los ciclos fueron necesarios crear nuevos campos tanto como simples derivaciones, así también como resultado de un análisis. De esta manera en esta sección se explicará únicamente la creación de datos como resultado de una simple derivación.

Los campos creados como una derivación implican una simple adecuación de un campo para que exprese lo mismo, pero escrito de manera más simple para su procesamiento. El campo 'noticia fue el único que se creó de esta manera. Como consecuencia todos los demás campos creados vienen del producto de un análisis de otros campos. El campo noticia es un campo estandarizado que indica el acontecimiento en el cual se escribió el tweet.

El campo 'noticia' se deriva directamente del campo 'fecha'. Una vez que el campo fecha fue formateado a un contexto más simple de tratar, se procedió a la creación del campo 'noticia'. Los registros al ser obtenidos en tres acontecimientos diferentes poseen tres rangos de fechas, uno para cada acontecimiento. Para encontrar el acontecimiento sucedido al momento de la descarga de los tweets se creó un script llamado *ClasificadorNoticia_2* cuya funcionalidad es la de identificar la fecha de un tweet y ubicarlo dentro de unos de los tres rangos disponibles. El primer rango referente a "Crisis Guayaquil" va desde 05/04/2020 a 10/05/2020, el segundo intervalo referente a "Navidad" va desde 01/11/2020 hasta 05/01/2021 y para el último intervalo referente a "Nuevo confinamiento" va desde 15/04/2021 hasta 20/05/2021. En resumen, el campo 'noticia' solo consta de tres posibles valores, "Crisis Guayaquil", "Navidad" y "Nuevo Confinamiento". En conclusión, el campo 'noticia' se crea a partir de la selección de la fecha de un tweet y su posterior clasificación dentro de los tres posibles acontecimientos.

2.3.4. Integrar los datos

La integración de los datos consiste en la creación de nuevos registros previo a un análisis de otros campos. En este sentido se han creado quince nuevos campos, once campos correspondientes a un análisis directo y 4 correspondientes a una derivación de estos campos. Los primeros hacen referencia a un análisis de emociones y sentimientos, mientras que los segundos expresan polaridad y emoción predominante del tweet. Estos nuevos campos junto con el campo 'noticia' permiten evaluar el estado emocional de una población en uno de los tres acontecimientos estudiados. La Tabla 4 muestra cómo se obtuvieron los nuevos campos y además presenta una breve descripción de estos. Los campos obtenidos de forma directa están representados de color celeste, mientras que los campos obtenidos como derivación están representados de color marrón.

La obtención de los primeros once campos se realizó gracias a la ayuda de la librería *NRCLex*, los siguientes cuatro campos son derivaciones y acciones secundarias de esta librería. La

librería *NRCLex* usa un diccionario en inglés obtenido del sitio http://saifmohammad.com para evaluar las emociones y sentimientos de un texto. En este proyecto se procedió a modificar esta librería y usar un diccionario en español, dicho diccionario traducido se obtuvo del sitio http://saifmohammad.com. De todas las funciones que ofrece *NRCLex* se han usado: *raw_emotion_scores* para determinar la cantidad de palabras que hacen referencia a una emoción y conocer la intensidad de una emoción, esta intensidad va de 0 a 1. La otra función utilizada fue *affect_dict* para obtener la lista de palabras que fueron analizadas en un tweet. A pesar de que se modificó la librería *NRCLex* no se la eliminó, simplemente se creó una nueva clase llamada *NRCLex_es* la cual tiene las modificaciones hechas, esta clase funciona como la clase principal de esta librería.

Campo	Tipo	Obtención de campo	Descripción
miedo	Entero	Directa	Cantidad de palabras que indican una emoción o sentimiento
ira	Entero	Directa	
expectación	Entero	Directa	
confianza	Entero	Directa	
sorpresa	Entero	Directa	
tristeza	Entero	Directa	
repulsion	Entero	Directa	
alegria	Entero	Directa	
positivo	Entero	Directa	
negativo	Entero	Directa	
emociones	String	Directa	Lista de emociones
polaridad	String	Derivada	Polaridad del tweet
palabra	String	Derivada	Palabra de mayor intensidad
emoción	String	Derivada	Emoción representativa del tweet
puntuación	Float	Derivada	Valor numérico de la intensidad de palabra

Tabla 4. Descripción de los campos a analizar

Los campos obtenidos por la función *raw_emotion_scores* se pueden visualizar en la Tabla 4 van desde 'miedo' hasta 'negativo; de igual forma la lista de emociones se obtuvo gracias a la función *affect_dict*. El campo polaridad se deriva de los campos positivo y negativo, indicando de acuerdo con la puntuación la polaridad del tweet. Los campos 'palabra', 'emocion' y 'puntuacion' están relacionados, pues los tres se derivan del análisis de intensidad de emoción de un tweet. El campo 'puntuacion' usa un diccionario diseñado para evaluar la intensidad de una palabra en diferentes emociones, dicho diccionario también se obtuvo de

la investigación Saif Mohammad (Mohammad & Turney, Emotions Evoked by Common Words and Phrases Using Mechanical Turk to Create an Emotion Lexicon, 2010) presente en su sitio web *http://saifmohammad.com*. Dado que este diccionario en formato CSV se encuentra traducido en más de 40 idiomas, se lo simplificó y adecuó para que reconozca únicamente palabras en español a este nuevo archivo se lo guardó con el nombre de *intensidadEmociones*.csv. La intensidad de emociones se obtuvo del análisis de cada palabra presente en el tweet.

El análisis de esta intensidad se lo realiza en el script *IntensidadPalabra*_3, donde a cada tweet se lo analiza en base al diccionario *intensidadEmociones*.csv. En conclusión, se usó un diccionario de intensidad para evaluar la emoción predominante en un tweet. Un aspecto a considerar es que en ciertos casos una palabra puede representar más de una emoción a la vez, en estas situaciones es necesario conocer primero la polaridad del tweet. La polaridad de un tweet representa emociones consideradas como positivas y negativas, (Singh, Singh, & Sohal, 2020). Tal como indica la Figura 10, conociendo la polaridad del tweet es posible conocer las emociones más representativas de este. Finalmente se analiza el tweet con sus emociones principales; la emoción que presenta un mayor valor de intensidad viene a ser la emoción que representa el tweet.

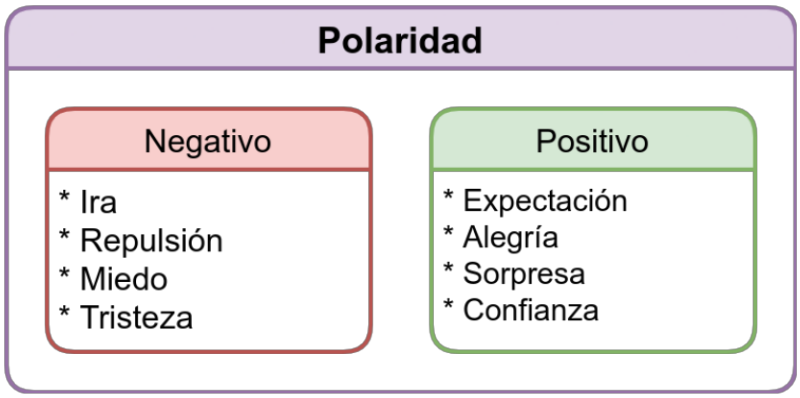


Figura 10. Clasificación de emociones por polaridad

2.3.5. Formatear los datos

La principal función de esta tarea es la adecuación sintáctica de los datos sin cambiar su significado (Gallardo, 2010), en este aspecto los campos a tratar fueron ‘fecha’ y ‘locacion’. De este modo para el campo ‘fecha’ se ha realizado una simple adecuación sintáctica eliminando elementos innecesarios. En cuanto a el campo ‘locacion’, además de una adecuación sintáctica se ha realizado un proceso de comparación con un diccionario. Esto implica que el campo ‘fecha’ se mantiene, pero con sus respectivos cambios sintácticos, mientras que en el análisis del campo ‘locacion’ este se mantiene hasta la creación de un nuevo campo denominado ‘pais’, una vez obtenido dicho campo el campo ‘locacion’ es

eliminado. En conclusión, el proceso de formateo para cada campo es diferente pues el análisis posterior de cada uno es distinto.

El campo 'fecha' original se encuentra en formato UTC extendido y fue cambiado a UTC básico. Este cambio se realizó con el objetivo de simplificar su análisis y así obtener uno de los tres eventos asociados a este proyecto; esto se detalla mejor en la fase de estructuración de datos. Por otro lado, para el campo 'locacion' se creó un diccionario el cual permite obtener el país al que hace referencia el campo 'locacion'. Este diccionario toma como datos el nombre de ciudades principales y las posibles formas de escribir el nombre de un país (incluyendo posibles errores ortográficos) devolviendo el país al que se hace referencia. Hay que recordar que los datos obtenidos son de países hispanohablantes incluyendo a EEUU por lo que el diccionario únicamente contiene países que cumplen esta característica, la Tabla 5 indica los países del estudio.

País	País
• Argentina	• Nicaragua
• Bolivia	• Panamá
• Chile	• Uruguay
• Colombia	• Perú
• Costa Rica	• Paraguay
• Ecuador	• Venezuela
• El Salvador	• España
• Guatemala	• EEUU
• Honduras	• Cuba
• México	• República Dominicana

Tabla 5. Lista de países a analizar

2.4. Fase de modelado

La fase de modelado consiste en la selección de las técnicas de modelado más idóneas para el proyecto. La o las técnicas adecuadas dependen del tipo de proyecto y de las características de los datos. (Gallardo, 2010). Este proyecto intenta analizar las emociones y sentimientos de comentarios en Twitter, por lo cual es necesario pensar en herramientas de procesamiento de lenguaje natural. Posterior a la selección de los modelos es necesario realizar un análisis de estos para determinar si son aptos o no para el proyecto.

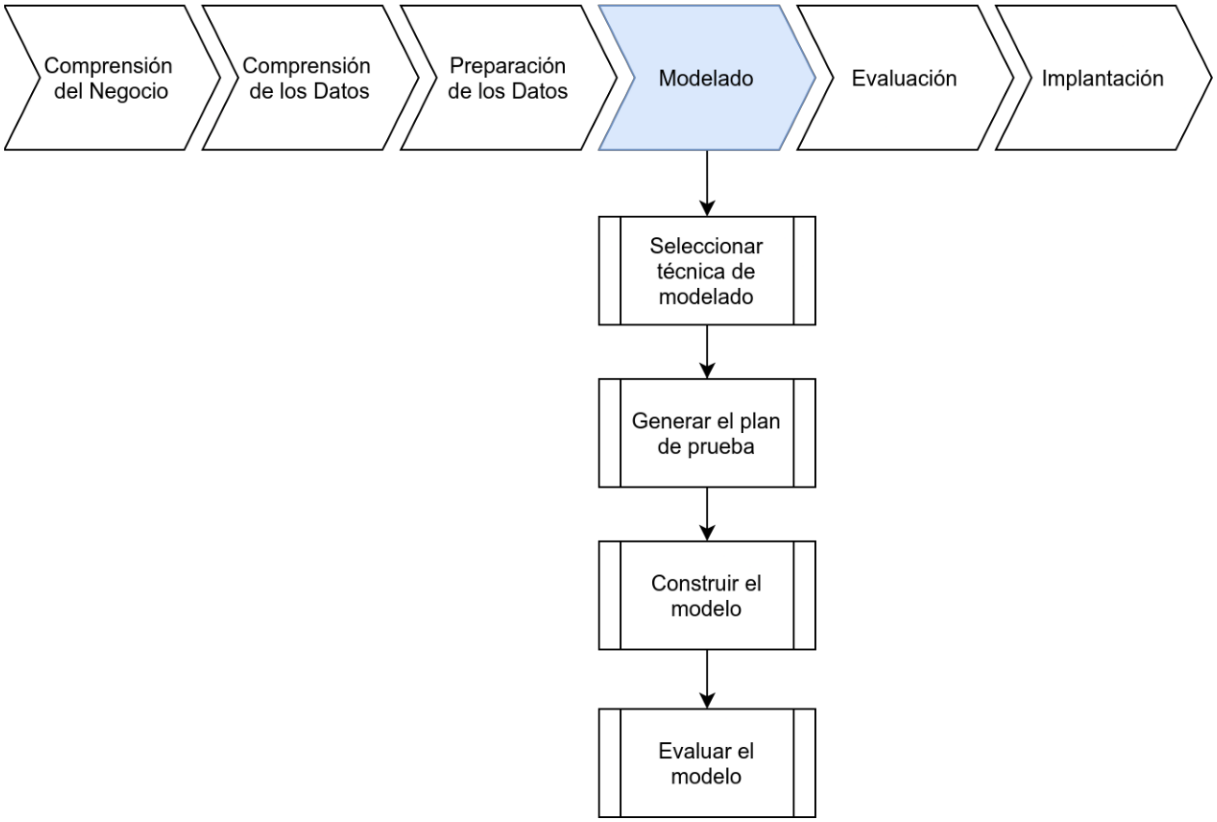


Figura 11. Fase de Modelado

2.4.1. Seleccionar técnica de modelado

El objetivo principal del proyecto es encontrar relaciones que den indicios de problemas psicológicos por lo cual se optó por modelos descriptivos, es decir se intenta encontrar patrones que permitan describir el comportamiento de una población. Asimismo, las librerías usadas para la clasificación de emociones y búsqueda de tópicos usan algoritmos de asociación. De esta manera se seleccionaron modelos descriptivos por asociación. Las librerías *NRCLex* y *Stanza* son las encargadas de realizar esta asociación, la primera asocia emociones valiéndose de un diccionario y la segunda asocia frases por medio de *part of speech* (PoS)

2.4.2. Generar el plan de prueba

Un plan de prueba para este proyecto no es necesario, debido a que las librerías que se van a usar (Stanza y NRCLex) son librerías ya entrenadas para el manejo de lenguaje natural. Esto implica que se trabajará con la totalidad de los datos para la construcción del modelo. Cabe mencionar que una vez obtenidos los modelos se contrastan con la opinión de un especialista en salud mental, de acuerdo con la opinión del especialista se puede aprobar o denegar la utilización de algún modelo.

2.4.3. Construir el modelo

El objetivo de esta fase es la ejecución del modelo en los datos; es aquí donde se describen las herramientas de minería de datos y los datos de salida. En consecuencia, en esta fase es donde se conseguirá tanto el análisis de emociones y sentimientos, como la asociación de tópicos en los datos. En este sentido se optó por dividir esta sección en dos partes, una para cada análisis.

- **Modelo de análisis de sentimientos y emociones (modelo 1)**

Este análisis intenta obtener las emociones y sentimientos del tweet, para esto hace uso de la librería *NRCLex* basándose en el diccionario producido por "*National Research Council Canada" (NRC)* y la librería *NLTK*. En este análisis se desea obtener las emociones que más representan a un tweet, las emociones presentes en la librería son: miedo, ira, expectación, confianza, sorpresa, tristeza, repulsión y alegría. Asimismo, esta librería permite analizar sentimientos (positivo y negativo), lista de emociones y la polaridad de un tweet. En cuanto a los campos que representan la intensidad de emociones (palabra, emoción, puntuación) se procedió a crear un script denominado *IntensidadPalabra_3.py* para que, junto con el diccionario Emolex sea posible determinar la emoción que más representa a un tweet. Los resultados de este análisis se pueden observar en la Figura 12.

G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
miedo	ira	expectante	confianza	sorpresa	tristeza	repulsion	alegria	depresion	positivo	negativo	emociones	polaridad	palabra	emocion	puntuacion
2	1	1	1	0	2	0	0	0	1	2	['miedo', 'triste	negativo	ansiedad	miedo	0.594
4	2	2	0	1	2	0	0	0	0	4	['miedo']	negativo	miedo	miedo	0.828
3	2	2	1	1	3	1	1	0	1	3	['miedo', 'triste	negativo	ansiedad	miedo	0.594
2	1	1	1	0	3	0	0	0	1	4	['tristeza']	negativo	angustia	tristeza	0.902
3	1	1	1	1	3	0	0	0	3	5	['miedo', 'triste	negativo	emergencia	sorpresa	0.828
2	1	2	2	0	2	0	2	0	6	4	['miedo', 'expec	positivo	ayudar	confianza	0.648
1	1	1	0	0	2	0	0	0	3	2	['tristeza']	positivo	angustia	sorpresa	0.625
1	1	1	0	0	1	0	0	0	0	1	['miedo', 'ira', '	negativo	ansiedad	miedo	0.594
4	1	4	1	2	3	0	3	1	4	5	['miedo', 'expec	negativo	miedo	miedo	0.828
3	1	2	1	0	3	0	0	0	1	4	['miedo', 'triste	negativo	depresión	tristeza	0.925
2	1	4	0	0	1	0	0	0	0	2	['expectante']	negativo	ansiedad	miedo	0.594
1	1	1	0	0	1	0	0	1	2	2	['miedo', 'ira', '	neutro	ansiedad	expectante	0.609
2	1	2	1	0	2	1	1	1	2	3	['miedo', 'expec	negativo	contagio	repulsion	0.719
3	1	3	1	1	3	2	1	0	2	5	['miedo', 'expec	negativo	obesidad	repulsion	0.695
2	1	1	0	0	2	0	0	0	2	3	['miedo', 'triste	negativo	aislamiento	tristeza	0.703
4	2	3	0	0	4	0	0	0	0	5	['miedo', 'triste	negativo	depresión	tristeza	0.925
1	1	1	0	0	1	0	0	0	0	1	['miedo', 'ira', '	negativo	ansiedad	miedo	0.594
1	1	1	0	0	2	0	0	0	0	2	['tristeza']	negativo	aislamiento	tristeza	0.703
2	1	2	0	0	2	0	0	0	0	3	['miedo', 'expec	negativo	depresión	tristeza	0.925
2	1	2	2	1	3	1	1	0	1	3	['tristeza']	negativo	angustia	tristeza	0.902

Figura 12. Resultados del análisis de emociones y sentimientos

Los nuevos campos obtenidos se adjuntaron a los datos preexistentes. Las uniones de estos campos tienen la finalidad de complementar la información inicial y facilitar la creación de los cuadros de mando (dashboards) que representen los resultados finales. La visualización de estos resultados debe reflejar la relación entre los datos y los resultados obtenidos del análisis. La descripción de cada campo se observa en la Tabla 6.

Campo	Script de Origen	Descripción
miedo	NRCLex_es	Cuentan la cantidad de palabras que representan una emoción en un tweet
ira	NRCLex_es	
expectacion	NRCLex_es	
confianza	NRCLex_es	
sorpresa	NRCLex_es	
tristeza	NRCLex_es	
repulsion	NRCLex_es	
alegria	NRCLex_es	
positivo	NRCLex_es	Cuenta la cantidad de palabras que representan un sentimiento en un tweet
negativo	NRCLex_es	
emociones	NRCLex_es	Lista de emociones presentes en el tweet
polaridad	3-ClasificadorSentimientos	Determina hacia que sentimiento tiene más inclinación un tweet
palabra	IntensidadPalabra_3	Palabra de mayor intensidad en un tweet
emoción	IntensidadPalabra_3	Indica la emoción predominante del tweet
puntuación	IntensidadPalabra_3	Indica numéricamente la intensidad de la palabra que posee la emoción predominante

Tabla 6. Descripción de los nuevos campos a analizar

La variación de cada emoción y sentimiento analizado en el total de datos se indica en la Figura 13. La figura claramente indica una predominancia de sentimientos y emociones negativas destacando el miedo y la tristeza. Esto resalta un miedo psicológico al COVID-19 entre los usuarios de Twitter.

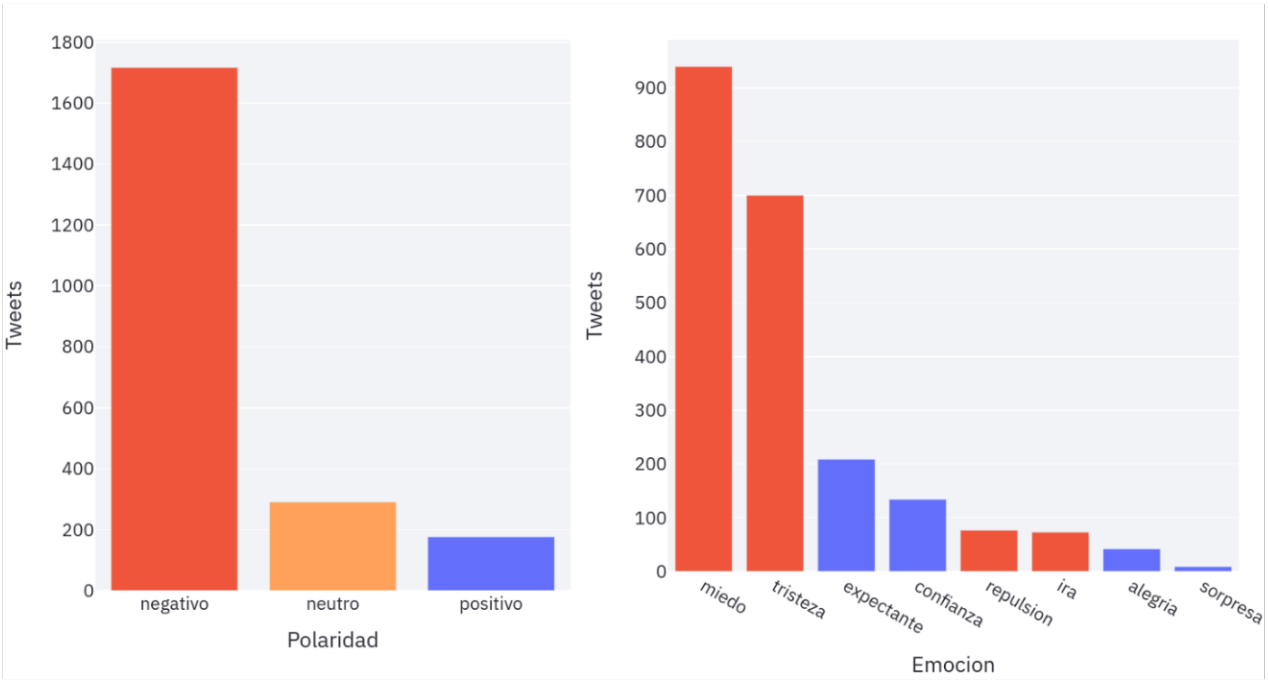


Figura 13. Relación entre emociones y sentimientos

- **Modelo de análisis de frases y palabras frecuentes (modelo 2)**

El objetivo de este análisis es encontrar tópicos o frases que se repitan con regularidad dentro de los datos, este proceso lo realiza la librería S*tanza*. La librería se basa en el procesamiento de lenguaje natural (NPL) usando técnicas de *Part of Speech (PoS)* y *dependency parsing*. De esta manera la librería busca en cada tweet las palabras más representativas de este. En cuanto a las palabras más frecuentes simplemente se usó librerías básicas de Python en combinación con la librería *Stopwords* para evitar que artículos o pronombres creen ruido. Al concluir cada uno de estos procesos generan diferentes archivos en formato CSV, uno con la lista de frases o tópicos frecuentes y el otro con una lista de palabras frecuentes.

La técnica de *PoS* divide cada elemento del tweet para obtener el núcleo, el objeto y un modificador en caso de ser necesario. Cada una de estas partes se complementan para obtener una frase que represente al tweet. En consecuencia, se usa la técnica de *PoS* para analizar cada tweet y encontrar frases y además se realiza un conteo de palabras buscando las más frecuentes. La Tabla 7 indica las 21 palabras más frecuentes asociadas al total de los datos. Dentro del análisis se ha excluido la palabra *COVID* y *ansiedad* debido a que fueron palabras clave para la búsqueda de tweets. De esta manera se observa que las palabras que más destacan son depresión, estrés, cuarentena, pandemia, salud y miedo; esto da indicios del sentir de una población referente al COVID19.

Palabra	**Frecuencia**	**Palabra**	**Frecuencia**	**Palabra**	**Frecuencia**
depresión	432	crisis	117	ataques	71
estrés	323	mental	108	tener	69
cuarentena	223	síntomas	99	confinamiento	66
pandemia	211	vida	85	insomnio	61
salud	197	tiempos	84	aislamiento	58
miedo	196	angustia	80	social	57
personas	125	gente	72	problemas	52

Tabla 7. Palabras frecuentes en tweets

La Tabla 7 puede ser complementada con la Figura 14 donde se muestra en porcentaje las diez palabras más frecuentes. La Figura 14 indica que la palabra que predomina es *depresión* con un 21% seguido de e*strés* con 15.9%, lo que da a interpretar que estas palabras son un pensamiento recurrente de una gran parte de los usuarios.

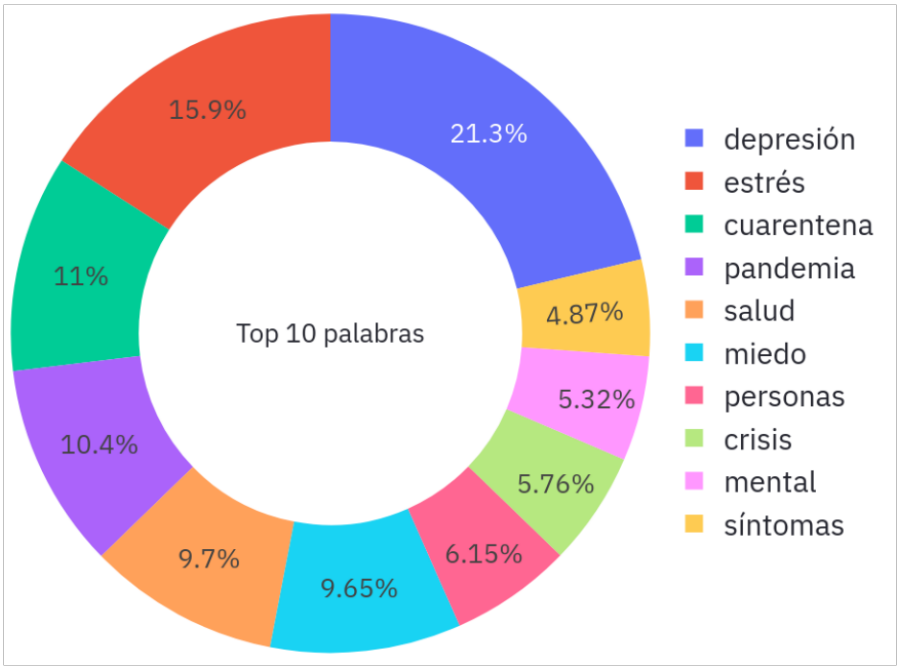


Figura 14. Porcentaje de las 10 palabras más frecuentes

El análisis de frases o tópicos entre los tweets da como salida una lista con los temas más recurrentes. La Tabla 8 muestra los 14 temas más recurrentes entre quienes comentan acerca del COVID 19. En esta tabla se observa que "salud mental", "ansiedad social", "estrés postraumático" son las temáticas más relevantes entre los usuarios asociadas a las palabras *COVID* y *ansiedad*. Estos tópicos pueden expresar pensamientos y emociones reales de como un grupo de personas se sienten respecto a la reciente pandemia.

Tema	**Frecuencia**	**Tema**	**Frecuencia**
Salud mental	89	Distanciamiento social	9
Ansiedad social	18	Ansiedad provocada	8
Estrés postraumático	12	Redes sociales	8
Emergencia sanitaria	10	Covid19 aislados	8
Ayuda psicológica	10	Seres queridos	8
Aislamiento social	10	Personal sanitario	8
Solución real	9	Ansiedad horrible	8

Tabla 8. Frases frecuentes en Tweets

Una representación más visual de los principales temas se encuentra en la Figura 15. La figura expresa claramente que el tópico más frecuente es *salud mental* con un porcentaje del 48.6%, este tema sin duda es el que más resalta, pues el siguiente tema más frecuente es *ansiedad social* llegando apenas al 9.84%. De igual forma los subsecuentes tópicos, aunque con porcentajes más bajos reflejan problemas con la salud mental.

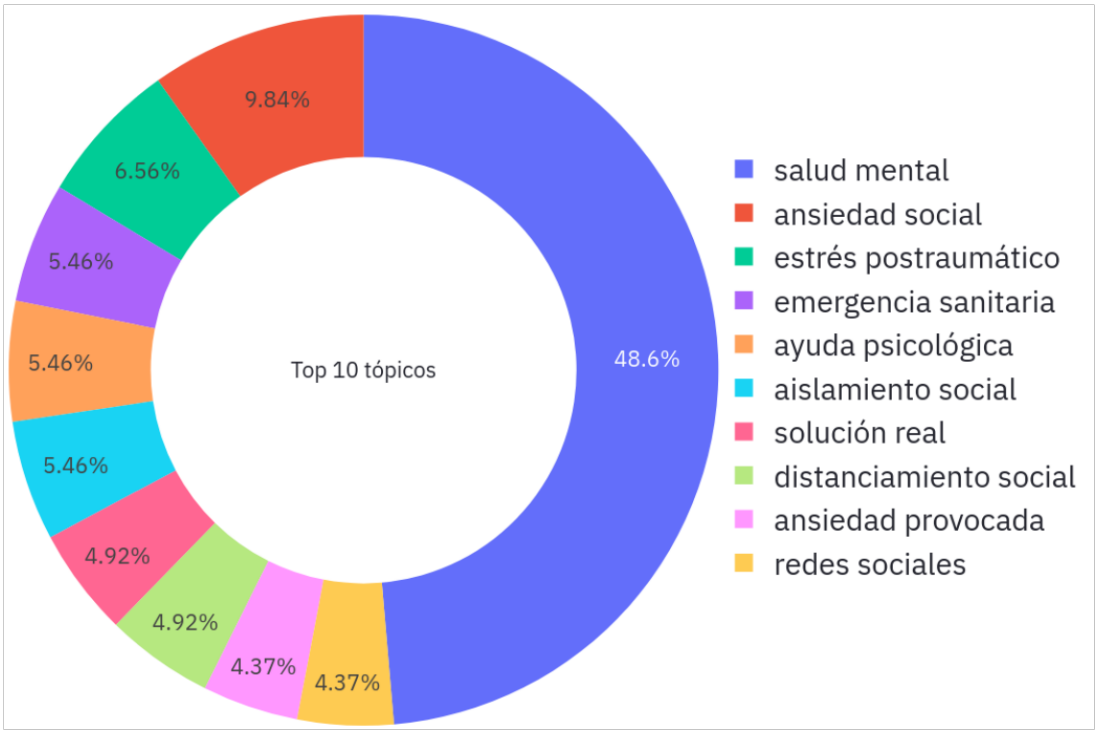


Figura 15. Porcentaje de los 10 temas más frecuentes.

2.4.4. Evaluar el modelo

La función de esta tarea es la evaluación del modelo con la finalidad de verificar que este se encuentre orientado a los objetivos del proyecto. Como consecuencia la evaluación se la realiza de acuerdo con el conocimiento y criterios de éxito preestablecidos en la subfase de: "*determinar los objetivos de minería de datos*" en la fase *"comprensión del negocio"*. Estos criterios son: encontrar temas recurrentes asociados a la salud mental, determinar cuáles emociones pueden afectar la salud mental y contrastar la información con un especialista en problemas psicológicos. Esto indica que en la evaluación se debe tener presente criterios de minería de datos, así como especialistas en el dominio, en este caso ciencias psicológicas. En conclusión, la evaluación del modelo está enfocado a la verificación del cumplimiento de los objetivos de minería de datos.

Para contrastar los resultados obtenidos se realizó una entrevista a un psicólogo familiarizado con la problemática del COVID19 y la salud mental. La entrevista tuvo la finalidad de conocer los problemas psicológicos más frecuentes asociados al COVID19. Por orden de frecuencia se indicó los siguientes problemas: depresión, ansiedad, trastorno ansioso depresivo, estrés, insomnio e irritabilidad; esto en parte provocado por emociones negativas como: tristeza miedo, ira, angustia y frustración. Cabe recalcar que los pacientes que presentaban estos trastornos no necesariamente contrajeron COVID19, sino que en algunos casos estos efectos se produjeron de manera indirecta.

El testimonio de quien ha tratado a pacientes con trastornos derivados de COVID19 junto con los resultados obtenidos fruto del análisis de NLP guardan similitudes. Estas similitudes van desde los principales problemas psicológico hasta las emociones que las pueden causar. Si bien el entrevistado no puede dar cifras exactas referentes a la cantidad de problemas tratados en su consultorio, el orden de frecuencia en que las menciona es bastante similar a la lista de palabras de mayor regularidad presentes en los datos como lo indica la tabla 7. Tomando en cuenta estas similitudes se puede estimar que los dos modelos cumplen con el objetivo de este trabajo.

2.5. Fase de evaluación

Esta fase consiste en la evaluación de los modelos generados desde el punto de vista de los objetivos establecidos. Una vez obtenida la evaluación se determina si los modelos cumplen o no con estos objetivos. (Gallardo, 2010)Si cumple con los objetivos se procede a la última fase, la implantación, de lo contrario se debe identificar el problema y repetir todo el ciclo.

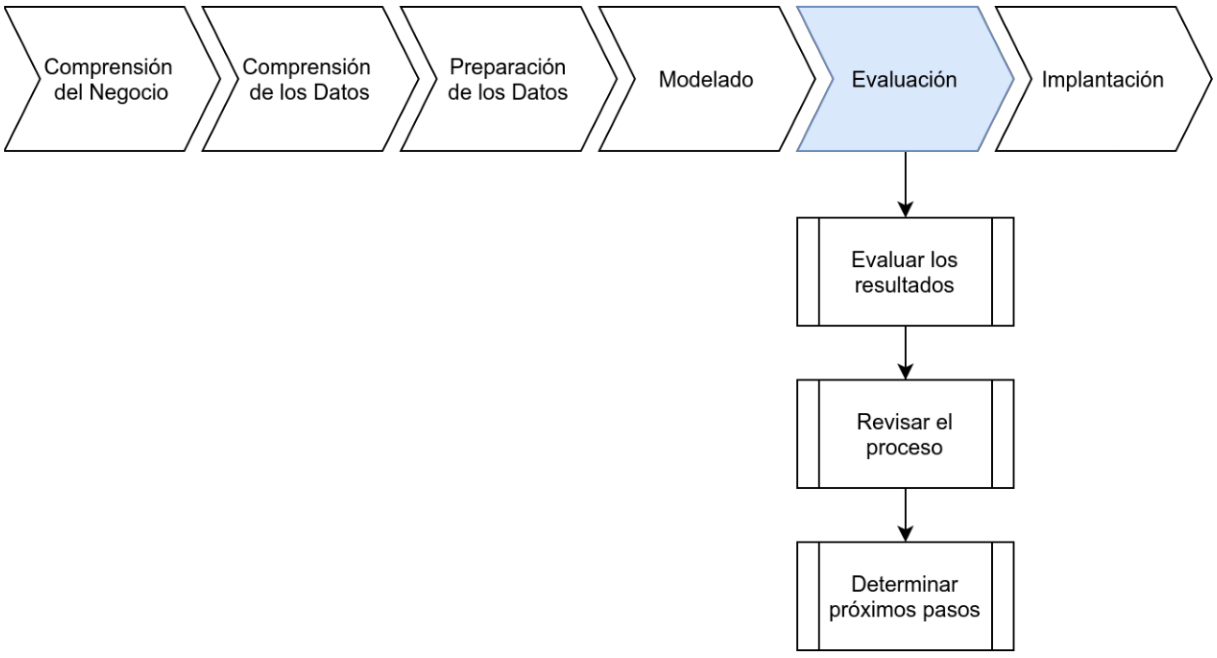


Figura 16. Fase de evaluación

2.5.1. Evaluar los resultados

Esta tarea se encarga de evaluar si los modelos establecidos cumplen con los objetivos del negocio. Desde esta perspectiva se crea un modelo que permita visualizar de manera más sencilla las afectaciones psicológicas producidas en la última pandemia. De igual manera también es necesario verificar los resultados con un especialista de la salud mental y corroborar que exista congruencia entre los resultados, el análisis del especialista y el objetivo del negocio. Considerando estos tres aspectos es posible realizar una evaluación para los modelos planteados y descartar aquellos que no cumplan con las expectativas del negocio.

Los resultados que se han obtenido marcan una preocupación con la salud mental de la población en Twitter. Esta preocupación se deriva principalmente en problemas de depresión, ansiedad, estrés y miedo; problemas frecuentes en un consultorio psiquiátrico. En relación con estas problemáticas es posible usar los datos de salida previamente analizados en la fase de modelado, para obtener dashboards que permitan manipular y visualizar emociones en diferentes momentos y lugares. Ambos modelos permiten analizar los sentimientos y emociones de la población, además presentan salidas que van acorde a casos tratados en un consultorio psiquiátrico, por lo cual son aptos para la generación de dashboards más complejos.

2.5.2. Revisar el proceso

Los procesos que se han realizado han funcionado de acuerdo con las estimaciones previstas cubriendo los objetivos del negocio. Si bien la cantidad inicial de registros contiene todos los tweets escritos en los intervalos de fecha adecuados para el estudio, el proceso de limpieza redujo una cantidad de registros notable. A pesar de que la cantidad de registros final es suficiente para realizar un análisis y conocer los problemas de una población, se desearía contar con una mayor cantidad de datos para obtener porcentajes y estimaciones más precisas. Una opción a esta problemática puede ser omitir el campo “pais”, con esto aumentarían los registros finales, pero se sacrificaría la situación de cada país respecto a la pandemia

2.5.3. Determinar próximos pasos

Los resultados que se han obtenido cumplen con los objetivos planteados tanto para la minería de datos como para las estimaciones de problemas psicológicos, por lo cual es posible continuar con la fase de implementación.

2.6. Fase de implantación

Esta es la última fase de la metodología CRISP-DM, consiste en definir el funcionamiento del proyecto al cliente (Gallardo, 2010). Asimismo, en esta fase se crean estrategias para el mantenimiento del proyecto y futuras mejoras. Para esto es necesario la creación de un informe donde se detalle los obstáculos al momento de la realización del proyecto junto con las posibles mejoras.

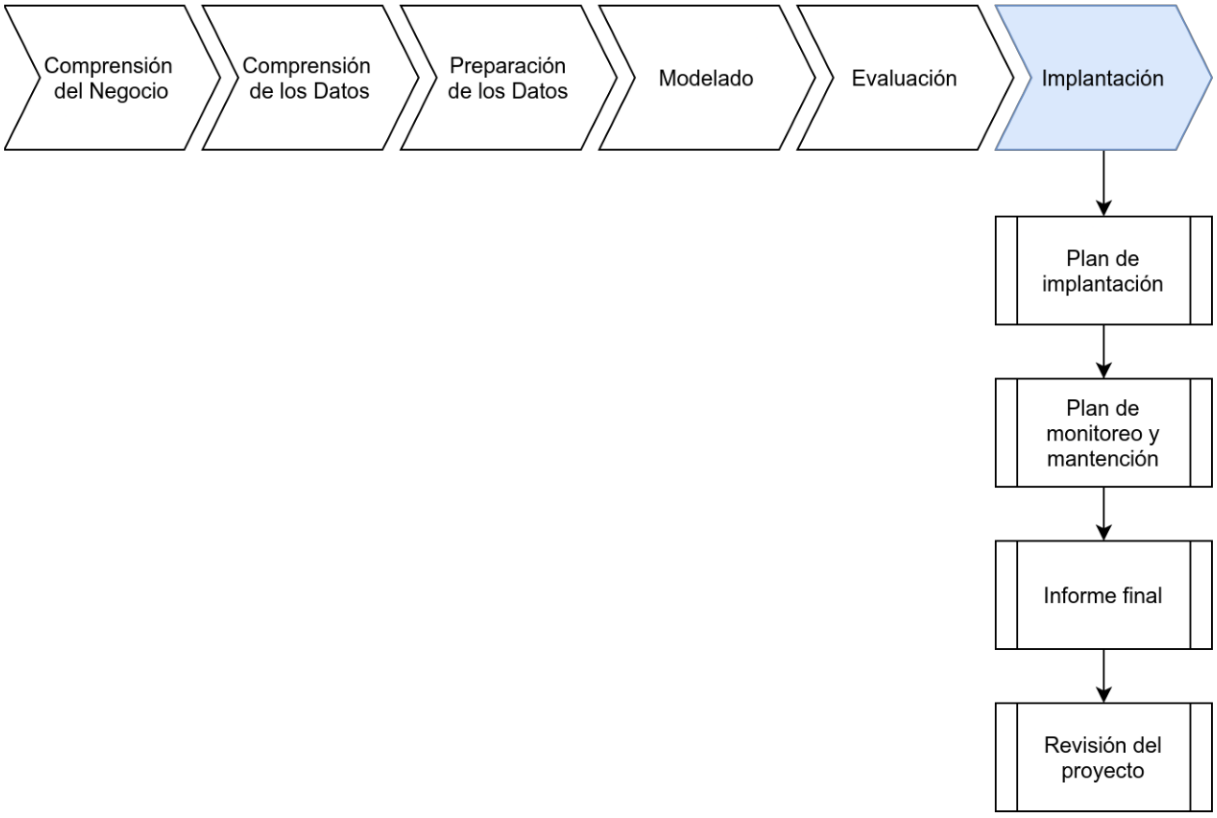


Figura 17. Fase de implantación

2.6.1. Plan de implantación

En esta tarea se intenta indicar los procedimientos a tomar para tratar el estado de salud mental de una población. En este sentido todos los datos obtenidos en las tres fechas tienen la misma emoción que sobresale sobre el resto, el miedo. Esta emoción es más notoria si se analiza por países, un gráfico de esto permite observar la variación de emociones para cada país, por lo cual a pesar de que todo un grupo de registros pertenece a un evento específico, se pueden tomar diferentes acciones dependiendo de la situación de cada país. De esta manera, en el caso de miedo lo primero a realizar es intentar reducirlo, esto puede hacerse con campañas informativas y tomando acciones que den confianza a la población.

2.6.2. Plan de monitoreo y mantención

El objetivo de esta tarea es conseguir una retroalimentación para conocer si los modelos se están utilizando de manera correcta. En este sentido los datos permiten conocer problemas psicológicos de una población, pero es necesario ir actualizando los datos con la finalidad de conocer cómo va cambiando la salud mental ya sea para mejorar o empeorar. Frente a esto la minería de datos debería realizarse cada vez que un evento considerado de impacto aparezca.

Como plan de monitoreo y mantención puede establecerse los siguientes procesos:

- Obtener de nuevos datos con cada noticia relevante a la opinión pública, como puede ser los procesos de vacunación, surgimientos de nuevas variantes de COVID, nuevos procesos de aislamiento social, etc.
- Verificar los cambios de emociones y sentimientos previo a haber aplicado medidas ya sea que ayuden a la mejora de la salud pública, o debido a algún evento negativo que haga referencia a la pandemia.
- Respaldar los datos con cada nuevo proceso de minería de datos, teniendo como finalidad asegurar los registros con posibles sobrescrituras y pérdida de información.
- Convertir los registros finales a hojas de cálculo con la finalidad de facilitar su respaldo y organización, además que podría facilitar las tareas de creación de nuevos gráficos y simplificar los procesos de visualización e interpretación para nuevos desarrolladores.

2.6.3. Informe final

En la actualidad las entidades de salud de cada país buscan maneras de contrarrestar los efectos del COVID. Con esta premisa el modelo puede ser usado para conocer el estado mental y la opinión pública respecto a la actual pandemia. De esta manera es posible conocer si las medidas que toma un país para combatir esta pandemia tienen impactos positivos, o negativos en su población. Asimismo, el modelo permite conocer si el estado de salud mental de una población mejora o empeora con las decisiones tomadas por las autoridades de cada país.

La realización de este proyecto de minería de datos siguió las directrices de la metodología CRISP-DM, especializada para el desarrollo de proyectos de minería de datos. Esta metodología consta de seis fases en las cuales se detalla el funcionamiento de cada una de estas, con la finalidad de generar resultados que vayan acorde al proyecto. Las fases van desde establecer los objetivos del negocio, pasando por la obtención de datos y el modelado de estos, hasta la obtención de resultados que resuelvan las problemáticas planteadas.

Los datos obtenidos fueron extraídos de Twitter mediante un proceso de raspado web o web scraping donde se filtró por idioma (español), palabras clave (ansiedad, covid) y por hechos de importancia (noticias). De esta forma, a todo este conjunto de registros se les aplicó técnicas de limpieza y adecuación de datos para su posterior modelado. Las técnicas de modelado empleadas (*Part of Speech y Dependency parsing*) han permitido conocer el estado de salud mental de los países hispanohablantes de América, dentro de este grupo se ha incluido a España y Estados Unidos. Finalmente es posible visualizar el impacto de cada noticia ya sea global o local en los países mencionados, con esto es posible supervisar el estado de una población con el surgimiento de noticias y hechos trascendentales.

2.6.4. Revisión del proyecto

La última tarea de la metodología CRISP-DM consiste en la evaluación de los aciertos y desaciertos y como estos se pueden mejorar. En este aspecto la cantidad de registros forma una premisa a mejorar, para esto además del tweet se puede extraer los comentarios que este tenga, aumentando así la cantidad de registros. Sin duda otro aspecto a considerar es la librería *NRCLex* la cual usa el diccionario NRC en inglés, pese a que su autor menciona que la traducción del inglés no presenta grandes variaciones con el resultado final, lo ideal sería conseguir un diccionario desarrollado en español.

Por otra parte, como aciertos cabe destacar la opinión dada por un especialista en la salud mental, quien específicamente ha tratado a pacientes con problemas derivados de la última pandemia. La situación de los pacientes dada por el especialista coincide con los resultados obtenidos fruto del análisis y modelado de los tweets obtenidos. Esto implica que a pesar de no poseer una cantidad de registros tan extensa como se desearía, los registros que se tienen son suficientes para conocer el estado de salud mental de diferentes poblaciones. Con una cantidad de datos más grande sería posible tener porcentajes más fiables y cercanos a la realidad de un país.

Se realizó una segunda entrevista con el psicólogo para conocer su criterio referente a los resultados del aplicativo y su experiencia tratando esta problemática. En esta entrevista menciona que, en su caso la emoción que más predominó fue el miedo seguido de la ansiedad y la tristeza. Si bien la ansiedad como emoción no analiza directamente el programa, al revisar el manual DSM-5 (psiquiatría, 2014) usado por psicólogos para generar un diagnóstico se observa que, en los postulados referentes a trastornos de ansiedad la emoción que más se marca es el miedo. El miedo sin duda es la emoción que sobresale en la sección de ansiedad.

Si se analiza la depresión dentro del manual DSM-5 se nota que en sus postulados no existe en sí una emoción que sobresalga notablemente sobre otras. Según el manual DSM-5 dependiendo de la causa del trastorno, dependerá la emoción predominante, aunque en gran parte será una combinación de múltiples emociones. Si se elimina los trastornos de depresión causados por el consumo de sustancias o por cambios hormonales normales como durante la menstruación. La emoción que sobresale sobre las demás es la tristeza seguida de nerviosismo, miedo, melancolía, irritabilidad, culpabilidad y desesperanza. Con estas emociones se puede contrastar y comparar las emociones de trastornos como la ansiedad y depresión con los obtenidos después del análisis de datos. Existe relación en cuanto a las emociones obtenidas y las emociones esperadas para casos de ansiedad y depresión.

3. DESARROLLO DE LA APLICACIÓN DE VISUALIZACIÓN

El modelo usado para el desarrollo del aplicativo web utilizado para la visualización fue el modelo en cascada. Se eligió este modelo por varios motivos: la simplicidad de la funcionalidad de la aplicación web, la facilidad que presenta para distinguir cada fase del proceso de desarrollo y la visión clara de los objetivos de la aplicación (crear gráficos que permitan encontrar patrones que puedan ser relacionados con problemas de salud mental).

El modelo en cascada propone el desarrollo del software de forma secuencial, de modo que cada fase empieza cuando termina la anterior. Para avanzar a la siguiente fase se procede verificar si la fase en curso cumple con los requerimientos del software. En síntesis, el modelo en cascada se define como una secuencia de fases en la que, en cada fase se verifica el cumplimento de los requisitos. (Sommerville, 2005). El modelo consta de cinco fases resumidas a continuación:

- Recolección de requerimientos: fase donde se reúnen todos los requisitos que debe cumplir el software.
- Diseño: fase donde se describe la arquitectura de la aplicación y la interfaz del software a desarrollar.
- Codificación: fase donde se pasa a la programación del software culminando con un software operativo.
- Pruebas: fase donde se prueba todos los componentes y módulos del software, de esta manera se trata de encontrar y corregir los errores
- Mantenimiento: fase donde se realizan actualizaciones, mantenimiento del software y cambios necesarios (de ser el caso).

3.1. Recolección de requerimientos

El levantamiento de los requerimientos se lo hizo en base a la entrevista a un psicólogo clínico en el Centro de Rehabilitación “La Dolorosa”. La Tabla 9 contiene los requerimientos identificados para la aplicación web.

Código	Nombre	Descripción
A01	Seleccionar análisis general o por eventos	La aplicación permitirá visualizar los resultados de manera general o por eventos
A02	Clasificar tweets por sentimientos	La aplicación permitirá visualizar el total de tweets clasificados por sentimientos
A03	Clasificar tweets por emociones	La aplicación permitirá visualizar el total tweets clasificados por emociones
A04	Indicar palabras más frecuentes	La aplicación permitirá visualizar una lista con las palabras más frecuentes en el total de tweets
A05	Indicar tópicos más frecuentes	La aplicación permitirá visualizar una lista con los tópicos más frecuentes en el total de tweets
A06	Indicar palabras frecuentes en cada emoción	La aplicación permitirá visualizar las palabras más usadas al expresar cada emoción en el total de tweets
A07	Indicar intensidad emociones por país	La aplicación permitirá visualizar la intensidad de emociones en cada país para cada evento
A08	Indicar palabras más frecuentes en cada evento	La aplicación permitirá visualizar las palabras más frecuentes para cada emoción para cada evento
A09	Comparar emociones por países	La aplicación permitirá realizar una comparación entre las emociones presentes en cada país para cada evento

Tabla 9. Tabla de requerimientos

3.2. Diseño

3.2.1. Arquitectura de la aplicación web

La aplicación web está diseñada para la visualización de los resultados producidos por el análisis de datos. Por este motivo está compuesta únicamente de tres niveles en su arquitectura: servidor web, estación de trabajo navegador web y cliente, tal como lo indica la Figura 18.

- Cliente: es quien realiza la petición al navegador web.
- Estación de trabajo: Es el lugar donde se tiene instalado el navegador web.

- Navegador web: es quien envía un mensaje de petición para conectarse con el servidor web.
- Servidor web: es quien recibe el mensaje por medio de la interfaz gráfica y lo envía a al fichero que contiene los datos.
- Interfaz gráfica: es quien envía un mensaje al archivo CSV que contiene los datos solicitando las consultas que se requieren obtener.
- Archivo CSV: contiene todos los datos de la aplicación.

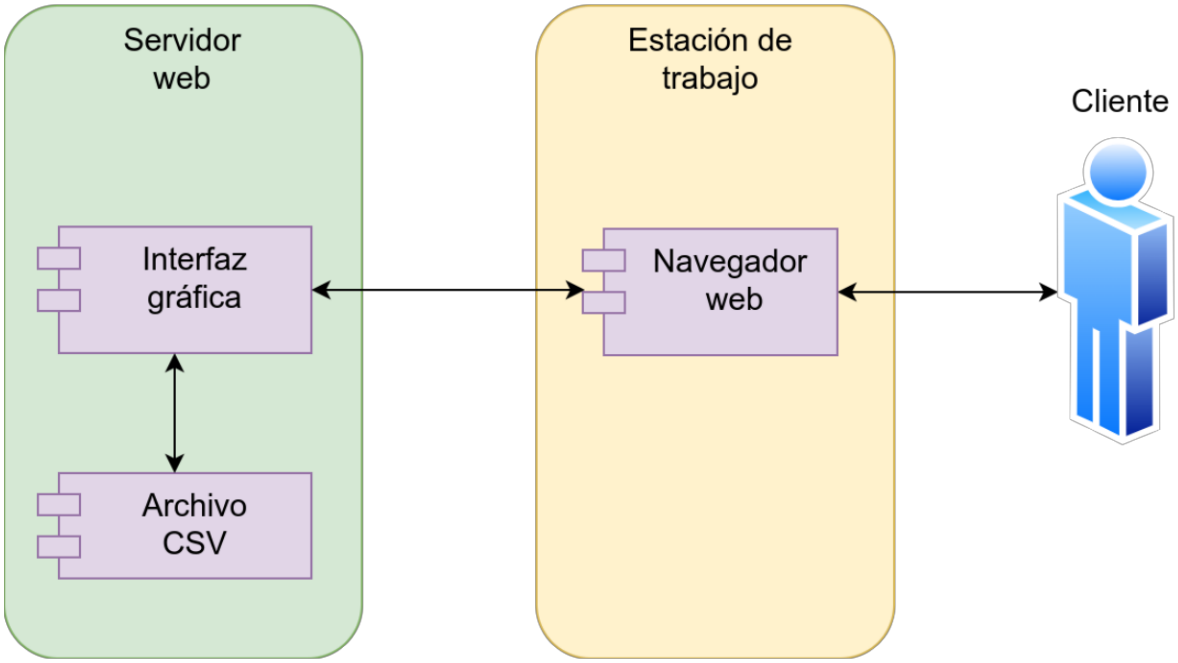


Figura 18. Arquitectura de la aplicación

La arquitectura de la aplicación web en su primer nivel usa un fichero CSV para almacenar los datos. De esta manera el fichero CSV funciona como una base de datos NoSQL. Al usar un texto plano como registro de datos la manipulación de los datos se la realizó mediante la librería *Pandas* de Python. Dicho archivo CSV previamente ha pasado por todo un proceso de limpieza, formateo y análisis. Este proceso se lo ha realizado en el lenguaje de programación Python mediante las librerías de *Pandas, NCRLex, Stopwords* y *Stanza*. El archivo CSV resultante fue diseñado para que pueda ser escalable, es decir se puedan agregar registros y estos no afecten en gran medida las consultas y análisis rápidos.

El propio servidor web aloja tanto el archivo CSV, como la interfaz gráfica quien por medio de dashboards, envían las instrucciones de consulta y visualización de los datos. La interfaz gráfica y dashboards se han programado en lenguaje Python mediante el uso principal de la librería *Streamlit*. Esta librería levanta un servidor, (en este caso local), para la presentación de la aplicación web en la URL*:* *http://localhost:8501* . Para la representación de los diferentes gráficos se usaron librerías como *Plotly* para los histogramas, diagramas de pastel y diagrama

dona, la librería *Wordcloud* para la visualización de nubes de palabras para las diferentes emociones, la librería *Folium* para la representación geográfica de los países.

En el segundo nivel, en la estación de trabajo se encuentra el navegador web, en el cual para acceder a la interfaz gráfica proveniente del servidor web se debe ingresar la dirección la dirección http://localhost:8501 . De esta manera el tercer nivel, es decir el cliente puede ingresar las instrucciones mediante el navegador web, quien a su vez envía las peticiones a la interfaz gráfica, y este transforma dichas peticiones a instrucciones de consulta en el archivo CSV.

3.2.2. Diseño de la aplicación web

La aplicación web está formada básicamente de dos secciones. La primera sección está compuesta de cuadros de mando (dashboards) y la segunda sección con la visualización de los resultados del análisis, una representación de esto se encuentra en la Figura 19. Los cuadros de mando, también llamado tablero de control es donde el usuario puede escoger el tipo de resultado que se desea observar (general o por eventos). En cuanto a la presentación de resultados, es aquí donde se muestra gráficamente los resultados del análisis de los datos. La presentación de los resultados puede ser modificada desde el tablero de control; los controles dependerán del tipo de gráfico. De esta manera es controlar desde la selección del modo de gráfico hasta la visualización por tipo de emoción.

Figura 19. Diseño de la aplicación web

3.2.3. Diseño de tablero de control y presentación de resultados

3.2.3.1. Diseño para el análisis general

El tablero de control permite seleccionar el tipo de resultados que se desea visualizar, ya sea un reporte general o por eventos. Dependiendo de la opción que se seleccione los controles cambiarán. De esta manera, si se desea observar el análisis general se obtendrán controles para los gráficos de *Emociones y sentimientos, Palabras y frases frecuentes* y finalmente *Nube de palabras por emociones* tal como se observa en la Figura 20. Estos controles a su vez permiten modificar la forma de visualización de los resultados de acuerdo con lo que se necesite.

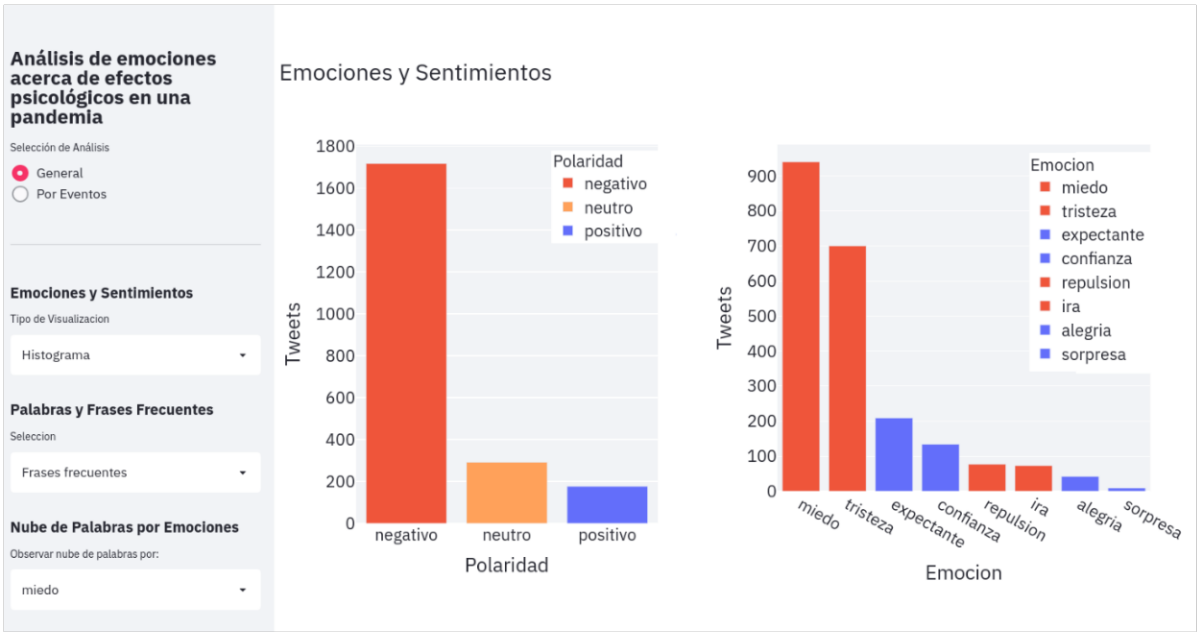


Figura 20. Tablero de control en Análisis General

3.2.3.1.1. Emociones y sentimientos

Este control permite escoger el tipo de gráfico con el que se desea observar las emociones y sentimientos. Las opciones de diagramas disponibles son: “Histograma” y “Diagrama de pastel”. Estos diagramas se los puede observar en a Figura 21 y en la Figura 22.

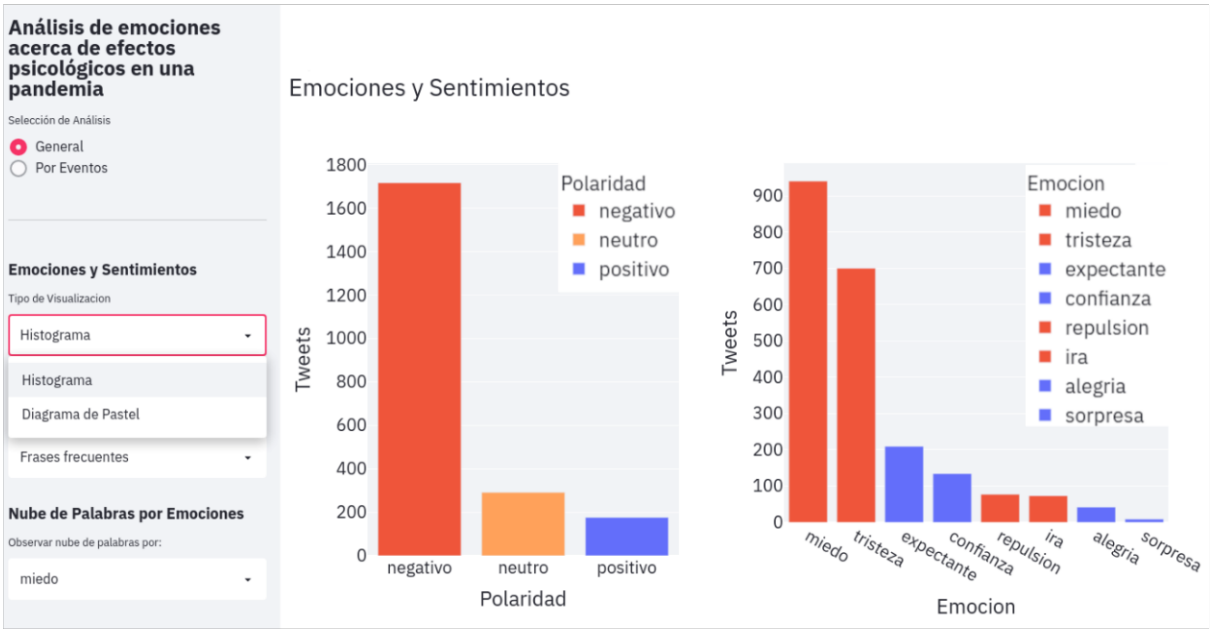


Figura 21. Histograma: sentimientos y emociones

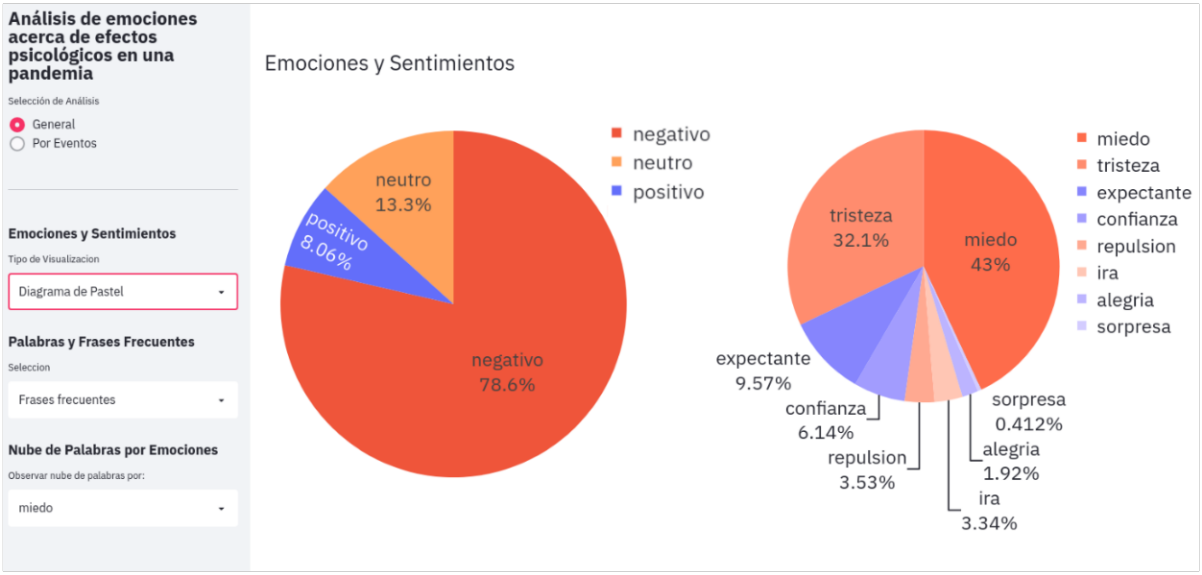


Figura 22. Diagrama de pastel: sentimientos y emociones

3.2.3.1.2. Palabras y frases frecuentes

El objetivo de este control es permitir la visualización de las palabras y frases que se usan con más regularidad. Para esto, muestra un diagrama dona con los porcentajes de las diez palabras o frases más frecuentes, además indica una tabla con todas las demás palabras y frases. El control permite seleccionar la vista de las frases o palabras frecuentes tal como lo indica la Figura 23 y la Figura 24.

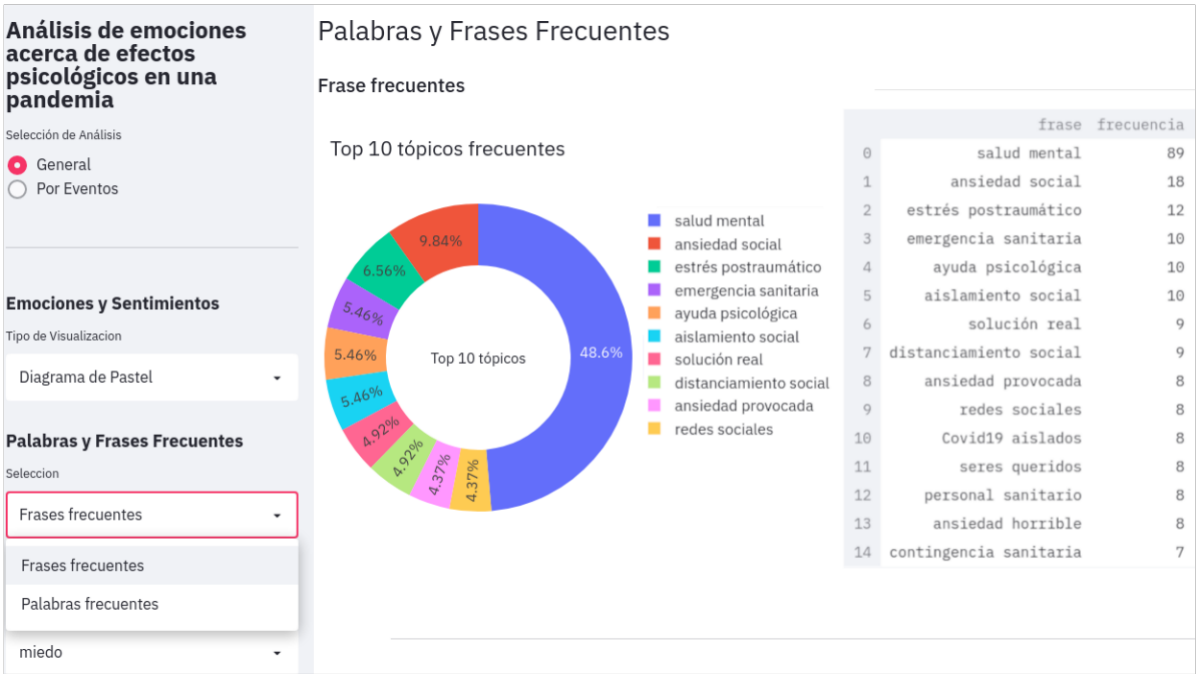


Figura 23. Frases frecuentes

Figura 24. Palabras frecuentes

3.2.3.1.3. Nube de palabras por emociones

En esta sección se puede visualizar una nube de palabras, aquellas con más repeticiones serán de mayor tamaño. El control permite seleccionar la emoción por la cual se desea visualizar la nube de palabras. La lista de emociones tiene las ocho emociones que se han analizado: miedo, tristeza, expectación, confianza, repulsión, ira, alegría y sorpresa. La representación de esto se lo puede apreciar en la Figura 25. La Figura 25 indica una nube de

palabras para la emoción de *miedo.* Para este ejemplo únicamente se seleccionó una emoción pues el objetivo es explicar la interfaz, en el análisis de resultados se detalla mejor la nube de palabras para más emociones.

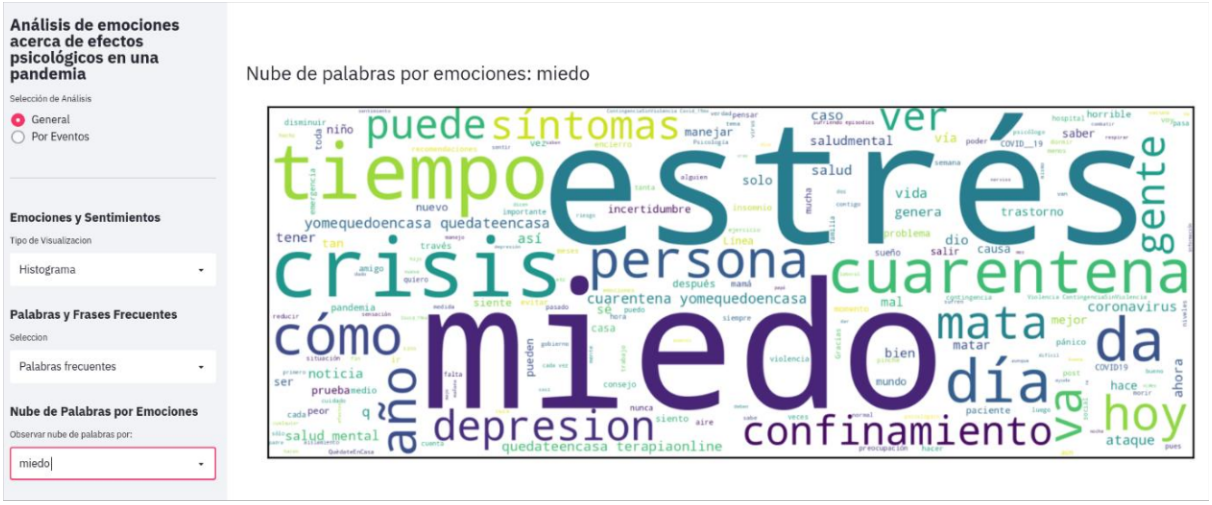


Figura 25. Nube de palabras por emociones

3.2.3.2. Diseño del análisis por eventos

El tablero de control para este análisis se puede apreciar en la Figura 26. En esta sección es posible observar el análisis de acuerdo con cada evento. Al igual que el análisis general, este consta de tres controles para tres diferentes gráficos, estos son: *Mapa de calor por emociones nube de palabras* y *Comparación de emociones por países.*

Figura 26. Tablero de control en análisis por eventos

3.2.3.2.1. Mapa de calor por emociones

El control permite seleccionar una de las tres opciones de eventos, además este control funciona en conjunto con una “casilla de selección” ubicado en la sección: presentación de

resultados. La casilla de selección permite elegir una emoción y junto con un evento muestra un mapa de calor con los países en estudio; entre más rojo se pinte un país, mayor porcentaje de una emoción tendrá. La Figura 27 indica el funcionamiento del mapa de calor.

Figura 27. Mapa de calor por eventos y emociones

3.2.3.2.2. Nube de palabras

El control en este apartado permite seleccionar la cantidad de palabras que se desea visualizar en la nube de palabras. La selección parte desde 20 hasta 100 palabras en intervalos de diez en diez. Este gráfico funciona en conjunto con el mapa de calor para visualizar las palabras más usadas en un determinado evento, además funciona con la casilla de selección en la presentación de resultados, para seleccionar también el tipo de emoción. En la Figura 28 se puede apreciar su funcionamiento, para esta explicación únicamente se seleccionó la emoción de “tristeza”.

Figura 28. Nube de palabras por noticias y emociones

3.2.3.2.3. Comparación de emociones por países

El objetivo de este control es seleccionar países para compararlos entre sí respecto a los tres diferentes eventos. La comparación se lo realiza con los porcentajes de todas las emociones presentes en un país. El gráfico presenta en una sola vista los porcentajes de todas las emociones para cada uno de los tres eventos. La Figura 29 indica el funcionamiento de este control.

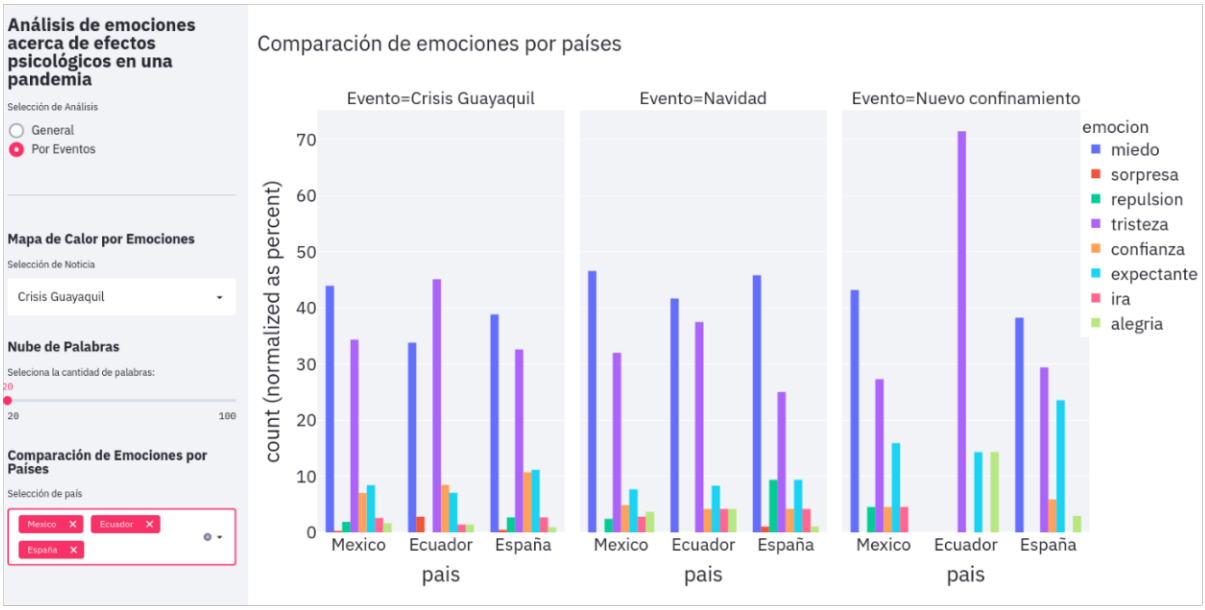


Figura 29. Comparación de emociones y noticias

3.3. Codificación

Todo el aplicativo web se programó en lenguaje Python usando como base la librería Streamlit. Dicha librería está diseñada para la creación de sitios web enfocados a la

visualización de datos. Los diagramas de pastel histogramas y el gráfico de comparación de emociones se lo realizó con la librería *plotly.express.* Los gráficos para observar palabras y frases frecuentes se los obtuvo mediante el uso de las librerías *plotly.graph_objects* y *stopwords.* Los gráficos de nubes de palabras se los obtuvo con la librería *wordCloud.* Por último, para la visualización del mapa de calor se usó la librería *steamlit_folium.* Los datos sobre los cuales se aplicó el código habían pasado todas las etapas de limpieza y preparación.

3.4. Pruebas

La fase de pruebas no presentó graves inconvenientes al evaluar la funcionalidad de la aplicación, sin embargo, hay que mencionar suscitados. Uno de los incidentes a mencionar fue el trabajar con streamlit en su última versión 1.11.0. La última versión presenta conflictos al crear gráficos por columnas (sobreposición de gráficos), para solucionar este problema se usó una versión más estable la versión 0.88.0. La librería stopwords permite eliminar palabras sin valor (artículos) al momento de crear nubes de palabras, aunque no evita el ruido de ciertas letras y caracteres recurrentes en Twitter como: “www”, “@”, “https”, “RT”, etc. Para solucionar esto se creó una lista con combinaciones de letras y caracteres a obviar, dicha lista trabaja en conjunto con la librería stopwords.

3.5. Mantenimiento

La aplicación web está diseñada para trabajar con un archivo de datos en formato CSV; siempre y cuando se respete la estructura del archivo, es posible agregar más registros sin afectar la funcionalidad del programa. De la misma manera el código de la aplicación web puede ser adaptado para generar nuevos gráficos en función de lo que se requiera.

4. ANÁLISIS DE RESULTADOS

4.1. Resultados generales

Al analizar toda la totalidad de los datos sin centrarse en un grupo específico es posible conocer el estado global de la población. Como consecuencia los tweets negativos son los que destacan con un 78.6% de todos los tweets, tal como lo marca la Figura 30. En este sentido las emociones más frecuentes son el miedo y la tristeza con 43% y 32.1% respectivamente, se puede visualizar esto en la Figura 31. Estos resultados indican la preocupación descrita en los tweets lo cual puede deberse y/o derivarse en problemas psicológicos.

Un referente para conocer tanto la opinión de una población, así como su salud mental es la frecuencia de palabras y tópicos. Este análisis intenta conocer los temas más recurrentes tal como se observa en la Figura 30 y la Figura 31, en estos gráficos se observa que los temas más frecuentes están inclinados a problemas de ansiedad, depresión y estrés. Aunque estos temas son los que prevalecen cabe destacar que también existen tópicos que hacen referencia a ayuda psicológica. De esta forma se puede decir que, aunque no se posea un diagnóstico psicológico y no se pueda afirmar si se posee o no algún problema de salud mental, una parte de la población siente la necesidad de ir con un especialista en ciencias psicológicas.

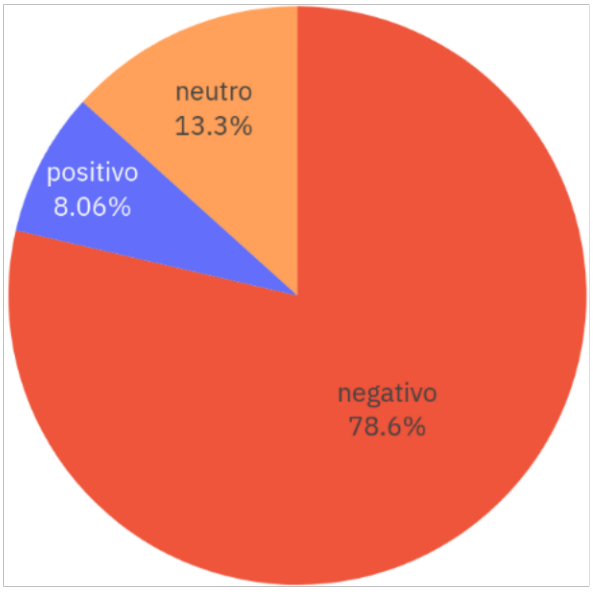


Figura 30. Porcentaje de sentimientos

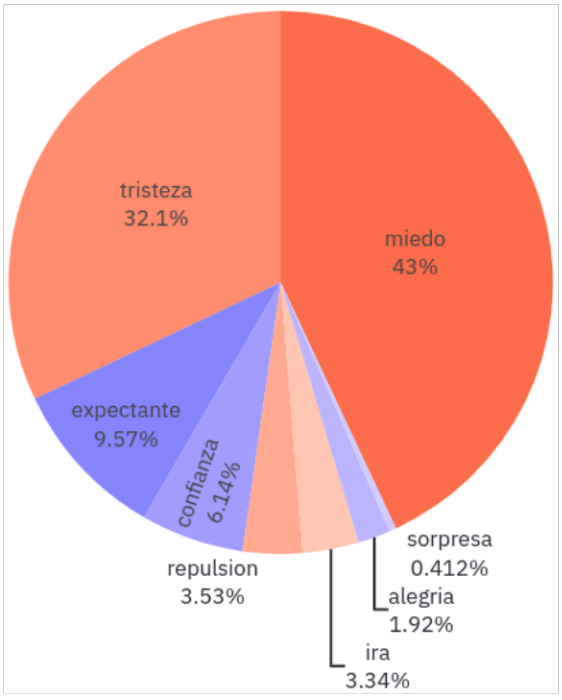


Figura 31. Porcentaje de emociones

Una nube de palabras permite tener una vista rápida de las palabras más usadas al expresar emociones tal como se observa en las Figuras 32 hasta la Figura 35. De esta manera la Figura

32 indica que si se habla de miedo las palabras más recurrentes son: crisis, estrés, depresión, etc. Si se habla de tristeza la Figura 33 indica que las palabras más frecuentes son: depresión, pandemia, angustia, etc. Al hablar de repulsión la Figura 34 muestra que las palabras más comunes son: falta, contagio, aire, etc. Si se desea conocer que palabras se usan más al hablar de ira, la Figura 35 indica que son: ataque, batalla, emociones, etc. Si bien la aplicación visualiza también las palabras más comunes al expresar emociones positivas, estas no se las ha considerado para este estudio.

Figura 32. Nube de palabras por miedo

Figura 33. Nube de palabras por tristeza

Figura 34. Nube de palabras por repulsión

Figura 35. Nube de palabras por ira

4.2. Resultados por eventos

Los siguientes resultados están centrados en los análisis de tres diferentes eventos considerados des interés (crisis sanitaria en Guayaquil, Navidad y año nuevo, Nuevo confinamiento en Ecuador). Estos análisis permiten visualizar las emociones de cada país respecto a cada evento y compararlos con otros.

4.2.1. Crisis sanitaria en Guayaquil

Los tweets obtenidos para esta noticia van desde el 5 de abril del 2020 hasta el 10 de mayo del 2020 y al igual que en el análisis general se tomarán únicamente las emociones consideradas como negativas. La Figura 36 permite conocer de manera global la intensidad de una emoción. Al observar la emoción de “miedo” y compararla con las demás emociones, claramente se entiende que es la emoción con mayor intensidad en este evento. Los países

que destacan con esta emoción son EEUU, Costa Rica, Nicaragua, Chile, Bolivia y Uruguay. A pesar de esto, el Ecuador no marca una intensidad tan fuerte como se esperaría. Esto puede ser entendido a las condiciones de cada país, por ejemplo, en el caso de EEUU en el mismo intervalo de tiempo que se está analizando, sucedieron las protestas por la muerte de George Floyd. De este modo a pesar de que, para Ecuador la crisis sucedida en Guayaquil fue de gran impacto, a nivel global muchos países tuvieron una respuesta al miedo más intensa que Ecuador.

Figura 36. Mapa de calor del miedo para noticias relacionadas con la crisis sanitaria en Guayaquil

Figura 37. Mapa de calor de la tristeza para noticias relacionadas con la crisis sanitaria en Guayaquil

Al hablar de tristeza, en la Figura 37 se observa que a nivel global es mucho menos marcada que el miedo, pero existen algunos países donde es más notable como son Guatemala y Paraguay. Al observar a Ecuador este sentimiento, no es tan intenso como en estos países, pero si lo es comparado con sus vecinos. El mapa de la Figura 38 muestra la intensidad de repulsión o aversión a nivel global, aquí resaltan los países de Venezuela y Guatemala. En relación a sus vecinos Ecuador presenta una menor intensidad, el país que resalta es Perú. Por último, el mapa de ira en la Figura 39 indica que el miedo se encuentra presente en la mayoría de los países, aunque con menor intensidad. Los países que resaltan son El Salvador, Venezuela y Perú; en este caso Ecuador presenta niveles bajos de ira.

Figura 38. Mapa de calor de aversión para noticias relacionadas con la crisis sanitaria en Guayaquil

Figura 39. Mapa de calor de la ira para noticias relacionadas con la crisis sanitaria en Guayaquil

Los mapas de calor presentan una vista comparativa rápida de las emociones, pero si se desea conocer los porcentajes de estas, un gráfico barras es una buena opción. Se ha realizado una comparación del Ecuador solo con países que destacan en sus emociones. Es así que la Figura 40 analiza las emociones para Colombia, Perú y Ecuador. Ecuador en comparación posee alrededor de un 15% más de “tristeza” que sus vecinos, al mismo tiempo presenta alrededor del 16% y 7% menos “miedo” respecto de sus vecinos. Asimismo, la Figura 41 compara Ecuador con EEUU y Chile; en este caso Ecuador tiene alrededor de un 20% menos “miedo” que EEUU y Chile. Si se compara la tristeza estos países tiene, Ecuador presenta alrededor 17% y 23% más “tristeza” que Chile y EEUU respectivamente.

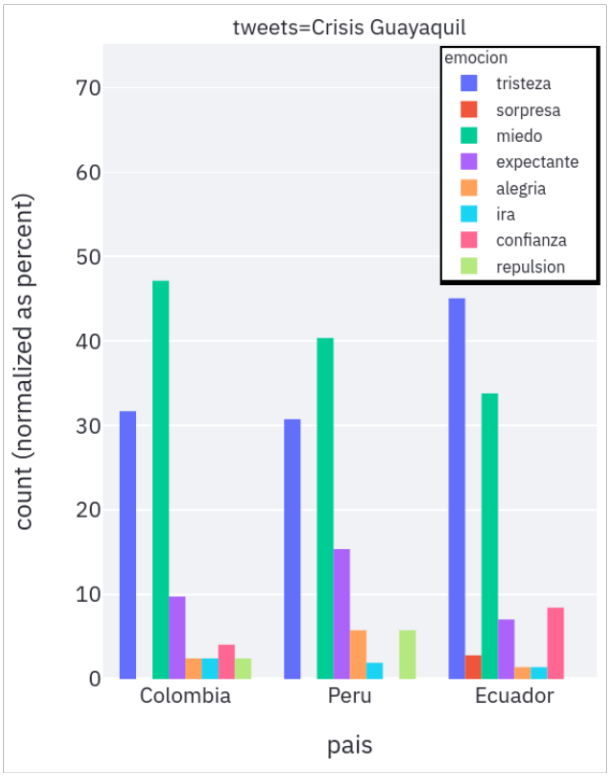


Figura 40. Comparación de emociones por países respecto a la crisis sanitaria en Guayaquil (1)

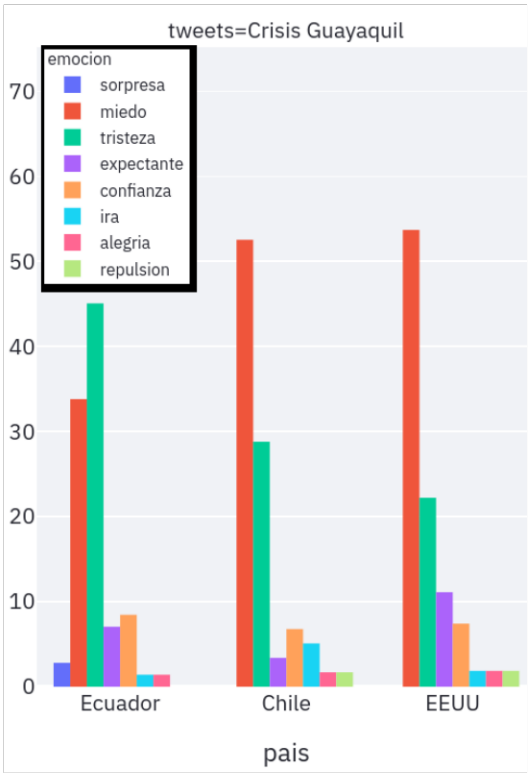


Figura 41. Comparación de emociones por países respecto a la crisis sanitaria en Guayaquil (2)

4.2.2. Navidad y año nuevo en pandemia

Los tweets de esta notica van desde el 3 de diciembre del 2020 hasta el 3 de enero del 2021, con el objetivo de obtener información en estas festividades. Al igual que “Crisis en Guayaquil” el miedo es la emoción que más resalta en esta noticia, pero esta vez con menor intensidad. La Figura 42 muestra que los países que presentan una mayor intensidad de esta emoción son: Honduras, Chile, Argentina y Paraguay. Ecuador respecto a sus vecinos no presenta una variación visible. Si se observa la Figura 43 se puede notar que la tristeza en relación con el miedo no presenta mucha diferencia, cambiando los focos de interés a Nicaragua y Uruguay.

El mapa de la Figura 43 representa la repulsión o aversión, este mapa representa grandes cambios, desde países con muy baja o nula intensidad en repulsión. De esta manera, los países a destacar son Venezuela, Honduras, Chile y España. Ecuador en este mapa presenta un bajo nivel de repulsión respecto a Colombia y Perú. El mapa de la Figura 45 representa la emoción de ira, aquí se observa que a nivel global esta emoción es muy baja con un solo país a destacar, Guatemala.

Figura 42. Mapa de calor del miedo para noticias relacionadas con navidad y año nuevo

Figura 43. Mapa de calor de la tristeza para noticias relacionadas con navidad y año nuevo

Figura 44. Mapa de calor de aversión para noticias relacionadas con navidad y año nuevo

Figura 45. Mapa de calor de la ira para noticias relacionadas con navidad y año nuevo

Al ver las diferencias en un diagrama de barra como el de la Figura 46, el miedo aún prevalece sobre las demás emociones seguido de la tristeza. En este aspecto, Ecuador mantiene un porcentaje similar entre miedo y tristeza que va del 41.6% y 37.5% respectivamente. Al comparar con otros países se visualiza que esta relación difiere como en el caso de Argentina con un 48.9% de miedo contra un 27.7% de tristeza o España con 45.8% de porcentaje de miedo y un 25% de tristeza. Estos dos países poseen una diferencia de alrededor del 20%, pero Ecuador apenas posee una diferencia del 4.1%; esta diferencia muestra que, a pesar de que otros países presentan mayor intensidad de "miedo" con respecto a Ecuador, este indica que su población se siente más angustiada con respecto a estos países. Al revisar otro país como en la Figura 47, Uruguay es uno de los pocos países que presenta la "tristeza" con mayor porcentaje de intensidad que el miedo, este se sobrepone con un porcentaje de 58.3% sobre el más cercano que es Ecuador con un 37.5%.

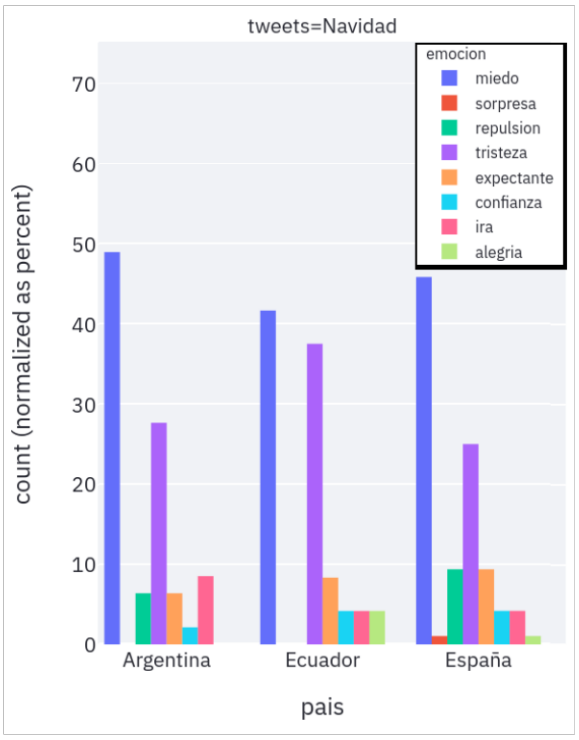


Figura 46. Comparación de emociones por países respecto a noticias de navidad y año nuevo (1)

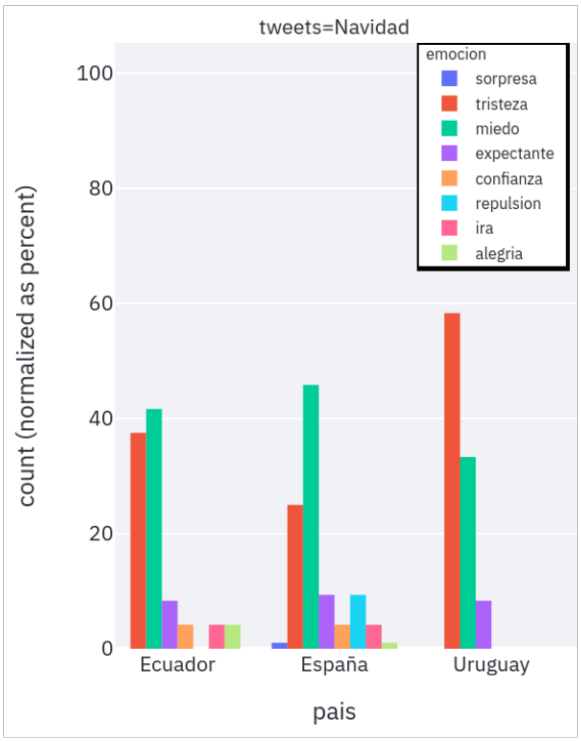


Figura 47. Comparación de emociones por países respecto a noticias de navidad y año nuevo (2)

4.2.3. Nuevo confinamiento en Ecuador

Los tweets de este evento van desde el 23 de abril del 2021 hasta el 16 de mayo del 2021 y cabe recalcar que estos tweets únicamente son de los fines de semana, los días en que se aplicó este nuevo confinamiento. Poco más de un año desde la declaración de pandemia por COVID19 la Figura 48 indica un cambio notorio en la intensidad de miedo con respecto al año anterior, aunque sigue prevaleciendo como la emoción más común a nivel general. El país que más destaca es Uruguay, es el único que presenta una intensidad fuerte de esta emoción. Por otro lado, al visualizar la Figura 49 que representa la tristeza, se tiene a países como

Nicaragua, Honduras, El Salvador, Bolivia y Paraguay, que poseen mayor intensidad de esta emoción. En esta emoción Ecuador destaca sobre sus vecinos.

Figura 48. Mapa de calor de miedo para noticias relacionadas con un nuevo confinamiento

Figura 49. Mapa de calor de tristeza para noticias relacionadas con un nuevo confinamiento

La Figura 50 representa la repulsión o aversión a nivel global, sin duda en este aspecto resalta Venezuela, Argentina y México, el resto de los países presentan un bajo porcentaje de esta emoción. El mapa en la Figura 51 muestra la intensidad de "ira" a nivel global, en esta emoción resalta Costa Rica, EEUU y España. Ecuador también posee una fuerte intensidad de esta emoción, pero a nivel de Sudamérica. Si se evalúa esto, Ecuador no presenta fuertes emociones negativas respecto a otros países, pero esto no indica necesariamente que el estado emocional del país sea mejor, también es necesario evaluar a nivel de porcentajes las demás emociones.

Al observar en el tiempo los mapas de calor para las tres noticias claramente presentan un cambio en su tonalidad. Si se toma por ejemplo la emoción de miedo se puede observar que a medida que pasa el tiempo la intensidad de este disminuye. A pesar de que las noticias seleccionadas pueden considerarse de interés general, dos de las tres son noticias negativas, pero aun así el "miedo" disminuye con respecto a la noticia anterior que se está evaluando. Para conocer si el estado emocional positivo de una población está aumentando, también se debe observar si emociones como la alegría o confianza van aumentando. Esto se lo puede observar más fácilmente con los gráficos de barras comparativos.

Figura 50. Mapa de calor de la aversión para noticias relacionas con un nuevo confinamiento

Figura 51. Mapa de calor de la ira para noticias relacionadas con un nuevo confinamiento

Al visualizar en porcentajes como lo indica la Figura 52, Ecuador presenta un porcentaje bastante elevado de tristeza alcanzando un 71.4%, si bien no es están alto como en otros países, esta emoción es básicamente es la única emoción negativa. Si se compara con otros países como México o Colombia esta emoción tiende a ser más del doble que en estos. Otro país a destacar es Argentina que similar a Ecuador posee una emoción con gran intensidad, pero en este caso es el miedo, alcanzando un 58.6%, a pesar de que la siguiente emoción en porcentaje es tristeza con un 13.8%, el miedo sin duda marca una gran brecha, esto se puede observar en la Figura 53.

Al visualizar los gráficos de barras para los tres eventos, es posible observar en una sola imagen la variación de las emociones en el transcurso del tiempo. De esta manera es posible conocer el desarrollo de una emoción a lo largo de los tres eventos. Si se desea conocer si el estado anímico de una población ha ido mejorando, se puede considerar si las emociones positivas han ido aumentando para un determinado país. Si se visualiza la Figura 53, se ha tomado de ejemplo a Ecuador, al evaluar la barra de alegría representada de color rosado, se puede distinguir que esta ha ido aumentando en el paso de los tres eventos. Por lo cual a pesar de que las emociones negativas aún tienen una fuerte presencia, la alegría del país ha ido aumentando, lo cual implica que su estado anímico ha ido mejorando.

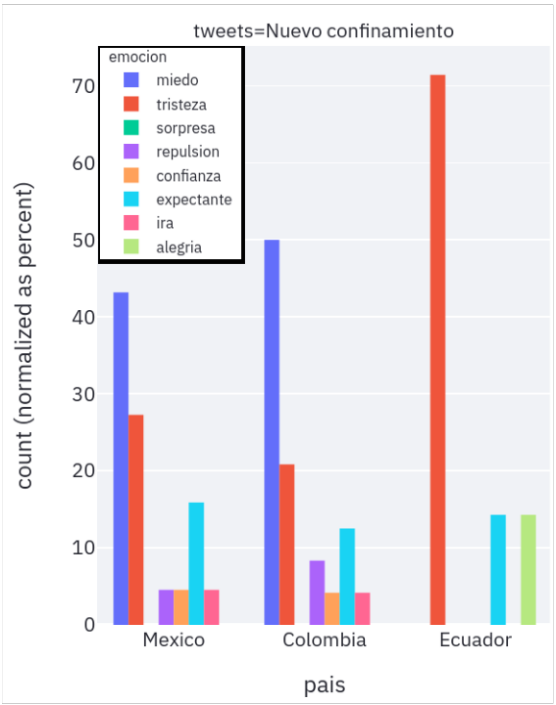


Figura 52. Comparación de emociones por países respecto a noticias acerca de nuevo confinamiento (1)

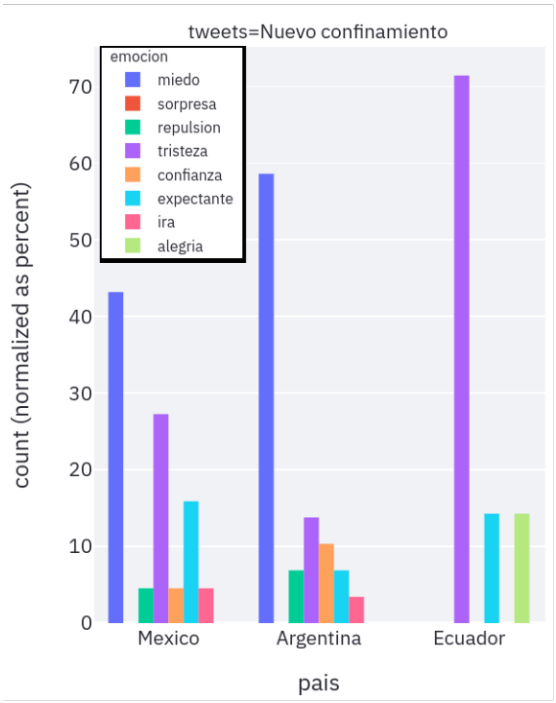


Figura 53. Comparación de emociones por países respecto a noticias acerca de nuevo confinamiento (2)

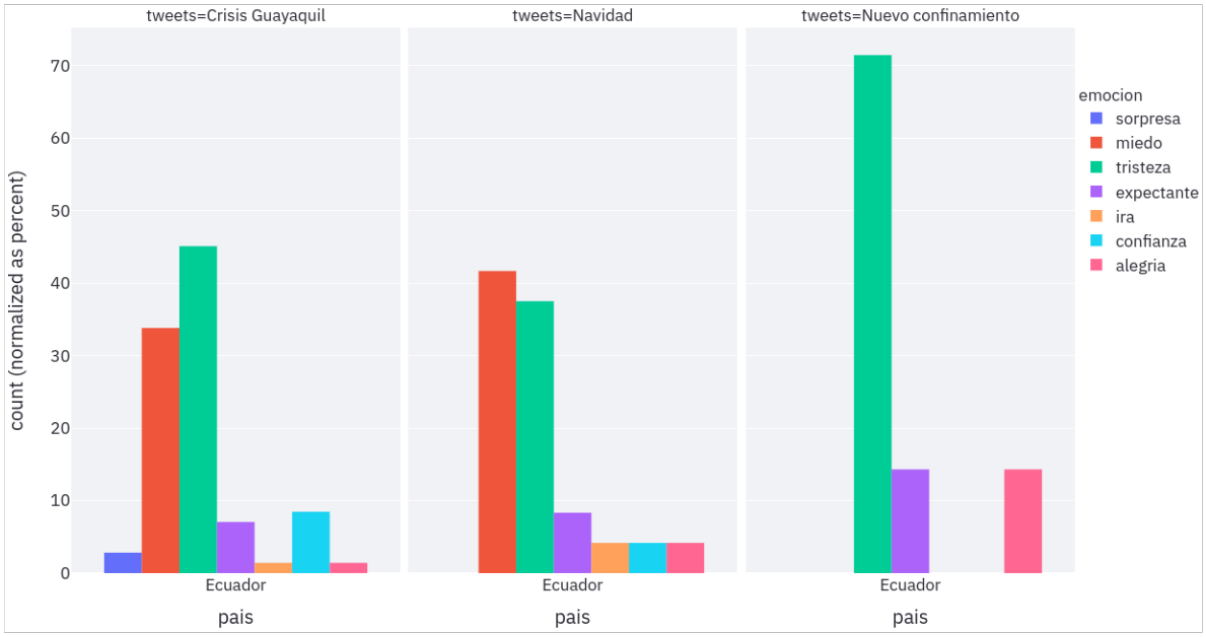


Figura 54. Comparación de emociones de Ecuador en los tres eventos analizados

5. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

- En el presente proyecto de titulación se desarrolló un sistema que permite monitorear emociones y sentimientos de una población al desatarse una pandemia. Esto con la finalidad de encontrar posibles efectos psicológicos. De esta manera se puede determinar cuáles son las afectaciones mentales más comunes.

- La adquisición de datos se la realizó desde Twitter por medio de raspado web o web scraping. No se usó la API de Twitter debido a su limitante, la cual solo permite descargar tweets de hasta dos semanas atrás. A pesar de que al realizar web scraping se obtiene menos información con respecto a la API, la información obtenida es suficiente para encontrar patrones que permitan observar algún problema psicológico.

- La desventaja de usar web scraping frente al uso de la API es el tiempo en la obtención de datos. La API es mucho más eficiente en la recolección de datos siendo hasta 13 veces más rápida comparada con la técnica de web scraping.

- Una manera de mejorar el tiempo en la obtención de registros puede ser con el proceso de hidratación de tweets. Este método consiste en poseer un archivo con los IDs (semillas) de tweets de interés. Los IDs de dicho archivo son usados para recuperar tweets con la información requerida (usuario, fecha, comentario, etc.). En este proceso es factible el uso de la API de Twitter pues al usarla de esta manera, no presenta restricciones de fecha. Sitios como https://zenodo.org y https://github.com/dh-miami/narratives_covid19 ofrecen semillas clasificadas por tema, idioma, fecha y país.

- El desarrollo de todo este aplicativo se realizó en el lenguaje de programación Python. Se lo diseño así, con la finalidad de usar software open source o en su defecto Creative Commos como es el caso de la librería Selenium utilizada para realizar web scraping.

- La cantidad total de registros a pesar de ser pequeña en relación a los estudios de Oladapo Oyebode (Oyebode, y otros, 2021) y Yinghui Haung (Huang, y otros, 2021) donde manejan una base superior a 10 mil registros, permite tener una visión general en el estado de la salud mental de una sociedad. Al comparar los resultados con la opinión del psicólogo entrevistado, se marca una similitud en las emociones obtenidas y las presentadas por el especialista. Cabe mencionar que este profesional trató en primera persona a pacientes con problemas psicológicos derivados del COVID19.

Según el especialista los problemas más comunes que llegaban a su consultorio como consecuencia de la pandemia eran problemas de depresión y ansiedad.

- La experiencia de un solo profesional puede no ser suficiente para asegurar que el modelo del proyecto funciona correctamente. Al tomar esto en consideración es oportuno destacar que, al preguntar sobre las emociones más comunes entre sus pacientes, su respuesta fue el miedo y la tristeza. Estas emociones concuerdan con las emociones que más se resaltan en el modelo. De esta manera, aunque no se puede tener una certeza sobre el estado psicológico y emocional de la población, el modelo cumple con dar una visión rápida y general de estos estados emocionales.

- Al comparar las emociones obtenidas y el criterio de un psicólogo se ve una relación entre las emociones de miedo y tristeza. El orden de importancia de cada una de estas emociones va acorde a la experiencia del psicólogo, sin embargo, cabe destacar que, a pesar de la aprobación del psicólogo respecto al resultado del análisis, hay emociones con mayor nivel de importancia que las analizadas. En la entrevista el psicólogo menciona que emociones como la ansiedad fueron mas recurrentes que la tristeza. Lamentablemente el diccionario base para este proyecto no toma en consideración esta emoción, por lo cual no es posible conocer de manera exacta el nivel de importancia que esta emoción tuvo en la población.

- El desarrollo de todo el proyecto se realizó de tal forma que, cada paso del proceso de minería de datos incluida la visualización de los dashboards pueda adaptarse con cambios menores hacia cualquier fichero de datos obtenido desde el proceso de web scraping. Los únicos cambios para realizar son las fechas y noticias que deberán ir acordes a la temática que se desee evaluar, estos cambios se los deberá hacer en la clase "*ClasificadorNoticia_2.py"* entre las líneas 55 y 65.

- Todos los datos presentes en el archivo CSV se obtuvieron desde Twitter, es decir provienen de textos de no más de 140 caracteres, lo cual puede presentar una limitante al analizar un texto. Para mejorar estos aspectos, se puede complementar la información de un tweet extrayendo también los comentarios que este pueda tener. A pesar de que en un inicio también se extrajeron emojis, estos no fueron tomados en consideración para su análisis, debido a la falta de una herramienta que permita su procesamiento.

- En este proyecto algunas subfases de la metodología CRISP-DM pueden ser combinadas con otras, con la finalidad de facilitar su entendimiento y el manejo de

datos. Las subfases *"verificar la calidad de los datos"* y *"seleccionar los datos"*, de las fases *"compresión de los datos" y "preparación de los datos"* respectivamente, pueden acoplarse en una sola tarea. De esta manera podría evitarse una iteración completa de la metodología.

- Al aplicar la metodología CRISP-DM en este proyecto, se presentaron retrasos en la fase de "*preparación de los datos*". Las subfases *"limpiar los datos"* y *"formatear los datos"* pueden ser integradas en un mismo proceso. De esta manera sería posible acortar el tiempo de procesamiento del programa. Cabe recalcar que estos cambios pueden ser posibles solo para los fines de este proyecto.

5.2. Recomendaciones

- Se recomienda realizar una segunda obtención de datos con los comentarios que posean mayor cantidad de retweets y mayor cantidad de respuestas. De este modo se puede aumentar la cantidad de tweets y los comentarios estarán más enfocados al sentir de la población respecto a la problemática en estudio.

- La extracción de toda la información se lo ha realizado hasta mayo del 2021, es decir casi al inicio de las campañas de vacunación. Una forma de ampliar este proyecto es evaluar las emociones de la población durante el proceso de vacunación. De esta manera es posible observar si existe una reducción en las emociones negativas. Para esto sería necesario evaluar las campañas de vacunación en cada país y adecuarlo al código del programa.

- Una manera de extender la cantidad de registros puede ser aplicando web scraping a otras redes sociales como Facebook o YouTube. El inconveniente con esta técnica es la creación de diversos scripts, uno para cada sitio web donde se desea realizar web scraping.

- Se recomienda aumentar la cantidad de tweets obtenidos en las diferentes fechas de estudio. De esta manera se podría realizar un filtrado más específico a nivel geográfico. Con una mayor cantidad de tweets es posible clasificarlos por provincias o ciudades, en lugar de países.

- Una manera de ampliar este proyecto es la obtención de datos después de los procesos de vacunación. Esto requerirá un estudio más detallado pues se necesitará

conocer las fechas de comienzo de los procesos de vacunación en cada país, en conjunto con la opinión ciudadana referente a los procesos de vacunación.

- Los dashborads empleados podrían ampliarse y modificarse con la finalidad de indicar datos estadísticos y así construir una línea de tiempo. De esta manera se podría visualizar el cambio de emociones en los países.

6. GLOSARIO DE TÉRMINOS

Application Programming Interfaces (API): Es un conjunto de protocolos que se usan para desarrollar e integrar el software de las aplicaciones, permitiendo así la comunicación entre dos aplicaciones.

Bot: Es un programa informático que realiza tareas de manera repetitiva.

Ciencias psicológicas: Hace referencia a todas aquellas áreas de la psicología y ciencias referentes a esta.

Creative commos: Es un conjunto de licencias que permiten compartir creaciones intelectuales especificando las condiciones de uso.

Comma Separated Values (CSV): Es un formato para representar datos en forma de tabla.

Dashboard: Es una herramienta que personalizable que facilita la visualización de datos.

Float: Es un tipo de dato utilizado para representar números decimales.

GeoJSON: Es un tipo de formato utilizado para representar elementos geográficos.

Librería: Es un conjunto de archivos ya sea código o dato, que se utilizan para el desarrollo de software

Minería de datos: Es un conjunto de técnicas y tecnologías que permiten el análisis de grandes cantidades de datos.

Open source: Es el término que se usa para describir el software que se distribuye mediante licencia de código abierto, es decir permite al usuario modificar y distribuir dicho software.

Procesamiento de lenguaje natural (NPL): Es una rama de inteligencia artificial que permite la interpretación de textos y datos mediante aprendizaje automático.

Raspado web (web scraping): Es una técnica utilizada para la extracción de información de sitios web.

Retweet: Consiste en volver a publicar un tweet ya sea propio o de otro usuario. Es posible agregar un comentario o una cita a dicho tweet.

Script: Son fragmentos de código que se almacenan en un archivo de texto para realizar tareas por medio de intérprete

String: Es un tipo de dato utilizado para representar texto

Tweet: Es un mensaje de texto, puede contener un máximo de 140 caracteres entre letras, signos y emojis.

UTC básico: Es un sistema para expresar una hora el formato básico tiene el siguiente formato: YYYYMMDDTHHMMSS (20220223T161205)

UTC extendido: Es un sistema para expresar una hora el formato extendido tiene el siguiente formato: YYYY-MM-DDTHH:MM:SS (2002-02-23T16:12:02)

7. REFERENCIAS

Bailey, M. (3 de Noviembre de 2020). *pypi.org*. Obtenido de PyPI: https://pypi.org/project/NRCLex/

Baviera, T. (2016). Técnicas para el análisis del sentimiento en Twitter: Aprendizaje. *Dígitos, 3(1)*, 33-50. orcid.org/0000-0002-2331-6628.

Brik, D. (23 de Abril de 2021). Parte de Ecuador vuelve al confinamiento por la covid, ahora en fin de semana. *Swissinfo*, págs. https://www.swissinfo.ch/spa/ecuador-coronavirus--previsión-_parte-de-ecuador-vuelve-al-confinamiento-por-la-covid--ahora-en-fin-de-semana/46561562.

Contributor, T. (Septiembre de 2005). *What is stop word*. Obtenido de WhatIs.com: https://whatis.techtarget.com/definition/stop-word

Conway, M., Hu, M., & Chapman, W. W. (2019). Recent advances in using natural language processing to address public health research questions using social media and consumer generated data. *Yearb Med Inform, 28(01)*, 208-217. dx.doi.org/10.1055/s-0039-1677918.

Cutting, D., Kupiec, J., & Pedersen. (1992). A practical part-of-speech tagger. *Third conference on applied natural language processing*, (págs. 133-140. doi.org/10.3115/974499.974523). Palo ALto.

Fortner, T. (11 de Diciembre de 2021). *Selenium*. Obtenido de selenium.dev: https://www.selenium.dev/documentation/

Gallardo, J. (24 de Enero de 2010). *oldemarrodriguez.* Obtenido de Metodología para el desarrollo de proyectos en Minería de Datos: http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037

Garcia, M. (2021). *How to Make a Twitter Bot in Python With Tweepy*. Obtenido de Real Python: https://realpython.com/twitter-bot-python-tweepy/#:~:text=Tweepy%20is%20an%20open%20source,Data%20encoding%20and%20decoding

Grimes, J. (3 de Diciembre de 2021). *Web Scraping*. Obtenido de BestProxy Reviews: https://www.bestproxyreviews.com/twitter-scraper/

Huang, Y., Liu, H., Zhang, L., Li, S., Wang, W., Ren, Z., . . . Ma, X. (2021). The psychological and behavioral patterns of online psychological help-seekers before and during COVID-19 Pandemic: A text mining-based longitudinal ecological study. *International Journal of Environmental Research and Public Health, 18(21)*, 11525. doi.org/10.3390/ijerph182111525.

Huarcaya, J. (2020). Consideraciones sobre la salud mental en la pandemia de COVID-19. *Rev Perú Med Exp Salud Pública, 37(2)*, 327-334. https://doi.org/10.17843/rpmesp.2020.372.5419.

Jurafsky, D., & Martin, J. (2021). Dependency Parsing. En D. Jurafsky, & J. Martin, *Speech and Language Processing* (págs. Chaper 14, 1-23). New Jersey: Pearson international edition.

Klein, S., & Simmons, R. (1963). A Computational Approach to Grammatical Coding of English Words. *ACM*, 334-347. doi.org/10.1145/321172.321180.

León, R. d. (23 de Mayo de 2020). *Category: Un conjunto de datos de Twitter para la narrativa digital*. Obtenido de Narrativas digitales de la COVID19: https://covid.dh.miami.edu/es/category/objetos-de-estudio/datos/

Li, S., Wang, Y., Xue, J., Zhao, N., & Zhu, T. (2020). The impact of COVID-19 epidemic declaration on psychological consequences: a study on active Weibo users. *Int. J. Environ. Res. Public Health, 17(6)*, doi.org/10.3390/ijerph17062032.

Liu, B. (2011). Web Data Mining. En B. Liu, *Web Data Mining* (págs. 459-514). Chicago: Springer.

Loria, S. (2020). *Simplified Text Processing*. Obtenido de TextBlob: https://textblob.readthedocs.io/en/dev/

Lozano, D. (2 de Abril de 2020). Cientos de cadáveres sobre el asfalto de las calles de guayaquil: "Estamos llenos de muertos por coronavirus". *El Mundo*, pág. https://www.elmundo.es/internacional/2020/04/01/5e84d472fdddffd4618b45c6.html.

Mohammad, S., & Turney, P. (2010). Emotions Evoked by Common Words and Phrases Using Mechanical Turk to Create an Emotion Lexicon. *HLT_NAACL 2010.*

Mohammad, S., & Turney, P. D. (2013). Crowdsourcing a Word-Emotion Association Lexicon. *CoRR abs 1308.6297*, 436-465. arxiv.org/abs/1308.6297.

Mottl, D. (11 de 2019). *Description GetOldTweets3*. Obtenido de Pypi: https://pypi.org/project/GetOldTweets3/

Muñoz, M., Recéndez, M., & Ñánez, A. (2021). Representaciones sociales de covid y ansiedad a través del microblogging. Análisis de contenido de tweets. *Encuentros: Revista de Ciencias Humanas, Teoría Social y Pensamiento Crítico (14)*, 176-188, doi.org/10.5281/zenodo.5205193.

Murillo, D., & Saavedra, D. (2017). Web Scraping de los Perfiles y Publicaciones de una Afiliación en Google Scholar utilizando Aplicaciones Web e implementando un Algoritmo en R. *Congreso Internacional AmITIC 2017*, (págs. 8-15. oai:revistas.utp.ac.pa:article/1465). Popayán.

Nekane Balluerka, J. G. (2020). *Las consecuencias psicológicas de la COVID-19 y el confinamiento.* Bilbao: Servicios de Publicaciones de la Universidad del País Vasco.

Osman, M. (3 de Enero de 2021). *Blog: Estadísticas Impresionantes de Twitter y Datos Importantes Sobre Nuestra Red Favorita*. Obtenido de Kinsta: https://kinsta.com/es/blog/estadisticas-twitter/

Oyebode, O., Ndulue, C., Adib, A., Mulchandani, D., Suruliraj, B., Orji, F. A., . . . Orji, R. (2021). Health, Psychosocial, and Social Issues Emanating From the COVID-19 Pandemic Based on Social Media Comments: Text Mining and Thematic Analysis Approach. *JMIR medical informatics, 9(4)*, doi.org/10.2196/22734.

Ozamiz Etxebarria, N., Dosil Santamaria, M., Picaza Gorrochategui, M., & Idoiaga Mondragon, N. (2020). Niveles de estrés, ansiedad y depresión en la primera fase del brote del COVID-19 en una muestra recogida en el norte de España. *Cadernos de Saúde Pública (36)4*, doi.org/10.1590/0102-311X00054020.

pandas, A. (2018). *About pandas*. Obtenido de Pandas: https://pandas.pydata.org/about/

Plaza, M., & Ureña, A. (2019). Improved emotion recognition in Spanish social media through incorporation of lexical knowledge. *Future Generation Computer Systems (110)*, 1000-1008. doi.org/10.1016/j.future.2019.09.034.

PRC, N. H. (5 de Abril de 2021). *National Health Commission of the People's Republic of China.* Obtenido de Distribution of COVID-19 Outbreak;: http://2019ncov.chinacdc.cn/2019-nCoV/

psiquiatría, A. a. (2014). *Guía de consulta de los criterios diagnóstios del DSM-5.* Washington: American psychiatric publishing.

Python. (25 de 1 de 2021). *Python*. Obtenido de https://docs.python.org/3.9/tutorial/index.html

Qi, P., Zhang, Y., Zhang, Y., & Bolton, J. (26 de Enero de 2021). *Stanza – A Python NLP Package for Many Human Languages*. Obtenido de Stanza: https://stanfordnlp.github.io/stanza/

Qiu, J., Shen, B., Zhao, X., Wang, Z., Xie1, B., & Xu, Y. (2020). A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: implications and policy recommendations. *General Psychiatry (33)*, doi: 10.1136/gpsych-2020-100213.

Singh, P., Singh, S., & Sohal, M. (2020). Psychological fear and anxiety caused by COVID-19: Insights from Twitter analytics. *Asian Journal of Psychiatry (54)*, doi.org/10.1016/j.ajp.2020.102280.

Sommerville, I. (2005). *Ingeniería del software.* Madrid: Pearson educación.

Streamlit. (2020). *Welcome to Streamlit*. Obtenido de Streamlit: https://www.streamlit.io

Twitter. (2021). *Twitter*. Obtenido de https://help.twitter.com/es/rules-and-policies/twitter-law-enforcement-support#1

Vargiu, E. (2013). Exploiting web scraping in a collaborative filtering- based approach to web advertising. *Artificial Intelligence Research, 2 (1)*, (44 -54) dx.doi.org/ 10.5430/ air.v2n1p44.

Wang, C., & Pan, R. (2020). Immediate Psychological Responses and Associated. *Environmental Research and Public Health, 17(1729)*, doi.org/10.3390/ijerph17051729.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (págs. 29-39). Ulm. Obtenido de http://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Xiang, J., & James, G. (2013). Epidemic Outbreak and Spread Detection System Based on Twitter Data. *Health Information Science*, (págs. 152-163, doi:10.1007/978-3-642-29361-0_19). New York.

Yao, H., Chen, J.-H., & Xu, Y.-F. (2020). Patients with mental health disorders in the COVID-19 epidemic. *The Lancet Psychiatry, 7(4)*, e21. doi:10.1016/s2215-0366(20)30090-0 .

Zhao, B. (2017). Web Scraping. En C. L. Laurie A. Schintler, *Encyclopedia of Big Data* (págs. 1-3, doi:10.1007/978-3-319-32001-4_483-1). Seattle: Springer International Publishing.

Zhao, B. (2017). Web Scraping. In: Schintler L., McNeely C. (eds) Encyclopedia of Big Data. *Springer, Cham*, doi.org/10.1007/978-3-319-32001-4_483-1.

8. ANEXOS

7.1. Anexo 1: descripción de problemas psicológicos causados por COVID-19 (Entrevista)

Enlace al documento de la transcripción de la entrevista realizado al especialista de la salud:

https://epnecuador-my.sharepoint.com/:f:/g/personal/julio_rosero_epn_edu_ec/Ep7AI3gceQ9Br_fWKEKoBboBF3wumM4eUWQppNUVfpdlXA?e=4SLK0u

7.2. Anexo 2: datos obtenidos desde Twitter

Enlace a la carpeta que contiene el archivo de datos original y el archivo de datos final:

https://epnecuador-my.sharepoint.com/:f:/g/personal/julio_rosero_epn_edu_ec/EnDMiIMawsVJrAjATgLA7isB8xufmJTrwj79eRAxvAqs9g?e=OrPgvj

7.3. Anexo 3: diccionario de emociones NRC

Enlace al diccionario NRC traducido al español:

https://epnecuador-my.sharepoint.com/:f:/g/personal/julio_rosero_epn_edu_ec/Ekpd4xQVBw1Dmr2ZjPx5OIcBjwyP-dB8NX9TSW6bOxTing?e=KPi90Q

7.4. Anexo 4: diccionario de intensidad de emociones Emolex

Enlace al diccionario de la intensidad de emociones traducido al español:

https://epnecuador-my.sharepoint.com/:f:/g/personal/julio_rosero_epn_edu_ec/EnuS102tHmZLi_yEhj3jQ80Bj26H3CLwglh3ZeuqWfMWow?e=PDDFsU

7.5. Anexo 5: archivo con datos geográficos para la representación de los mapas en formato GeoJSON

Enlace al archivo GeoJSON con el mapa de los países en estudio:

https://epnecuador-my.sharepoint.com/:f:/g/personal/julio_rosero_epn_edu_ec/EpNPW3PX511FvNWKQPnEWTkBroptvLmfABqZvF-krvDBVw?e=bQEm3R

7.6. Anexo 6: opinión de psicólogo respecto a los resultados (Entrevista)

Enlace al documento de la transcripción de la entrevista realizado al especialista de la salud:

https://epnecuador-my.sharepoint.com/:f:/g/personal/julio_rosero_epn_edu_ec/EhaqjjA4KSBHrpcunyAakoMBkEVeIovNNvJtg3okZv19yQ?e=mR5StD