UTMACH

FACULTAD DE INGENIERÍA CIVIL

CARRERA DE INGENIERÍA DE SISTEMAS

MINERÍA DE DATOS Y ANÁLISIS DE SENTIMIENTOS EN REDES SOCIALES, CASO DE ESTUDIO: PERCEPCIÓN DEL COVID-19 EN EL ECUADOR.

SIMBAÑA CRUZ HENRY VINICIO
INGENIERO DE SISTEMAS

MACHALA
2021

UTMACH

FACULTAD DE INGENIERÍA CIVIL

CARRERA DE INGENIERÍA DE SISTEMAS

MINERÍA DE DATOS Y ANÁLISIS DE SENTIMIENTOS EN REDES SOCIALES, CASO DE ESTUDIO: PERCEPCIÓN DEL COVID-19 EN EL ECUADOR.

SIMBAÑA CRUZ HENRY VINICIO
INGENIERO DE SISTEMAS

MACHALA
2021

UTMACH

FACULTAD DE INGENIERÍA CIVIL

CARRERA DE INGENIERÍA DE SISTEMAS

TRABAJO TITULACIÓN
PROPUESTAS TECNOLÓGICAS

MINERÍA DE DATOS Y ANÁLISIS DE SENTIMIENTOS EN REDES SOCIALES, CASO DE ESTUDIO: PERCEPCIÓN DEL COVID-19 EN EL ECUADOR.

SIMBAÑA CRUZ HENRY VINICIO
INGENIERO DE SISTEMAS

MAZÓN OLIVO BERTHA EUGENIA

MACHALA, 27 DE SEPTIEMBRE DE 2021

MACHALA
2021

MINERÍA DE DATOS Y ANÁLISIS DE SENTIMIENTOS EN REDES SOCIALES, CASO DE ESTUDIO: PERCEPCIÓN DEL COVID-19 EN EL ECUADOR

INFORME DE ORIGINALIDAD

6%	5%	3%	3%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	rios.tecnm.mx Fuente de Internet	1%
2	www.redalyc.org Fuente de Internet	1%
3	repositorio.ulima.edu.pe Fuente de Internet	<1%
4	ijercse.com Fuente de Internet	<1%
5	repositorio.utmachala.edu.ec Fuente de Internet	<1%
6	María Cecilia Johnson, Lorena Saletti-Cuesta, Natalia Tumas. "Emociones, preocupaciones y reflexiones frente a la pandemia del COVID-19 en Argentina", Ciência & Saúde Coletiva, 2020 Publicación	<1%
7	docplayer.es Fuente de Internet	<1%

CLÁUSULA DE CESIÓN DE DERECHO DE PUBLICACIÓN EN EL REPOSITORIO DIGITAL INSTITUCIONAL

El que suscribe, SIMBAÑA CRUZ HENRY VINICIO, en calidad de autor del siguiente trabajo escrito titulado MINERÍA DE DATOS Y ANÁLISIS DE SENTIMIENTOS EN REDES SOCIALES, CASO DE ESTUDIO: PERCEPCIÓN DEL COVID-19 EN EL ECUADOR., otorga a la Universidad Técnica de Machala, de forma gratuita y no exclusiva, los derechos de reproducción, distribución y comunicación pública de la obra, que constituye un trabajo de autoría propia, sobre la cual tiene potestad para otorgar los derechos contenidos en esta licencia.

El autor declara que el contenido que se publicará es de carácter académico y se enmarca en las dispociones definidas por la Universidad Técnica de Machala.

Se autoriza a transformar la obra, únicamente cuando sea necesario, y a realizar las adaptaciones pertinentes para permitir su preservación, distribución y publicación en el Repositorio Digital Institucional de la Universidad Técnica de Machala.

El autor como garante de la autoría de la obra y en relación a la misma, declara que la universidad se encuentra libre de todo tipo de responsabilidad sobre el contenido de la obra y que asume la responsabilidad frente a cualquier reclamo o demanda por parte de terceros de manera exclusiva.

Aceptando esta licencia, se cede a la Universidad Técnica de Machala el derecho exclusivo de archivar, reproducir, convertir, comunicar y/o distribuir la obra mundialmente en formato electrónico y digital a través de su Repositorio Digital Institucional, siempre y cuando no se lo haga para obtener beneficio económico.

Machala, 27 de septiembre de 2021

Henry Simbaña Cruz
HENRY

SIMBAÑA CRUZ HENRY VINICIO
0704265099

DEDICATORIA

Conseguir este peldaño de mi formación académica es bastante gratificante, razón por la cual deseo dedicar el presente trabajo a un ser supremo y sobrenatural que es Dios, el cual fue mi refugio.

A mis padres por apoyarme e impulsarme a superar mis expectativas propias además de su esfuerzo que se ve reflejado en lo que soy en la actualidad, todos los logros que he cosechado y éste, el más relevante en la actualidad.

A mis demás parientes, por sus palabras de aliento, por su apoyo incondicional, que Dios con su infinita grandeza los bendiga constantemente.

A mis amigos y compañeros de clase que constantemente me han colaborado en lo cual he necesitado su ayuda además de sus tips que en su tiempo no los tome presente, empero después de haber enfrentado los inconvenientes me han servido de mucho.

Además, dedico este trabajo a mis docentes por sus consejos motivadores y sus enseñanzas dadas, que influyen mucho para desarrollar cada una de las ocupaciones que se muestran en mi vida académica.

Henry Vinicio Simbaña Cruz

AGRADECIMIENTO

A mis padres por su apoyo incondicional durante todos los años. Gracias, a su cariño todo ha sido más sencilla, a mi tutora Ing. Bertha Eugenia Mazón Olivo por su gran ayuda y colaboración en cada momento de consulta y soporte en este trabajo de titulación y las clases que me ha impartido durante la carrera académica, a Dios y al mundo por ayudarme a mantenerme firme y no decaer a pesar de las adversidades presentadas durante este gran esfuerzo y dedicación.

A la Universidad Técnica de Machala, Facultad de Ingeniería Civil, Carrera de Ingeniería de Sistemas por aceptarme en su establecimiento y ayudarme en mi desarrollo profesional.

RESUMEN

El análisis de sentimientos (SA: Sentiment Analysis) es de gran importancia para determinar las actitudes, emociones, opiniones y comentarios de las personas. Hay varios casos de aplicación de SA; por ejemplo, en el caso de las empresas, puede ser útil para determinar las preferencias de sus clientes a un determinado producto o servicio; en el sector gubernamental SA puede usarse para estudiar la satisfacción de los ciudadanos respecto a su gestión; otro caso de aplicación de SA es en la industria del ocio y espectáculo como el cine, y puede servir para analizar las opiniones del público. El análisis de sentimientos comúnmente se realiza mediante comentarios de redes sociales como: Facebook, Instagram, Twitter, YouTube, etc. El propósito de este trabajo consistió en el análisis de comentarios recuperados de la red social Twitter en el contexto de COVID-19 en Ecuador, desde abril del 2020 hasta julio del 2021. La metodología aplicada para desarrollar este proyecto se basó en las mejores prácticas de Cross Industry Standard Process for Data Mining (CRISP-DM) de IBM y Team Data Science Process (TDSP) de Microsoft; las fases y actividades desarrolladas son: 1) Comprensión empresarial, que consiste en el establecimiento de herramientas y repositorios de comentarios a obtener de Twitter. 2) Adquisición y comprensión de datos, que se refiere al preprocesamiento y limpieza de los datos. 4) El modelado, implica la exploración, procesamiento, análisis y visualización de los datos. 5) La evaluación, consiste en la comparación de modelos mediante el cálculo y análisis de las métricas. Y finalmente, 6) El despliegue que trata de tomar la decisión de elegir el método más adecuado de análisis de sentimientos. Los primeros resultados obtenidos fueron de las palabras más frecuentes; se analizó por segmentos de tiempo: 1) En el periodo comprendido entre abril y julio del 2020, las palabras más frecuentes fueron “COVID” y “miedo”, dando a entender que la población estaba entrando en una etapa de pánico. 2) Luego en el periodo entre agosto y noviembre del 2020, se destacaron las palabras: “Pandemia”, “Casos”, “Funerarias”, “desbordadas”, “putrefactos”, lo que denota que se estaba viviendo momentos muy difíciles de pérdidas humanas y una fuerte incertidumbre en el país. 3) Posteriormente, el periodo analizado entre diciembre del 2020 y marzo del 2021, las palabras más relevantes fueron: “Ecuador”, “Suma”, “Casos”, al comparar con las estadísticas en ese período, efectivamente se elevaron los casos de COVID-19 en el país. 4) Finalmente en el periodo de abril a julio del 2021, las palabras que más mencionadas fueron: “vacuna”, “COVID”, lo que coincide con el proceso de vacunación que se llevó a cabo en este periodo. Los métodos de análisis de sentimientos utilizados fueron: Valence Aware Dictionary and sEntiment Reasoner (VADER), TEXTBLOB Y AFFIN, desarrollados y ejecutados en el lenguaje Python. Luego de evaluar los métodos, VADER

superó a los demás, con una exactitud del 65,67%, una precisión de 68,93% y, una sensibilidad de 66,44%. La polaridad de los sentimientos durante el periodo analizado, según AFFIN mayoritariamente fueron Negativos (-); según VADER fueron Positivos (+), sin embargo, en septiembre del 2020 y entre marzo y abril del 2021 hubo más sentimientos negativos (-); y, para TEXTBLOB, los sentimientos fueron mayoritariamente positivos (+). También se realizó modelos de predicción mediante el lenguaje R, aplicando series temporales como: Autoregressive Integrated Moving Average (ARIMA), Seasonal-Trend decomposition using LOESS (STL) y HOLT-WINTERS; según estos modelos, hay una tendencia a disminuir las intervenciones de los usuarios en las redes sociales acerca del COVID-19 en vista que ya no se lo menciona con mucha frecuencia en muchos tweets.

PALABRAS CLAVES:

Minería de Datos, Análisis de Sentimientos, Red Social Twitter, Covid-19, Series Temporales

ABSTRACT

Sentiment Analysis (SA: Sentiment Analysis) is of great importance in determining people's attitudes, emotions, opinions and comments. There are several cases of application of SA; For example, in the case of companies, it can be useful to determine the preferences of their customers for a certain product or service; In the government sector SA can be used to study the satisfaction of citizens with respect to its management; Another case of application of SA is in the entertainment and entertainment industry such as the cinema, and it can be used to analyze the opinions of the public. Sentiment analysis is commonly done through comments from social networks such as: Facebook, Instagram, Twitter, YouTube, etc. The purpose of this work consisted in the analysis of comments retrieved from the social network Twitter in the context of COVID-19 in Ecuador, from April 2020 to July 2021. The methodology applied to develop this project was based on the best practices of Cross Industry Standard Process for Data Mining (CRISP-DM) from IBM and Team Data Science Process (TDSP) from Microsoft; The phases and activities developed are: 1) Business understanding, which consists of establishing tools and repositories of comments to be obtained from Twitter. 2) Data acquisition and understanding, which refers to the pre-processing and cleaning of data. 4) Modeling involves the exploration, processing, analysis and visualization of data. 5) The evaluation consists of the comparison of models through the calculation and analysis of the metrics. And finally, 6) The deployment that tries to make the decision to choose the most appropriate method of sentiment analysis. The first results obtained were for the most frequent words; It was analyzed by time segments: 1) In the period between April and July 2020, the most frequent words were “COVID” and “fear”, implying that the population was entering a stage of panic. 2) Then in the period between August and November 2020, the words: "Pandemic", "Cases", "Funeral homes", "overflowing", "rotten" stood out, which denotes that very difficult moments of losses were being experienced human rights and a strong uncertainty in the country. 3) Subsequently, the period analyzed between December 2020 and March 2021, the most relevant words were: "Ecuador", "Sum", "Cases", when comparing with the statistics in that period, COVID cases actually rose -19 in the country. 4) Finally, in the period from April to July 2021, the most frequently mentioned words were: “vaccine”, “COVID”, which coincides with the vaccination process that was carried out in this period. The sentiment analysis methods used were: Valence Aware Dictionary and sEntiment Reasoner (VADER), TEXTBLOB and AFFIN, developed and executed in the Python language. After evaluating the methods, VADER outperformed the others, with an accuracy of 65.67%, a precision of 68.93%, and a sensitivity of 66.44%. The polarity of feelings during the analyzed period, according to

AFFIN, were mostly Negative (-); According to VADER they were Positive (+), however, in September 2020 and between March and April 2021 there were more negative feelings (-); and, for TEXTBLOB, sentiments were mostly positive (+). Prediction models were also made using R language, applying time series such as: Autoregressive Integrated Moving Average (ARIMA), Seasonal-Trend decomposition using LOESS (STL) and HOLT-WINTERS; According to these models, there is a tendency to decrease comments on social networks, because COVID-19 is also decreasing.

KEYWORDS:

Data Mining, Sentiment Analysis, Twitter Social Network, Covid-19, Time Series

ÍNDICE GENERAL

DEDICATORIA 4

AGRADECIMIENTO 5

RESUMEN 6

ABSTRACT 8

GLOSARIO 14

INTRODUCCIÓN 16

1 CAPÍTULO I. DIAGNÓSTICO DE NECESIDADES Y REQUERIMIENTOS 18

1.1 ÁMBITO DE APLICACIÓN: CONTEXTUALIZACIÓN Y DESCRIPCIÓN DEL PROBLEMA OBJETO DE INTERVENCIÓN 18

1.2 ESTABLECIMIENTO DE REQUERIMIENTOS 18

1.3 JUSTIFICACIÓN DEL REQUERIMIENTO A SATISFACER 20

2 CAPÍTULO II. DESARROLLO DEL PROYECTO 21

2.1 DEFINICIÓN DEL PROTOTIPO TECNOLÓGICO 21

2.2 FUNDAMENTACIÓN TEÓRICA DEL PROTOTIPO 21

2.2.1 HISTORIA COVID-19 21

2.2.2 EVOLUCIÓN DEL COVID-19 EN ECUADOR 22

2.2.3 MINERÍA DE TEXTO 23

2.2.4 PROCESAMIENTO DE LENGUAJE NATURAL 27

2.2.5 ANÁLISIS DE SENTIMIENTOS 30

2.2.6 METODOLOGÍA CRISP-DM (CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING) 36

2.2.7 METODOLOGÍA TEAM DATA SCIENCE PROCESS 36

2.3 OBJETIVOS DEL PROTOTIPO 37

2.3.1 OBJETIVO GENERAL 37

2.3.2 OBJETIVOS ESPECÍFICOS 37

2.4 DISEÑO DEL PROTOTIPO. 37

2.4.1 COMPRENSIÓN DEL NEGOCIO 38

2.4.2 ESTUDIO Y COMPRENSIÓN DE LOS DATOS 38

2.4.3 MODELADO 39

2.4.4 EVALUACIÓN 39

2.4.5 DESPLIEGUE 39

2.5 ENSAMBLE Y EJECUCIÓN DEL PROTOTIPO 39

2.5.1 CONOCIMIENTO DEL NEGOCIO 39

2.5.2 ESTUDIO Y COMPRENSIÓN DE LOS DATOS 42

2.5.3 MODELADO 44

3 EVALUACIÓN DEL PROTOTIPO 56

3.1 PLAN DE EVALUACIÓN 56

3.1.1 MATRIZ DE CONFUSIÓN 56

3.2 RESULTADOS DE LA EVALUACIÓN 58

3.2.1 MÉTRICAS DE EVALUACIÓN 60

3.3 CONCLUSIONES 64

3.4 RECOMENDACIONES 64

BIBLIOGRAFIA 65

ÍNDICE DE FIGURAS

Figura 1:Arquitectura de Propuesta Tecnológica.....19

Figura 2: Comparativa tasa de letalidad por COVID-19 mundial - Ecuador del 13 al 31 de marzo del 2020..... 21

Figura 3: Diagrama esquemático del enfoque propuesto..... 28

Figura 4: Evaluación de modelo propuesto en el articulo..... 29

Figura 5: Resumen de cantidad de síntomas de enfermedades populares..... 30

Figura 6: Técnicas de Clasificación de Sentimientos..... 31

Figura 7: Arquitectura Metodología Propuesta..... 36

Figura 8: Tweets adquiridos y unificados..... 40

Figura 9: Nube de palabras de abril a julio del 2020..... 42

Figura 10: Nube de Palabras de agosto a noviembre del 2020..... 42

Figura 11: Nube de Palabras de diciembre 2020 a marzo del 2021..... 43

Figura 12: Nube de Palabras de abril 2021 a julio del 2021..... 44

Figura 13: Nube de Palabras de abril 2020 a julio 2021..... 44

Figura 14: Clasificación de Sentimientos de cada Método..... 45

Figura 15: Cantidad de Tweets por Método de Análisis de Sentimientos..... 46

Figura 16: Serie Temporal de cantidad de Comentarios Método AFFIN..... 47

Figura 17: Serie Temporal de cantidad de Comentarios Método VADER..... 47

Figura 18: Serie Temporal de cantidad de Comentarios Método TEXTBLOB..... 48

Figura 19: Serie Temporal de los sentimientos en tweets..... 49

Figura 20: Predicción de cantidad de tweets Positivos modelo Holt-Winters..... 50

Figura 21: Predicción de cantidad de tweets Negativos modelo Holt -Winters..... 51

Figura 22: Predicción de cantidad de tweets Positivos modelo STL..... 51

Figura 23: Predicción de cantidad de tweets Negativos modelo STL..... 52

Figura 24: Clasificación por Método aplicado a la Muestra..... 52

Figura 25: Matriz de Confusión de Método AFFIN..... 55

Figura 26: Matriz de Confusión de Método VADER..... 56

Figura 27: Matriz de Confusión de Método TEXTBLOB..... 56

Figura 28: Métricas de Evaluación de Métodos de Análisis de Sentimientos..... 60

ÍNDICE DE CODIGOS

Código 1: Script de Normalización de los datos 40

Código 2: Script de limpieza de texto_ 41

Código 3: Script de Generación de Matriz de Confusión_ 55

Código 4: Script de Arreglo para Matriz de Confusión_ 55

Código 5: Script de Cálculo de Precisión_ 57

Código 6: Script de Cálculo de Sensibilidad_ 58

Código 7: Script de Cálculo de Exactitud_ 59

Código 8: Script de Cálculo de Medida F 59

ÍNDICE DE TABLAS

Tabla 1: Características de los Tweets 37

Tabla 2: Atributos claves de Twitter 38

Tabla 3: Representación de Valores para Métricas de Evaluación_ 53

Tabla 4: Representación de Valores para Métricas de Evaluación_ 57

Tabla 5: Métricas de Acurracy de Método Holt-Winters 60

Tabla 6: Métricas de Acurracy de Método STL ARIMA_ 60

Glosario

Análisis de Sentimientos: Detección de opiniones, posturas, enfoques, intensidades, sentimientos expresados en textos, apasionamientos y emociones en funcionalidad de un contenido, producto, marca o servicio y su siguiente categorización las cuales tienen la posibilidad de ser positivo, negativo o neutro.

BigData: tiene relación con un grupo de datos cuyo enorme tamaño provoca que sea complejo examinar y laborar con los mismos para lo que se hace el procesamiento y pre-procesamiento de los mismos.

Comentarios: Un comentario es una opinion o escrito sobre cualquier cosa puesta en estudio. además, es el título que se da a alguna historia redactada con brevedad, los cuales en redes sociales pueden ser en gran volumen.

Covid-19: Es una enfermedad infecciosa provocada por el virus SARS-CoV-2; la mayor parte de los individuos que sufren COVID-19 padecen indicios de magnitud leve a moderada y se recuperan sin necesidad de tratamientos especiales. No obstante, varias personas desarrollan casos graves y requieren atención médica.

Matriz de Confusión: Es un instrumento de Machine Learning la cual contabiliza las predicciones comparativamente con los valores reales; para comprender de mejor forma que tan bien se comporta el modelo.

Minería de Datos: La minería de datos o investigación de datos es un campo de la estadística y las ciencias de la computación referido al proceso que aspira hallar patrones en grandes volúmenes de conjuntos de datos. Usa los procedimientos de la ia (inteligencia artificial), aprendizaje automático, estadística y sistemas de bases de datos.

Procesamiento de Datos: Es, generalmente, la acumulación y manipulación de recursos de datos para generar información significativa. El procesamiento de datos trata de un subconjunto del procesamiento de la información, el cambio de la información de todas maneras detectable por un observador.

Script: Es un grupo de instrucciones escritas en código que sirven para llevar a cabo distintas funcionalidades en un programa.

Textblob: Es una librería de Python la cual utiliza native bayes y reglas de lexico para el análisis de sentimientos.

Twitter: Es un servicio de micro blogueo la cual posibilita mandar mensajes de escrito plano de corta longitud, con un mayor de 280 letras y números (originalmente 140), denominados tuits o tweet.

Tweet: Es una publicación o actualización de estado elaborada en Twitter la cual procede del inglés, y podría traducirse al español como trino, pío o gorjeo, en referencia al ruido que realizan los pájaros.

Vader: Es una librería de Python la cual utiliza reglas de léxico para el análisis de sentimientos.

Introducción

Las opiniones y comentarios que se generan en las redes sociales pueden ser analizados con la finalidad de orientar la toma de decisiones. Por ejemplo, las empresas pueden determinar si su marca o producto está teniendo la acogida esperada o no; o para un sondeo de mercado antes del lanzamiento de un nuevo producto o servicio. También puede ser muy útil, para los gobiernos nacionales y sectoriales, analizar el impacto de nuevos decretos, proyectos de ley, nuevas regulaciones de impuestos, entre otros. En política y época de campaña electoral, se puede analizar la polaridad de los sentimientos que tienen los ciudadanos por un determinado candidato. En los sectores de ocio, cine, turismo, periodismo, entre otros casos, puede ser útil analizar los comentarios u opiniones de los usuarios o clientes. Las técnicas más utilizadas para analizar textos, como comentarios u opiniones son: la minería de datos, minería de textos, el procesamiento del lenguaje natural y, más específicamente el análisis de sentimientos.

La Minería de Datos (DM: Data Mining) [1]–[5], permite la búsqueda de patrones, tendencias, o información útil que a simple vista está oculta en data sets, aplicando algoritmos específicos. Para [6], DM es una técnica nacida a finales de la década de 1990, que permite entender la información y descubrir las posibles respuestas buscadas; además, explica los elementos primordiales y la importancia que tiene para los investigadores. Varios científicos han utilizado o lo están utilizando DM, para obtener resultados de sus investigaciones; sin embargo, destacan la importancia de contar con datos de calidad; para eso es necesario realizar un preprocesamiento que implica manejar las inconsistencias, ruido, datos erróneos o faltantes, datos inusuales, datos de gran tamaño, etc. La minería de texto (TM: Text Mining) es una rama específica de DM y consiste en la utilización de métodos y técnicas automatizadas de identificación de patrones y correlaciones de palabras, a fin de obtener información útil y que no es explícita, a partir de un amplio conjunto de textos [7].

Para [8], el Procesamiento del Lenguaje Natural (PLN) es una subdisciplina de la inteligencia artificial, que funciona a través del aprendizaje automático (ML: Machine

Learning) y se complementa con la minería de textos. PLN involucra el diseño e implementación de sistemas y algoritmos que pueden interactuar a través del lenguaje humano, con el propósito de realizar las tareas deseadas. Una de las áreas que posee PLN es el Análisis de Sentimientos (SA: Sentiment analysis); ésta es de mucha utilidad para a el monitoreo de las opiniones de un cierto sector de la población acerca de algún tema y, permite determinar tendencias y poder tomar decisiones. En la actualidad, es importante el uso de SA, ya que se puede encontrar gran volumen de información subjetiva en Internet.

Existen varios trabajos que aplican PLN; por ejemplo en [9], se plantea un modelo computacional que consiste en mostrar los pensamientos subjetivos de las personas, tratando de comprender su personalidad y la ideología en base a la subjetividad que poseen sus textos. En trabajos realizados recientemente indican sobre los daños que causa el COVID-19 y el aislamiento social en la población en lo que corresponde a salud mental. Según [10], en un estudio realizado a la población China el 53,8% de personas se consideró en con un impacto psicológico del COVID-19 como moderado o severo. También se identificaron a personas infectadas y miembros que tengan vínculos cercanos con ellas, además de personas con problemas mentales y personal de salud como grupos más vulnerables y que requieren atención y apoyo psicoterapéuticos.

Debido al COVID-19, los gobiernos han detectado la necesidad de someter cuarentenas obligatorias y protocolos de bioseguridad, para evitar contagiar más vidas humanas, y nuestro país no ha sido la excepción. El encierro ha provocado un mayor incremento en el uso de las redes sociales por la población, mayoritariamente para comunicación con sus familiares y amigos; pero también para opinar sobre temas variados y relacionados con la pandemia. Además, es notorio que las redes sociales desempeñan un papel fundamental para las autoridades gubernamentales, cuando han tenido que informar a los ciudadanos sobre las estrategias a implementar en respuesta a la emergencia sanitaria que comenzó en marzo del año 2020 y que aún lo estamos viviendo.

Con los antecedentes antes mencionados, en este trabajo se planteó el siguiente problema: ¿Qué sentimientos ha generado la pandemia del COVID 19 en la población ecuatoriana? Para dar respuesta a esta interrogante, se planteó el siguiente objetivo: realizar análisis sentimientos de comentarios recuperados de la red social Twitter en el contexto de COVID-19 en Ecuador, desde abril del 2020 hasta julio del 2021, para indagar la polaridad de los sentimientos de población respecto a la pandemia.

Este documento está organizado por: Capítulo 1: diagnóstico de necesidades y requerimientos; se describe las necesidades de la minería de datos y análisis de sentimientos, así como de su importancia. Capítulo 2: desarrollo del prototipo, se detallan los objetivos, el procedimiento de la metodología a utilizar y la ejecución de cada una de las fases de la metodología utilizada. Capítulo 3: evaluación del prototipo; se realizan las distintas pruebas a los métodos utilizados en el análisis de sentimientos, para comprobar la efectividad de los resultados obtenidos.

CAPÍTULO I. DIAGNÓSTICO DE NECESIDADES Y REQUERIMIENTOS

1.1. ÁMBITO DE APLICACIÓN: CONTEXTUALIZACIÓN Y DESCRIPCIÓN DEL PROBLEMA OBJETO DE INTERVENCIÓN

Debido a la pandemia causada por COVID-19, las personas reaccionan de una forma positiva, negativa o neutra. Dado que se desea analizar dicha información para determinar cuál es la tendencia en base a la pandemia actual, y de esta forma brindar esta información a los expertos. Obteniendo como evidencia un efecto negativo en cuanto a los tweets posteados por los usuarios de Twitter en cuanto a la temática del COVID-19.

Por lo tanto, es necesario crear herramientas capaces de describir cual es la situación actual relacionados a la temática en cuestión, para así predecir el índice de sentimientos negativos o positivos determinados en rangos de fecha aplicando métodos de análisis de sentimientos como VADER, TEXTBLOB y AFFIN desarrollados y ejecutados en el lenguaje Python.

Para explicar el impacto en el estado mental de una persona, en el 2006 la Unidad de Salud Mental de la Organización Panamericana de la Salud realizo un escrito técnico con el fin de orientar las acciones en el contexto de epidemias en lo correspondiente a salud mental y emocional. Recientemente, ante la pandemia suscitada a finales del 2019 la OMS con el objetivo de orientar a diferentes grupos desarrollo una serie de mensajes para afianzar el bienestar mental y psicosocial de las personas. Por ejemplo, recomienda la búsqueda de información de fuentes confiables, la realización de rutinas diarias y la indagación de historias positivas de personas que se ha recuperado del COVID-19, con la finalidad de

obtener un estado emocional y mental positivo y estable en medio de la situación vivida [11] .

La finalidad de la presente propuesta tecnológica tiene el objetivo de que el análisis de comentarios recuperados de la red social Twitter en el contexto de COVID-19 en Ecuador desde abril del 2020 hasta julio del 2021 que nos permiten determinar la tendencia y evolución de los pensamientos acerca de la pandemia posteados en la red social Twitter en las fechas mencionadas, mediante la aplicación de series temporales se determina la cantidad te tweets con sentimientos positivos, negativos y neutros en lo referente al COVID-19 para su posterior análisis predictivo con la utilización de modelos de series temporales como: autoregressive integrated moving average (ARIMA), Seasonal-Trend decomposition using LOESS (STL) y HOLT-WINTERS.

1.2 ESTABLECIMIENTO DE REQUERIMIENTOS

Es una necesidad el análisis de sentimientos ya que nos permite monitorizar cual es la tendencia que tiene en cuanto al COVID-19 en los diferentes sectores del país ya que estos indicadores permiten a los gobiernos establecer cuál es el estado de ánimo de sus habitantes y así abordar rápidamente medidas para contrarrestar síntomas depresivos, la idea del presente proyecto es analizar al COVID-19 en términos de sentimientos y expectativas en Ecuador durante la etapa de la pandemia en la red social Twitter aplicando minería de datos, utilizando la API de Twitter para la obtención de los datos que serán sometidos al procesamiento y aplicación de algoritmos de clasificación de texto para la obtención de los resultados de las predicciones de la cantidad de tweets posteados por sentimiento a lo largo del tiempo.

Los datos generados en tweets generalmente son utilizados para fines investigativos debido a la facilidad de obtención de los mismos utilizando la API de Twitter lo cual nos facilita una excelente herramienta al momento de analizar la información. En la actualidad los usuarios de esta red social la han utilizado ya sea para twittear acerca de información del COVID-19 u otro tipo de información por lo cual genera grandes volúmenes de información.

La extensión de la enfermedad y la angustia emocional y social depende de las reacciones psicológicas, este trastorno se da durante y después del brote, a pesar de esto no se desarrolla ni se establece recursos con el objetivo de disminuir los efectos dañinos que causa la pandemia en términos de estado mental y bienestar diario [12] .

La información originada en Twitter ya sea de medios oficiales respecto al COVID-19 o medios de comunicación importantes, así como también de otros usuarios lo cual genera una gran cantidad de información.

La alternativa para aprovechar los datos obtenidos de los tweets es tener una herramienta con la capacidad de generar conocimiento a partir del hallazgo de los patrones o semejanzas que guardan los conjuntos de datos, que aportarán y producirán recursos que deben ser aplicados a una determinada área, logrando así información validada y adecuada para la utilización de los profesionales en el área de la salud mental con el fin de ayudar a las personas a sobrellevar la situación emocional. A parte del óptimo uso de recursos que se logra con dicha herramienta, se puede predecir la situación emocional que pasará con las personas en un determinado tiempo.

Dada la problemática se planteó como objetivo el análisis de comentarios recuperados de la red social Twitter en el contexto de COVID-19 en Ecuador desde abril del 2020 hasta julio del 2021.

1.3 JUSTIFICACIÓN DEL REQUERIMIENTO A SATISFACER

Actualmente el análisis de sentimientos es de gran importancia al momento de examinar cuáles son las emociones que causa el COVID-19 la investigación [13], explica que la estadística es una herramienta de gran utilidad en el ámbito empresarial, gubernamental debido a que su implementación apoya a investigadores a recolectar y reconocer las emociones y opiniones del público para la toma de decisiones; para el desarrollo del presente trabajo se eligió a Twitter debido a que es una de las redes sociales más activas, con millones de tweets enviados diariamente y muchos usuarios comentando sobre viajes, asuntos económicos, decisiones políticas y más [14], como tal, es una valiosa fuente de información para su opinión y al existir una diversidad de comentarios almacenados en dicha red social nos obliga a buscar técnicas que nos permitan analizar de mejor manera los datos obtenidos que se presentan en el transcurso del tiempo; el éxito de cómo se lo implemente radica en posterior trata y limpieza de la data.

El trabajo [15] , nos señala que el análisis de sentimientos nos permite la realización de sistemas de predicción, así como la generación de información útil sobre el cuidado de la salud, la cual es muy necesaria para la observación continua de las opiniones de la ciudadanía para así tener una mejor adaptación a las intervenciones y disposiciones del gobierno [16] ; la información obtenida permite mejorar los sistemas de salud pública debido a que proporciona suficiente información en base a falencias [17] .

La finalidad del presente trabajo es el análisis de comentarios recuperados de la red social Twitter en el contexto de COVID-19 en Ecuador desde abril del 2020 hasta julio del 2021, que será de gran utilidad como herramienta para los expertos en salud mental y las autoridades pertinentes para el tratamiento de síntomas depresivos.

CAPÍTULO II. DESARROLLO DEL PROYECTO

2.1 DEFINICIÓN DEL PROTOTIPO TECNOLÓGICO

El prototipo de esta propuesta tecnológica trata del análisis de sentimientos de los comentarios de Twitter, el cual es de gran utilidad ya que permite indagar expectativas y sentimientos acerca del COVID-19 en Ecuador. Para determinar la polaridad de los sentimientos y tendencias en lo que respecta al COVID-19, en este trabajo se utilizó el lenguaje Python. Primero se codificaron los scripts de recolección de los datos de la red social Twitter referente a comentarios elaborado por usuarios con localización en Ecuador y los temas: COVID-19 y pandemia. Luego se limpió los datos y se procedió a realizar el análisis de sentimientos mediante la aplicación de tres métodos (AFFIN, VADER y TEXBLOB). Finalmente se procedió a realizar el análisis predictivos y presentación de los resultados (ver F igura1).

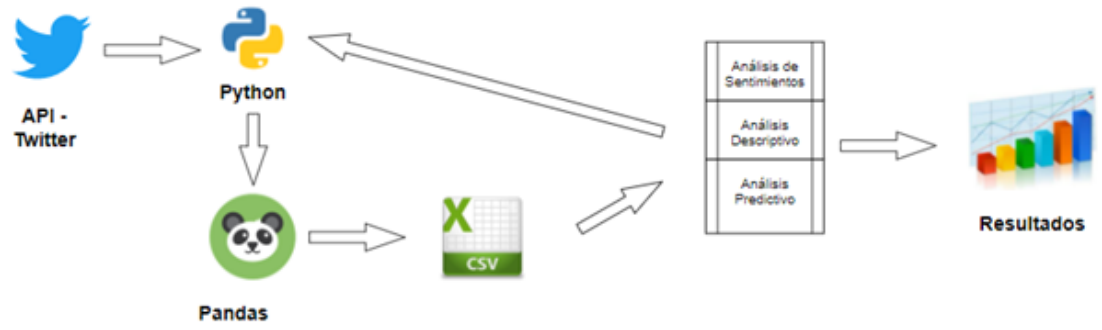

Figura 1:Arquitectura de Propuesta Tecnológica

Fuente: Elaboración propia

2.2 FUNDAMENTACIÓN TEÓRICA DEL PROTOTIPO

2.2.1 Historia COVID-19

A inicios de diciembre del 2019 se detectaron un gran número de pacientes asilados en hospitales con una enfermedad que presentaba síntomas de insuficiencia respiratoria y neumonía, a causa de un nuevo coronavirus (SARS-CoV-2), su nombre COVID-19 como tal lo estableció la Organización Mundial de la Salud el día 11 de febrero del 2020 en la provincia de Hubei China. Desde entonces a pesar de las duras medidas tomadas por los gobiernos la epidemia ha seguido extendiéndose y afectando a otros continentes como Asia, Oriente Medio y Europa. Su declaración como pandemia fue dada en conferencia de prensa el 11 de marzo del 2020 por Tedros Adhanom Ghebreyesus, director de la Organización Mundial de la Salud [18] .

El 29 de febrero del 2020 se reportó el primer caso de COVID-19 en Ecuador luego que se dio el retorno desde España de un compatriota de 71 años el 1 de febrero del 2020, a partir de esa fecha los casos de COVID-19 fueron replicándose a nivel nacional [19] .

2.2.2 Evolución del covid-19 en Ecuador

El índice de casos de COVID-19 en Ecuador se establecía con un porcentaje de 10.94 por cada 100000 habitantes, medida que supera a la detallada a nivel mundial que es de 7.33 por cada 100000. Los casos empiezan a multiplicarse en el país a partir del décimo día de haberse reportado el primer caso. El aumento acelerado del número de casos preocupa a las autoridades y al país entero lo cual debe llevar a reflexionar a las autoridades, con el fin de reforzar principalmente a las actividades de prevención y compromiso comunitario.

Figura 2: Casos de COVID-19 en Ecuador desde marzo del 2020 hasta septiembre del 2021

Fuente: Obtenido de [20]

La letalidad del COVID‑19 es uno de los temas de preocupación a nivel mundial y ha evolucionado en forma diferente en cada país. La tasa de letalidad por COVID‑19 en Ecuador es de 6.44%, por encima del promedio mundial que está en el 2.05%, al 19 de septiembre del 2021 [21]. La evolución de los casos de COVID en Ecuador se puede observar en la Figura 2 [20].

2.2.3 Minería de Texto

En [22], la minería de texto se deriva de la minería de datos, por lo que tiene similitudes en su arquitectura de alto nivel. Por ejemplo, ambos sistemas se basan en capas de elementos de presentación, incluidos procedimientos de preprocesamiento, algoritmos de detección de patrones y herramientas de visualización para mejorar la navegación de los conjuntos de respuestas.

La minería de texto se la cataloga en el campo de la investigación como el procesamiento automático de la información donde se da el proceso de encuentro de patrones interesantes y nuevos conocimientos a partir de una colección de texto, quiere decir que una secuencia de pasos que se da para la obtención de conocimientos que no existían explícitamente en ningún texto de la colección pero que se originan al enlazar el contenido de varios de ellos [22], [23].

La minería de textos implementa diversas tecnologías y métodos como: procesamiento de lenguaje natural, la recuperación de información, métodos de clasificación y agrupamiento

de datos, entre otros con el fin de lograr el objetivo de encontrar patrones a partir de una colección de textos.

2.2.3.1 Etapas de la minería de Texto

Para lograr el objetivo del procesamiento de texto y la obtención de datos/ conocimientos la minería de texto proporciona una serie de pasos [24]:

Primero es precisar el estudio de la minería de texto.

El segundo paso es encontrar, reconocer, obtener y dar validez a la información.

En el tercer paso se realiza la eliminación de información que no aporta en el propósito de la minería de texto, mediante la implementación de algunas de las siguientes acciones: análisis léxico, tratamiento y separación de palabras vacías (artículos, preposiciones, conjunciones), tratamiento de términos flexionados (términos relacionados morfológicamente, número o tiempo verbal), normalización de palabras , variaciones de género, tratamiento de palabras compuestas, obtención de las raíces de las palabras y etiquetado de palabras, además de corregir algunos problemas que presenten los documentos como: los problemas de formato, polisemia, homonimia, sinonimia.

En el cuarto paso se realiza el análisis de clases y la extracción, asociaciones o secuencias, interrelaciones con el objetivo de conocer pruebas de conceptos y de construcciones encontradas [24].

La minería de textos es el proceso de examinar colecciones de documentos de texto no estructurado, para capturar los temas y conceptos clave, descubriendo interrelaciones ocultas y tendencias existentes entre los textos sin necesidad de conocer los vocablos precisos que los autores han usado para manifestar estos conceptos, llevado a cabo a usando un grupo de herramientas de estudio [25].

El procesamiento de textos suele ser complejo. Su falta de estructura, su contenido y naturaleza heterogénea presentes en enormes bases de datos, hace que se requiera de técnicas que los trate convenientemente, con el objetivo de exponer el proceso de manipulación de textos y aplicar un análisis exploratorio, se realiza un ejemplo con la variable. Para ello, se utiliza principalmente el paquete TM ya que ofrece la cantidad de

funcionalidades necesarias para gestionar y manipular documentos de texto en R, las técnicas aplicadas son las siguientes [26]:

- Creación del Corpus.
- Preprocesamiento y limpieza de textos:
- Conversión a minúsculas.
- Eliminamos las puntuaciones.
- Eliminando números.
- Removemos tildes.
- Se eliminan los stopwords.
- Se eliminan espacios vacíos.
- Stremming.

Matriz de Documentos-Término

Es ideal para medir la similitud entre palabras. Como se señaló, la medida de similitud más utilizada es el coseno del ángulo entre diferentes vectores en la matriz de contexto de la palabra. Hay muchas aplicaciones basadas en la similitud de palabras, como la generación de oraciones de similitud, la generación de sinónimos y las rupturas semánticas [27] .

2.2.3.2 Técnicas de minería de Texto

Modelo Booleano

Se trata de un modelo en el que las palabras clave están interconectados basan en la teoría de conjuntos y el álgebra de Boole. Debido a su simplicidad inherente y el formato ordenado, que ha recibido una gran atención y ha sido adoptado por muchos de los sistemas bibliográficos comerciales más tempranas. Su estrategia se basa en la recuperación determinantes binarios (vinculados o no), sin concepto de escala o el concepto de la correspondencia parcial de las condiciones de la demanda [28] .

El presente modelo se origina de las reglas clásicas de la lógica de conjuntos y sus operadores lógicos como unión, intersección y negación; dado esto es muy común la introducción de alguna combinación de las cláusulas anteriores.

En este tipo de modelos un documento es relevante o no lo es, es decir su relevancia es binaria cuando se da la consulta una palabra el modelo le establece de gran importancia si y solo si se la encuentra dentro de la composición del documento; en las consultas que utilizan “Y” deben tener dentro de su composición todas las palabras detalladas en la consulta; en las consultas que son de tipo “O” los documentos deben precisar algunas de las palabras y por ultimo en consultas A pero no B los documentos deben ser relevantes solo para A pero no para B.

Algunas de las desventajas de este enfoque son:

- No clasifica entre documentos de mayor y menor relevancia.
- Establece el mismo resultado si el documento contiene una o cien veces las palabras de la consulta.
- Da lo mismo que cumpla una o todas las cláusulas de un OR.
- No considera una clase parcial de un documento (Ej., que cumpla con casi todas las cláusulas de un “Y”).

El modelo booleano nos permite identificar la relación entre términos en base a la algebra booleana; el presente modelo al ser considerado primitivo implementa un método de relevancia simple es decir un documento se lo puede catalogar como relevante o no relevante, con lo cual al momento de hacer una consulta lo primero que se requiere es un índice con el contenido de las palabras relevantes para una consulta, es decir primero se requiere tratar el documento con la finalidad de eliminar palabras irrelevantes sean estos números, preposiciones, conjunciones, pronombres.

Análisis de semántica latente (Latent Semantic Analysis; LSA)

La semántica latente (LSA) es un enfoque basado en corpus que se usa ampliamente para evaluar la similitud del texto sobre la base de relaciones semánticas entre palabras. Las LSA se han aplicado con éxito a varios sistemas lingüísticos para calcular la similitud semántica de los textos. LSA ignora el patrón de oración. Es decir, es sintácticamente ciego. LSA no distingue entre oraciones que contienen palabras semánticamente similares, pero con significados opuestos [29] .

Los modelos LSA nos permite analizar un corpus de grandes dimensiones el cual contiene una gran cantidad de frases y palabras los cuales nos permite modelar el conocimiento semántico y las relaciones entre los distintos términos almacenados en el corpus.

Probabilistic Latent Semantic Analysis (PLSA).

PLSA (Análisis semántico latente probabilístico) es una técnica de modelado de materias popular ampliamente aplicada en aplicaciones de minería de texto para descubrir los temas subyacentes incrustados en el contenido de datos. Sin embargo, el crecimiento de los datos varía, por lo que debe investigar argumentos dinámicos y procesar grandes conjuntos de datos en etapas [30] .

El PLSA es, ante todo, una continuación del LSA, en el sentido de que atribuye palabras a tópicos o conceptos latentes con base en la frecuencia ponderada de unos términos en algunos documentos en vez de en otros, aunque interpreta esas frecuencias en términos de probabilidad. El PLSA sigue siendo una técnica de estadística descriptiva de tipo frecuentista en la medida en que la probabilidad de que un término forme parte de clusters de términos pertenecientes a un tópico o concepto dado depende de parámetros arrojados por un conteo de frecuencias dadas en una matriz de bolsas de palabras basadas en un

cálculo de probabilidad multinomial [31] .

Latent Dirichlet Allocation (LDA)

Es considerado un modelo probabilístico no supervisado y parametrizado que potencia a PLSA, induciendo priors en las distribuciones que enlaza cada documento como un tema. Este modelo fue publicado por D. Blei, A. Ng y M. Jordan en el año 2003 y se cataloga como el primer algoritmo de Toping Modeling. La solución al problema del PLSA de aumento de parámetros como una opción aleatoria, con esta distribución se obtiene el modelo sea considerado plenamente generativo [32] .

El modelo Latent Dirichlet Allocation es considerado de tópicos generativos, este modelo infiere que dentro de un documento cada palabra parte de un tópico que es seleccionado a partir de una distribución de tópicos para cada documento. Es decir que la distribución de tópicos se da en primera instancia por una distribución de Dirichlet, esto significa que LDA da acceso a que un documento pueda pertenecer a varios tópicos cada uno con peso diferente [33].

Modelo de Espacio Vectorial Semántico

Los modelos vectoriales se constituyen como documentos de texto que parten de un vector de términos, pero no incluyen enlaces semánticos entre palabras. Los espacios vectoriales semánticos se dan origen a partir de la idea de que el significado de las palabras se logra aprender del entorno lingüístico y existen dos enfoques: semántica de distribución y semántica sintética. El primer enfoque considera el significado de las palabras individuales mientras que el segundo enfoque considera el significado de las palabras y párrafos [27].

2.2.4 Procesamiento de Lenguaje Natural

El procesamiento del lenguaje natural es un área de la inteligencia artificial que utiliza el lenguaje natural para estudiar la comunicación entre humanos y computadoras. El lenguaje natural es el lenguaje que la gente usa para la comunicación verbal o escrita. Las aplicaciones típicas de PLN incluyen búsqueda de respuestas, revisión ortográfica, reconocimiento de voz, generación automática de resúmenes, traducción automática y análisis de sentimientos [34] .

Existen diversos conceptos implicados en el procesamiento del lenguaje que es necesario entender y que se usarán a lo largo de este documento.

Lexicón: Se lo considera como una secuencia abstracta de palabras desordenadas que se atribuyen a un lenguaje, una persona o una región de la misma forma las reglas que permiten enlazar las mismas.

Etiquetado gramatical (Part-Of-Speech Tagging): También conocido como POS Tagging es catalogado como un conjunto de palabras perteneciente a un texto seccionados gramaticalmente. Puede proyectarse de dos diferentes maneras, a medida que se lo haga en concepto a la definición propia de la palabra o en función que desempeña en su contexto [35] .

2.2.4.1 Utilidad y Aplicaciones de Procesamiento de Lenguaje Natural

GPT2

Es un modelo de lenguaje basado en transformadores con 1.500 millones de parámetros entrenados en un conjunto de datos de 8 millones de páginas web. GPT2 se entrena con el simple objetivo de predecir la siguiente palabra, asumiendo todas las palabras anteriores en el texto. La variación del conjunto de datos significa que este sencillo objetivo contiene evidencia natural para muchas tareas en diferentes campos. GPT2 es una escala directa de

GPT, con más de 10 por parámetros y resultando en un volumen de datos 10 veces mayor [36] .

GPT3

Es un modelo de lenguaje autocurativo que se entrena con 175 mil millones de parámetros y se prueba en un "entorno de aprendizaje de baja probabilidad" (las nuevas tareas del lenguaje se pueden realizar con solo unos pocos ejemplos). El modelo de lenguaje predice automáticamente el siguiente elemento (normalmente una palabra) del texto basándose en el texto en lenguaje natural anterior [37] .

Este modelo suele ser utilizado para la redacción de textos similares a los de una persona, así como del uso para traducciones y utilización para generaciones de códigos de programación como es el caso de GitHub Copilot el cual fue entrenado con los códigos almacenados en los repositorios de GitHub.

2.2.4.2 Herramientas de Procesamiento de Lenguaje Natural

NLTK

Natural Language Toolkit (NLTK) es una biblioteca de Python de código abierto ampliamente utilizada para NLP (Proyecto NLTK, 2018). Se pueden utilizar varios algoritmos para el cifrado de texto, la derivación, la eliminación de palabras vacías, la clasificación, la agrupación, el cifrado de PoS, el análisis y la inferencia semántica. También proporciona envoltorios para otras bibliotecas de PNL [38] .

NLTK fue diseñado teniendo en cuenta cuatro objetivos principales:

- Simplicidad.
- Consistencia.
- Extensibilidad.
- Modularidad.

TreeTagger Wrapper

El módulo TreeTagger es un motor de texto procesado que funciona en 20 idiomas en total y se puede adaptar a nuevos idiomas si se proporcionan diccionarios y corpus de texto adecuados.

En términos de capacidades, las características de esta biblioteca son bastante limitadas, con solo raíces de texto y etiquetado gramatical y raíces de palabras cuando se trabaja en inglés, alemán, francés y español [39] .

Stanford CoreNLP

Este es un marco de canalización de anotaciones de Java (o al menos basado en JVM) incluye componentes para manejar marcado, división de oraciones, análisis de punto de venta (determinación de forma básica), NER, análisis, análisis de anáforas y otras anotaciones como análisis, género y emoción actualmente admite 6 idiomas: árabe, chino, inglés, francés, alemán y español [40] .

Ixa Pipes

La biblioteca de tuberías IXA Consta de un grupo de herramientas desarrolladas en el lenguaje de Java con el fin de procesamiento del lenguaje natural que se basa en un aprendizaje automático y diseñado para su implementación con tuberías.

Ofrece una anotación lingüística robusta y eficiente tanto a investigadores como a expertos que no son expertos en PNL con el objetivo de reducir las barreras del uso de la tecnología de PNL, ya sea con fines de investigación o para pequeños desarrolladores industriales y pymes. Las tuberías IXA se pueden utilizar o aprovechar su modularidad para seleccionar y cambiar diferentes componentes [41] .

2.2.5 Análisis de Sentimientos

El análisis de sentimientos tiene como finalidad establecer cómo se expresan las emociones en el texto. El enfoque pionero hace que el análisis se base únicamente en el texto entregado, lo que dificulta la detección de las emociones subyacentes e ignora el papel de la transmisión sensorial. En el sector financiero, esta opinión puede extenderse a diversas plataformas como analistas e informes corporativos, artículos y microblogging [42] .

La investigación [43] , tiene como finalidad examinar grandes datos sociales y la utilización de CNN y RNN a partir del modelo Co-LSTM para la clasificación de sentimientos, para lo cual se obtuvo información de redes sociales.

El enfoque presentado pasa por las tres capas, como la incrustación de palabras, la convolución y la capa LSTM. El diagrama esquemático del enfoque propuesto para el análisis de sentimientos.

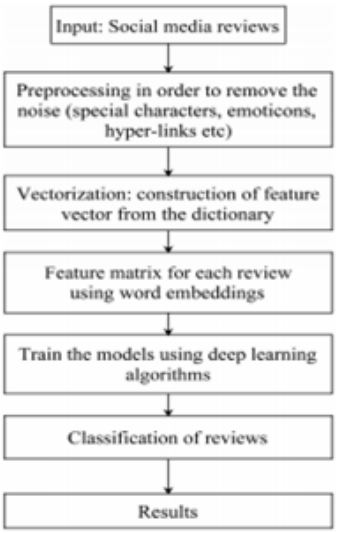

Figura 3: Diagrama esquemático del enfoque propuesto.

Fuente: Obtenido de [42]

Como se muestra en la Figura 3, en la primera capa, se aplica la incrustación de palabras para incrustar las palabras en la revisión, lo que erradica la dependencia del dominio de las características de la revisión. La segunda fase utiliza la capa de convolución y el proceso de agrupación para identificar las características locales y profundas importantes en la oración. La tercera capa aplica la red LSTM en la salida obtenida de la segunda capa para capturar su dependencia secuencial de izquierda a derecha.

La combinación de tres capas ayuda a comprender el comportamiento de la oración. La salida del LSTM se suministra luego a la capa sigmoidea completamente conectada para evaluar el resultado considerando la entropía cruzada binaria como la función de pérdida.

Confusion matrix, evaluation parameters for movie review dataset.

Models		Confusion matrix		Evaluation parameter			
		Predicted Yes	Predicted No	Precision	Recall	F-Measure	Accuracy
SVM	Actual Yes	329	66	0.8329	0.8266	0.8298	0.8311
	Actual No	69	336				
		Predicted Yes	Predicted No				
Naive Bayes	Actual Yes	355	40	**0.8987**	0.7230	0.8014	0.7800
	Actual No	136	269				
		Predicted Yes	Predicted No				
Linear Regression	Actual Yes	318	77	0.8051	0.8010	0.8030	0.8050
	Actual No	79	326				
		Predicted Yes	Predicted No				
Random Forest	Actual Yes	302	93	0.7646	0.6028	0.6741	0.6350
	Actual No	199	206				
		Predicted Yes	Predicted No				
CNN	Actual Yes	316	79	0.8000	0.8294	0.8144	0.8200
	Actual No	65	340				
		Predicted Yes	Predicted No				
RNN	Actual Yes	296	99	0.7494	0.7810	0.7649	0.7725
	Actual No	83	322				
		Predicted Yes	Predicted No				
Co-LSTM	Actual Yes	330	65	0.8354	**0.8350**	**0.8302**	**0.8313**
	Actual No	70	335				

Figura 4: Evaluación de modelo propuesto en el articulo

Fuente Obtenido de [43]

Se puede observar en la Figura 4 la precisión y la Medida F para el modelo Co-LSTM propuesto producen mejores resultados en comparación con otros algoritmos. Los modelos Naive Bayes y CNN tienen mejor precisión y valor de recuperación respectivamente para el conjunto de datos de reseñas de películas, ya que están más sesgados hacia sentimientos positivos. Se encontró que los tres modelos principales para conjuntos de datos de reseñas de películas en términos de precisión son Co-LSTM, SVM y CNN con 83.13%, 83.11% y 82% respectivamente.

El propósito de [23] , es la implementación de la minería de textos en los documentos médicos de la gente de mar para generar un mejor conocimiento de los problemas médicos que a menudo ocurren a bordo del barco.

Se han utilizado datos de pacientes de tres años (2018-2020) para el análisis. Se adoptó el léxico y los algoritmos de Naïve Bayes para realizar análisis sentimentales y se llevaron a cabo experimentos sobre la herramienta estadística R. La visualización de la información sintomática se realizó a través de nubes de palabras y se logró el 96% de la correlación entre los problemas médicos y el resultado del diagnóstico. Validamos el análisis de sentimiento con más del 80% de exactitud y precisión.

Se realizo minería de texto de documentos de pacientes con problemas médicos de la gente de mar. En todo el corpus, iniciamos la tokenización, la eliminación de palabras vacías y espacios en blanco, y la conversión de minúsculas para extraer solo conjuntos de datos sintomáticos individuales; se consideró que la retroalimentación del paciente lleva a cabo una búsqueda de opiniones para definir sentimientos positivos, negativos y neutrales. A partir de entonces, se evaluó el análisis de frecuencia de síntomas para identificar términos de enfermedades populares y crear su léxico preensamblado.

Summary of collected patient symptoms among retrieved documents.

Symptom group	Total number of documents Retrieved	Total number of patients symptoms
2018	1002	937
2019	1136	951
2020	974	889

Figura 5: Resumen de cantidad de síntomas de enfermedades populares

Fuente Obtenido de [23]

Podemos observar en la Figura 5, que el análisis de sentimientos se la realiza mediante el uso de técnica de redes neuronales convolucionales y el uso de minería de texto de lo cual en ambos casos se requieren de un listado de palabras para el posterior aprendizaje o en el caso de la minería de texto para la clasificación de las palabras obtenidas en este caso de Twitter del cual se obtuvieron patrones para la predicción de sentimientos en un futuro dividiéndolos en sentimientos positivos, negativos y neutros.

2.2.5.1 Técnicas de Clasificación de Sentimientos

Las técnicas de categorización de sentimientos tienen la posibilidad de dividir a grandes aspectos en un enfoque de machine learning como lo categoriza la Figura 6; un enfoque con base en el léxico y un enfoque híbrido; los enfoques de machine learning (ML) aplica los clásicos algoritmos ML y usa propiedades lingüísticas [44] .

Machine Learning Aproach

El aprendizaje automático (ML) es un subcampo de la inteligencia artificial que posibilita que la máquina aprenda patrones usando datos históricos sin estar programada explícitamente para realizarlo [45] .

Árbol clasificador de decisiones: es una manera de aprendizaje inductivo. Para un grupo de datos dado, el propósito es construir un modelo que capture el mecanismo que origino los datos [46] .

Máquina de Soportes Vectoriales: son un grupo de procedimientos de aprendizaje supervisado involucrados, que son famosas para hacer estudio de categorización y regresión por medio de la investigación de datos y el reconocimiento de patrones la cual construye un hiperplano o un grupo de hiperplanos para clasificar cada una de las entradas en un lugar de alta magnitud o inclusive infinito [47] .

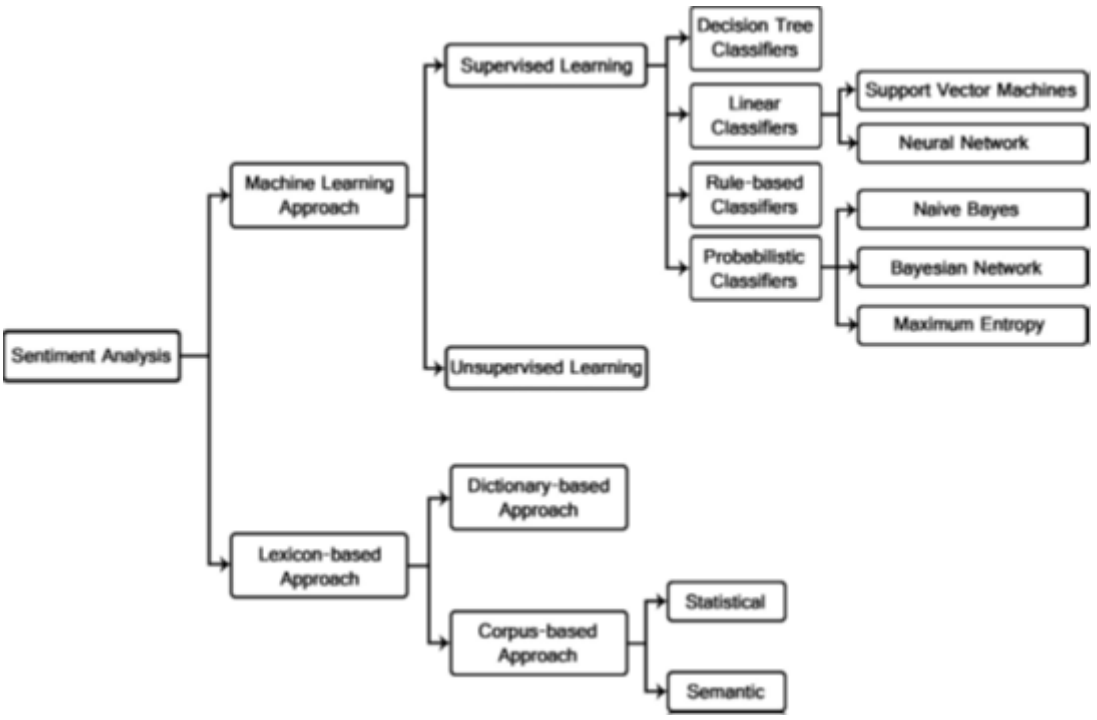

Figura 6: Técnicas de Clasificación de Sentimientos

Fuente Obtenido de [44]

Redes neuronales: una red neuronal se lo cataloga como un modelo que proyecta el modo con el cual el cerebro humano realiza el procesamiento de la información: Funciona a la par de un numero alto de unidades interconectadas de procesamiento que son similares a variantes abstractas de neuronas.

El estudio [48] , nos indica que las redes neuronales otorgan una poderosa maquinaria de aprendizaje que es bastante llamativa para su uso en inconvenientes de lenguaje natural. Un elemento fundamental de las redes neuronales para el lenguaje es la utilización de una capa de incrustación, un mapeo de símbolos discretos a vectores seguidos en un lugar dimensional subjetivamente bajo.

Native Bayes: Es un procedimiento de categorización estadística que se puede usar para pronosticar la posibilidad de pertenencia a una clase. Se demostró que el clasificador Naïve Bayes (NBC) tiene una alta exactitud y rapidez una vez que se aplica a la base de datos con datos [49] .

Redes Bayesianas: Una red bayesiana es un instrumento analítico para calcular el reparto de posibilidad siguiente de cambiantes no observadas condicionadas a las cambiantes observadas. El BN tiene numerosas ventajas, como la función de combinar diversas fuentes de información, la probabilidad de aprendizaje estructural y el procedimiento explícito de la incertidumbre [50] .

Máxima Entropía: El principio de máxima entropía instituye que la distribución probabilística menos sesgada que se le puede atribuir a un sistema estadístico es aquella en la que dadas unas ciertas condiciones estáticas maximiza la entropía, S, en otras palabras, aquella en la que la desinformación es máxima; Nace de la física estadística y la teoría de la información, que piensan el tamaño de la incertidumbre o entropía de la información [51] .

Lexicon-based Aproach

En el estudio realizado por [52] , nos explica que la utilización de enfoques basados en el léxico, permite detectar el contenido de un mensaje y las emociones y emociones en estas palabras. Luego, se puede hacer la categorización sobre la base de la información obtenida.

Métodos basados en diccionarios: en dichos procedimientos, el diccionario de léxico se usa para hallar los vocablos de crítica positiva y las palabras de crítica negativa.

Métodos basados en corpus: en dichos procedimientos se usa un gran corpus de palabras y, basándose en patrones sintácticos, tienen la posibilidad de hallar otras palabras de crítica dentro del contexto [53] .

2.2.5.2 Herramientas de Análisis de Sentimientos

TextBlob: el trabajo [54] , nos define a TEXTBLOB como una librería de Python que aporta al procesamiento de datos de texto. Ofrece una API simple para explorar tareas similares en cuanto a procesamiento de lenguaje natural (NLP), como marcado parcial del habla, extracción de oraciones nominales, análisis de sentimientos, clasificación y traducción.

La investigación [55], nos da a conocer que TEXTBLOB utiliza la clasificación Naive Bayes, dando los mensajes una valoración entre 1 a -1, teniendo a 0 como valor para los neutros.

La forma de aplicación se reúne en el componente de polaridad en una escala de muy negativo dándole el valor de (-1) hasta muy positivo otorgándole el valor de (1), dejando a (0) como valor neutro.

Vader: VADER (Valence Aware Dictionary and Sentiment Reasoner) es una librería de Python de análisis de opinión léxica y basada en reglas especialmente diseñada para las emociones expresadas en las redes sociales, pero se puede utilizar en fuentes relacionadas como se describe en este estudio. VADER utiliza una combinación de palabras que a menudo se etiquetan como positivas o negativas de acuerdo a su orientación semántica [54] .

El artículo académico [40] , nos da a saber que VADER realiza el estudio de sentimientos con base en léxico y normas; este módulo nos otorga 4 tipos de resultados (positivo, neutro, negativo, compuesto) para cada archivo o tuit. Una puntuación compuesta representa la emoción total de un tuit, donde –1 es el más negativo y 1 el más positivo. Mediante la sumatoria de valencia de cada palabra en el léxico se origina la puntuación compuesta, se acomoda conforme a las normas y después se normaliza para estar entre el intervalo de –1 hasta +1 siendo más extremo negativo y más extremo negativo respectivamente. Los valores umbrales clásicos son:

- sentimiento positivo: compuesto >= 0,05
- sentimiento neutral: compuesto > –0,05 y compuesto < 0,05
- sentimiento negativo: compuesto <= – 0,05

AFFIN. Este método se lo implementa con el objetivo de remplazar las palabras que contiene un diccionario por sus respectivas puntuaciones, en el caso de no existir las palabras en el diccionario se le otorga un valor de 0 [57] .

Como nos señala [58] , AFINN es una lista de palabras en inglés clasificada por valencia con un número completo entre -5 (negativo) y 5 (positivo). Los vocablos en Afinn fueron etiquetadas manualmente por Finn Arup Nielsen en 2009-2011.

2.2.6 METODOLOGÍA CRISP-DM (CROSS Industry Standard Process for Data Mining)

CRISPDM es un modelo de proceso independiente de la industria de facto para implementar proyectos de minería de datos resultantes que surge en respuesta a la falta de estandarización. Relaciona las diferentes etapas del proceso para que se integren procesos iterativos y recíprocos. Se ofrece como una metodología justa o neutral para las herramientas utilizadas para el desarrollo de almacenamiento de datos y minería de datos y se distribuye de forma gratuita [59] .

La metodología CRISP-DM CRISPDM consta de seis fases. La primera fase tiene como finalidad conocer sus objetivos comerciales, evaluar su situación actual y planificar su proyecto. La segunda fase tiene como objetivo conocer los datos iniciales disponibles para usted para la realización de su proyecto. La tercera fase prepara los datos recopilados en el paso anterior, y el cuarto paso aplica técnicas de modelado para adaptarse a los objetivos comerciales y generar información nueva y relevante. En la quinta fase se analizan los resultados obtenidos, se evalúa su calidad y finalmente se presentan las opciones para la entrega del proyecto [60] .

2.2.7 Metodología Team Data Science Process

El proceso de ciencia de datos en (TDSP) es una metodología perteneciente a la ciencia de datos con el fin de dar como resultado estudios predictivos eficientes. TDSP Estas resoluciones integran ia (inteligencia artificial) y aprendizaje automático. Inicia la velocidad del plan de ciencia de datos, el trabajo en grupo y el aprendizaje por medio del trabajo de las superiores prácticas y construcciones famosas de Microsoft [61] .

Es catalogada una metodología de estudio de datos expedito y repetitiva diseñada para crear tecnologías de estudio predictivo eficaces y aplicaciones capaces. TDSP incluye lo mejor prácticas y construcciones de Microsoft para contribuir a llevar a cabo las iniciativas de ciencia de datos exitosamente.

TDSP da el periodo de vida para estructurar el proceso de desarrollo de proyectos de ciencia de datos.

El ciclo de vida se ha elaborado para la inclusión en proyectos orientados a la ciencia de datos y se puede usar como parte de aplicaciones capaces que implementan modelos de ia (inteligencia artificial) o aprendizaje automático. Además, este procedimiento puede

favorecer a la ciencia de datos en aspecto exploratorios o de estudios improvisados. El periodo de vida explica las primordiales fases que los proyectos acostumbran llevar a cabo, comúnmente de manera iterativa [62] .

Aun cuando el gráfico del ciclo de vida parece bastante distinto; El periodo de vida del plan de TDSP es semejante a CRISP-DM e incluye 5 fases iterativas:

- Comprensión empresarial
- Adquisición y comprensión de datos
- Modelado
- Despliegue
- Aceptación

2.3.1 Objetivo general

Aplicar minería de datos y análisis sentimientos a comentarios recuperados de la red social Twitter, en el contexto de COVID-19 en Ecuador, desde abril del 2020 hasta julio del 2021, para indagar la percepción y los sentimientos de población respecto a la pandemia.

2.3.2 Objetivos específicos

- Definir los métodos y técnicas de obtención de datos de Twitter mediante librerías de Python y data set de tuits de fuentes académicas o auténticas.
- Recolectar datos de la red social Twitter relacionados al COVID-19 en Ecuador.
- Preprocesar los datos obtenidos mediante algoritmos de limpieza de datos.
- Programar y ejecutar scripts de métodos de análisis de sentimientos a la data de comentarios de COVID-19 en Python.
- Interpretar y evaluar los resultados de cada método de análisis de sentimiento.

2.4 DISEÑO DEL PROTOTIPO.

La metodología a utilizar es una combinación entre la metodología TDSP y CRISP-DM la cual en el caso de TDSP es una metodología ágil e iterativa para brindar resultados en los que interviene el uso de aplicaciones inteligentes y análisis predictivo de modo eficiente. TDSP aporta a optimizar la ayuda y el aprendizaje del equipo al proponer cómo los roles del equipo funcionan mejor juntos además de contener las mejores prácticas y estructuras de Microsoft y otros líderes de la industria para ayudar a lograr una ejecución exitosa de las decisiones de ciencia de datos [63]. Mientras que CRISP-DM es una metodología enfocada

a la minería de datos la cuenta con un ciclo de vida flexible y personalizable lo cual nos permite retroceder a cualquiera de sus faces cuando se requiera. De lo cual se establecieron fases de CRISP-DM para lo que respecta al proceso de minería de datos y TDSP para lo que respecta a los procesos colaborativos. A continuación, se detallan las faces utilizadas en la Figura 7.

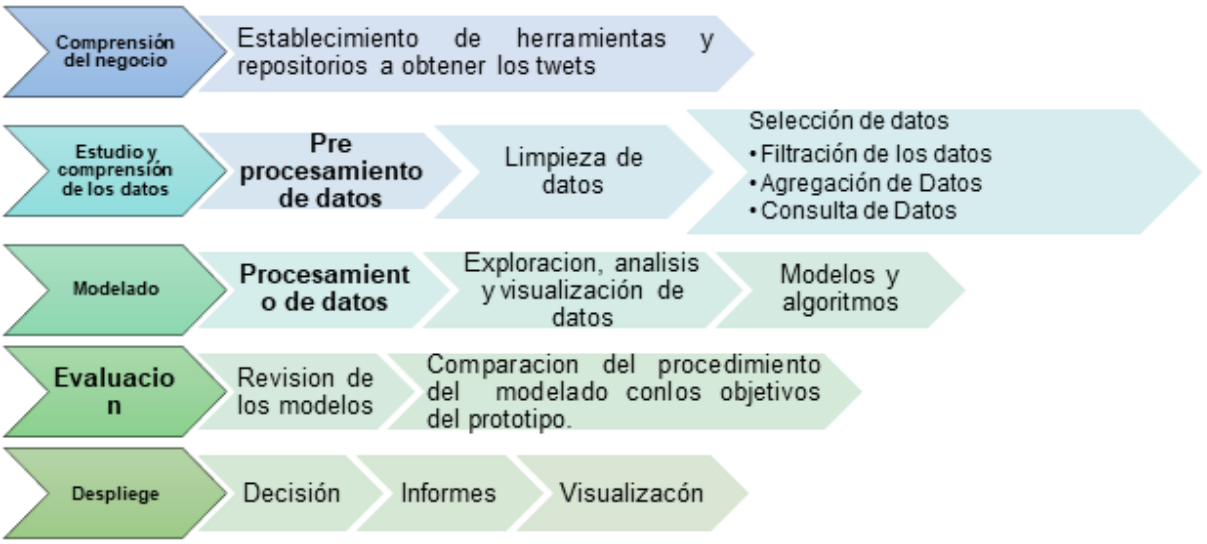

Figura 7: Arquitectura Metodología Propuesta

Fuente: Elaboración Propia

2.4.1 Comprensión del Negocio

Esta fase inicial se establecerá la segmentación del Twitter en este caso será las personas pertenecientes a Ecuador, luego se realizó un análisis de los requisitos del proyecto en este caso se utilizará el api de Twitter para la obtención de los datos mediante el lenguaje de programación Python, así como de fuentes de datos de Universidades.

2.4.2 Estudio y Comprensión de los datos

En esta fase se realizará la obtención de los datos de Twitter, así como la descripción de los datos los cuales se realizó la selección de los datos de Twitter, para el posterior

almacenamiento en un archivo csv el cual se efectuará una limpieza de los datos mediante algoritmos de limpieza de datos.

2.4.3 Modelado

Se procedió a la limpieza manual de los datos eliminando comentarios sin sentido y segmentación de fechas de dichos comentarios además de la creación de data referente a la cantidad de comentarios positivos, negativos, neutros de los cuales se realizó un modelo de series temporales que nos permite analizar a lo largo del tiempo y tipo de sentimientos además de su predicción. Se realizo la obtención de las palabras más frecuentes encontradas en los datos obtenidos.

2.4.4 Evaluación

En esta fase se realizaron las respectivas evaluaciones del modelo generado de las series temporales, así como de los diferentes modelos de análisis de sentimientos utilizados para la generación de resultados.

2.4.5 Despliegue

Se realizó la visualización de las series temporales de los diferentes modelos de análisis de sentimientos utilizados además de la predicción de datos en fechas posteriores.

2.5 ENSAMBLE Y EJECUCIÓN DEL PROTOTIPO

A continuación, se presenta el ensamblaje y ejecución del prototipo en función de la metodología que se explicó en la sección anterior.

2.5.1 Conocimiento del negocio

Se obtuvieron los tweets más actuales mediante la utilización de la API de Twitter para lo cual se obtuvieron las respectivas credenciales como nos muestra [64] , para poder acceder a la API hay que registrar la aplicación, las cuales por defecto tendrán acceso a la información Pública de Twitter.

Los tweets se los cataloga a los mensajes que envían los usuarios que poseen un registro en la red social Twitter, se los conoce principalmente por tener un contenido de hasta 140 caracteres, se pueden implementar caracteres especiales que se interpretan por el sistema de los cuales podemos observar en la Tabla 1:

Tabla 1: Características de los Tweets

Características de los Tweets	
Característica	**Descripción**
Mención @nombre usuario	Si se lo implementa en un Tweet esto permite mencionar a otro usuario que posea cuenta en la red social.
Retweet RT @nombre usuario Tweet usuario	Si se detecta algo con esta estructura tiene significado que el usuario ha Retwitteado el mensaje de nombre usuario y el Tweet usuario.
Hashtag #palabra	Da acceso a unirse a tendencias pertenecientes alguna noticia de relevancia e impacto en tiempo real. Cualquier usuario puede unirse a una conversación de un hashtag dado.

Enlaces a recursos	Facilita a la implementación de enlaces a recursos que están fuera de la red social como documentos, fotos en los tweets. A causa de la limitación establecido de 140 caracteres establecido por la red social es necesario precisar servicios con la finalidad de acortar la longitud de la web ya que si no sería imposible su inclusión en los Tweets.

Fuente: Obtenido de [64].

Tabla 2: Atributos claves de Twitter

Atributos clave en Tweets	
Text, full_text	El texto atribuido al tweet.
created_at	La fecha de origen del tweet.
favorite_count, retweet_count	Número de favorito y los retweets.

favorited, retweeted	Valor booleano que permite conocer si el usuario autenticado ha retweeteado.
lang	Abreviatura de lenguaje (por ejemplo, en de inglés.)
id	Reconocedor del tweet.
place, coordinates, geo:	Información de geo-localización que otorga la red social si es que se encuentra disponible.
user:	Perfil de usuario completo y detallado perteneciente al autor del tweet.
entities	Lista de entidades como URLS, menciones, hashtags y símbolos.

in_reply_to_user_id	Reconocedor de usuario si el tweet es una replica a un usuario específico.
in_reply_to_status_id	Identificador de estado del id del tweet

Fuente: Elaboración Propia

Mediante la utilización del api de Twitter para lo cual mediante el lenguaje de programación Python se obtuvo 900 tweets filtrándolos por idioma en este caso en español y se los procedió a almacenar en un archivo csv. mediante la librería de Pandas de la cual en la Tabla 2 nos explica cada uno de los atributos de Twitter.

Se obtuvieron los tweets de data set de Narrativas digitales de Covid-19, la cual recopila conjunto de datos el 24 de abril de 2020; dicho data set contiene un corpus relacionado con los tweets de Covid-19 y está organizado de la siguiente manera:

- Por idioma: español, inglés
- Por día
- Por geolocalización: EE. UU. (Sur de Florida), Latinoamérica (Argentina, Colombia, México, Perú, Ecuador), España
- Por hashtags

Luego se unificaron los datos manualmente de lo cual se obtuvo el siguiente resultado:

	id	location	created_at	full_text
0	1373426895652544512	Caracas, cuna de Bolívar	2021-03-21 00:11:03	Lamentamos profundamente la partida del camara...
1	1373741606881406982	Venezuela	2021-03-21 21:01:36	#EnVivo ☎ | Jornada de trabajo para hacer bala...
2	1373422316219207680	Venezuela	2021-03-20 23:52:51	#EnVideo☎| Vicepdte. Sectorial de Comunicación...
3	1373812375036567554	NaN	2021-03-22 01:42:49	RT @DAO_AMB: Ante la variante brasileña P-1 de...
4	1373812371085529088	Quibor	2021-03-22 01:42:48	RT @NicolasMaduro: #EnVivo ☎ | Jornada de trab...

Figura 8: Tweets adquiridos y unificados

Fuente: Elaboración Propia

2.5.2 Estudio y comprensión de los datos

para la optimización del análisis textual se realizó la respectiva limpieza de los datos mediante el siguiente código:

2.5.2.1 Normalización

En este proceso se realizó la respectiva eliminación de tildes y conversión de los tweets a minúscula (ver código 1)

```
def normalize(s):
    s = str(s)
    replacements = (
        ("á", "a"),
        ("é", "e"),
        ("í", "i"),
        ("ó", "o"),
        ("ú", "u"),
        ("ñ","n"),
    )

    for a, b in replacements:
        s = s.replace(a, b).replace(a.upper(), b.upper())
    return s.lower()
```

```
def eliminar_emojis(cadena):
    return re.sub(u"[^a-zA-Z0-9áàéíóúÁÉÍÓÚñ: ]"," ",cadena)
```

Código 1: Script de Normalización de los datos

Fuente: Elaboración Propia

2.5.2.2 Limpieza de texto

Para la limpieza del texto se establecieron las reglas de limpieza de hashtag, nombre de las fuentes de información por temas de privacidad de los datos, también se realizó la eliminación de números y el remplazo de palabras incompletas por su significado correspondiente (ver Código 2).

```python
def limpiar_texto(cadena):
    nuevo_texto = normalize(str(cadena))
    # Eliminación de páginas web (palabras que empiezan por "http")
    nuevo_texto = re.sub('http\S+', ' ', nuevo_texto)
    nuevo_texto = re.sub("rt | amp", ' ', nuevo_texto)
    #Eliminacio de palabras que empiezan por @ y #
    caracteres = ["@#"]
    nuevo_texto = re.sub(f'{caracteres}\S+', '', nuevo_texto)
    # Eliminación de números
    #nuevo_texto = re.sub("\d+", ' ', nuevo_texto)
    #Eliminacion de emojis
    nuevo_texto = eliminar_emojis(nuevo_texto)
    #Remplazo de abrebiaturas
    nuevo_texto = re.sub(" q  ", ' que ', nuevo_texto)
    nuevo_texto = re.sub(" sr  ", ' señor ', nuevo_texto)
    nuevo_texto = re.sub(" x ", ' por ', nuevo_texto)
    nuevo_texto = re.sub(" p ", ' por ', nuevo_texto)
    nuevo_texto = re.sub(" d ", ' de ', nuevo_texto)
    nuevo_texto = re.sub(" xq ", ' porque ', nuevo_texto)
    # Eliminación de espacios en blanco múltiples
    nuevo_texto = re.sub("\\s+", ' ', nuevo_texto)
    sin_puntuacion = [c for c in nuevo_texto if c not in string.punctuation]
    sin_puntuacion= ''.join(sin_puntuacion)
    return sin_puntuacion
```

Código 2: Script de limpieza de texto

Fuente: Elaboración Propia

2.5.3 Modelado

Se realizó la respectiva limpieza de texto innecesario de los datos, así como de la eliminación de mensajes innecesarios y datos de otros países, se procedió a la segmentación de fechas los cuales se los categorizo de la siguiente manera:

S1: Abril -julio 2020 (4 meses)

S2: agosto-noviembre 2020 (4 meses)

S3: diciembre 2020-marzo 2021 (4 meses)

S4: abril-julio 2021 (4 meses)

2.5.3.1 Palabras frecuentes

Se realizo una nube de palabras en la cual se muestran las palabras más importantes de la cual se obtuvieron los siguientes resultados (ver Figura 9):

S1 (abril 2020 – julio 2020)

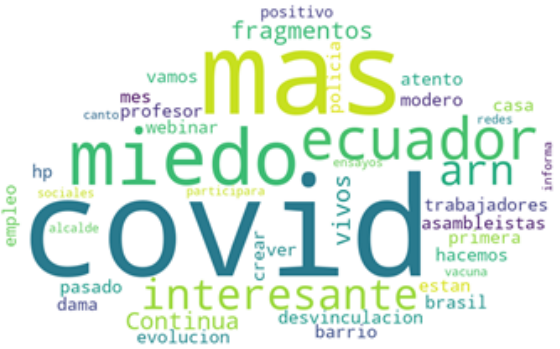

Figura 9: Nube de palabras de abril a julio del 2020

Fuente: Elaboración Propia

Dentro de las palabras más comunes vemos las palabras: COVID-19, más, miedo, Ecuador, interesante, ARN, fragmentos, vivos, continua, desvinculación, de lo cual podemos analizar que en los primeros 4 meses existía miedo en cuanto a la pandemia y debido a que se publicaba más noticias y cifras de muertes en Ecuador se las resaltas como frecuentes.

S2 (agosto 2020 –noviembre 2020)

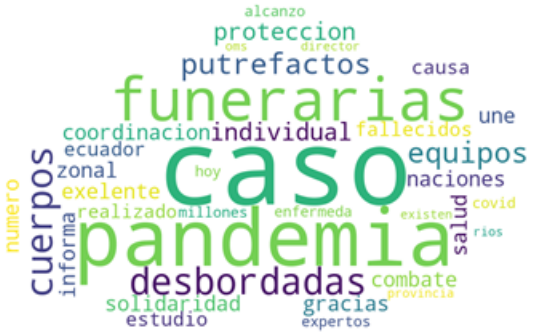

Figura 10: Nube de Palabras de agosto a noviembre del 2020

Fuente: Elaboración Propia

En la Figura 10 se puede observar que las palabras más frecuentes en los meses de agosto a noviembre del 2020 son: caso, pandemia, funerarias, cuerpos, desbordadas, cuerpos, putrefactos, equipos, protección, individual, coordinación; de los cuales se analiza que existen noticias y comentarios en cuanto al incremento de la mortalidad de las personas

en el año 2020 además de que existía un desbordamiento de hospitalizados, además podemos observar que esta como palabra frecuente protección lo cual se refiere a que la gran parte de la población comenzó a implementar medidas de bioseguridad ya sea individualmente como en los negocios.

S3(diciembre 2020 – marzo 2021)

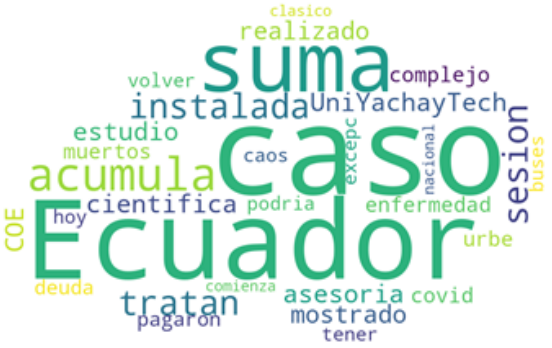

Figura 11: Nube de Palabras de diciembre 2020 a marzo del 2021

Fuente: Elaboración Propia

En la Figura 11 podemos visualizar que las palabras más comunes en los meses de diciembre del 2020 a marzo del 2021 son: caso, ecuador, suma, acumula, sesión, instalada, tratan UNIYACHAYTECH, asesoría, científica de lo cual tenemos que se habla de los casos de covid-19 se van sumando a los del mes anterior, así como de las colaboraciones científicas de la Universidad Yachay.

S4(abril 2021 – julio 2021)

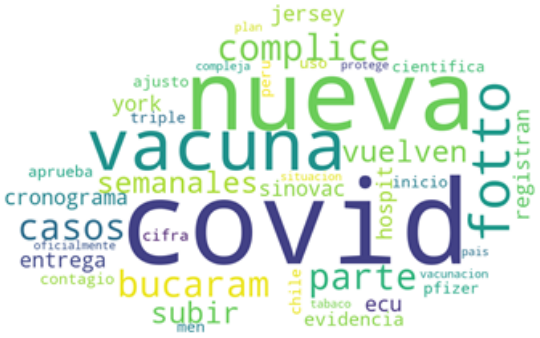

Figura 12: Nube de Palabras de abril 2021 a julio del 2021

Fuente: Elaboración Propia

En la Figura 12 comprendida en los meses de abril a julio del 2021 las palabras más recurrentes son: nueva, vacuna, COVID, casos, cómplices, vuelven, semanales, Bucaram, parte, fotto de lo cual analizamos que debido a que ya comienza el plan de vacunación contra el covid-19 denominado **Plan 9/100,** el cual empezó en mayo del 2021 además los usuarios de Twitter han comentado acerca de los casos en los que está involucrado el expresidente Abdála Bucaram y su familia; a lo largo de este mes también se ha discutido acerca de los casos de covid-19.

Abril 2020 – julio 2021

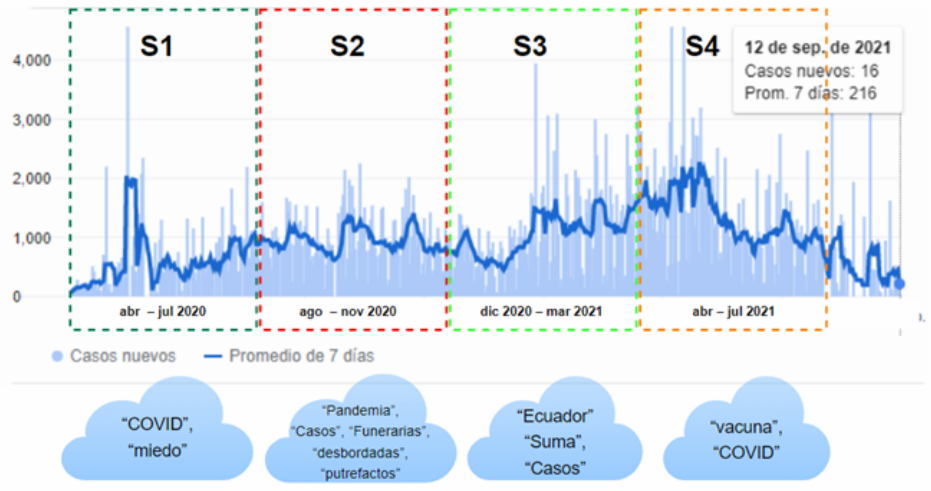

Figura 13: Nube de Palabras de abril 2020 a julio 2021

Fuente: Elaboración Propia

Dentro de las estadísticas categorizadas en 4 periodos de tiempo la palabras más destacadas de los tweets analizados son: COVID, miedo, pandemia, funerarias, desbordadas, putrefactos, Ecuador, suma, casos, vacuna, dé los cuales existen palabras coloquiales tales como fotto lo cual hace referencia al ex vicepresidente Otto Sonnenholzner el cual en el 2020 se tomaba fotos frecuentemente en los hospitales además podemos visualizar que la palabra que más resalta es COVID-19 debido a que es un tema con mayor frecuencia durante la pandemia, la palabra vacuna se dieron en los meses de mayo hasta julio debido a la llegada de el plan de vacunación denominado 9/100 en Ecuador.

2.5.3.2 Análisis de los sentimientos en los tweets

Mediante el usos de librería VADER, TEXTBLOB y el corpus AFFIN se obtuvo lo que respecta al corpus para la posterior utilización en lo que respecta a analizar sentimientos expresado en las frases en este caso se realizó el análisis de cada uno de los tweets almacenados en el archivo csv generado y limpiado anteriormente de lo cual los valores mayores a 0 son considerados como positivos, 0 como neutros y valores menores a 0 negativos; de lo cual se procedió a analizar y almacenar en las columnas de AFFIN, VADER y TEXTBLOB de lo cual se obtuvo la siguiente data (ver Figura 14):

	Fecha	Ano	Mes	Nombre_mes	Dia Semana	Segmento	location	comentario	afinn	textblob	vader	afinn_sen	vader_sen	textblob_sen
0	2021-07-01	2021	7	Julio	jueves	C4	Pichincha	fotto complice de los bucaram fue parte del at...	0.0	0.111905	-0.8020	0	-1	1
1	2021-07-01	2021	7	Julio	jueves	C4	Pichincha	casos semanales de covid vuelven a subir en el...	0.0	0.000000	0.0000	0	0	0
2	2021-07-01	2021	7	Julio	jueves	C4	Pichincha	cronograma de entrega de vacunas sinovac a ecu...	0.0	0.000000	0.0000	0	0	0
3	2021-07-01	2021	7	Julio	jueves	C4	Pichincha	nueva york y nueva jersey registran las hospit...	0.0	0.224545	0.1779	0	1	1
4	2021-07-01	2021	7	Julio	jueves	C4	Pichincha	evidencia cientifica de contagio de covid 19 p...	0.0	0.500000	-0.7579	0	-1	1

Figura 14: Clasificación de Sentimientos de cada Método

Fuente: Elaboración Propia

De lo cual se realizó la visualización del análisis de sentimientos de las herramientas AFINN, VADER y TEXTBLOB de lo cual obtenemos el siguiente gráfico:

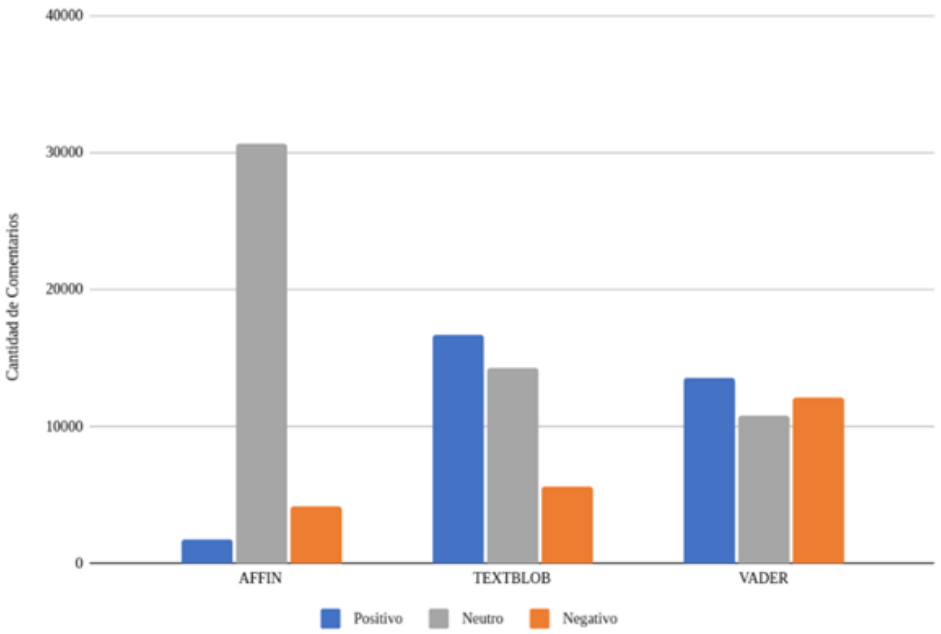

Figura 15: Cantidad de Tweets por Método de Análisis de Sentimientos

Fuente: Elaboración Propia

La herramienta AFFIN nos expresa en la Figura 15, que los sentimientos neutros son los más frecuentes mientras que TEXTBLOB y VADER nos muestra una tendencia a los sentimientos positivos, a pesar de esto podemos observar que los datos negativos analizados por VADER existen una pequeña diferencia entre los sentimientos positivos y negativos, mientras que las herramientas TEXTBLOB y AFFIN nos muestra diferencias muy distantes de dichos sentimientos.

2.5.3.3 Series Temporales

Una vez realizados los análisis por los diferentes métodos de análisis de sentimientos se procedió a realizar las respectivas series temporales de las cuales obtuvimos los siguientes resultados:

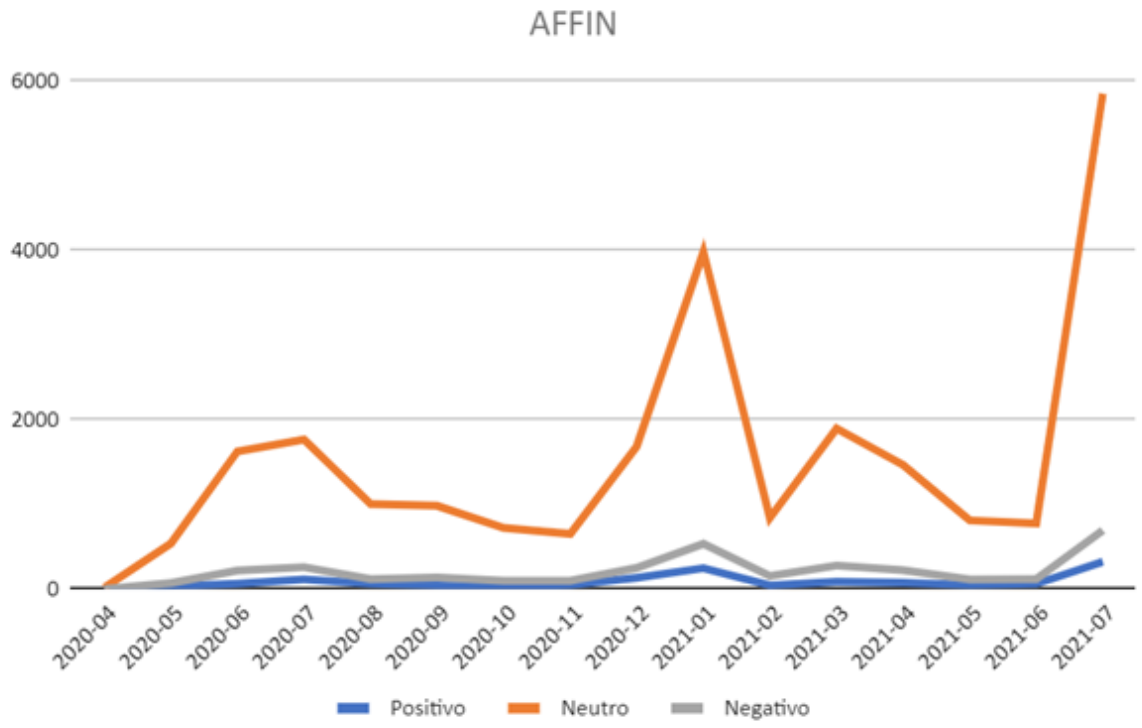

Figura 16: Serie Temporal de cantidad de Comentarios Método AFFIN

Fuente: Elaboración Propia

Según los análisis de sentimientos expuestos en la Figura 16; podemos comparar que el método AFFIN existe una mayor cantidad de comentarios neutrales en los últimos meses por lo que se puede analizar que no existe un estado depresivo en la población

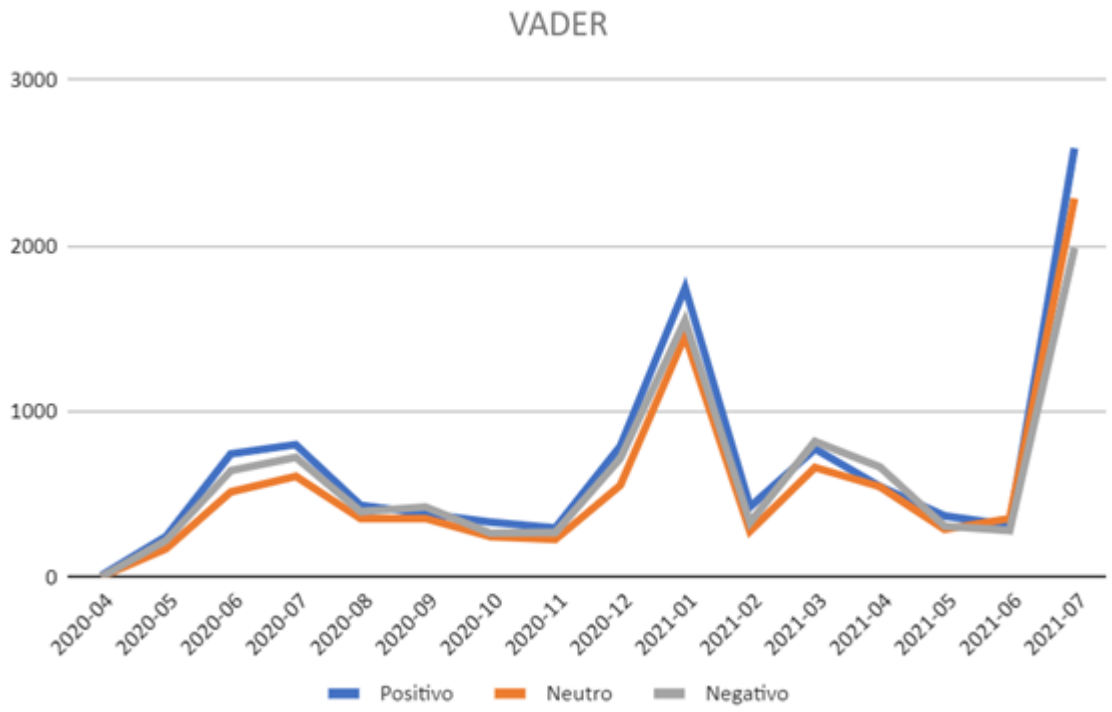

Figura 17: Serie Temporal de cantidad de Comentarios Método VADER.

Fuente: Elaboración Propia

Según los análisis de sentimientos realizados con el método VADER en la Figura 16 nos expone la existencia de una mayor cantidad de comentarios positivos en los últimos meses por lo que se puede analizar que al no estar tan distantes existe una variedad de sentimientos en la población sin embargo a diferencia de TEXTBLOB existe más tendencia a subir la cantidad de sentimientos positivos.

Según los análisis de sentimientos visualizados en la Figura 18 nos indica que el método TEXTBLOB existe una mayor cantidad de comentarios positivos en los últimos meses por lo que se puede analizar que al no estar tan distantes existe una variedad de sentimientos en la población.

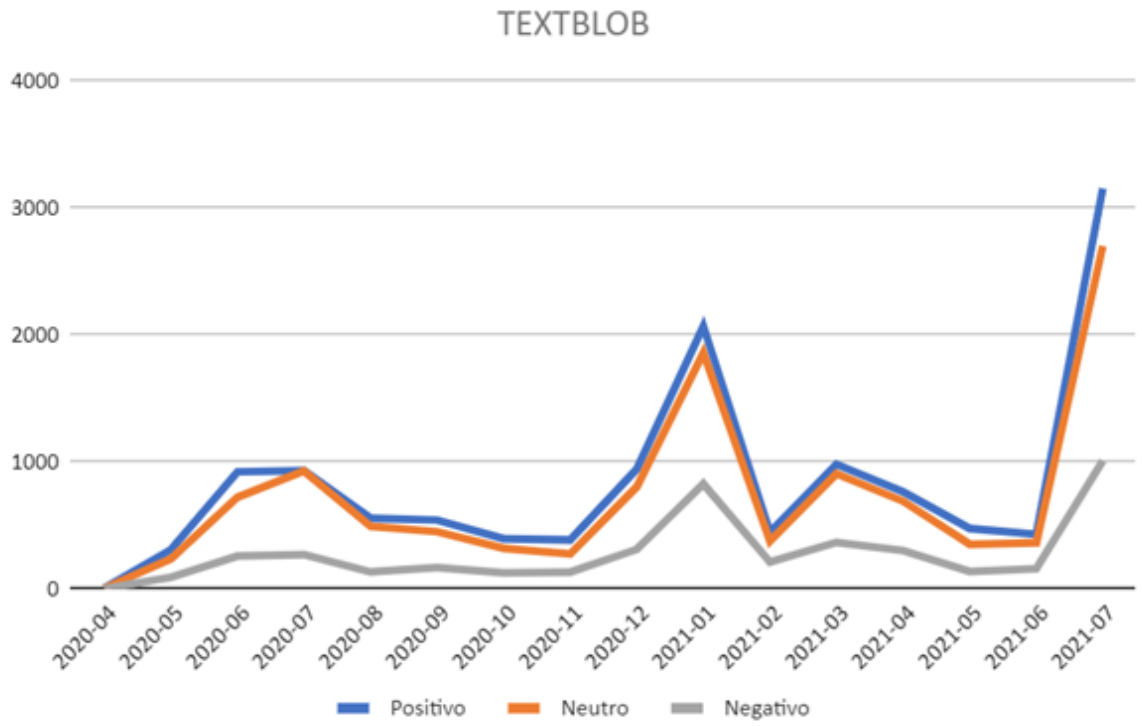

Figura 18: Serie Temporal de cantidad de Comentarios Método TEXTBLOB.

Fuente: Elaboración propia

Según los análisis de sentimientos realizados en R mostrados en la Figura 19, existe una mayor cantidad de comentarios positivos en los últimos meses por lo que se puede analizar que al no estar tan distantes existe una variedad de sentimientos cabe destacar que se ha tomado en consideración por periodos de quincena los cuales van desde el periodo 10 al 15 teniendo en cuenta que del periodo 10 al 12,5 pertenecen al 2020 y del 12,6 al 15 pertenecen al año 2021.

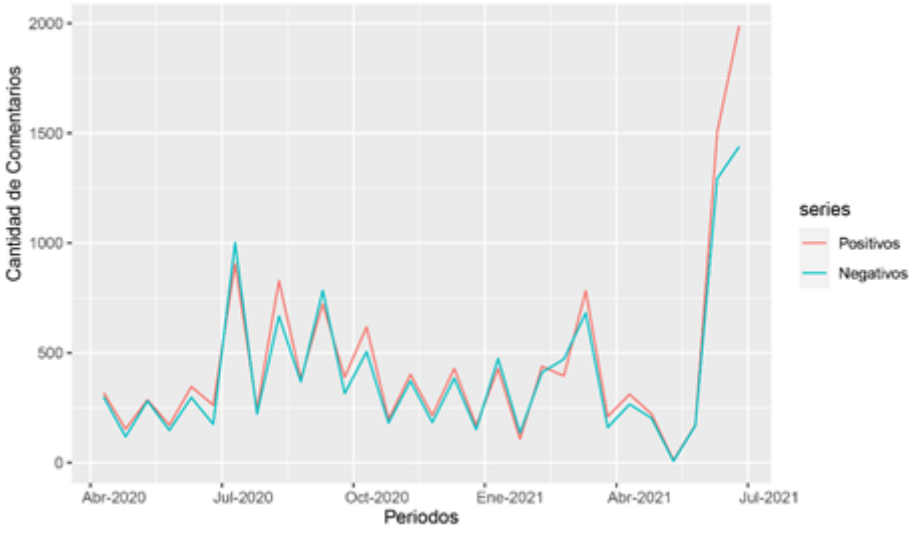

Figura 19: Serie Temporal de los sentimientos en tweets

Fuente: Elaboración Propia

2.5.3.4 Predicciones

Una vez realizado el preprocesamiento de los datos y haber ejecutado el análisis de sentimientos mediante los métodos de VADER, AFFIN Y TEXTBLOB, se procedió a realizar las respectivas predicciones aplicando modelos de series temporales: autoregressive integrated moving average (ARIMA), Seasonal-Trend decomposition using LOESS (STL) y HOLT-WINTERS.

Holt – Winters

Se escogió este modelo para realizar las predicciones porque nos entrega la ventaja de ajustarse a los datos. El procedimiento Holt- Winters se considera como una expansión del procedimiento Holt que da lugar a la estimación de 2 exponentes suavizantes. Estima grado, tendencia y estacional de una cierta serie de tiempos. Este procedimiento contiene 2 modelos primordiales que dependen del tipo de estacionalidad; el modelo multiplicativo estacional y el modelo aditivo estacional [65] .

Positivos

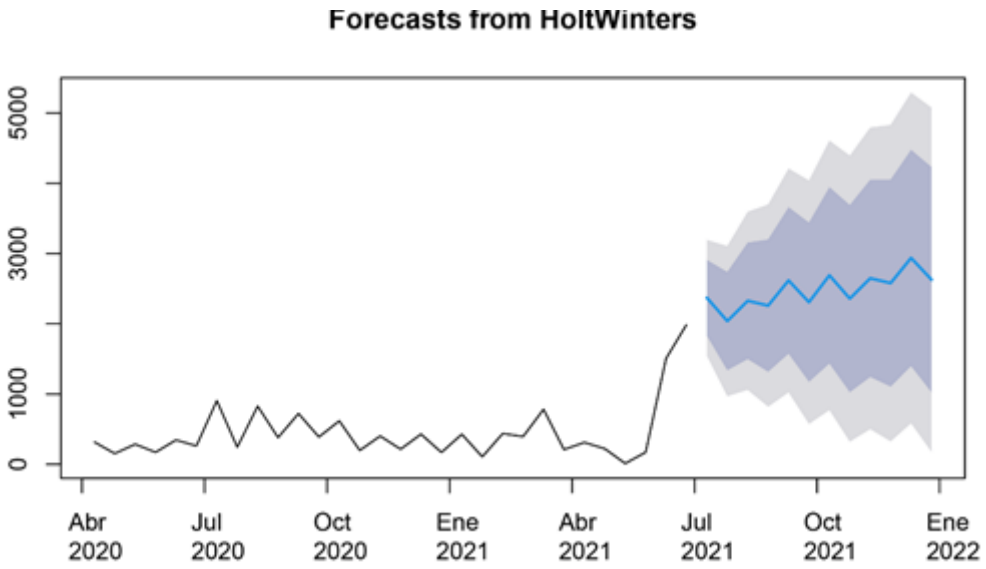

Figura 20: Predicción de cantidad de tweets Positivos modelo Holt-Winters

Fuente: Elaboración Propia

Podemos visualizar en la Figura 20 que, debido a eventos favorecedores y a la aplicación de las vacunas en contra al COVID-19 los sentimientos positivos han llegado a sus niveles más altos. Analizando el modelo de Holt-Winters, la cantidad de sentimientos positivos no será constante e incrementará.

Negativos

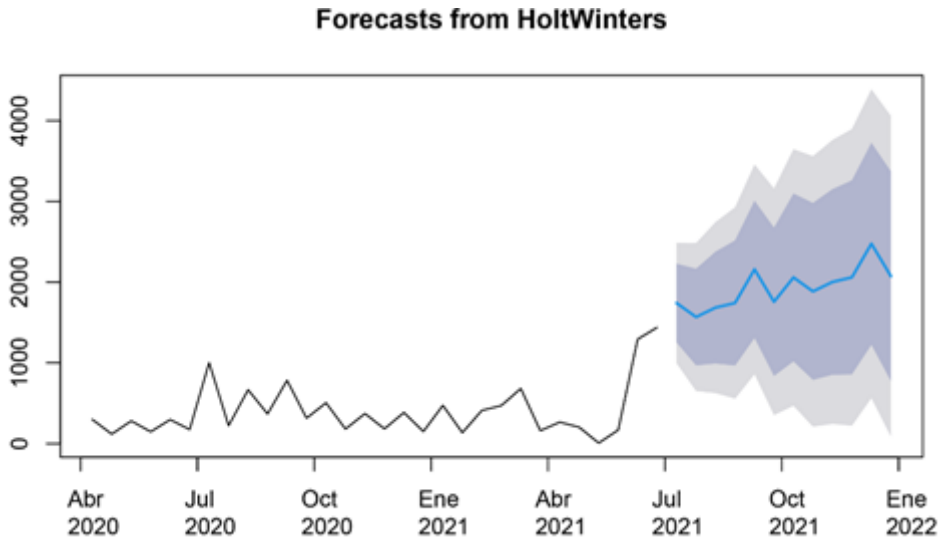

Figura 21: Predicción de cantidad de tweets Negativos modelo Holt -Winters

Fuente: Elaboración Propia

En la Figura 21 podemos visualizar que existe una tendencia a disminuir los sentimientos negativos sin embargo también existen variaciones en los demás periodos en los que se analizó.

Seasonal-Trend decomposition using LOESS (STL) + Autoregressive integrated moving average (ARIMA)

Positivos

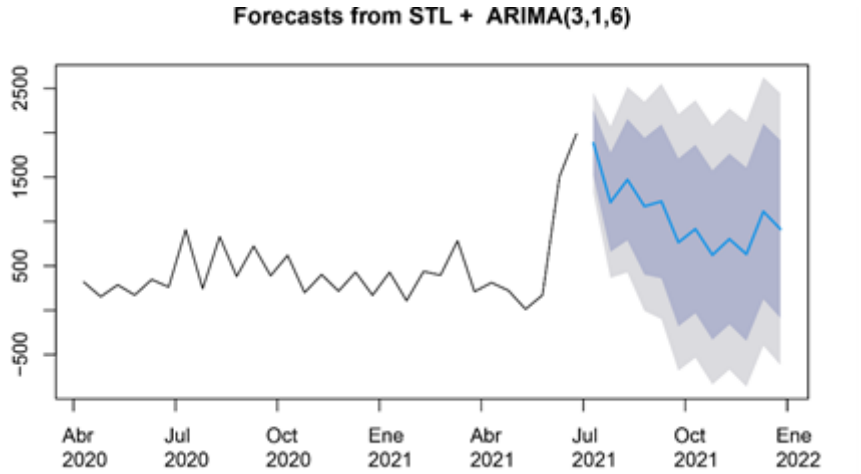

Figura 22: Predicción de cantidad de tweets Positivos modelo STL

Fuente: Elaboración propia

Podemos visualizar en la Figura 22 que, debido a eventos favorecedores y el plan de vacunación 9/100 en contra al COVID-19 los sentimientos positivos han llegado a sus niveles más altos. Analizando el modelo de STL + ARIMA, la cantidad de sentimientos positivos no será invariable y disminuirá.

Negativos

Como podemos visualizar en la Figura 23 existe una tendencia a disminuir la cantidad de sentimientos negativos existiendo también variaciones en los demás periodos de tiempo en los que se analizó.

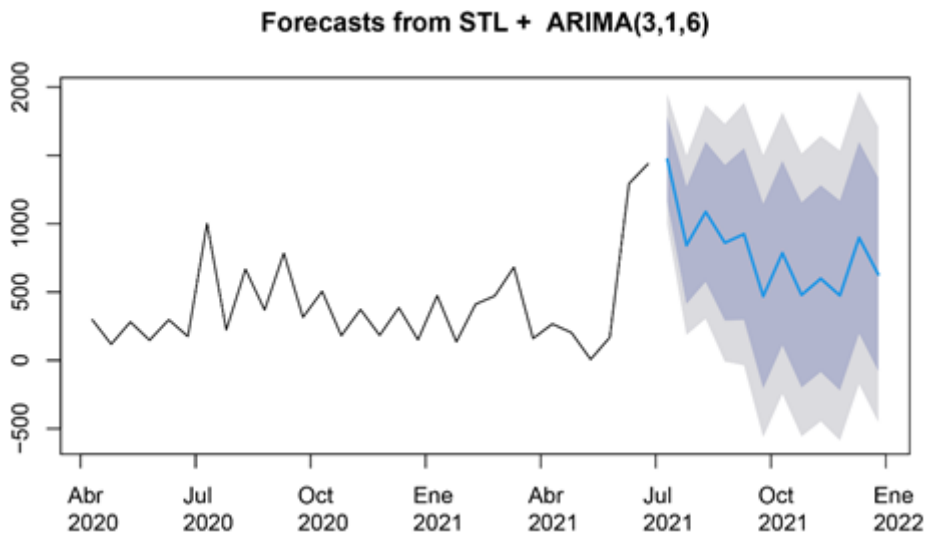

Figura 23: Predicción de cantidad de tweets Negativos modelo STL

Fuente: Elaboración Propia

CAPÍTULO III. EVALUACIÓN DEL PROTOTIPO

3.1 Plan de evaluación

Una vez procesada la data, se continúa con la elaboración de la muestra que se dio origen por medio del 30% de la data preprocesada, lo cual equivale a 10958 tweets; a continuación, se procedió a realizar el análisis manual, por consiguiente, se les asigno el siguiente puntaje: –1 para sentimiento negativos, 0 para sentimiento neutros y 1 para sentimiento positivos, luego se realizó el análisis de sentimiento utilizando los métodos: VADER, TEXTBLOB Y AFFIN arrojando el siguiente resultado:

	comentario	Sentimiento	afinn_sen	vader_sen	textblob_sen
0	el riesgo de coagulos de la vacuna astrazeneca...	-1	0	-1	1
1	hoy una chica hospitalizada por hematemesis y ...	-1	0	-1	-1
2	empresa privada se suma al plan de vacunacion ...	1	0	0	0
3	estimado doc Esteban Ortiz tal vez hay alguna ...	0	-1	1	1
4	para este domingo 11 de julio se seguira admin...	1	0	-1	0

Figura 24: Clasificación por Método aplicado a la Muestra

La clasificación de sentimientos analizados se realizó de la siguiente manera: los valores entre –1 y menores que cero son negativos los equivalentes a 0 son neutros y mayores que cero positivos dado esto se los procedió a clasificar de la misma manera que la puntuación manual.

3.1.1 Matriz de Confusión

La matriz de confusión y sus estadísticas similares son un instrumento bien establecido en el aprendizaje automático para evaluar la exactitud de un clasificador [66] .

El procedimiento para realizar la matriz de confusión es el siguiente:

- Elija una de ambas clases como positiva y la otra como negativa
- Esta elección cambia de un dominio a otro y es subjetiva para un sujeto o los equipamientos involucrado.
- Los datos de la realidad elemental (el grupo de datos) ahora tienen la posibilidad de clasificar en: positivos reales y negativos reales

· Las predicciones del modelo ahora tienen la posibilidad de clasificar en: predicciones positivas y predicciones negativas

· Desde dichos conjuntos de puntos de vista de datos, tienen la posibilidad de decidir los puntos de vista en común para conformar una matriz de confusión [67] .

En la tabla 3 se representa el formato de la matriz de confusión utilizada.

Tab

		Predicción		
		Comentarios Positivo	Comentario Negativo	Comentario Neutro
Realidad	Comentarios Positivo	TP	FP 1	FP 2
	Comentario Negativo	FN 1	TN	FN 2
	Comentario Neutro	FNE 1	FNE 2	TNE

la 3: Representación de Valores para Métricas de Evaluación

Para el cálculo de las métricas de evaluación se tomó en cuenta cada una de las categorías de la matriz de confusión, las cuales se abrevian de la siguiente manera:

TP: Verdaderos Positivos

TN: Verdaderos Negativos

TNE: Verdaderos Neutros

FP1: Falsos Positivos 1

FP2: Falsos Positivos 2

FN1: Falsos Negativos 1

FN2: Falsos Negativos 2

FNE1: Falsos Neutros 1

FNE2: Falsos Neutros 2

Total = (TP + TN + TNE + FP1+ FP2 + FN1 + FN2 +FNE1 + FNE2)

3.1.1.1 Métricas de Rendimiento

Exactitud (Accuracy)

Es el porcentaje de instancias del grupo de datos de prueba que son clasificados de manera correcta, se puede calcular tomando el promedio de los valores que se hallan en la diagonal primordial de la matriz de confusión [68] .

$$AC = \frac{\text{total de clasificación correcta}}{total}$$

Sensibilidad o Exhaustividad (Recall)

Es una medida que nos ayuda a conocer el porcentaje de casos positivos que permanecen de manera correcta clasificados. En el modelo completo, la convocatoria es 1 para cada clase. Analíticamente, los estudiosos tenían como fin incrementar la memoria sin alterar el costo de exactitud [69] .

$$RC = \frac{\text{clasificación correcta}}{\text{categoría actual}}$$

Precisión (Precision)

Mide como de exactas son las predicciones de los modelos por medio de un porcentaje de predicciones primordiales [70] .

$$P = \frac{prediciones\ correctas}{total\ de\ Prediciones}$$

F1-Score (Valor F)

Se define como una media armónica de Precisión y Recall esta métrica tiene la capacidad de tener en cuenta tanto los falsos positivos como los falsos negativos, y así, un F1-Score de 1 indicará una precisión y un Recall perfectos, y, por lo tanto, un clasificador perfecto al momento de discriminar las clases. Si bien la intuición de esta métrica es un poco más compleja de entender que el Accuracy, Precisión y el Recall, esta resulta realmente útil en los casos donde existen clases distribuidas de forma desigual, para la aplicación del método de clasificación de Machine Learning, como ocurre en el caso de este [71] .

$$\boldsymbol{MF} = \frac{2 * RC * P}{RC + P}$$

3.2 Resultados de la evaluación

Para calcular la matriz de confusión se procedió a utilizar la librería de Python sklearn específicamente confusion_matrix el cual se lo realizó mediante el siguiente script:

```
def matriz_confusion(metodo_sentimiento):
    new_confusion_matrix=[]
    fn,fne,fp,ftn,ftne,ftp,tn,tne,tp = confusion_matrix(sentimientos,metodo_sentimiento).ravel()
    new_confusion_matrix.append([tp,tn,tne])
    new_confusion_matrix.append([fp,fn,fne])
    new_confusion_matrix.append([ftp,ftn,ftne])
    return new_confusion_matrix
```

Código 3: Script de Generación de Matriz de Confusión

Posteriormente se procedió calcular para los diferentes métodos mediante el siguiente código:

```
new_confusion_matrix_a = matriz_confusion(affin)
new_confusion_matrix_v = matriz_confusion(vader)
new_confusion_matrix_tb = matriz_confusion(textblob)
```

Código 4: Script de Arreglo para Matriz de Confusión

Una vez realizada la matriz de confusión se obtuvieron los siguientes resultados:

AFFIN

	Comentarios Positivos	Comentarios Negativos	Comentarios Neutros
Comentarios Positivos	219	322	2773
Comentarios Negativos	170	635	3760
Comentarios Neutros	142	275	2662

Figura 25: Matriz de Confusión de Método AFFIN

Los resultados generados por el método AFFIN podemos observar que existe mayor cantidad de comentarios Negativos clasificados como Comentarios Neutros.

VADER

	Comentarios Positivos	Comentarios Negativos	Comentarios Neutros
Comentarios Positivos	2298	396	620
Comentarios Negativos	833	3075	657
Comentarios Neutros	993	167	1919

Figura 26: Matriz de Confusión de Método VADER

Como podemos visualizar en el método de VADER que existe una mayor cantidad de Comentarios Negativos clasificados como Comentarios Negativos es decir a simple vista se muestra una mejor precisión en los resultados analizados por Vader.

TEXTBLOB

	Comentarios Positivos	Comentarios Negativos	Comentarios Neutros
Comentarios Positivos	1817	360	1137
Comentarios Negativos	1880	1098	1587
Comentarios Neutros	1254	302	1523

Figura 27: Matriz de Confusión de Método TEXTBLOB

Podemos visualizar que existe mayor cantidad de Comentarios Negativos clasificados como Comentarios Positivos por lo que existe una pequeña diferencia en cuanto al análisis realizado por VADER.

3.2.1 Métricas de evaluación

Para la realización de las métricas de evaluación se tomó en cuenta cada una de las categorías de la matriz de confusión (ver tabla 3)

Precisión (PRECISION)

```
def presision(df):
    #Total de Columnas

    p= sumaa = df['Comentarios Positivos'].sum()
    n = df['Comentarios Negativos'].sum()
    ne= sumac = df['Comentarios Neutros'].sum()


    pf  = df.iloc[:,0:3].sum(axis=1)['Comentarios Positivos']
    nf  = df.iloc[:,0:3].sum(axis=1)['Comentarios Negativos']
    nef = df.iloc[:,0:3].sum(axis=1)['Comentarios Neutros']

    #Prediciones corectas

    tp  = df['Comentarios Positivos']['Comentarios Positivos']
    tn  = df['Comentarios Negativos']['Comentarios Negativos']
    tne = df['Comentarios Neutros']['Comentarios Neutros']

    pp  = tp/p
    png  = tn/n
    pnt = tne/ne

    #Peso

    total = p+n+ne
    pr =  (pf/total)* pp + (nf/total) *png + (nef/total)*pnt

    return pr
```

Código 5: Script de Cálculo de Precisión

En el siguiente script basándose en la fórmula general de precisión se realizaron las siguientes funciones:

$$P_P = \frac{\text{TP}}{\text{TP} + \text{FN1} + \text{NE1}}$$

$$P_{NG} = \frac{\text{TN}}{\text{TN} + \text{FP1} + \text{FNE 2}}$$

$$P_{NT} = \frac{\text{TNE}}{TNE + FP2 + FN2}$$

$$P = \left(\frac{(TP + FP1 + FP2)}{Total}\right) * P_P + \left(\frac{(FN1 + TN + FN2)}{Total}\right) * P_{NG} + \left(\frac{(FNE1 + FNE1 + TNE)}{Total}\right) * P_{NT}$$

Sensibilidad (RECALL)

```python
def recall(df):

    p  = df.iloc[:,0:3].sum(axis=1)['Comentarios Positivos']
    n  = df.iloc[:,0:3].sum(axis=1)['Comentarios Negativos']
    ne = df.iloc[:,0:3].sum(axis=1)['Comentarios Neutros']

    #Prediciones corectas

    tp  = df['Comentarios Positivos']['Comentarios Positivos']
    tn  = df['Comentarios Negativos']['Comentarios Negativos']
    tne = df['Comentarios Neutros']['Comentarios Neutros']

    rcp  = tp/p
    rcng  = tn/n
    rcnt = tne/ne

    #Peso

    total = p+n+ne
    rc = (p/total)* rcp + (n/total) *rcng + (ne/total)*rcnt


    return rc
```

Código 6: Script de Cálculo de Sensibilidad

Para la realización del cálculo de la sensibilidad se realizaron las siguientes fórmulas basándonos en la fórmula general de sensibilidad:

$$RC_P = \frac{TP}{TP + FN1 + FNE1}$$

$$RC_{NG} = \frac{TN}{TN + FN1 + FN2}$$

$$RC_{NT} = \frac{TNE}{TNE + FNE1 + FNE2}$$

$$RC = \left(\frac{(\mathrm{TP} + \mathrm{FP1} + \mathrm{FP2})}{Total}\right) * RC_P + \left(\frac{((\mathrm{FN1} + \mathrm{TN} + \mathrm{FN2})}{Total}\right) * RC_{NG} + \left(\frac{(\mathrm{FNE1} + \mathrm{FNE1} + \mathrm{TNE})}{Total}\right) * RC_{NT}$$

Exactitud (ACCURACY)

```
def acuracy(df):
    tp = df['Comentarios Positivos']['Comentarios Positivos']
    fp1 = df['Comentarios Positivos']['Comentarios Negativos']
    fp2 = df['Comentarios Positivos']['Comentarios Neutros']

    fn1 = df['Comentarios Negativos']['Comentarios Positivos']
    tn = df['Comentarios Negativos']['Comentarios Negativos']
    fn2 = df['Comentarios Negativos']['Comentarios Neutros']

    fne1 = df['Comentarios Neutros']['Comentarios Positivos']
    fne2 = df['Comentarios Neutros']['Comentarios Negativos']
    tne= df['Comentarios Neutros']['Comentarios Neutros']

    total=tp+tn+tne+fp1+fp2+fn1+fn2+fne1+fne2
    t= tp+tn+tne
    ac= t/total
    return ac
```

Código 7: Script de Cálculo de Exactitud

Para la realización del cálculo de exactitud se realizó mediante la siguiente fórmula:

$$AC = \frac{(TP + TN + TNE)}{(TP + TN + TNE + FP1 + FP2 + FN1 + FN2 + FNE1 + FNE2)}$$

MEDIDA F

```
def mf(df):
    mf=(2*recall(df)*presision(df))/(recall(df)+presision(df))
    return mf
```

Código 8: Script de Cálculo de Medida F

Una vez realizado los cálculos correspondientes se procedió a almacenarlos en una nueva matriz la cual es la siguiente:

	RECALL	PRECISION	ACCURACY	MEDIDA F
VADER	0.665450	0.689354	0.665450	0.677191
AFFIN	0.320861	0.420796	0.320861	0.364096
TEXBLOB	0.405001	0.471647	0.405001	0.435791

Figura 28: Métricas de Evaluación de Métodos de Análisis de Sentimientos

En la matriz de la Figura 28, observamos que el método de VADER es el más eficaz ya que posee exactitud de 0.665, una precisión de 0.6893, seguido del método de TEXTBLOB con una exactitud del 0.4050 y una precisión de 0.4716 y el método menos eficaz es AFFIN con una exactitud de 0.3208 y una precisión de 0.4716.

Tabla 5: Métricas de Acurracy de Método Holt-Winters

		Métricas						
		ME	**RMSE**	**MAE**	**MPE**	**MAPE**	**MASE**	**ACF1**
Sentimiento	**Positivo**	29.0592	412.1695	304.2094	-205.3801	278.2857	0.9548819	0.2251837

	Negativo	4.945996	372.1692	260.1937	-268.0051	320.8618	0.86647	0.2042295

En la matriz de la Tabla 5, analizamos que existe un valor MASE (error en escala absoluta media) de 0.955 para positivos y 0.866 para negativos.

Tabla 6: Métricas de Acurracy de Método STL ARIMA

		Métricas						
		ME	**RMSE**	**MAE**	**MPE**	**MAPE**	**MASE**	**ACF1**
Sentimiento	**Positivo**	44.526559	222.0687	188.4818	-31.74285	88.00715	0.5916248	-0088201003
	Negativo	44.526559	222.0687	188.4818	-31.74285	88.00715	0.5916248	-0088201003

En la matriz de la Tabla 6, analizamos que el valor MASE para el método de STL ARIMA en la predicción de la cantidad de comentarios positivos en el contexto del COVID-19 es de 0.591 tanto para positivos como para negativos.

En conclusión, se puede destacar que el método con menor MASE es STL-ARIMA, es más eficiente al momento de realizar las predicciones de cantidades de comentarios positivos y negativos en el contexto del COVID-19.

3.3 Conclusiones

- La recolección de los datos de la red social Twitter es un poco más sencillo de realizar ya que se puede obtener las credenciales de utilización del API de manera rápida y se lo debe de realizar mediante un script que vaya recolectando

automáticamente cada día los Tweets y en caso de los más recientes sólo se los puede obtener datos de los últimos 15 días, esto facilitó en el proceso de desarrollo e implementación del trabajo.

- La recolección de los datos más antiguos se los realizó mediante la utilización de un data set denominado Narrativas del COVID-19 aportando en desarrollo de la data utilizada en el presente estudio.
- Para la realización del preprocesamiento primero se realizó la utilización de scripts de limpieza en Python para posteriormente eliminar frases innecesarias manualmente y generar las nubes de palabras que están divididas por periodos con el objetivo de determinar el comportamiento de las personas durante el periodo de la pandemia.
- Se realizó el análisis de los sentimientos mediante la utilización de los métodos de VADER, AFFIN y TEXTBLOB en el contexto de COVID-19 en Ecuador; el método VADER presenta los mejores resultados siendo el más eficiente que los demás métodos analizados.
- Se realizó modelos de predicción mediante el lenguaje R, aplicando series temporales como: Autoregressive Integrated Moving Average (ARIMA), Seasonal-Trend decomposition using LOESS (STL) y HOLT-WINTERS; según estos modelos, hay una tendencia a disminuir los comentarios en las redes sociales, en base a que a que el COVID-19 también va disminuyendo. El método con menor MASE es STL-ARIMA, y es más eficiente al momento de realizar las predicciones de cantidades de comentarios positivos y negativos en el contexto del COVID-19 en función del tiempo.

3.4 Recomendaciones

El Análisis de sentimientos es recomendable para el uso de procedimientos en los cuales exista la necesidad de determinar la polaridad de los sentimientos mediante el análisis automático de los comentarios en redes sociales o de páginas web de venta de productos o servicios al cliente. SA es de gran importancia para identificar los sentimientos acerca de un producto, servicio o situación social, económica, política, religiosa entre otras.

BIBLIOGRAFÍA

[1] S. D. Páez Juka, “Análisis comparativo de herramientas Open Source para Data Mining sobre datos públicos del Ministerio de Educación de la República del Ecuador”, Pontificia Universidad Católica del Ecuador, 2019. Consultado: ago. 19, 2021. [En línea]. Disponible en: http://repositorio.puce.edu.ec/handle/22000/17060

[2] B. Mazon-Olivo, W. Rivas-Asanza, J. Novillo-Vicuña, y C. Flores-Cabrera, “Análisis de producción avícola mediante técnicas de inteligencia de negocios y minería de datos”, *Alternativas*, vol. 19, núm. 2, pp. 80–88, ago. 2019, doi: 10.23878/alternativas.v19i2.203.

[3] B. Mazon-Olivo, M. Pinta, y F. Redrovan, “Desarrollo de competencias en Minería de Datos, una experiencia didáctica”, en *Sistematización de experiencias educativas innovadoras*, 1a ed., Universidad Técnica de Machala, 2020, pp. 383–406. [En línea]. Disponible en: http://repositorio.utmachala.edu.ec/handle/48000/15219

[4] B. Mazón-Olivo, M. Jaramillo-Paredes, O. Romero-Hidalgo, A. Borja-Herrera, M. Aguirre-Benalcazar, y M. Contento-Segarra, “Business Intelligence and Data Mining Technologies for the analysis of cocoa production and commercialization”, *Espacios*, vol. 39, núm. 32, p. 6, 2018.

[5] I. Ramírez-Morales, B. Mazon-Olivo, y A. Pan, “Ciencia de datos en el sector agropecuario”, en *Análisis de Datos Agropecuarios*, 1a ed., I. Ramírez-Morales y B. Mazón-Olivo, Eds. Machala-Ecuador: Universidad Técnica de Machala, 2018, pp. 12–44. [En línea]. Disponible en: http://repositorio.utmachala.edu.ec/handle/48000/13324

[6] C. Valdiviezo-Abad y T. Bonini, “Uso de big data y data mining en los procesos de automatización de la comunicación de las organizaciones”, *GIGAPP Estud. Work. Pap.*, vol. 8, núm. 190–212, Art. núm. 190–212, feb. 2021.

[7] H. Benhar, A. Idri, y J. L. Fernández-Alemán, “Data preprocessing for heart disease classification: A systematic literature review”, *Comput. Methods Programs Biomed.*, vol. 195, p. 105635, oct. 2020, doi: 10.1016/j.cmpb.2020.105635.

[8] I. Lauriola, A. Lavelli, y F. Aiolli, “An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools”, *Neurocomputing*, jul. 2021, doi: 10.1016/j.neucom.2021.05.103.

[9] J. C. Sobrino, “Análisis de sentimientos en Twitter”, Universidad Oberta de Cataluña, 2018.

[10] A. Lozano-Vargas, “Impacto de la epidemia del Coronavirus (COVID-19) en la salud mental del personal de salud y en la población general de China”, *Rev. Neuro-Psiquiatr.*, vol. 83, núm. 1, pp. 51–56, ene. 2020, doi: 10.20453/rnp.v83i1.3687.

[11] M. Taquet, J. Quoidbach, E. I. Fried, y G. M. Goodwin, “Mood Homeostasis Before and During the Coronavirus Disease 2019 (COVID-19) Lockdown Among Students in the Netherlands”, *JAMA Psychiatry*, vol. 78, núm. 1, pp. 110–112, ene. 2021, doi: 10.1001/jamapsychiatry.2020.2389.

[12] “Emociones, preocupaciones y reflexiones frente a la pandemia del COVID-19 en Argentina”, *Ciênc. Saúde Coletiva*, vol. 25, pp. 2447–2456, jun. 2020, doi: 10.1590/1413-81232020256.1.10472020.

[13] V. Zamora y D. Alonso, “La importancia de la estadística aplicada para la toma de decisiones en Marketing”, *Rev. Investig. Negocios*, vol. 12, núm. 20, pp. 31–44, oct. 2019.

[14] M. Birjali, M. Kasri, y A. Beni-Hssane, “A comprehensive survey on sentiment analysis: Approaches, challenges and trends”, *Knowl.-Based Syst.*, vol. 226, p. 107134, ago. 2021, doi: 10.1016/j.knosys.2021.107134.

[15] S. Zervoudakis, E. Marakakis, H. Kondylakis, y S. Goumas, “OpinionMine: A Bayesian-based framework for opinion mining using Twitter Data”, *Mach. Learn. Appl.*, vol. 3, p. 100018, mar. 2021, doi: 10.1016/j.mlwa.2020.100018.

[16] R. Singh y R. Singh, “Applications of sentiment analysis and machine learning techniques in disease outbreak prediction – A review”, *Mater. Today Proc.*, may 2021, doi: 10.1016/j.matpr.2021.04.356.

[17] D. T. Huerta, J. Hawkins, J. Brownstein, y Y. Hswen, “Exploring discussions of health and risk and public sentiment in MA during COVID-19 pandemic mandate implementation: A Twitter analysis”, *SSM - Popul. Health*, vol. 15, p. 100851, jun. 2021, doi: 10.1016/j.ssmph.2021.100851.

[18] R. F. Roca, “Pandemia por COVID-19: el mayor reto de la historia del intensivismo”, *Med. Intensiva*, vol. 44, núm. 6, Art. núm. 6, 2020.

[19] L. E. Fernández-Garza y A. Marfil, “Neurological aspects that should not be forgotten during the COVID-19 pandemic”, *Interam. J. Med. Health*, vol. 3, abr. 2020, doi: 10.31005/iajmh.v3i0.89.

[20] “Covid ecuador”, *Google*, sep. 16, 2021. https://www.google.com/search?q=covid+ecuador (consultado sep. 19, 2021).

[21] “Ecuador Coronavirus (Live): 505.860 Cases”, *The Coronavirus App*. https://coronavirus.app/tracking/ecuador (consultado sep. 19, 2021).

[22] A. F. G. Viera y A. F. G. Viera, “Técnicas de aprendizaje de máquina utilizadas para la minería de texto”, *Investig. Bibl.*, vol. 31, núm. 71, pp. 103–126, abr. 2017, doi: 10.22201/iibi.0187358xp.2017.71.57812.

[23] N. Chintalapudi, G. Battineni, M. D. Canio, G. G. Sagaro, y F. Amenta, “Text mining with sentiment analysis on seafarers’ medical documents”, *Int. J. Inf. Manag. Data Insights*, vol. 1, núm. 1, p. 100005, abr. 2021, doi: 10.1016/j.jjimei.2020.100005.

[24] M. C. Barrera, “Minería de texto en la clasificación de documentos digitales”, *Biblios J. Librariansh. Inf. Sci.*, núm. 64, Art. núm. 64, nov. 2016, doi: 10.5195/biblios.2016.309.

[25] “Acerca de la minería de textos”, mar. 02, 2021. https://prod.ibmdocs-production-dal-6099123ce774e592a519d7c33db8265e-0000.us-south.containers.appdomain.cloud/docs/es/spss-modeler/SaaS?topic=guide-introduction-crisp-dm (consultado ago. 10, 2021).

[26] Jomayra Ramírez Figuereo, “Aplicación de Técnicas de Minería de Textos en Inteligencia de Clientes”, Trabajo Fin de M´aster, Universidad de Coruña, España, 2018. [En línea]. Disponible en: http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1575.pdf

[27] C. Torres López y L. Arco García, “Representación textual en espacios vectoriales semánticos”, *Rev. Cuba. Cienc. Informáticas*, vol. 10, núm. 2, pp. 148–180, jun. 2016.

[28] Julio Alexander Bernal-Chávez y Ruth Yanira Rubio López, *Introducción a la lingüística computacional*, 1a.ed. Bogota: Ediciones de la U, 2016.

[29] R. M. Suleman y I. Korkontzelos, “Extending latent semantic analysis to manage its syntactic blindness”, *Expert Syst. Appl.*, vol. 165, p. 114130, mar. 2021, doi: 10.1016/j.eswa.2020.114130.

[30] N. Li, W. Luo, K. Yang, F. Zhuang, Q. He, y Z. Shi, “Self-organizing weighted incremental probabilistic latent semantic analysis”, *Int. J. Mach. Learn. Cybern.*, vol. 9, núm. 12, pp. 1987–1998, dic. 2018, doi: 10.1007/s13042-017-0681-9.

[31] Dra.Luz Marina Barreto, “Estudio de dos Paradigmas de Modelado de Tópicos en un Corpus de Documentos Tomados de una Red Social”, Maestria, Universidad Central de Venezuela, Caracas-Venezuela, 2018. Consultado: jun. 13, 2021. [En línea]. Disponible en: http://hdl.handle.net/10872/20686

[32] María Silvestre Gómez, “Implementación de Asignación Jerárquica Latente de Dirichlet para Modelado de Temas”, Universidad de Sevilla, Sevilla, 2018.

[33] Hammoe, Luciano, “Detección de tópicos utilizando el modelo LDA”, Postgrado, Instituto Tecnológico de Buenos Aires – ITBA, BUENOS AIRES, 2018. Consultado: jun. 13, 2021. [En línea]. Disponible en: https://ri.itba.edu.ar/bitstream/handle/123456789/1250/TFI_Hammoe.pdf?sequence=1&isAllowed=y

[34] L. Talamé, A. Cardoso, y M. Amor, “Comparación de herramientas de procesamiento de textos en español extraídos de una red social para Python”, Salta, sep. 2019, pp. 53–67. [En línea]. Disponible en: http://sedici.unlp.edu.ar/handle/10915/87854

[35] Bernabé González Isabel Ivagnes Joaquín Lejtreger, “Construcción de herramientas para soporte a la enseñanza de lenguas”, Postgrado, Universidad de la Republica de Urugay, Urugay, 2019. [En línea]. Disponible en: https://www.colibri.udelar.edu.uy/jspui/bitstream/20.500.12008/20585/1/tg-gonzales-ivagnes-lejtreger.pdf

[36] “Better Language Models and Their Implications”, *OpenAI*, feb. 14, 2019. https://openai.com/blog/better-language-models/ (consultado ago. 20, 2021).

[37] D. M. Korngiebel y S. D. Mooney, “Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery”, *Npj Digit. Med.*, vol. 4, núm. 1, pp. 1–3, jun. 2021, doi: 10.1038/s41746-021-00464-x.

[38] V. N. Gudivada y K. Arbabifard, “Chapter 3 - Open-Source Libraries, Application Frameworks, and Workflow Systems for NLP”, en *Handbook of Statistics*, vol. 38, V. N. Gudivada y C. R. Rao, Eds. Elsevier, 2018, pp. 31–50. doi: 10.1016/bs.host.2018.07.007.

[39] J. G. M. Reyes, “Procesamiento de Lenguaje Natural y su aplicación en servicios de hostelería”, Trabajo de Fin de Grado, Universidad de La Laguna, Latacunga, 2016. [En línea]. Disponible en: http://riull.ull.es/xmlui/handle/915/2912

[40] E. Nazaruka, J. Osis, y V. Griberman, “Extracting Core Elements of TFM Functional Characteristics from Stanford CoreNLP Application Outcomes”, en *Topological UML modeling: an improved approach for domain analysis and software development*, Czech Republic, ene. 2019, pp. 591–602. doi: 10.5220/0007831605910602.

[41] Aitor Soroa, German Rigau, Jordi Porta, Jordi Atserias, Xavier Gómez Guinovart, y Horacio Saggion, “Plataformas y sistemas de procesamiento lingüístico de alto rendimiento.”, Ministerio de Energia, turismo y Agenda digital de España, Madrid, jun. 2017.

[42] T. Daudert, “Exploiting textual and relationship information for fine-grained financial sentiment analysis”, *Knowl.-Based Syst.*, vol. 230, p. 107389, ago. 2021, doi: 10.1016/j.knosys.2021.107389.

[43] R. K. Behera, M. Jena, S. K. Rath, y S. Misra, “Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data”, *Inf. Process. Manag.*, vol. 58, núm. 1, p. 102435, ene. 2021, doi: 10.1016/j.ipm.2020.102435.

[44] W. Medhat, A. Hassan, y H. Korashy, “Sentiment analysis algorithms and applications: A survey”, *Ain Shams Eng. J.*, vol. 5, núm. 4, pp. 1093–1113, dic. 2014, doi: 10.1016/j.asej.2014.04.011.

[45] “Artificial intelligence: New age of transformation in petroleum upstream”, *Pet. Res.*, jul. 2021, doi: 10.1016/j.ptlrs.2021.07.002.

[46] L. Vanfretti y V. S. N. Arava, “Decision tree-based classification of multiple operating conditions for power system voltage stability assessment”, *Int. J. Electr. Power Energy Syst.*, vol. 123, p. 106251, dic. 2020, doi: 10.1016/j.ijepes.2020.106251.

[47] R. Gove y J. Faytong, “Chapter 4 - Machine Learning and Event-Based Software Testing: Classifiers for Identifying Infeasible GUI Event Sequences”, en *Advances in Computers*, vol. 86, A. Hurson y A. Memon, Eds. Elsevier, 2012, pp. 109–135. doi: 10.1016/B978-0-12-396535-6.00004-1.

[48] Y. Goldberg, “Neural Network Methods for Natural Language Processing”, *Synth. Lect. Hum. Lang. Technol.*, vol. 10, núm. 1, pp. 1–309, abr. 2017, doi: 10.2200/S00762ED1V01Y201703HLT037.

[49] W. B. Zulfikar, M. Irfan, C. N. Alam, y M. Indra, “The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter”, en *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, ago. 2017, pp. 1–5. doi: 10.1109/CITSM.2017.8089231.

[50] R. Ojha, A. Ghadge, M. K. Tiwari, y U. S. Bititci, “Bayesian network modelling for supply chain risk propagation”, *Int. J. Prod. Res.*, vol. 56, núm. 17, pp. 5795–5819, sep. 2018, doi: 10.1080/00207543.2018.1467059.

[51] T. Rabczuk, J.-H. Song, X. Zhuang, y C. Anitescu, “Chapter Five - Extended meshfree methods”, en *Extended Finite Element and Meshfree Methods*, T. Rabczuk, J.-H. Song, X. Zhuang, y C. Anitescu, Eds. Academic Press, 2020, pp. 161–313. doi: 10.1016/B978-0-12-814106-9.00011-5.

[52] R. Kamal, M. A. Shah, C. Maple, M. Masood, A. Wahid, y A. Mehmood, “Emotion Classification and Crowd Source Sensing; A Lexicon Based Approach”, *IEEE Access*, vol. 7, pp. 27124–27134, 2019, doi: 10.1109/ACCESS.2019.2892624.

[53] S. Taj, B. B. Shaikh, y A. Fatemah Meghji, “Sentiment Analysis of News Articles: A Lexicon based Approach”, en *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, ene. 2019, pp. 1–5. doi: 10.1109/ICOMET.2019.8673428.

[54] H. Newman y D. Joyner, “Sentiment Analysis of Student Evaluations of Teaching”, en *Artificial Intelligence in Education*, Cham, 2018, pp. 246–250. doi: 10.1007/978-3-319-93846-2_45.

[55] A. Vizcaino-Verdu y I. Aguaded, “Análisis de sentimiento en Instagram: polaridad y subjetividad de cuentas infantiles”, *ZER Rev. Estud. Comun. Komunikazio Ikasketen Aldizkaria*, vol. 25, núm. 48, Art. núm. 48, may 2020, doi: 10.1387/zer.21454.

[56] E. E. Condor-Tinoco, J. A. Rojas-Cusi, A. Zevallos-Rodríguez, y C. Y. Castro-Buleje, “Minería de datos: análisis de sentimiento en Twitter basado en lexicones sobre el uso de dióxido de cloro para el tratamiento del COVID-19”, *Repos. Inst. - Ulima*, 2021, Consultado: ago. 24, 2021. [En línea]. Disponible en: https://hdl.handle.net/20.500.12724/13898

[57] S. Salinas y J. Ulises, “Análisis de sentimientos en los mensajes recibidos en el entorno virtual de aprendizaje de la modalidad abierta y a distancia de la UTPL /”, Trabajo de Titulación de Magíster en Ciencias y Tecnologías de la Computación, UTPL, Loja, 2020. Consultado: ago. 08, 2021. [En línea]. Disponible en: http://dspace.utpl.edu.ec/handle/20.500.11962/26503

[58] K. Z. Aung y N. N. Myo, “Sentiment analysis of students’ comment using lexicon based approach”, en *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, may 2017, pp. 149–154. doi: 10.1109/ICIS.2017.7959985.

[59] C. Schröer, F. Kruse, y J. M. Gómez, “A Systematic Literature Review on Applying CRISP-DM Process Model”, *Procedia Comput. Sci.*, vol. 181, pp. 526–534, ene. 2021, doi: 10.1016/j.procs.2021.01.199.

[60] R. Nishizaki Fernandes, “Análisis de datos sanitarios aplicando metodología CRISP-DM”, Trabajo de fin de grado, Universidad Autónoma de Madrid, Madrid, 2017. Consultado: jun. 13, 2021. [En línea]. Disponible en: https://repositorio.uam.es/handle/10486/680056

[61] marktab, “What is the Team Data Science Process? - Azure Architecture Center”. https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview (consultado ago. 10, 2021).

[62] I. Karamitsos, S. Albarhami, y C. Apostolopoulos, “Applying DevOps Practices of Continuous Automation for Machine Learning”, *Information*, vol. 11, núm. 7, Art. núm. 7, jul. 2020, doi: 10.3390/info11070363.

[63] marktab, “¿Qué es el Proceso de ciencia de datos en equipo (TDSP)? - Azure Architecture Center”. https://docs.microsoft.com/es-es/azure/architecture/data-science-process/overview (consultado ago. 20, 2021).

[64] “Información sobre las API de Twitter”. https://help.twitter.com/es/rules-and-policies/twitter-api (consultado jul. 02, 2021).

[65] E. J. M. Vásquez y S. G. Chávez, “Predicción del consumo de energía eléctrica residencial de la Región Cajamarca mediante modelos Holt -Winters”, *Ing. Energética*, vol. XL, núm. 3, pp. 181–191, dic. 2019.

[66] I. Düntsch y G. Gediga, “Indices for rough set approximation and the application to confusion matrices”, *Int. J. Approx. Reason.*, vol. 118, pp. 155–172, mar. 2020, doi: 10.1016/j.ijar.2019.12.008.

[67] K. Yeturu, “Chapter 3 - Machine learning algorithms, applications, and practices in data science”, en *Handbook of Statistics*, vol. 43, A. S. R. Srinivasa Rao y C. R. Rao, Eds. Elsevier, 2020, pp. 81–206. doi: 10.1016/bs.host.2020.01.002.

[68] M. C. Pesantez-Chuqui, “Comparativa de técnicas Machine Learning sobre comportamiento de pago de clientes con cuentas por cobrar”, Tesis de Maestria, 2019. Consultado: ago. 24, 2021. [En línea]. Disponible en: https://reunir.unir.net/handle/123456789/9734

[69] W. Drzewiecki, “Thorough statistical comparison of machine learning regression models and their ensembles for sub-pixel imperviousness and imperviousness change mapping”, *Geod. Cartogr.*, vol. 66, dic. 2017, doi: 10.1515/geocart-2017-0012.

[70] A. Simón Rodríguez, R. Ruz Gómez, A. Simón Rodríguez, y R. Ruz Gómez, “Evaluación de algoritmos de machine learning para conducción”, Trabajo Fin de Grado, Universidad Complutence Madrid, 2020. doi: 10/1/Ruz_Gomez_Entrega_memoria_TFG_Evaluacion_de_algoritmos_para_la_conduccion_4398577_1208955137.pdf.

[71] W. Ossa Giraldo y V. Jaramillo Marin, “Machine Learning para la estimación del riesgo de crédito en una cartera de consumo”, Tesis de Maestria, Universidad EAFIT, 2021. Consultado: ago. 24, 2021. [En línea]. Disponible en: http://repository.eafit.edu.co/handle/10784/29589