LaSalle Universidad

Una obra de los Hermanos de La Salle

+300 años de experiencia en educación

Innovación y Software

Vol. 5 N° 1 Marzo – Agosto 2024

Revista de la Facultad de Ingeniería de ULASALLE

Universidad La Salle, Arequipa, Perú
facin.innosoft@ulasalle.edu.pe
https://revistas.ulasalle.edu.pe/innosoft

ARK: ark:/42411/s15
PURL: 42411/s15
DOI: 10.48168/innosoft.s15

Vol. 5 Nº. 1 2024 Marzo - Agosto

ISSN: 2708-0935

DOI: 10.48168/innosoft.s15

ARK: ark:/42411/s15

PURL: 42411/s15

Depósito Legal: 2023-08884

Periodicidad: Semestral

Publicado: 30/03/2024

Editado por:

Universidad La Salle

RUC: 20456344004

Ave. Alfonso Ugarte No. 517, Cercado, Arequipa, Perú

COMITÉ EDITORIAL

EDITORIAL

ARTÍCULOS ORIGINALES

ARTÍCULOS DE REVISIÓN

Es un honor para nosotros dar la bienvenida a todos nuestros lectores a este primer número del quinto Volumen de la Revista Innovación y Software de la Facultad de Ingeniería en la Universidad La Salle. En esta ocasión, nos complace presentar una selección de artículos que reflejan el continuo avance y la innovación en el campo de las Tecnologías de la Información y las Comunicaciones, así como en el ámbito de la Inteligencia Artificial y la Ingeniería de Software.

Este número destaca por la diversidad y la relevancia de los temas abordados por nuestros autores, quienes han compartido sus investigaciones y desarrollos más recientes con nuestra comunidad académica. Desde el diseño y construcción de una Estación Terrena de Bajo Costo para CubeSat con Tecnología IoT-LoRa, hasta el análisis de sentimiento en Twitter en relación a la tecnología IA para generación de imágenes, pasando por la clasificación de comentarios tóxicos en los Videojuegos y la concientización sobre la obesidad en Latinoamérica en los centros de salud utilizando un árbol de decisión, entre otros temas de gran relevancia.

La presencia de la Inteligencia Artificial como hilo conductor en varios de estos artículos resalta la importancia de esta disciplina en la actualidad y su impacto en diversos aspectos de nuestra sociedad. Además, no podemos pasar por alto el papel fundamental de la Ingeniería de Software en la creación de sistemas y aplicaciones cada vez más sofisticados y eficientes. Desde el desarrollo de un clasificador de reseñas de videojuegos hasta la implementación de un sistema web para mejorar la gestión comercial y de talento humano utilizando la metodología Scrum, los avances en este campo son clave para el progreso tecnológico.

Agradecemos sinceramente a todos los autores por su dedicación y su contribución a esta edición, así como a nuestros revisores y al comité editorial por su riguroso trabajo en la revisión y producción de estos artículos.

Esperamos que los lectores encuentren en estos artículos inspiración y conocimiento, y que se unan a nosotros en este apasionante viaje hacia el futuro de la tecnología.

Comité Editorial

Diseño, Construcción y Pruebas de una Estación Terrena de Bajo Costo para CubeSat con Tecnología IoT-LoRa

Design, Construction and Testing of a Low Cost Ground Sta-tion for CubeSat with IoT-LoRa Technology

Gary Fernando Flores Cadena
Escuela Politécnica Nacional. Quito, Ecuador.
gary.flores@epn.edu.ec
https://orcid.org/0000-0003-3815-7866

Pablo Anibal Lupera Morillo
Escuela Politécnica Nacional. Quito, Ecuador.
pablo.lupera@epn.edu.ec
https://orcid.org/0000-0002-0416-4980

Darwin Antonio Mena
Escuela Politécnica Nacional. Quito, Ecuador.
darwin.mena@epn.edu.ec
https://orcid.org/0000-0003-1186-4448

David Benalcazar Rojas
Escuela Politécnica Nacional. Quito, Ecuador.
david.benalcazar@epn.edu.ec
https://orcid.org/0000-0001-5174-9482

Henry Paul Llumiquinga Loya
Escuela Politécnica Nacional. Quito, Ecuador.
henry.llumiquinga@epn.edu.ec

Santiago Sandobalin Guaman
Escuela Politécnica Nacional. Quito, Ecuador.
santiago.sandobaling@epn.edu.ec

Ericson Daniel Lopez Izurieta
Escuela Politécnica Nacional. Quito, Ecuador.
ericsson.lopez@epn.edu.ec

ARK: ark:/42411/s15/a116
DOI: 10.48168/innosoft.s15.a116
PURL: 42411/s15/a116

RECIBIDO 10/08/2023 • ACEPTADO 05/11/2023 • PUBLICADO 30/03/2024

RESUMEN

En esta investigación se presenta el diseño y construcción de una estación terrena de bajo costo, compuestas por una antena Yagi , Amplificador de bajo ruido (LNA) y un nodo receptor LoRa que permite recepción de señales de CubeSats basadas en tecnología LoRa, para lo cual se emplearon

conceptos de diseño general de antenas y amplificadores electrónicos y se comprobó el diseño mediante pruebas de laboratorio de características eléctricas, radiación de señal y respuesta en frecuencia con lo que se pudo recibir datos de telemetría de los satélites CubeSat LoRa.

Palabras claves: Bajo costo, estación terrena LoRa, CubeSat, recepción LoRa, satelital.

ABSTRACT

This research article presents the design and construction of a low-cost ground station, consisting of a Yagi antenna, Low Noise Amplifier (LNA), and a LoRa receiver node, enabling the reception of CubeSat signals utilizing LoRa technology. The design incorporates general antenna and electronic amplifier design concepts and was validated through laboratory tests assessing electrical characteristics, signal radiation, and frequency response. This setup successfully received telemetry data from LoRa-based CubeSat satellites.

Keywords: *CubeSat, low cost,LoRa ground station, LoRa reception, satellite.*

INTRODUCCIÓN

Los últimos avances en materia espacial y satelital emplean nuevas plataformas de comunicación [1], permitiendo que satélites investigativos y de desarrollo, como son los CUBESAT y los PICOSAT, usen dichas plataformas de largo alcance y empleadas en internet de las cosas (IOT). Una de estas plataformas es Lo-Ra cuyo nombre significa cobertura de largo alcance (LOng RAnge) y que es un estándar de red inalámbrica basado en la modulación de espectro ensanchado de CHIRP (Compressed High Intensity Radar Pulse) en una señal cuya frecuencia aumenta o disminuye con el tiempo [2].

En la actualidad varios desarrollos espaciales hacen uso de estas tecnologías, uno de los más recientes es el ideado por Julián Fernández, un adolescente que propuso la idea de enviar un PicoSat de 250 gramos que emplea la tecnología LoRa y de código abierto, para logar comunicaciones con satélites pequeños de experimentación. Para ello creó la empresa FOSSAAT [3] y con el apoyo de start ups y crowfunding, permitió poner en el espacio a FossaSat-1, el 12 de junio de 2019, el primer satélite LoRa basado en el estándar PocketQube para CubeSat que tienen un tamaño de 5x5x5cm. Ésta primera experiencia permitió que miles de datos de telemetría sean recibidos por estaciones terrenas en muchas partes del planeta [4].

Se ha constatado en investigaciones previas sobre comunicaciones satelitales [5] que la utilización de componentes económicos, previamente validados en entornos de laboratorio y posteriormente implementados en la recepción de señales satelitales, ha arrojado resultados sumamente efectivos. Basándonos en estas premisas, se ha concebido un sistema que reúna

tanto las especificaciones requeridas para la recepción de datos LoRa como la ventaja de emplear componentes de bajo costo.

En la actualidad existen varios satélites LoRa que se encuentran rotando a la tierra en una órbita baja LEO (acrónimo del inglés Low Earth Orbit) donde sus distancias oscilan cerca de los 400 km. [6] y sus tiempos de rotación por los polos es cercana a los 92 minutos, por lo que diariamente se pueden obtener varios pases sobre la estación terrena instalada en tierra, sin embargo, no todos los pases tienen las mejores características para poder recibir los datos, por lo que se limita a uno o dos pases diarios del satélite que permiten recibir correctamente los datos.

Para poder captar a esos satélites, una de las cosas a tomar en cuenta es saber los pases del satélite con sus horarios exactos para apuntar las antenas directivas y con ello captar los datos que se transmiten desde más de 400 km de distancia. Adicionalmente la señal a recibir suele ser muy débil debido a las atenuaciones atmosféricas, las perdidas por distancia, zona de Fresnel y apuntamiento.

Para lograr que estos paquetes LoRa sean recibidos con éxito en tierra se requiere de una estación terrena [7] [8]receptora, que cuente con un sintonizador que procese las señales con protocolo LoRa y además de un amplificador de bajo ruido (LNA) conectado a una antena de alta ganancia, que esté diseñada para trabajar a la frecuencia de transmisión del satélite receptado.

En el presente estudio se propone el diseño, construcción y pruebas de una estación terrena para CubeSat que apliquen tecnología IoT- LoRa con dispositivos de bajo costo, por cuanto cada vez son más instituciones de estudio e investigación al igual que universidades las que requieren contar con estaciones terrenas para CubeSat.

La estación consta de un módulo LoRa de desarrollo que posee un procesador ESP32[9] con varios pines de entrada y salida multiprósito y cuyo precio no supera los 45 USD, de igual manera se emplea un LNA[10] de diseño propio y cuyo costo en materiales es menor a los 30 USD y se complementa con una antena tipo Yagi diseñada específicamente para trabajar a las frecuencias de los CubeSat con tecnología LoRa y que es de construcción propia, hecha con tubos de aluminio y materiales comunes.

Materiales y métodos

El presente estudio se empleó el diseño electrónico al igual que el diseño y simulación de antenas, también con los diseños planteados se hizo las pruebas de laboratorio y en campo para determinar el acertado desempeño de la estación terrena LoRa, con lo que se puedo construir la primera estación terrena para CubeSat-LoRa de Ecuador.

Para entender la manera como se envían las señales desde estos satélites LoRa, es necesario definir de que se trata el protocolo LoRa.
LoRa, que se aplica a comunicaciones con Largo Alcance emplea un tipo de modulación en radiofrecuencia patentado por Semtech [6][10], que, por emplear una modulación de espectro ensanchado, permite atravesar obstáculos y que en transmi-siones abiertas ofrece grandes coberturas por su mayor sensibilidad de recepción con ni-veles muy bajos de señal [2]que entre sus principales características se encuentran:

- Alta tolerancia a las interferencias
- Alta sensibilidad para recibir datos (-168dB)
- Basado en modulación chirp
- Bajo consumo para nodos, hasta 10 años con una batería
- Largo alcance, 10 a 20 km para enlaces en tierra
- Baja transferencia de datos (hasta 255 bytes)
- Conexión punto a punto
- Frecuencias de trabajo: 915MHz- América, 868MHz – Europa,433MHz - Asia (en el caso de los CubeSat es la frecuencia más empleada).

Todo esto hace a la tecnología LoRa ideal para conexiones a grandes distancias y para redes de IoT que se pueden utilizar en ciudades inteligentes, lugares con poca cober-tura celular o redes privadas de sensores o actuadores.

LoRaWAN es protocolo de red que usa la tecnología LoRa, para redes de baja potencia y área amplia, LPWAN (Low Power Wide Area Network) empleado para comunicar y administrar dispositivos LoRa con uso del internet [11]. El protocolo LoRa-WAN se emplean dispositivos denominados gateways y nodos:

Los gateways con sus antenas son los encargados de recibir y enviar información a los nodos y a su vez permitir la comunicación en la nube o internet.

Nodos o dispositivos: son los elementos finales que envían y reciben información hacia el Gateway haciendo uso de la antena [12].
Con la comunión de gateways y nodos sumados al uso de internet se tiene un sistema que ofrece muchas prestaciones y aplicaciones, así como bondades en comunicación a grandes distancias, con bajo consumo en energía y con niveles de señal muy pequeños [12].

Uso Del Protocolo Lora En Picosatélites

En las comunicaciones CubeSat se envían de pequeños volúmenes de información y se requiere que se puedan recibir y decodificar cuando las señales presenten niveles muy bajos, por lo que se emplea LoRa[4],[1] y de acuerdo a los primeros resultados obtenidos en recepción a nivel mundial, se han logrado comunicaciones exitosas.
Con el primer lanzamiento de un Picosatélite LoRa en el 2019, varios investigadores entusiastas se organizaron para recibir esos datos, y crearon la plataforma TinyGS, que es una red abierta de estaciones terrestres distribuidas por todo el mundo para recibir y operar satélites LoRa, sondas meteorológicas y otros objetos voladores.
Inicialmente TinyGS nació bajo el nombre ESP32 Fossa Groundstation, desarrollado con la finalidad de recibir las señales para el satélite FossaSAT-1 LoRa, en noviembre del 2019. Actualmente, la red está abierta a cualquier satélite LoRa y también para otros objetos voladores que transmitan con una modulación en radio como FSK, GFSK, MSK, GMSK, LoRa y OOK. con lo que el proyecto pasó a llamarse TinyGS.

La plataforma TinyGS cuenta, a octubre del 2023, con 4391 miembros de los cuales existen 1377 estaciones activas alrededor del mundo. La plataforma cuenta con una página web donde se muestran las estaciones terrenas y sus ubicaciones en el mapa, como se muestra en la Figura 1.

Figura 1. Mapa de estaciones satelitales LoRa a nivel mundial a octubre 2023 (TinyGS).

Componentes del sistema receptor LoRa

El sistema de recepción como indica el diagrama de la Figura 2, está formado por una antena de recepción directiva y de alta ganancia, un amplificador de bajo ruido LNA, un nodo LoRa que en esta investigación emplea el microprocesador ESP32 y que a través de su conexión WIFI sube los datos usando el internet, hacia la plataforma de TinyGS.

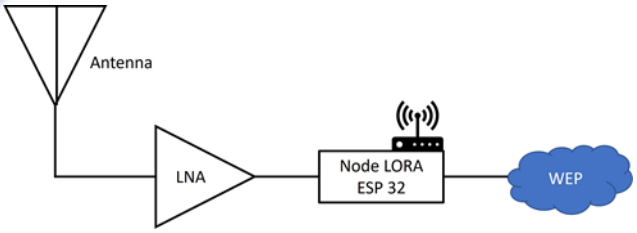

Figure 2. Diagrama de elementos necesarios en la Estación Terrena LoRa.

Para determinar la ubicación de los diferentes satélites LoRa con sus efemérides en tiempo real y las predicciones de los futuros pases con las mejores características para recepción de señal con sus ángulos de elevación , azimut y distancias, se hace uso de predicciones online como del sitio web https://www.n2yo.com, donde se puede obtener los datos del satélite LoRa de interés, en tiempo real, el tracking y la predicción de los mejores pases indicando el nivel de señal esperada, como se muestra en el ejemplo de la Figura 3 para el satélite LoRa ruso, llamado NORBI.

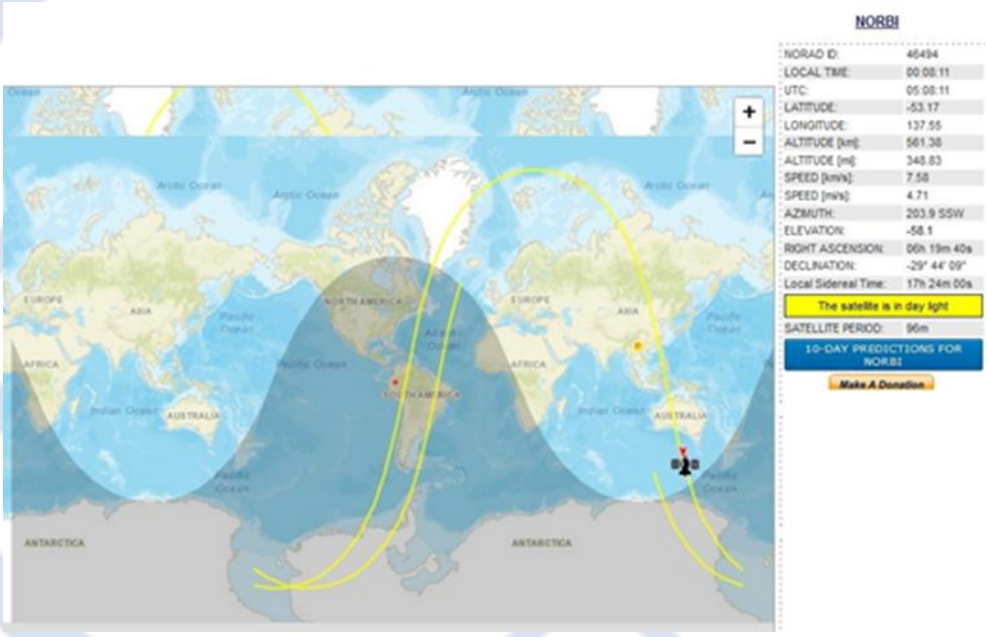

Figura 3. Tracking Sat. NORBI de www.n2yo.com

Se realizó un análisis de varios satélites LoRa activos desde su fecha de lanzamiento, la potencia con la que transmiten sus señales a tierra y la frecuencia que emplean para transmitir los datos en los canales LoRa, tal como se indica en la Tabla 1 de resumen.

Tabla 1. Satélites LoRa, frecuencias y Potencias de Downlink (obtenidos de la plataforma TinyGS)

Fecha de lanzamiento	***Satélites LoRa***		
	Nombre del Satélite	***Frecuencia (MHz)***	***Potencia (mW)***
27-09-2020	NORBY	436.703	2000
22-04-2021	FEES	437.2	500
13-01-2022	SATLLA-2B	437.25	63
13-01-2022	FossaSat-2E2	401.7	158

Se pudo determinar que el satélite con mayor potencia de transmisión es el satélite Ruso NORBI, que la mayoría del tiempo transmite con una señal de 2000 mW y a veces eleva su potencia hasta a los 7000 mW, según se publica en la plataforma TinyGS.

Antena Yagi en la Banda de los 430 Mhz

Para garantizar que las señales enviadas por los satélites LoRa sean captadas, se propuso el diseño de una antena Yagi [14] directiva, que es la antena que posee mejores características directivas que una dipolo o cuarto de lambda, que ayude a captar estas señales muy débiles provenientes del espacio desde el CubeSat y que a su vez tenga una gran ganancia.

La antena Yagi [14][15]propuesta y diseñada, está compuesta por 5 directores y posee una ganancia teórica de 16.8 dBi. Está sintonizada a la frecuencia central de 436.7 MHz y presenta una impedancia de 50 Ohms [15]. Esta frecuencia coincide con la de transmisión del satélite ruso NORBI, que ostenta la mayor potencia de emisión entre los satélites activos. Esta elección asegura la recepción efectiva al orientar la antena hacia el satélite que emite la señal más robusta.

La antena se simuló en el software MMANA-GAL, para obtener los datos teóricos y optimizados con los que se la construyó. Posteriormente se hizo las pruebas en el laboratorio de Alta Frecuencia de la Escuela Politécnica Nacional de Quito, con el equipo analizador de redes de marca KEYSIGHT modelo N9916A, de donde se obtuvieron los resultados descritos en la Tabla 2.

Tabla 2. Magnitudes de la Antena Yagi diseñada a las frecuencias de transmisión de varios CubeSat-LoRa.

FRECUENCIA [MHz]	***PARÁMETRO***	***SIMULADO***	***MEDIDO***
436.703	R [Ohms]	43.71	53.3
	ROE (a 50 Ohms)	1.18	2.1
437.2	R [Ohms]	43.96	51.5
	ROE (a 50 Ohms)	1.16	1.79
401.7	R [Ohms]	29.01	30.6
	ROE (a 50 Ohms)	3.8	1.86

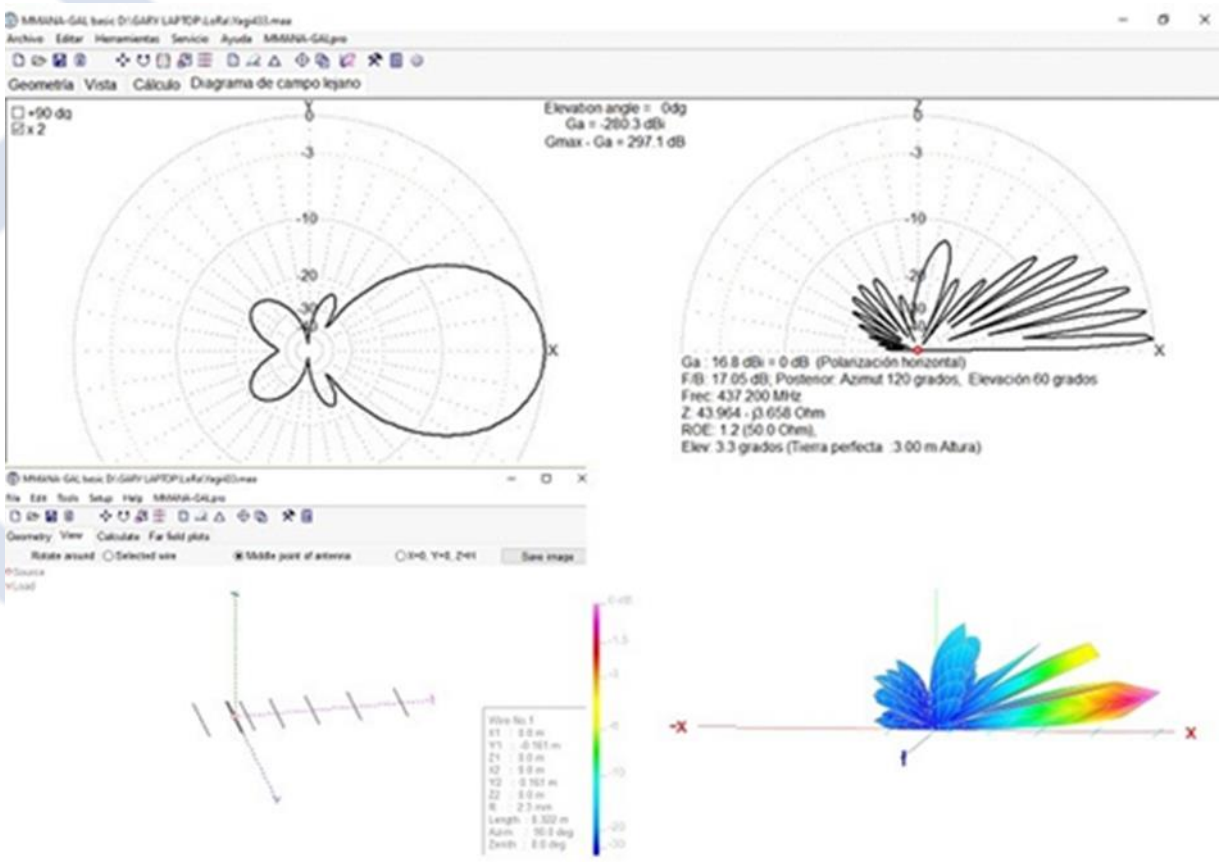

Figura 4. Diagrama de radiación de la antena Yagi diseñada y simulada.

Los diagramas de radiación obtenidos mediante la simulación ofrecen una visualización clara de los lóbulos y las pautas de radiación correspondientes a los cortes realizados en los planos horizontal y vertical. Estos aspectos se detallan en la Figura 4. En lo que respecta al campo lejano

de la antena, se destacan las características más óptimas en términos de directividad y ganancia para la aplicación diseñada.

Amplificador de Bajo Ruido LNA

Para garantizar que la débil señal que envían los CubeSat LoRa sea receptada correctamente, se requiere un amplificador de bajo ruido (LNA), para alcanzar los niveles mínimos que permitan decodificar las señales.

Estos amplificadores de bajo ruido (LNA) están disponibles en versiones comerciales que prometen amplificaciones superiores a los 23 dB. Sin embargo, la calidad y el ancho de banda no están garantizados en dichas versiones, y además, se debe considerar el costo adicional relacionado con la importación. Adicionalmente, si se opta por un LNA de una marca comercialmente reconocida, el precio se incrementa aún más. Dado que el objetivo del proyecto era desarrollar un sistema de bajo costo, se tomó la decisión de diseñar el circuito por completo en lugar de utilizar componentes comerciales. Como elemento principal del circuito LNA se empleó el chip amplificador PSA4-5043+ del fabricante Mini-circuits, que posee excelentes características para las frecuencias de trabajo en la banda de los 400 a 500 MHz [10]. Con una Ultra Low Noise Figure 0.65 dB, Gain 22.1 dB, trabajando a +5V tiene un consumo de corriente de 56mA.

Se añadieron al diseño del circuito LNA, un regulador de voltaje y un circuito BIAS-TEE [7] así como un led indicador de polarización y los filtros de aterrizaje y pasa banda [7], el diagrama completo se muestra en la Figura 5.

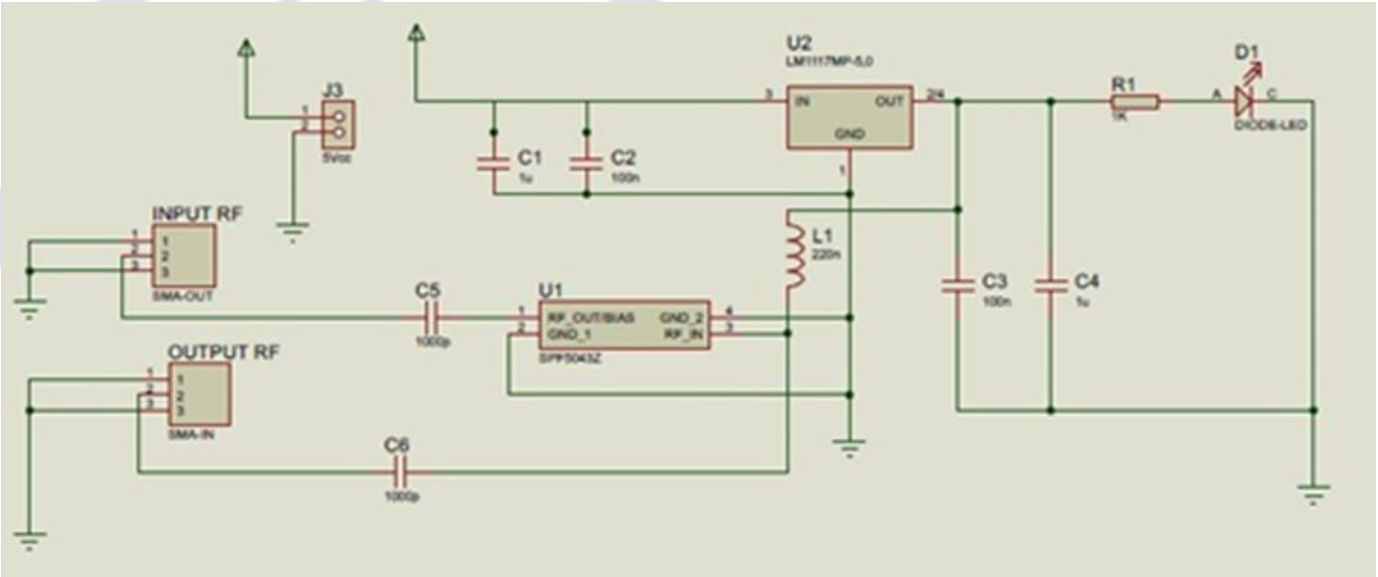

Figura 5. Circuito LNA diseñado usando el dispositivo PSA4-5043+

Nodo LoRa a 433 MHz

Se utilizó un módulo de desarrollo denominado LoRa HETEC 32 WIFI. Este módulo presenta una placa que incorpora un procesador ESP32 y una pantalla OLED de 0,96 pulga-das. Además, dispone de conectividad WiFi y un conector SMA para la antena. El módulo opera con la modulación LoRa en la banda de frecuencia de 433 MHz, tal como se señala en la Figura 6. Es importante mencionar que este módulo tiene un coste inferior a los 45 dólares estadounidenses.

Para habilitar el funcionamiento del dispositivo ESP32 y permitir el envío de datos a la plataforma TinyGS, se utiliza el software de código abierto disponible en el siguiente enlace: https://github.com/G4lile0/tinyGS. Este software proporciona la funcionalidad necesaria para que el ESP32 se integre con la plataforma TinyGS y pueda transmitir datos de manera efectiva.

Figura 6. Módulo ESP32-OLED- LoRa.

Resultados y discusión

Los datos que se reciben son de telemetría de los satélites LoRa e indican la potencia de transmisión del satélite, la temperatura interna y externa del satélite en su vuelo, consumo de energía, las caras del satélite que reciben energía solar en sus paneles, número de paquetes de datos enviados, numero de resets del computador interno , entre otras cosas, los datos recibidos y decodificados de la estación terrena son mostrados en la plataforma de TinyGS, donde son graficado tal como se indica en la Figura 7.

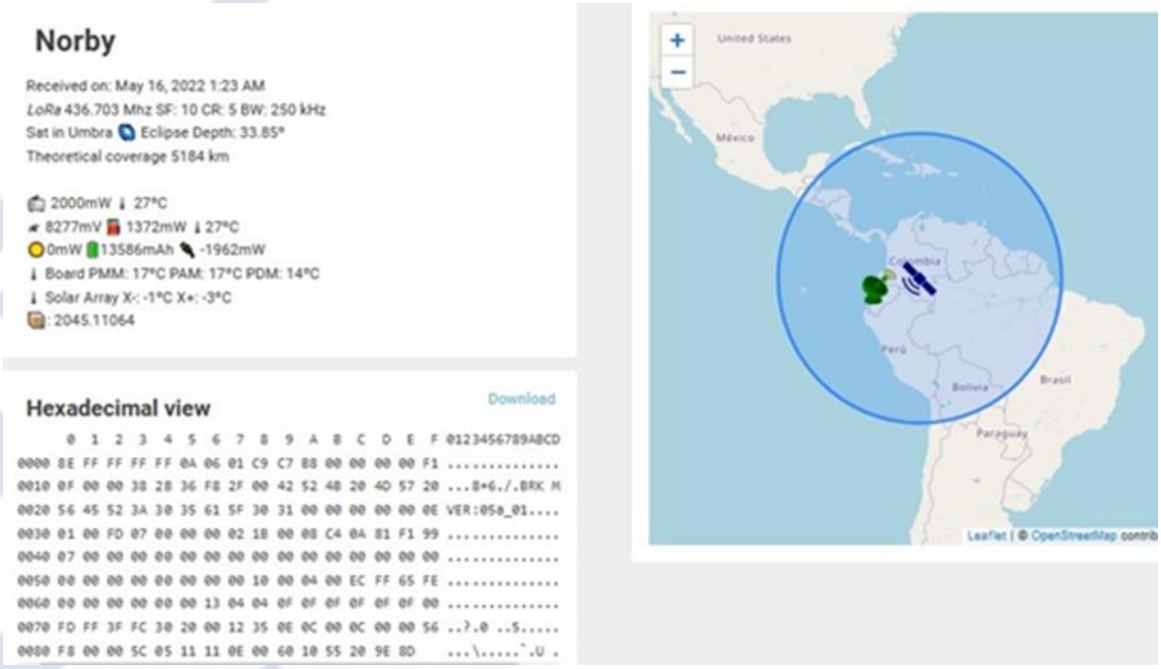

Figura 7. Ejemplo datos de telemetría recibida en la estación Terrena LoRa

La estación terrena instalada en Quito- Ecuador ha estado trabajando desde junio del 2021 y lleva 1153 paquetes recibidos de varios satélites LoRa, que se encuentra a una distancia promedio de 400 km. y se ha logrado una distancia máxima en la recepción de paquetes correctos de 1900 km y corresponde a la señal del satélite Norby cuando transmite a una potencia de 7000 mW.

Además, en el caso del satélite FossaSat-2E2, que transmite a una frecuencia de 401,7 MHz con una potencia de emisión de tan solo 158 mW. A pesar de estas cifras relativamente bajas de potencia de transmisión, se logró recibir la señal a una distancia de 804 km. Esto destaca la eficacia de la comunicación, incluso con niveles de potencia de transmisión moderados y refleja el buen desempeño de la estación terrena construida.

Conclusiones

De acuerdo a los resultados obtenidos se puede determinar que pese al bajo costo de los implementos del sistema, los resultados son bastante óptimos si los comparamos con los de sistemas profesionales que usan otras estaciones terrenas. Se pudo observar que la mayoría de las recepciones exitosas de paquetes de datos LoRa provenientes del CubeSat ocurrieron cuando estos pasaban en un ángulo de elevación relativo superaba los 20 grados con respecto a la posición de la estación terrena.

Este fenómeno se debe a la geografía de la ciudad de Quito, que está rodeada de montañas. Estas montañas limitan la recepción de señales en ángulos de elevación más bajos. La estación terrena forma parte de las 4391 estaciones instaladas mundialmente, de las cuales sólo 1377 son activas, es decir sólo el 31,4% recibe datos, por lo que hay estaciones que pese a contar con todos los equipos no han logrado recibir ningún paquete.

Hay que tomar en cuenta que varios paquetes no se han recibido correctamente, por cuanto, en esta primera parte del proyecto se diseñó una antena que no cuenta con un sistema de seguimiento automático, por lo que a futuro se planea hacer el apuntamiento automático.

La comunicación LoRa resultó ser muy efectiva en uso en satélites CubeSat, pues las distancias alcanzadas superan los 1000 km y con una potencia de transmisión que no supera los 7000 mW. Es decir, se reciben datos con un nivel de señal de -128dBm, en comparación con los sistemas de recepción satelital tradicional, sería imposible recibir una señal exitosa con esos niveles de señal.

Contribución de Autoría

Gary Fernando Flores Cadena: Conceptualización, Análisis formal, Investigación, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Pablo Anibal Lupera Morillo:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original. **Darwin Antonio Mena:** Conceptualización, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos. **David Benalcazar Rojas:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos. **Henry Paul Llumiquinga Loya:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Henry Paul Llumiquinga Loya:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original. **Santiago Sandobalin Guaman:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Ericson Daniel Lopez Izurieta:** Visualización, Software, Validación, Redacción - borrador original, Curación de datos. Conceptualización, Investigación, Metodología, Análisis formal, Supervisión.

Referencias

[1] N. Saeed, A. Elzanaty, H. Almorad, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "Cubesat communications: Re-cent advances and future challenges," IEEE Commun. Surv. \& Tutorials, vol. 22, no. 3, pp. 1839–1862, 2020.

[2] D. Kjendal, "LoRa-Alliance regional parameters overview," J. ICT Stand., pp. 35–46, 2021.

[3] P. Lepcha et al., "Assessing the Capacity and Coverage of Satellite IoT for Developing Countries Using a CubeSat," Appl. Sci., vol. 12, no. 17, p. 8623, 2022.

[4] E. Bashir and M. Luštrek, "Low power LoRa transmission in low earth orbiting satellites," Intell. Environ., p. 233, 2021.

[5] G. Flores, E. López, L. Tituaña, and P. Lupera, "Low Cost Multiband Receiver for the Reception of Images from Me-teorological Satellites and SSTV," Rev. Politécnica, vol. 40, no. 2, pp. 25–30, 2018.

[6] M. Aref and A. Sikora, "Free space range measurements with Semtech LoRaTM technology," in 2014 2nd interna-tional symposium on wireless systems within the conferences on intelligent data acquisition and advanced computing sys-tems, 2014, pp. 19–23.

[7] J. B. Hagen, Radio-frequency electronics: circuits and applications. Cambridge University Press, 2009.

[8] J. A. M. Lara, J. A. R. Agredo, and M. P. M. Atencia, "Sistema de monitoreo de señales en tierra usando la Estación Terrena Satelital UPTC," INGE CUC, vol. 15, no. 1, pp. 36–44, 2019.

[9] A. Maier, A. Sharp, and Y. Vagapov, "Comparative analysis and practical implementation of the ESP32 microcon-troller module for the internet of things," in 2017 Internet Technologies and Applications (ITA), 2017, pp. 143–148.

[10] H. Daryanavard and others, "A Low Noise Figure Rail-to-Rail Variable-Gain LNA for 900-MHz LoRa Application in 65nm CMOS Technology," 2022.

[11] C. A. Trasviña-Moreno, R. Blasco, R. Casas, and A. Asensio, "A network performance analysis of LoRa modulation for LP.WAN sensor devices," in Ubiquitous computing and ambient intelligence, Springer, 2016, pp. 174–181.

[12] M. A. Moya Quimbita, "Evaluación de pasarela LoRa/LoRaWAN en entornos urbanos," 2018.

[13] M. A. Gaybor Murillo, M. D. Maridueña Chunga, and others, "Diseño de un sistema de adquisición de datos de una red de sensores inalámbricos que miden variables oceanográficas en el perfil costanero de Santa Elena, usando tecnología LoRa," Espol, 2018.

[14] P. P. Viezbicke, Yagi antenna design, vol. 688. US Government Printing Office, 1976.

[15] W. Tomasi, Sistemas de comunicaciones electrónicas. Pearson education, 2003.

Reconocimiento y clasificación de comentarios de productos de Amazon

Recognition and rating of Amazon product reviews

Luisfelipe Rodrigo Mamani Arosquipa
Universidad La Salle. Arequipa, Perú.
lmamania@ulasalle.edu.pe

Frank Jhoseph Duarte Oruro
Universidad La Salle. Arequipa, Perú.
fduarteo@ulasalle.edu.pe

ARK: ark:/42411/s15/a119
DOI: 10.48168/innosoft.s15.a119
PURL: 42411/s15/a119

RECIBIDO 26/08/2023• ACEPTADO 25/10/2023• PUBLICADO 30/03/2024

RESUMEN

El flujo de información surge día a día mediante internet de manera continua gracias a las constantes interacciones presentes entre los usuarios, estas interacciones se presentan en comentarios que pueden ser positivos o negativos. Esto puede ayudar mucho al servicio que ofrece Amazon en sus productos para poder comprender si esta´ en buen estado o no, para que sus usuarios de la plataforma se puedan convencer al momento de comprar un producto, y es que, si estos son un gran número, un análisis hecho por una sola persona no es suficiente. Para ello es necesario el uso de herramientas que operan con grandes cantidades de datos como (nombre del procesamiento de datos), que es un modelo que ayuda al análisis de clasificación de comentarios basados en lo que expresan los usuarios. En este trabajo se usará´ este modelo para la clasificación de comentarios de productos de Amazon, valorando estos comentarios según su descripción. Se harán además uso de métricas y de sugerencias futuras para la propuesta mencionada en este trabajo. El análisis de los comentarios ayudara´ a entender cómo es que las personas clasifican estas diferentes situaciones de su vida cotidiana. Los datos de las redes sociales se utilizan durante todo el proceso de análisis y clasificación, que consiste en datos de texto. Utilizando las redes sociales, se puede monitorizar o analizar los comentarios. En este trabajo de investigación clasificaremos los datos de los comentarios que se realizan en Amazon relativos a su calificación en cada comentario.

Palabras claves: Sentimiento del consumidor, análisis de comentarios, minería de opiniones, clasificación de datos, Amazon, Inteligencia artificial.

ABSTRACT

The flow of information arises every day through the internet continuously thanks to the constant interactions between users, these interactions are presented in comments that can be positive or negative. This can help a lot to the service offered by Amazon on their products to understand if this' in good condition or not, so that its users of the platform can be convinced when buying a product, and is that, if these are a large number, an analysis made by one person is not enough. This requires the use of tools that operate with large amounts of data such as (name of data processing), which is a model that helps the analysis of classification of comments based on what users express. In this paper we will use' this model for the classification of Amazon product reviews, rating these reviews based on their description. It will also make use of metrics and future suggestions for the proposal mentioned in this paper. The analysis of comments will help to understand how people classify these different situations in their daily lives. Social network data is used throughout the analysis and classification process, which consists of text data. Using social networks, comments can be monitored or analyzed. In this research work, we will classify the data of comments made on Amazon relating to their rating on each comment.

Keywords: *Consumer sentiment, comment analysis, opinion mining, data classification, Amazon, IA.*

INTRODUCCIÓN

La capacidad de adaptación de la tecnología a diferentes contextos ha transformado diversos aspectos de la vida cotidiana, incluyendo la educación y la interacción en línea. La enseñanza y el aprendizaje se han vuelto más accesibles gracias a la tecnología. Sin embargo, esta adaptabilidad tecnológica se ha expandido a diversas áreas de interacción en línea, siendo las redes sociales uno de los ejemplos más prominentes.

Las redes sociales representan un amplio y complejo en- torno de interacción en línea, donde usuarios de todo el mundo comparten información, intereses y opiniones sobre diversos temas de manera remota. Esta interacción se lleva a cabo principalmente a través de los comentarios en los productos, publicaciones, vídeos, tendencias, etc., de otros usuarios, lo que a menudo proporciona una visión general de las opiniones sobre un tema específico. Sin embargo, esta visión puede no ser siempre precisa debido a la diversidad de información presente en los comentarios.

Entre las plataformas más destacadas para este intercambio de opiniones se encuentran Facebook, Twitter y YouTube, pero el objetivo de este proyecto de aplicación NLP es detectar los diferentes tipos de comentarios expresados en los productos de Amazon. Sabemos que en la actualidad las redes sociales se han convertido en una parte integral de la vida diaria de muchas

personas y pueden tener un impacto significativo. Muchas personas utilizan las redes sociales para conectarse con otros, pero también pueden exponerse a una gran cantidad de información negativa, información positiva. Amazon es una de las plataformas de redes sociales más populares y la gente la utiliza a menudo, la plataforma ofrece una amplia gama de opciones para compras en línea y es conocida por su conveniencia y entrega rápida. El proyecto desarrolla un modelo de aprendizaje automático que pueda analizar los comentarios para identificar patrones lingüísticos asociados a los tipos de comentarios, nos enfocaremos en obtener opiniones de comentarios que reaccionan a los productos de Amazon. El objetivo es determinar su utilidad y clasificar los comentarios en categorías como ”positivos”, ”negativos”, ”neutros”, ”muy buenos” y ”muy malos”. Para lograr esto, primero realizaremos un preprocesamiento de la información, seleccionando los comentarios que expresen sentimientos de manera explícita, y luego utilizaremos algoritmos de procesamiento de lenguaje natural para determinar la polaridad predominante, a su vez puede mejorar la calidad de los productos en venta en Amazon.

Motivación

Los comentarios y opiniones de otros compradores pueden influir significativamente en las decisiones de compra. Al clasificar y analizar estos comentarios de manera eficaz, se puede ayudar a los consumidores a tomar decisiones informadas y, al mismo tiempo, mejorar su experiencia de compra.

A. ¿En qué dominio del conocimiento está trabajando?

Consumidores de Amazon que desean obtener información sobre productos antes de realizar una compra.

B. ¿Quiénes son los usuarios objetivo?

Vendedores, compradores, anunciantes de Amazon interesa- dos en comprender las opiniones y reacciones de los clientes hacia sus productos.

C. ¿Por qué es interesante el tema que proponen?

Investigadores y profesionales en el campo de la inteligencia artificial, el procesamiento de lenguaje natural y el análisis de sentimientos.

D. ¿Cuáles son las preguntas que su proyecto de NLP intenta responder?

Personas interesadas en cómo las opiniones en línea influyen en las decisiones de compra y la calidad de los productos en una plataforma como Amazon.

Problema

El problema central que aborda este proyecto es la creciente complejidad de gestionar y analizar grandes conjuntos de datos generados por compras masivas en línea.

A medida que el comercio electrónico se expande, surge la necesidad de comprender profundamente los patrones de compra, las preferencias del consumidor y los comportamientos de compra en escenarios de compras masivas, lo que presenta un desafío significativo tanto para las empresas como para los investigadores. Además, la gestión de datos masivos requiere soluciones eficaces de procesamiento de datos y técnicas de aprendizaje automático para extraer información valiosa y tomar decisiones estratégicas basadas en datos precisos y confiables.

Objetivo

El objetivo principal de este estudio es investigar y desarrollar soluciones efectivas para la gestión y análisis de grandes volúmenes de datos de compras en línea en Amazon. Se pretende abordar la creciente demanda de herramientas que permitan comprender los patrones de comportamiento del consumidor en el entorno de compras en línea a gran escala.

Datos

A. ¿Qué datos necesitará?

Datos de transacciones de compras en línea, que incluyen detalles como productos comprados, fechas y montos. Datos de navegación en línea, que pueden incluir información sobre páginas visitadas, tiempo de navegación y acciones realizadas en el sitio web. Datos de usuarios, que pueden incluir información demográfica, preferencias, historial de compras previas y otros datos relevantes.

B. ¿Cómo recolectarán los datos?

Los datos de transacciones se pueden recopilar a través de los sistemas de registro de compras en línea de empresas asociadas o a través de acuerdos de colaboración. Los datos de navegación en línea pueden ser obtenidos mediante el seguimiento de la actividad del usuario en el sitio web de la empresa o mediante la implementación de cookies y seguimiento web. Los datos de usuarios pueden recopilarse a través de formularios de registro, cuentas de usuario en línea o incluso mediante encuestas voluntarias.

C. ¿Dónde planea obtenerlos?

Los datos de transacciones pueden obtenerse directamente de las empresas de comercio electrónico o de terceros que recopilen y proporcionen datos de transacciones en línea. Como: Kaggle, Hugging Face, Google.

D. ¿Cómo planea almacenarlos?

Los datos se almacenarán en un excel con extensión .xls para ser analizados más fácilmente.

E. ¿Cómo accederá a ellos para utilizarlos en su proyecto?

Ya que trabajaremos con Python en Google Colab usaremos la librería Pandas Numpy que nos permitirá la lectura de la data y su obtención en una estructura de datos como diccionarios y matrices, para su posterior procesamiento.

Se podrían utilizar lenguajes de programación como Python o herramientas de análisis de datos como SQL para realizar consultas y análisis de los datos almacenados.

Revisión Literaria

Con el propósito de abordar los desafíos inherentes al análisis de comentarios de productos en Amazon, llevamos a cabo una revisión exhaustiva de artículos disponibles en la plataforma de Google Scholar. En este proceso, presentamos de manera sucinta una descripción detallada de cada enfoque aplicado en diversas investigaciones relacionadas con este tema.

El primer paper se enfoca en el análisis de sentimientos en reseñas de productos de Amazon. Utiliza técnicas de aprendizaje profundo y aprendizaje automático para clasificar la polaridad de las palabras, frases y documentos en positiva, negativa o neutral. Se comparan nueve algoritmos diferentes, y Bert se destaca con un rendimiento excepcional, logrando una precisión del 0.94. Además, se aplica el modelo Bert a un gran conjunto de reseñas adicionales para evaluar su eficacia. El estudio demuestra que Bert es muy efectivo para la clasificación de sentimientos en reseñas de productos. [?]

El enfoque del segundo paper se centra en la presentación de un recurso valioso para la investigación en procesamiento de lenguaje natural: el ”Multilingual Amazon Reviews Corpus (MARC)”. El texto describe la creación de este corpus que contiene reseñas de productos de Amazon en varios idiomas y resalta su importancia para abordar tareas de clasificación de texto multilingüe. También se proporcionan resultados basales para la clasificación supervisada y la transferencia cero, subrayando la relevancia del corpus en la comunidad de investigación. En resumen, el enfoque principal es informar sobre la disponibilidad y utilidad del corpus MARC para avanzar en investigaciones relacionadas con el procesamiento de lenguaje natural en un contexto multilingüe. [?]

El enfoque del tercer paper es evaluar la eficiencia de tres enfoques de aprendizaje automático (Support Vector Machines, Naive Bayes y Maximum Entropy) para clasificar reseñas de productos en Amazon. Se centra en el uso de técnicas de análisis de sentimientos para analizar el contenido de las reseñas y comprender las opiniones de los usuarios sobre los productos. El texto también explora cómo los aspectos y la polaridad de las opiniones pueden influir en la clasificación de las reseñas. Además, se utiliza una variedad de características, como unigrams y weighted unigrams, para entrenar a los clasificadores y se presentan los resultados de precisión para cada método. En resumen, el enfoque se concentra en la clasificación de reseñas en Amazon y la comparación de diferentes enfoques de aprendizaje automático para lograrlo. [1]

El cuarto paper presenta un enfoque de detección de sarcasmo en reseñas de productos de Amazon utilizando técnicas de procesamiento de lenguaje natural (NLP) y clasificación de

sentimientos. Comienza destacando la creciente importancia de analizar las opiniones en línea debido al gran número de reseñas en Amazon y la necesidad de comprender y analizar estos datos para realizar mejoras en productos. Luego, se centra en el desafío específico de detectar el sarcasmo en estas reseñas y analiza una serie de características y enfoques utilizados en la literatura previa, incluyendo el uso de algoritmos basados en reglas y clasificadores de machine learning. El trabajo también menciona la importancia de etiquetar la polaridad de las palabras antes del proceso de clasificación y evalúa el rendimiento de varios clasificadores, como SVM, Random Forest y K Vecinos más Cercanos (K Nearest Neighbors), proporcionando las tasas de precisión respectivas de cada uno de ellos. [2]

Diseño

A. Propuesta

Para este trabajo se utilizó un dataset ”Amazon multilingual Corpus” que están previamente categorizadas en cinco niveles. Que son correspondidas al número de estrellas que se les da en la reseña. Como ventaja de este dataset es que se encuentra en varios idiomas que son Español, Inglés, Francés, Japones, Alemán y Chino que son los idiomas más populares a nivel mundial. Además de que el dataset cuenta con 1.2 millones de datos en total. Es importante señalar que, específicamente para la sección de comentarios en inglés, se utilizó un subconjunto de 50,000 datos para llevar a cabo el proceso de entrenamiento. Este enfoque permite centrarse en un conjunto específico de comentarios en un idioma particular, facilitando así la tarea de modelado y análisis de sentimientos en el ámbito del inglés.

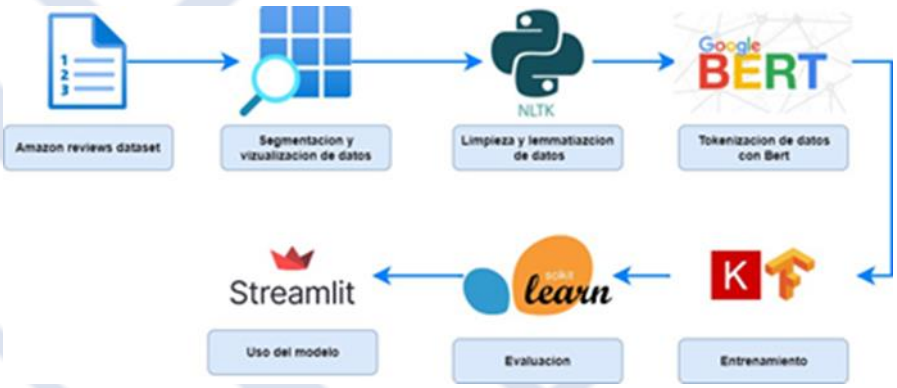

Figura 1. Pipeline del proceso de clasificación.

B. Preprocesamiento de los datos

El preprocesamiento consiste en preparar el conjunto de datos para que tenga una alta calidad. Se trata de tareas importantes que deben antes de utilizar un conjunto de datos para el entrenamiento de modelos. Utilizamos la biblioteca NLTK para realizar tareas de limpieza, incluyendo la eliminación de caracteres especiales, stop words y otros elementos no relevantes. Además, aplicamos lematización para reducir las palabras a su forma base, simplificando así la representación léxica de los comentarios.

- Eliminación de etiquetas HTML.
- Eliminación de datos NULL y repetidos del conjunto de datos.
- Filtrado de todos los emojis.
- Eliminación de palabras vacías
- Corrección ortográfica.

C. Tokenización

La tokenización desempeña un papel vital en la preparación de datos para modelos basados en BERT. Implementamos un tokenizador específico de BERT para dividir los comentarios en unidades semánticas, preservando así la información contextual. Se aplica padding para ajustar todas las secuencias a la misma longitud, permitiendo un procesamiento eficiente durante el entrenamiento del modelo.

D. Entrenamiento del Modelo

Para el entrenamiento del modelo, se emplea una arquitectura basada en BERT, aprovechando las capacidades preentrenadas del modelo en tareas generales del lenguaje. Añadimos capas específicas para la tarea de análisis de sentimientos. Se experimenta con diferentes hiperparámetros para optimizar el rendimiento del modelo en el conjunto de entrenamiento y se valida en el conjunto de validación.

E. Evaluación del Modelo

La evaluación del modelo se realiza en un conjunto de prueba independiente, utilizando métricas como precisión, recall y F1-score. Se analiza la capacidad del modelo para generalizar a datos no vistos y para manejar la variabilidad lingüística presente en el corpus multilingüe de comentarios de Amazon.

F. Uso del Modelo con Streamlit

La implementación práctica del modelo se logra mediante la integración con Streamlit, una biblioteca de Python para la creación rápida de aplicaciones web interactivas. Describimos la interfaz de usuario diseñada para la aplicación, que permite a los usuarios finales ingresar comentarios y obtener la clasificación en tiempo real a través de la web.

Principios básicos de diseño que nos guió para crearlo

- **Fortalezas**

Nuestro diseño ofrece varias ventajas significativas para el procesamiento de lenguaje natural (NLP) en el re- conocimiento de mensajes de odio. Aquí se mencionan algunas:

  - Adaptabilidad a diferentes idiomas La capacidad de adaptarse a diversos idiomas es una característica clave de nuestro diseño. Aunque se entrena y aplica inicialmente en español, la presencia de conjuntos de datos que abarcan varios idiomas permite que se adapte para reconocer y clasificar textos en diferentes lenguajes. Esta versatilidad es particular- mente valiosa en el reconocimiento y clasificación de mensajes y comentarios, dado que la información esclarecedora puede presentarse en variados idiomas y contextos culturales. Estas ventajas posicionan nuestra propuesta como una elección flexible para el procesamiento de lenguaje natural en la identificación y clasificación de comentarios. No obstante, es esencial tener en cuenta que ningún modelo es infalible, y se deben considerar siempre las limitaciones y posibles sesgos inherentes en los datos y el proceso de entrenamiento.
  - Contextualización bidireccional: Nuestra estructura de diseño logra una captura bidireccional del con- texto y la relación entre las palabras en una oración. En otras palabras, tiene la capacidad de comprender tanto las palabras que preceden como las que siguen en un texto, lo que eleva la comprensión del significado y la intención subyacentes en la clasificación de comentarios. Esta mejora en la contextualización facilita la identificación más precisa de clasificación de comentarios.

- **Debilidades**

  - Manejo de texto no estructurado: La clasificación de comentarios, comúnmente expresados en forma de texto no estructurado como publicaciones en redes sociales o foros, representa un desafío para nuestro enfoque. Aunque logra captar el contexto y el significado gramatical más informal, su eficacia se ve afectada en términos de prueba y entrenamiento, ya que se basa en un conjunto de datos estructurado y previamente clasificado.

- Requisitos computacionales y de recursos: Nuestra propuesta emplea un modelo de lenguaje profundo y complejo, como BERT, que demanda una cantidad significativa de recursos computacionales y memoria tanto durante el entrenamiento como la aplicación. Esta característica dificulta su implementación en entornos con recursos limitados, lo que restringe su accesibilidad y uso en ciertos casos.
- Tamaño del modelo: La eficacia del modelo de reconocimiento y clasificación de comentarios puede verse desafiada por el tamaño del modelo y el conjunto de datos utilizado, especialmente en casos donde la relación es de aproximadamente 10,000 palabras.
- Dependencia de datos etiquetados: Aunque BERT puede adaptarse a tareas específicas con datos etiquetados limitados, sigue siendo necesario contar con una cantidad significativa de datos anotados para obtener un rendimiento óptimo. La recopilación y etiquetado de grandes conjuntos de datos pueden resultar costosos y consumir tiempo, especialmente en el ámbito de la clasificación de comentarios, que a menudo requiere una revisión manual exhaustiva.

Resultados

La siguiente tabla muestra la cantidad de datos, el número de categorías y las métricas de precisión, ofreciendo una comparación del rendimiento entre los modelos previos. También se incluyó una comparación con un modelo de Bert llamado 'Bert antiguo,' que emplea la misma Base de Datos. Durante el proceso de entrenamiento, se aplicó la ley de Paretto.

Tabla 1. Resultados detallados del modelo

Clase	Precision	Recall	F1-Score	Soporte
0	0.71	0.64	0.67	2043
1	0.49	0.44	0.47	1993
2	0.44	0.57	0.50	2016
3	0.57	0.47	0.51	1998
4	0.68	0.74	0.71	1950
Accuracy	0.57			
Macro Avg	0.58	0.57	0.57	10000
Weighted Avg	0.58	0.57	0.57	10000

Trabajos Relacionados

El estudio "Amazon Fine Food Reviews se enmarca en la creciente importancia de las reseñas en línea para consumidores y empresas, especialmente en la industria del comercio electrónico. Las reseñas, generalmente compuestas por una puntuación general y una descripción de texto, son cruciales para que los consumidores comprendan intuitivamente los productos y servicios antes

de realizar una compra. Este trabajo se centra en desarrollar un modelo basado en BERT, una técnica de Procesamiento del Lenguaje Natural (NLP), para predecir la puntuación general de reseñas basándose en las descripciones de texto [7].
La investigación aborda la influencia masiva de las opiniones implícitas en las revisiones de clientes en las decisiones de compra. Se exploran diversas técnicas de representación vectorial, como Bag-Of-Words, Tf-Idf y Glove, para transformar las revisiones en datos procesables. Luego, se emplean varios algoritmos de aprendizaje automático, como Regresión Logística, Bosques Aleatorios, Na¨ıve Bayes, Memoria a Corto y Largo Plazo Bidireccional (Bi-LSTM) y BERT, para clasificar el sentimiento [8].
El artículo titulado ”Amazon products reviews classification based on machine learning, deep learning methods and BERT” aborda la creciente tendencia de las compras en línea y la importancia de las reseñas de clientes en las plataformas de comercio electrónico. Los autores proponen un sistema automatizado para el análisis de sentimientos de las reseñas de productos de Amazon utilizando diversas técnicas de aprendizaje automático (ML) y aprendizaje profundo (DL), incluido BERT [9].

Conclusiones

En resumen, la calidad y la adecuada preparación de los conjuntos de datos utilizados para entrenar modelos, como BERT, son factores críticos que influyen en los resultados de la evaluación y las métricas obtenidas. La importancia de contar con un conjunto de datos equilibrado y someter los datos a una exhaustiva limpieza, centrándose en los campos pertinentes para el entrenamiento, se revela como un aspecto crucial para lograr resultados óptimos. Un conjunto de datos equilibrado asegura que el modelo se entrene con una representación justa de las diversas clases o categorías presentes en los datos, evitando sesgos y mejorando su capacidad para clasificar correctamente las muestras de prueba.

La limpieza de datos también desempeña un papel funda- mental. Al eliminar ruido, datos irrelevantes o redundantes, y garantizar la integridad de los campos necesarios para el entrenamiento, se eleva la calidad y la coherencia del conjunto de datos. Este proceso se traduce en un aprendizaje más preciso por parte del modelo, lo que se refleja en una evaluación más confiable y en métricas más sólidas.

Trabajos futuros

En el contexto de futuras investigaciones, se identifican diversas áreas clave para el perfeccionamiento del pipeline propuesto en la clasificación de comentarios de productos de Amazon. Se sugiere la exploración de técnicas de pre- procesamiento más avanzadas y

personalizadas para abordar desafíos específicos en el ámbito de las reseñas de productos, como la adaptación de la limpieza de datos para considerar peculiaridades lingüísticas y contextuales. Además, se sugiere la exploración de la sensibilidad del modelo a la variación en el número de épocas de entrenamiento, considerando la posibilidad de incrementar este parámetro para evaluar su impacto en el rendimiento y la convergencia del modelo. La experimentación con otras arquitecturas de modelos, aparte de BERT.

Contribución de Autoría

Luisfelipe Rodrigo Mamani Arosquipa: Conceptualización, Análisis formal, Investigación, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición, Visualización, Metodología, Software, **Frank Jhoseph Duarte Oruro:** Conceptualización, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos.

Referencias

[1] P. Keung, Y. Lu, G. Szarvas y N. A. Smith, "The Multilingual Amazon Reviews Corpus", en Proc. 2020 Conf. Empirical Methods Natural Lang. Process. (EMNLP), Online. Stroudsburg, PA, USA: Assoc. Comput. Linguistics, 2020. Accedido el 21 de octubre de 2023. [En línea]. Disponible: https://doi.org/10.18653/v1/2020.emnlp-main.369

[2] M. V. Rao and S. C., "Detection of Sarcasm on Amazon Product Reviews using Machine Learning Algorithms under Sentiment Analy- sis," 2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 2021, pp. 196-199, doi: 10.1109/WiSPNET51692.2021.9419432.

[3] Jain, K. (2021). Amazon reviews [Data set].

[4] asin. Geophysical research letters, 25(2), 155-158.

[5] Rorato, A. C., Dal'Asta, A. P., Lana, R. M., Dos Santos, R. B., Escada, M. I. S., Vogt, C. M.,Codec¸o, C. T. (2023). Trajetorias: a dataset of environmental, epidemiological, and economic indicators for the Brazilian Amazon. Scientific Data, 10(1), 65.

[6] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee, "How opinions are received by online communities: A case study on Amazon.com helpfulness votes," arXiv (Cornell University), Jun. 2009, [Online]. Available: https://arxiv.org/pdf/0906.3741.pdf

[7] X. Zhao y Y. Sun, "Amazon Fine Food Reviews with BERT Model", Procedia Comput. Sci., vol. 208, pp. 401-406, 2022. Accedido el 26 de noviembre de 2023. [En línea]. Disponible: https://doi.org/10.1016/j.procs.2022.10.056

[8] A. Verma, C. Rawat y M. S. Gupta, "Sentiment Analysis for Amazon Product Reviews", Int. J. Recent Technol. Eng. (IJRTE), vol. 11, n.º 2, pp. 109-112, julio de 2022. Accedido el 26 de noviembre de 2023. [En línea]. Disponible: https://doi.org/10.35940/ijrte.b7099.0711222

[9] S. Iftikhar, B. Alluhaybi, M. Suliman, A. Saeed y K. Fatima, "Ama- zon products reviews classification based on machine learning, deep learning methods and BERT", TELKOMNIKA (Telecommunication Comput. Electron. Control), vol. 21, n.º 5, p. 1084, octubre de 2023. Accedido el 26 de noviembre de 2023. [En línea]. Disponible: https://doi.org/10.12928/telkomnika.v21i5.24046

Análisis de sentimiento en Twitter en relación a la tecnología IA para generación de imágenes

Sentiment analysis on Twitter in relation to AI technology for image generation

Antony Pyero Rosales Espinoza
Universidad Católica Sedes Sapientiae.
Junin-Tarma, Perú.
2015101350@ucss.pe

Juan Carlos Gonzales Suarez
Universidad Católica Sedes Sapientiae.
Junin-Tarma, Perú.
jgonzaless@ucss.pe

ARK: ark:/42411/s15/a125
DOI: 10.48168/innosoft.s15.a125
PURL: 42411/s15/a125

RECIBIDO 07/09/2023 • ACEPTADO 18/11/2023 • PUBLICADO 30/03/2024

RESUMEN

El avance en la tecnología de inteligencia artificial (IA) ha llevado a mejoras significativas en la generación de imágenes en términos de velocidad y calidad. Sin embargo, se ha generado preocupación e incertidumbre entre los artistas, quienes temen ser reemplazados por la IA en su campo de trabajo. En este contexto, se tuvo como objetivo el análisis de los Tweets donde se define el impacto de la inteligencia artificial (IA) en la adopción de tecnologías de generación de imágenes. Para ello, se llevó a cabo la recopilación, creación y evaluación de una red neuronal convolucional que clasifique los datos según un análisis de sentimiento entre positivo y negativo. Finalmente, la investigación se determinó la tasa de pérdida de un 63%, la precisión con un 61% y la curva ROC alrededor de un 64% de una red neuronal convolucional para la predicción de Tweets.

Palabras claves: Inteligencia artificial, Análisis de sentimiento, Red neuronal convolucional, Ámbito artístico, Twitter.

ABSTRACT

Advances in artificial intelligence (AI) technology have led to significant improvements in image generation in terms of speed and quality. However, it has generated concern and uncertainty among artists, who fear being replaced by AI in their field of work. In this context, the objective was to analyse Tweets defining the impact of artificial intelligence (AI) on the adoption of imaging technologies. For this purpose, the collection, creation and evaluation of a convolutional neural network that classifies the data according to a sentiment analysis between positive and negative

was carried out. Finally, the research determined the loss rate of 63%, the accuracy with 61% and the ROC curve around 64% of a convolutional neural network for predicting Tweets.

Keywords: *Artificial intelligence, Sentiment analysis, Convolutional neural network, Artistic field, Twitter.*

INTRODUCCIÓN

Las plataformas de redes sociales se han transformado en un espacio virtual equivalente a una plaza pública contemporánea, donde un gran número de personas debaten, discuten y comparten sus experiencias y puntos de vista.

Twitter ocupa una posición excepcional dentro del conjunto de las redes sociales más ampliamente utilizadas, ya que su número de usuarios diarios supera los 436 millones, que lo convierte en una plataforma de elección para la expresión y difusión pública de opiniones.

Un tema que se discute entre artistas actualmente en Twitter es la utilización de IA para la generación de imágenes que causa preocupación e incertidumbre por ser reemplazados por tecnología en el campo artístico y, por otro lado, las empresas que realizan estas herramientas no saben exactamente que mejorar o cambiar para satisfacer a las demandas de sus clientes.

Para ello se realiza un análisis de sentimiento donde se lleva a cabo la recopilación, diferenciación, análisis y obtención de la información necesaria del usuario para la mejoría del software.

Para realizar esta investigación se ha revisado fuentes que incluyen las redes neuronales profundas, convolucionales y recurrentes donde se ha decidido utilizar una red neuronal convolucional con Word2Vec que según Kalluri [1] es el más optimo como solución para obtener y analizar la información con los procesos de recopilación de datos, preprocesado de datos, incrustación y modelación.

En la sección 2 se realizó una revisión conceptual que abarcó la definición de diversos conceptos. La sección 3 se hace una comparación entre modelos DNN, CNN y RNN según lo propuesto por Kalluri[1]. La sección 4 se define los pasos clave de la solución. La sección proporciona una descripción detallada de la ejecución de la solución. La sección 6 se muestran los resultados y finalmente en la sección 7 se realiza las conclusiones.

Revisión conceptual

Para elegir el mejor modelo para la solución se describió diferentes conceptos y analizó una valorización de los criterios con la comparativa que realizo Kalluri [1] entre DNN, CNN y RNN con TF-IDF y Word2vec.

- Análisis de sentimiento: El análisis de sentimiento es la técnica que consiste en obtener información sobre una entidad y determinar sus subjetividades de forma automática. El propósito es identificar si el material generado por el usuario expresa sentimientos favorables, negativos o neutrales. La clasificación del sentimiento puede realizarse en tres niveles de extracción: nivel de aspecto o rasgo, nivel de frase y nivel de documento.
- Word Embedding (Incrustación de palabras): La mayoría de los algoritmos de aprendizaje automático no pueden interpretar cadenas o texto plano en su forma básica. Por el contrario, necesitan entradas numéricas para funcionar. Al convertir las palabras en vectores, la incrustación de palabras permite procesar grandes cantidades de datos de texto y adaptarlos a los algoritmos de aprendizaje automático. En este apartado se usa Word2vec [14] y TF-IDF [15,25] como técnicas de incrustación de palabras [8]. Se examina todo el texto y el procedimiento de construcción del vector se lleva a cabo identificando las palabras con las que la palabra objetivo aparece con más frecuencia. De este modo, también se muestra la proximidad semántica de las palabras, a diferencia de otras técnicas se lleva a cabo un procedimiento de aprendizaje no supervisado mediante redes neuronales artificiales, se utilizan datos no etiquetados para entrenar el modelo Word2Vec que crea vectores de palabras.
- Red neuronal profunda (DNN): Esta red neuronal tiene la característica distintiva de permitir captar automáticamente las funciones necesarias. El modelo de aprendizaje profundo es una función matemática f: X, Y. El aprendizaje profundo es el desarrollo de una red neuronal artificial (ANN) que emplea más de una capa oculta para modelar un conjunto de datos [7].
- Red neuronal Convolucional (CNN): consisten en múltiples capas de convoluciones con funciones de activación no lineales, como ReLU o tanh, aplicadas a los resultados. En una red neuronal feedforward convencional, cada neurona de entrada está conectada a cada neurona de salida de la capa siguiente. Esto también se conoce como una capa totalmente conectada [6]. Finalmente, los datos se preprocesan para la matriz de incrustación, pasan por capas y filtros para obtener un resultado de análisis de sentimiento entre positivo o negativo [23,24].
- Red neuronal recurrente (RNN): La red neuronal recurrente es una extensión de la red neuronal directa con una memoria interna ya que realiza la misma función para cada entrada de datos, mientras que el resultado de la entrada actual depende del cálculo anterior [11]. Una vez generada la salida, se duplica y se vuelve a introducir en la red recurrente. Para la toma de decisiones, evalúa tanto la entrada actual como el resultado

de la entrada anterior de la que ha aprendido. La red de memoria a corto plazo es una RNN avanzada, una red secuencial, que permite la persistencia de la información. Es capaz de resolver el problema de gradiente, pero son incapaces de reconocer las dependencias a largo plazo por eso se usan las LSTM [16,19] que significa memoria a corto plazo para evitar dificultades.

El preprocesamiento de los datos de entrada en parecida al CNN, pasando por celdas y funciones para la disminución de vectores de salida con el fin de dar un resultado (positivo o negativo).

Comparación de modelos

Ahora que sabemos cada concepto se pasa a la valorización de los criterios usando la Tabla 1 para determinar el puntaje según el rango.

Tabla 1. Tabla de evaluación de rangos para los criterios.

Rango de evaluación	Puntaje
0-6	1
6-8	2
8-10	3

Según el resultado de las pruebas que realizo Kalluri [1], lo reemplazaremos con un puntaje para determinar el modelo óptimo según la Tabla 1.

Tabla 2. Comparativa entre modelos según Kalluri[1].

Metrics	TF-IDF			Word2vec		
	DNN	CNN	RNN	DNN	CNN	RNN
Accuracy	0.7548	0.7563	0.5432	0.7702	0.8001	0.815
Recall	0.7423	0.7321	0.7623	0.7865	0.8012	0.8241
Precision	0.748	0.7366	0.7635	0.7845	0.8023	0.8269
F Score	0.764	0.7542	0.6412	0.7888	0.8074	0.818
AUC	0.746	0.754	0.7557	0.7875	0.8006	0.8214

Se observa la Accuracy (Exactitud), Recall (Rellamada), Precision, F Score y AUC (Area Under The Curve) para cada red neuronal y se realiza una comparación de los datos [17,21] según la Tabla 2.

Tabla 3. Comparativa entre métodos basado en la valorización del criterio.

Metricas	TF-IDF			Word2vec		
	DNN	CNN	RNN	DNN	CNN	RNN
Accuracy	2	2	1	2	3	3
Recall	2	2	2	2	3	3
Precision	2	2	2	2	3	3
F Score	2	2	2	2	3	3
AUC	2	2	2	2	3	3
Total	10	10	9	10	15	15

El análisis del resultado da un total de 15 para CNN y RNN por parte de Word2Vec como los más óptimos dando un empate entre los modelos según la Tabla 3.

Tabla 4. Comparativa entre métodos basado en el tiempo de procesamiento según Kalluri [1].

Metricas	TF-IDF			Word2vec		
	DNN	CNN	RNN	DNN	CNN	RNN
Accuracy	2	2	1	2	3	3
Recall	2	2	2	2	3	3
Precision	2	2	2	2	3	3
F Score	2	2	2	2	3	3
AUC	2	2	2	2	3	3
Total	10	10	9	10	15	15

Para hacer el desempate, se realizó una siguiente comparación, pero basado en el tiempo de procesamiento de cada red neuronal que según el tiempo nos da para CNN 36 minutos y 12 segundos que es mucho menor que el RNN que tarda 1 hora y 32 segundos según la Tabla 4.

Modelo de la solución

Para realizar el modelo se necesita diferentes pasos para llevar los mensajes o tweets a una clasificación por sentimiento.

- Recopilación de datos, se extrae los datos de twitter sin ninguna corrección o modificación, en otras palabras, son los datos brutos.
- Preprocesado de datos, se limpia los datos de toda clase de ruido, emojis, caracteres especiales, mayúsculas, entre otras.
- Preparación de la capa de incrustación, convierte los datos textuales en forma numérica para que se pueda interpretar para la clasificación.
- Red neuronal convolucional, los datos son clasificados por medio de la red neuronal dando un resultado numérico que se interpreta como positivo, neutro o negativo.

En la Figura 1, se muestra el esquema del flujo de trabajo que planeo Paredes [2] con el objetivo de brindar una comprensión más clara de la metodología empleada para el desarrollo del análisis de sentimiento.

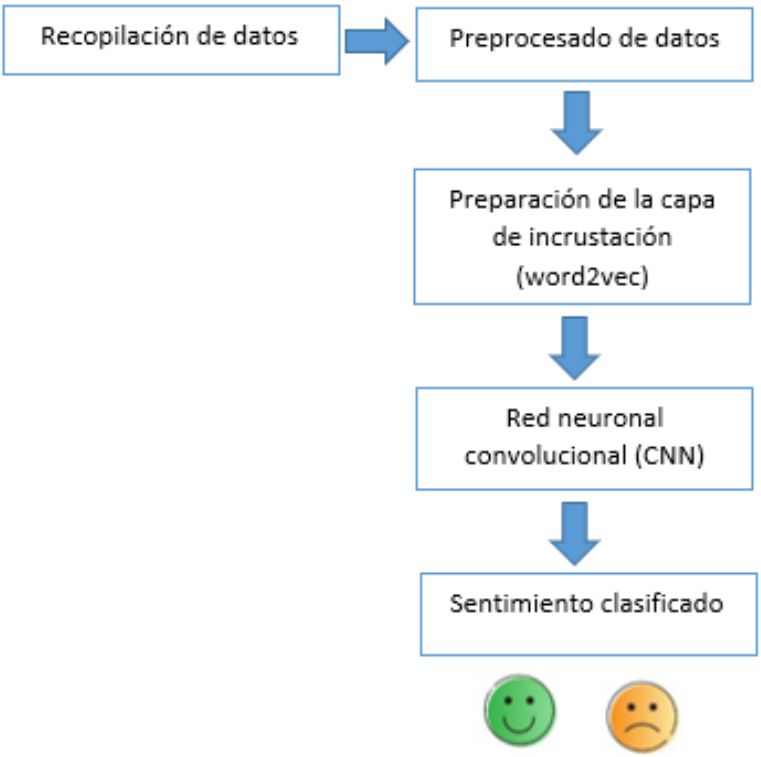

Figura 1. Esquema del flujo de trabajo según Paredes [2]

Desarrollo de la solución

Se realizo la solución siguiendo los pasos del modelo de la solución, en este apartado se dividió en 4 módulos:

Módulo de recopilación de datos de Twitter se desarrolló con la ayuda de la librería de Selenium que permite automatizar las acciones en el navegador web para recopilar los datos y ChromeDriver que facilita la interacción de Selenium con el navegador.

Tabla 5. Datos recopilados

Tweet
"Foxy Heaven IA cool #AIArt"
"GM, X Posting some tech-inspired stuff for #TechTuesday. Feel free to share your creations. #AIart #AIArtwork #AIArtCommunity #AIArtGallery #AIã,¤ãƒ©ã,¹ãƒˆ"
"A girl pilot in a blue pilot suit AI is cool. #BluePilotSuit
gooood morning! #AIart #AIã,¤ãƒ©ã,¹ãƒˆ #AIã,°ãƒ©ãƒ"ã,¢ #AIå½¼å¥³
Blending some styles this AMâ€¦ #AIart #AIArtwork #midjourneyart #midjourneyartwork
Starry Dreams! #AIArtGallery #AIArtworks #AIartists #AIart
"Out of all the vibrant, colorful raincoats out there, you choose transparent? This is beyond frustrating." #AIArtwork #AIart #pixai #AIgirl #AIã,¤ãƒ©ã,¹ãƒˆ #AIã,°ãƒ©ãƒ"ã,¢
All right people of X! How about some action. Itâ€™s a technical drawing time. Prompt: Patent drawing. Technical drawing of a [your fav object], [your fav colour] eyes colour. Letâ€™s tag friends into this.
Welcome To The Best MMORPG #IA
#AIart #nijijourney #fabricfolklore #AIArtwork fabric folklore 106
Silver knight #AIart

En la tabla 5, se muestra los datos recopilados en bruto que aún no han sido procesados. Módulo de preprocesado de datos, en esta etapa se realiza la limpieza de los tweets, este proceso se realiza de dos maneras: utilizando software o de forma manual. En este caso, se optó por la última opción con el objetivo de proporcionar coherencia a los datos. Para lograrlo, se siguieron indicaciones específicas, tales como:

- No debe contar con caracteres especiales, emojis, entre otros.
- No debe existir espacios entre vacíos entre filas.
- No se considera los caracteres especiales incluyendo las palabras que contengan "#" o "@" solo en caso de que sea una palabra dentro de la oración.
- No se considera elementos que tengan URL.
- No se considera frases sin sentido o ilegibles.
- Debe estar relacionado al tema artístico o IA.
- Debe existir coherencia las frases y el sentimiento.

Tabla 6. Datos recopilados y preprocesados

Tweet
"Foxy Heaven IA cool #AIArt"
"GM, X Posting some tech-inspired stuff for #TechTuesday. Feel free to share your creations. #AIart #AIArtwork #AIArtCommunity #AIArtGallery #AIã,¤ãƒ©ã,¹ãƒ"
"A girl pilot in a blue pilot suit AI is cool. #BluePilotSuit
gooood morning! #AIart #AIã,¤ãƒ©ã,¹ãƒ #AIã,°ãƒ©ãƒ"ã,c #AIå½¼å¥³
Blending some styles this AMâ€¦ #AIart #AIArtwork #midjourneyart #midjourneyartwork
Starry Dreams! #AIArtGallery #AIArtworks #AIartists #AIart
"Out of all the vibrant, colorful raincoats out there, you choose transparent? This is beyond frustrating." #AIArtwork #AIart #pixai #AIgirl #AIã,¤ãƒ©ã,¹ãƒ #AIã,°ãƒ©ãƒ"ã,c
All right people of X! How about some action. Itâ€™s a technical drawing time. Prompt: Patent drawing. Technical drawing of a [your fav object], [your fav colour] eyes colour. Letâ€™s tag friends into this.
Welcome To The Best MMORPG #IA
#AIart #nijijourney #fabricfolklore #AIArtwork fabric folklore 106
Silver knight #AIart

En la tabla 6, los datos han sido recopilado, preprocesados y clasificados con una librería de python llamada textblob aun así el error persiste por lo cual se realizó una revisión de forma manual.

Módulo de preparación de la capa de incrustación, en esta etapa se usa el word2vec para la incrustación de palabras, esta herramienta implementa el modelo de bolsa continua de palabras (CBOW) y el modelo de para calcular representaciones vectoriales de palabras [3,9].

La incrustación de las palabras es esencial en la estructura de las CNN, ya que posibilitan la obtención de información tanto sintáctica (estructura gramatical) como semántica (significado y contexto) de los tweets. Esta característica resulta crucial en el proceso de clasificación de sentimientos.

Antes de crear el word2vec, se dividió los datos entre: datos de entrenamiento y datos de prueba, por lo cual se usó los datos de entrenamiento para etiquetarlos que consiste en dividir las palabras en más pequeñas para facilitar el manejo y procesamiento como se muestra en la Figura 2.

```
[TaggedDocument(words=['AI',
'opens', 'the', 'door', 'for',
'far', 'more', 'people', 'to',
'create', 'art.', 'Good',
'artists', 'who', 'embrace', 'it',
'will', 'be', 'able', 'to', 'get',
'results', 'that', 'were', 'only',
'attainable', 'by', 'great',
'artists', 'before', 'AI,', 'and',
'great', 'artists', 'will',
'reach', 'into', 'the', 'realms',
'of', 'artistic', 'possibilities',
'that', 'we', 'still', "can't",
'imagine', 'yet.', 'AI', 'art...'],
```

Figura 2. Palabras tokenizadas de los Tweets

Luego se creó el word2vec con el uso de la librería de Python llamado Gensim y en base a los datos etiquetados se realizó el entrenamiento.

Módulo de red neuronal convolucional (CNN), en esta etapa se crea la CNN para clasificar los tweets en positivos y negativos. Su arquitectura requiere vectores de palabras concatenadas del texto como entrada.

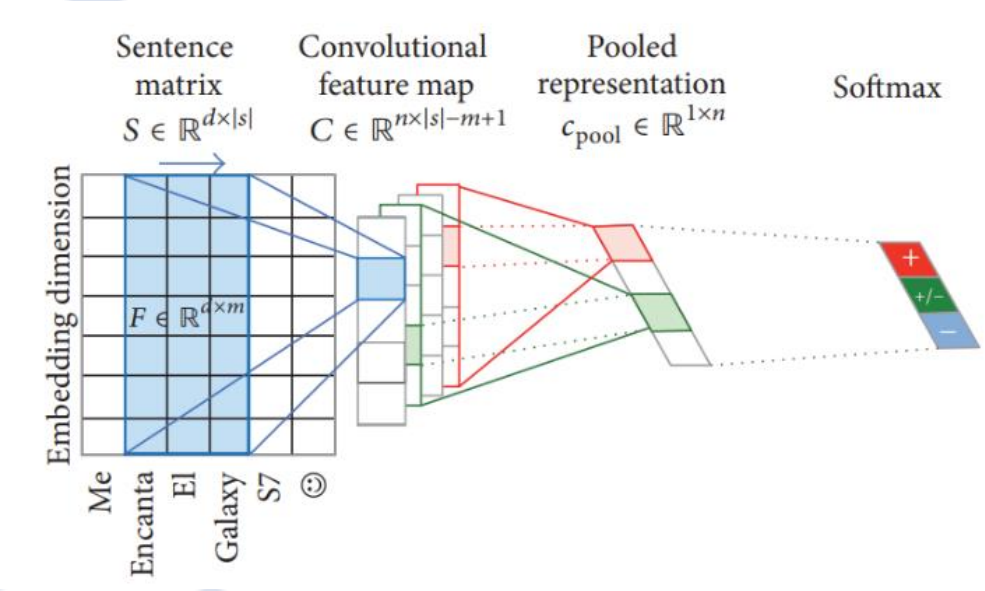

Figura 3. Modelo de aprendizaje para la clasificación de sentimientos según Severyn [2]

En la figura 3, se muestra la arquitectura de una red neuronal para la clasificación de sentimiento. Para realizar la CNN se utilizó la librería de Python Keras y se utilizó varias capas como:

- Capa de entrada, para el ingreso de los datos de los tweets.
- Capa de Embedding, para convertir a cada palabra a un vector [22].
- Capa de Conv1D, donde se hace las convoluciones de brigram, trigram, fourgram [12,13].
- Capa de GlobalMaxPool, donde se realiza la operación max pooling para reducir la dimensionalidad y mejorar el modelo [26].
- Capa de salida, se obtiene el resultado mostrando los valores de 0 (negativo) o 1 (positivo).

Además, se utilizó hiperparametros para ajuste del modelo como la tasa de aprendizaje (learning rate) que determina los ajustes de los pesos en cada iteración con un valor de 0.0001 y una tasa de perdida (dropout) para evitar el sobreajuste con un valor de 0.2.

Resultados

Los datos obtenidos son divididos en 2 partes:

- Datos para el entrenamiento con un 45.16% negativos y un 54.84% positivos.
- Datos para prueba con un 54.55% negativos y un 45.45% positivos.

Los datos tienen casi un 50% entre las dos partes dando una igualdad casi pareja entre positivo y negativo.
Por parte, en la red neuronal convolucional se tiene los siguientes resultados para determinar la eficiencia actual con los datos de entrenamiento:

Tasa de Perdida (Dropout): 0.6318122148513794 = 63%
Precisión: 0.6181818246841431 = 61%
ROC = 0.643 = 64%

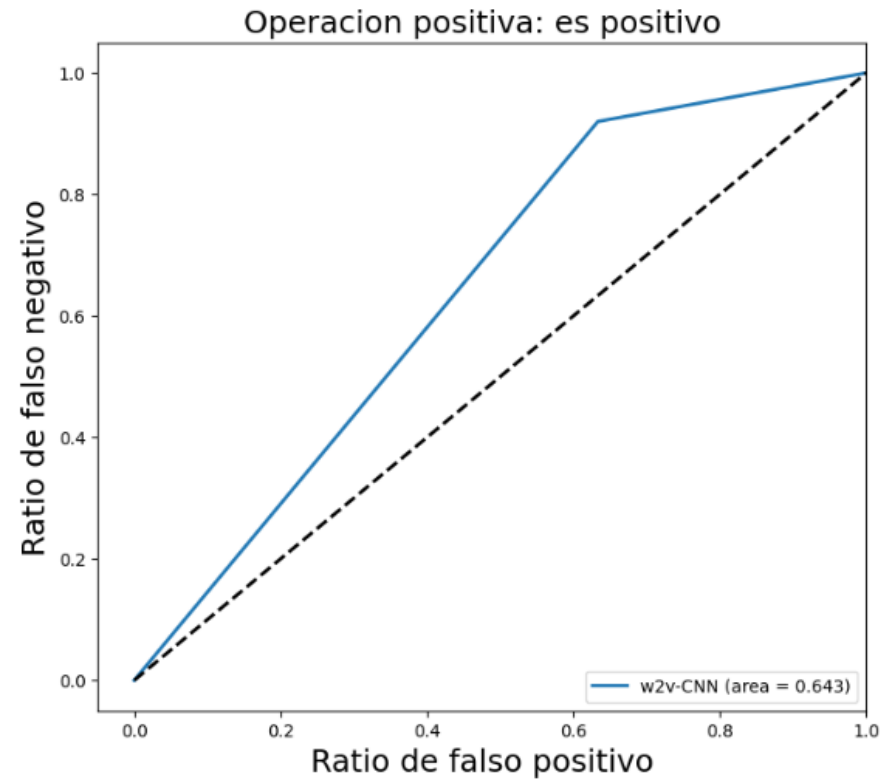

Figura 4. Curva ROC

En la Figura 4, se muestra gráficamente la curva ROC (Característica Operativa del Receptor) con los datos de eficiencia obtenidos de la CNN.

Tabla 7. Datos recopilados y preprocesados

Tabla de prediccion de los datos con la red neuronal

	Frases	Pred.	Pred.Red
121	Come Master. Let's finish quickly these meetin...	1	1
122	A picture says it all, increible	1	1
16	SUNSPROUT Knight IA is incredible	1	1
229	Mediocre artists see the writing on the wall. ...	0	1
149	If we're so mediocre, maybe y'all should stop ...	0	0
41	I feel cruelly cheated when I see an interesti...	0	1
249	That's not AI coloring, that's running your sk...	0	0
262	You can consume bad writing and good writing. ...	0	1
265	IA you suck	0	1
209	What I have said is that I’m an artist and I l...	1	1

En la Tabla 7, se muestra la predicción de los datos de prueba comparando con la predicción que realiza la CNN.

Conclusiones

En este estudio, se realizó un análisis de sentimiento con una red neuronal convolucional para clasificar los mensajes de tweets entre positivo y negativo dando los siguientes resultados:

- La precisión es un 61%, el porcentaje esperado era entre 70-80%, esto sugiere que la CNN no está llegando a la eficiencia deseada en la predicción.
- La tasa de perdida es de un 63%, el porcentaje esperado era de menos de un 15%, esto sugiere que la CNN tiene dificultades para minimizar el error durante el entrenamiento.
- ROC es de 64%, el porcentaje esperado era entre 70-80%. esto sugiere que la CNN no logra separar las clases adecuadamente entre positivo y negativo.

Al realizar la clasificación con los datos de prueba con la red neuronal ocurre errores con la predicción que sucede por la insuficiencia de datos para mejorar la capacidad predictiva de la red neuronal, en este caso se podría considerar la obtención de datos adicionales.También se puede mejorar agregando más especificaciones en el preprocesado como convertir las mayúsculas a minúsculas o seguir ajustando los hiperparametros para obtener un resultado más óptimo.

Contribución de Autoría

Antony Pyero Rosales Espinoza: Conceptualización, Análisis formal, Investigación, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición, Visualización, Metodología, Software, **Juan Carlos Gonzales Suarez:** Conceptualización, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Redacción - borrador original.

Referencias

[1] Kalluri, S (2023). Deep Learning Based Sentiment Analysis. Faculty of Faculty, Blekinge Institute of Technology, Karlskrona,Sweden. https://www.diva-portal.org/smash/get/diva2:1741487/FULLTEXT02.pdf

[2] Paredes-Valverde, M. A., Colomo-Palacios, R., Salas-Zárate, M. D. P. & Valencia-García, R. (2017). Sentiment Analysis in Spanish for Improvement of Products and Services: A Deep Learning Approach. Scientific Programming, 2017. https://doi.org/10.1155/2017/1329281

[3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS '13), pp. 3111-3119, December 2013.

[4] A. Severyn and A. Moschitti, "UNITN: training deep convolutional neural network for twitter sentiment classification," in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval '15), pp. 464-469, 2015.

[5] A. Severyn and A. Moschitti, "UNITN: training deep convolutional neural network for twitter sentiment classification," in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval '15), pp. 464-469, 2015.

[6] A. Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. WIREs Data Min. Knowl. Discov. 2018, 8, e1253.

[7] Kraus, M.; Feuerriegel, S. Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. Expert Syst. Appl. 2019, 118, 65-79.

[8] Ain, Q.T.; Ali, M.; Riaz, A.; Noureen, A.; Kamran, M.; Hayat, B.; Rehman, A. Sentiment analysis using deep learning techniques: A review. Int. J. Adv. Comput.Sci. Appl. 2017, 8, 424

[9] N. F. F. Da Silva, E. R. Hruschka, and E. R. Hruschka, "Tweet sentiment analysis with classifier ensembles," Decision Support Systems, vol. 66, pp. 170-179, 2014.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS '13), pp. 3111- 3119, December 2013.

[11] Britz, D. Recurrent Neural Networks Tutorial, Part 1-Introduction to Rnns. https://dennybritz.com/posts/wildml/recurrent-neural-networks-tutorial-part-1/

[12] M. del Pilar Salas-Zarate, M. A. Paredes-Valverde, J. Limon- ´ Romero, D. Tlapa, and Y. Baez-Lopez, "Sentiment classification of Spanish reviews: an approach based on feature selection and machine learning methods," Journal of Universal Computer Science, vol. 22, no. 5, pp. 691-708, 2016.

[13] P. Smith and M. Lee, "Cross-discourse development of supervised sentiment analysis in the clinical domain," in Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, pp. 79-83, 2012.

[14] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." arXiv preprint arXiv: 1301.3781 (2013). https://doi.org/10.1145/3388218.3388229

[15] Djoerd Hiemstra. 2000. A probabilistic justification for using tf× idf term weighting in information retrieval. International Journal on Digital Libraries 3, 2 (2000), 131-139.

[16] M. Cliche, "BB twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs," Apr. 2017.

[17] M. D. P. Salas-Zarate, R. Valencia-Garc ´ ´?a, A. Ruiz-Mart´?nez, and R. Colomo-Palacios, "Feature-based opinion mining in financial news: an ontology-driven approach," Journal of Information Science, 2016.

[18] Z. Wang, H. Wang, Z. Liu & J. Liu, "Rolling Bearing Fault Diagnosis

[19] Using CNN-based Attention Modules and Gated Recurrent Unit", Global Reliability and Prognostics and Health Management 7(2020) 6.

[20] Umarania, V., Juliana, A., & Deepab, J. (2023). Sentiment Analysis using various Machine Learning and Deep Learning Techniques. Journal of Computational Intelligence, 7(3), 245-260. https://doi.org/10.46481/jnsps.2021.308

[21] S. Md and S. Krishnamoorthy, "Student performance prediction, risk analysis, and feedback based on context-bound cognitive skill scores,"Educ. Inf. Technol., vol. 27, no. 3, pp. 3981-4005, 2022, doi: 10.1007/s10639-021-10738-2.

[22] D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou, "Coooolll: A Deep Learning System for Twitter Sentiment Classification," in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval '14), pp. 208-212, Dublin, Ireland, 2014.

[23] Bhavitha, B.; Rodrigues, A.P.; Chiplunkar, N.N. Comparative study of machine learning techniques in sentimental analysis. In Proceedings of the 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 10-11 March 2017; pp. 216-221.

[24] Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. WIREs Data Min. Knowl. Discov. 2018, 8, e1253.

[25] Sohangir, S.; Wang, D.; Pomeranets, A.; Khoshgoftaar, T.M. Big Data: Deep Learning for financial sentiment analysis. J. Big Data 2018, 5, 3

[26] D. Britz, "Understanding Convolutional neural networks for NLP," in WildML, WildML. http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

Clasificación de comentarios tóxicos en los Videojuegos

Classification of toxic comments in video games

Luis Fernando Luque Nieto

Universidad La Salle. Arequipa, Perú.
lluquen@ulasalle.edu.pe
https://orcid.org/0009-0008-2937-9620

Elmerson Ramith Portugal Carpio

Universidad La Salle. Arequipa, Perú.
eportugalc@ulasalle.edu.pe

ARK: ark:/42411/s15/a124
DOI: 10.48168/innosoft.s15.a124
PURL: 42411/s15/a124

RECIBIDO 21/09/2023• ACEPTADO 05/12/2023• PUBLICADO 30/03/2024

RESUMEN

La toxicidad puede tener un gran impacto en el compromiso y la satisfacción del jugador. Se trata de un fenómeno complejo que tiene causas y consecuencias diversas. Entre las causas más comunes se encuentran la anonimidad, la competitividad y la frustración. Las consecuencias pueden ser graves, como el acoso, el abandono del juego y el daño psicológico. Las empresas de juegos están trabajando para encontrar formas de abordar las formas de toxicidad en sus plataformas. Una de las interacciones más comunes con la toxicidad se produce en las ventanas de chat o en los sistemas de mensajería del juego. El trabajo propuesto es sacar algunos mensajes de chat que se dan en estos "lobby" o sacarlos de internet para así poder clasificarlos y determinar si el jugador que escribió en el chat cometió una infracción y dependiendo de la categoría tomar acciones en el caso.

Palabras claves: Aprendizaje automático, Chat tóxico, Procesamiento de lenguaje natural, Videojuegos.

ABSTRACT

Toxicity can have a major impact on player engagement and satisfaction. It is a complex phenomenon that has diverse causes and consequences. Among the most common causes are anonymity, competitiveness and frustration. The consequences can be serious, such as harassment, abandonment of the game and psychological damage. Gaming companies are working to find ways to address forms of toxicity on their platforms. One of the most common interactions with toxicity occurs in chat windows or in-game messaging systems. The proposed

work is to pull some chat messages that occur in these "lobbies" or take them offline so that they can be categorized to determine if the player who wrote in the chat committed an infraction and depending on the category take action on the case.

Keywords: *Machine learning, Toxic chat, Natural language processing, Video games.*

INTRODUCCIÓN

La toxicidad en los videojuegos, un fenómeno destacado en la era digital, se manifiesta a través de comportamientos agresivos y comunicación perjudicial en comunidades de jugadores en línea. Desde comentarios despectivos hasta amenazas y expresiones discriminatorias, estos comportamientos afectan negativamente la experiencia de juego, generando un ambiente hostil. La toxicidad se caracteriza por la agresión hacia adversarios y compañeros, contribuyendo a emociones desagradables y dificultades comunicativas. Factores como la anonimidad en línea, competitividad extrema y falta de moderación influyen en su prevalencia. Evaluar y abordar esta problemática de manera eficiente es un desafío; sin embargo, la aplicación del Procesamiento del Lenguaje Natural (NLP) ofrece una solución, permitiendo a usuarios y desarrolladores obtener calificaciones que indican el grado de toxicidad en los comentarios, agilizando el proceso y mejorando la experiencia de juego.

Motivación

Nuestro trabajo busca ayudar y fomentar un ambiente de tranquilidad y serenidad a los jugadores de la gran mayoría de juegos que tienen chat de texto incluido, para que les dé una clasificación general si un usuario tiene muchas quejas de comentarios tóxicos podría ser tachado como persona “no grata”. Las preguntas que busca responder este proyecto son:

P1: ¿Cuál es la proporción de cada categoría sobre un comentario de chat?
P2: ¿Cuál es la palabra que más se repite en cada categoría?
P3: ¿Cómo afecta la competitividad a la toxicidad en los videojuegos en términos de mental?
P4: ¿Qué características del lenguaje se pueden utilizar para identificar comentarios tóxicos?
P5: ¿Cómo se puede utilizar la educación para combatir la toxicidad en los videojuegos?
P6: ¿Cómo se puede utilizar el NLP para ayudar a los jugadores a identificar y responder a los comentarios tóxicos en los lobbys?

Trabajos Relacionados

El NLP se utilizará para analizar los mensajes en los chats de los juegos. El NLP se utilizará para identificar palabras y frases que se asocian con la toxicidad, como insultos, amenazas y acoso. En [2], este trabajo aborda el problema de la toxicidad en los juegos multijugador en línea. Se utiliza un novedoso marco de procesamiento del lenguaje natural para detectar malas palabras y clasificar los comentarios tóxicos en función de su gravedad. Los resultados muestran que la toxicidad está relacionada de manera no trivial con el éxito del juego, y cómo es que impacta a la hora de tomar decisiones importantes en la partida.

En el trabajo [3], habla del trolling es un comportamiento tóxico que se observa con frecuencia en las plataformas de comunicación en línea, pero sigue siendo un fenómeno poco claro. Esto se debe a que hay poca investigación empírica sobre él y a que no hay consenso entre los expertos sobre su definición. Este artículo tiene como objetivo proporcionar una visión general del trolling presentando una revisión de la literatura anterior sobre los comportamientos tóxicos que se consideran trolling en las comunidades en línea y en el contexto de los juegos en línea. Además, se realizó un cuestionario a 83 participantes para observar cómo perciben el trolling. El estudio también encontró que la percepción del trolling varía de persona a persona. Algunos participantes perciben el trolling como un comportamiento divertido e inofensivo, mientras que otros lo perciben como un comportamiento dañino y malicioso. El estudio concluye que el trolling es un problema complejo que requiere más investigación. Se necesitan más estudios para comprender mejor las motivaciones de los trolls. En el estudio [4], habla de cómo los juegos en línea son populares entre los jugadores para comunicarse, discutir estrategias y hacer amigos. Sin embargo, a menudo se enfrentan a discursos abusivos y de acoso.

Para abordar este problema, se utilizó un conjunto de datos de ciberbullying recopilados de los foros de World of Warcraft (WoW) y League of Legends (LoL) para entrenar modelos de clasificación, como Toxic-BERT, con el objetivo de detectar automáticamente comentarios abusivos. Los resultados muestran que el modelo Toxic-BERT logra una puntuación macro F1 del 82,69% para el foro LoL y del 83,86% para el foro WoW en el conjunto de datos de ciberbullying. Esto ayuda a mantener los foros de juegos limpios y amigables al identificar y eliminar comentarios ofensivos de manera automática En el trabajo [5] se aborda el crecimiento significativo de la interacción social en línea, que a menudo está plagada de comportamientos hostiles o agresivos, como el ciberacoso. Se desarrolla un clasificador de lenguaje tóxico basado en audio utilizando Redes Neuronales Convolucionales (CNN) auto-atentas. A diferencia de depender de términos de léxico individuales, se toma un enfoque más general para identificar expresiones tóxicas que considere el contexto acústico completo de las frases cortas o expresiones. La arquitectura propuesta utiliza el mecanismo de autoatención para capturar la dependencia temporal del contenido verbal al resumir toda la información relevante de diferentes partes de la expresión. El modelo de CNN auto-atentas basado en audio se evalúa en un conjunto

de datos público y otro interno, logrando un 75% de precisión, un 79% de recall y un 80% de recuperación en la identificación de grabaciones de discursos tóxicos.

Propuesta

Primero se buscó un dataset ya clasificado por gente experta en la materia de clasificación de textos, donde se evaluaron diferentes métricas para seleccionar uno. Antes de realizar el procesamiento propiamente dicho, es necesario llevar a cabo el pre procesamiento de los datos. Esto implica eliminar palabras vacías, puntuación, caracteres especiales y otros elementos innecesarios que no aportan información relevante al análisis. Después de haber preprocesador los datos, se procede al análisis de la toxicidad. En este paso, se calculan las probabilidades de las categorías de toxicidad que tiene el dataset, así como la categoría predominante en cada reseña. Esto permitirá comprender el tipo de texto que es de acorde a la categoría. Una vez clasificadas las reseñas, se exportan los datos en un archivo CSV para su posterior análisis. Esta exportación facilita el procesamiento y la manipulación de los resultados obtenidos. Finalmente, se procede a la visualización gráfica de los datos, donde las reseñas se muestran agrupadas por categorías del producto.

Esto permite identificar patrones y tendencias en las opiniones de los usuarios sobre los productos analizados. (ver Fig. 1).

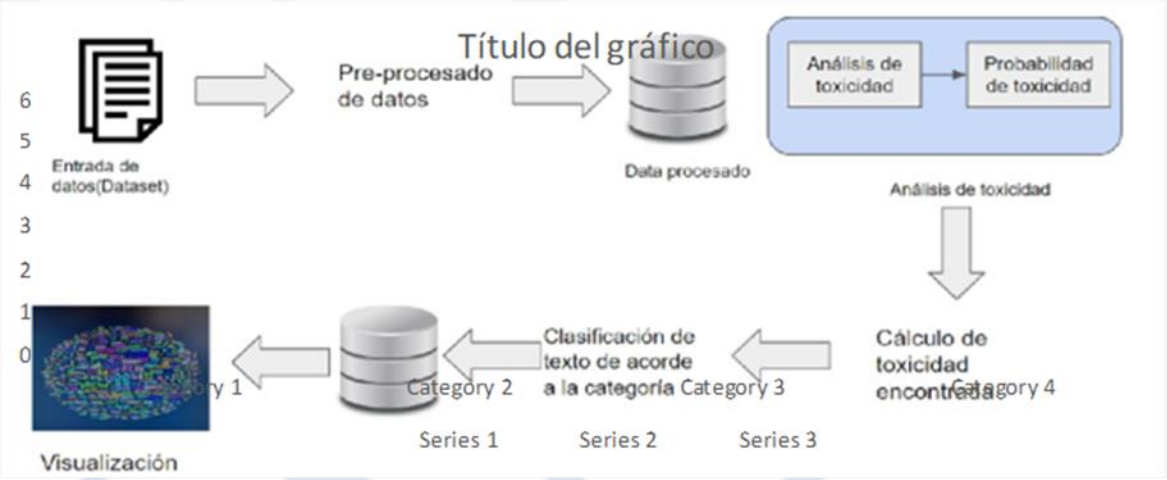

Figura 1. Pipeline Propuesto

Descripción de Data

El datasat consiste en más de 500 mil registros en csv, con su id, el comentario y la clasificación, siendo las clases “toxic”,“severe-toxic”, “obscene”, “threat”, “insult”,"identity_hate".

Tabla 1. Atributos y Descripción

Atributo	**Descripción**
Id	Identificador del comentario
Comment_text	El comentario como tal
classification	Tipo de toxicidad en el cual clasificado

Implementación

Para este trabajo, seguimos los pasos descritos en el algoritmo de la Tabla 2, implementado en Python, para abordar la clasificación de mensajes de chat en entornos de videojuegos con respecto a su toxicidad.

Tabla 2. Algoritmo Entrenamiento.

Algorithm 1: Entrenamiento del Modelo

def train_model(model, train_loader, val_loader, optimizer, device, max_steps_per_epoch, num_epochs):

1.train_loss_history = [] 5

2.val_loss_history = []

3.val_accuracy_history = []

4.for epoch in range(num_epochs):

5.model.train()

6.total_loss = 0

```
7.steps_taken = 0
8.for batch in train_loader:
9.if steps_taken >= max_steps_per_epoch:
10.break
11.input_ids, attention_mask, labels = batch
12.input_ids,attention_mask,labels=input_ids.to(device), attention_mask.to(device),
labels.to(device) 13.optimizer.zero_grad()
14.outputs = model(input_ids, attention_mask=attention_mask, labels=labels)
15.loss = outputs.loss
16.total_loss += loss.item()
17.loss.backward()
18.optimizer.step()
19.train_loss_history.append(loss.item())
20.steps_taken += 1
21.if (steps_taken + 1) % 10 == 0:
22.print(f"Training Step {steps_taken + 1}: Loss = {loss.item()}")
23.model.eval()
24.val_loss = 0
25.correct_predictions = 0
26.total_predictions = 0
27.with torch.no_grad():
28.for val_batch in val_loader:
29.val_input_ids, val_attention_mask, val_labels = val_batch
30.val_input_ids,val_attention_mask,val_labels =
val_input_ids.to(device),val_attention_mask.to(device),val_labels.to(device)
31.val_outputs = model(val_input_ids, attention_mask = val_attention_mask, labels =
val_labels)
32.val_loss += val_outputs.loss.item()
33.predictions = torch.argmax(val_outputs.logits, dim=1)
34.correct_predictions += torch.sum(predictions == val_labels).item()
```

```
35.total_predictions += len(val_labels)
36.val_accuracy = correct_predictions / total_predictions
37.val_loss_history.append(val_loss / len(val_loader))
38.val_accuracy_history.append(val_accuracy)
39.print(f"Epoch {epoch + 1}/{num_epochs}: "
40.f"Train Loss = {total_loss / steps_taken}, "
41.f"Val Loss = {val_loss / len(val_loader)}, "
42.f"Val Accuracy = {val_accuracy}")
```

Debido a que no contamos con los recursos físicos necesarios para poder hacer el decidimos utilizar Google Colab para agilizar el procesamiento de cada archivo CSV en la etapa de búsqueda de cada palabra de un comentario en los diccionarios. Nos centramos en la clasificación de comentarios tóxicos en videojuegos utilizando el modelo BERT.
Para reducir el tiempo de procesamiento, implementamos la biblioteca nativa multiprocessing de Python.

Resultados

El entrenamiento fue realizado en google Colab, con una TPU proporcionado por la misma google. El tamaño del dataset era de unos 40 MB aprox. Tuvimos algunas limitaciones. ya que Colab solo nos daba un tiempo límite de 12 horas (algunas veces solo eran 8). Generamos un Word-cloud, para cada categoría en la cual se puede apreciar cuáles son las palabras más usadas por cada categoría.

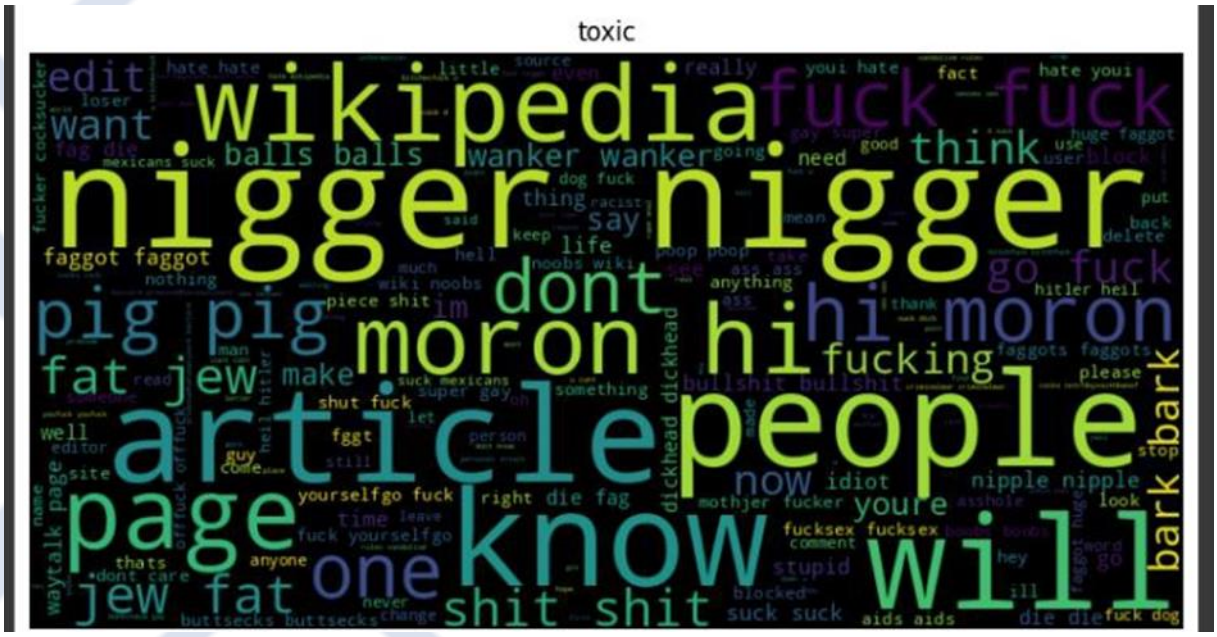

Figura. 2. WordCloud

Dadas las limitaciones de tiempo y al no contar con los recursos físicos necesarios, nuestro entrenamiento no pudo ser el esperado, el entrenamiento fue de 8 épocas y con 50 pasos por épocas. Con una función que evalúa el modelo se logró sacar los siguientes datos: Recall: F1 = Verdadero Positivo Verdadero Positivo + Falso Negativo Donde el F1 score se calculó de la siguiente manera: F1 = 2 * precisión * recall / (precision + recall)

Tabla. 2. Resultados de Bert.

Precisión	***Recall***	***F1 Score***
0.8325	*0.8900*	*0.8590*

Conclusiones y trabajos futuros

A lo largo de este proyecto, las restricciones temporales y de recursos han tenido un impacto significativo en el entrenamiento del modelo de Procesamiento del Lenguaje Natural (PLN). Aunque los indicadores de F1 Score y Precisión proporcionan resultados alentadores, es evidente que la duración limitada del entrenamiento ha afectado la capacidad predictiva del modelo.

A pesar de estas limitaciones, hemos adquirido conocimientos sustanciales en el campo del Aprendizaje Automático y el funcionamiento general de los modelos de Procesamiento del Lenguaje Natural. Reconocemos la necesidad de una inversión adicional de tiempo y recursos para lograr un entrenamiento más exhaustivo, lo que, a su vez, mejoraría la capacidad de clasificación del sistema.

Este proceso de aprendizaje ha sido invaluable, proporcionando una comprensión más profunda de cómo, incluso bajo condiciones subestimas, un algoritmo respaldado por un conjunto de datos sólido puede aprender y mejorar sus capacidades. Aunque las circunstancias actuales han impuesto limitaciones, aspiramos a contar con los recursos necesarios para llevar a cabo un entrenamiento más prolongado y detallado en el futuro.

Contribución de Autoría

Luis Fernando Luque Nieto: Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Elmerson Ramith Portugal Carpio:** Conceptualización, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura.

Referencias

[1] https://apdev.org.pe/la-toxicidad-en-los-esports/

[2] Märtens, M., Shen, S., Iosup, A., & Kuipers, F. (2015, December). Toxicity detection in multiplayer online games. In 2015 International Workshop on Network and Systems Support for Games (NetGames) (pp. 1-6). IEEE.

[3] Komaç, G., & Çağıltay, K. (2019, November). An overview of trolling behavior in online spaces and gaming context. In 2019 1st International Informatics and Software Engineering Conference (UBMYK) (pp. 1-4). IEEE.

[4] Vo, H. H. P., Tran, H. T., & Luu, S. T. (2021, August). Automatically detecting cyberbullying comments on online game forums. In 2021 RIVF International Conference on Computing and Communication Technologies (RIVF) (pp. 1-5). IEEE.

[5] Yousefi, M., & Emmanouilidou, D. (2021, August). Audio-based toxic language classification using self-attentive convolutional neural network. In 2021 29th European Signal Processing Conference (EUSIPCO) (pp. 11-15). IEEENewspaper Article from the Internet

Clasificación de comentarios suicidas en Reddit

Ranking Suicidal Comments on Reddit

Aron Josue Hurtado Cruz

Universidad La Salle. Arequipa, Perú.
ahurtadoc@ulasalle.edu.pe
https://orcid.org/0009-0002-0423-1730

Isabel Karina Ttito Campos

Universidad La Salle. Arequipa, Perú.
ittitoc@ulasalle.edu.pe
https://orcid.org/0009-0008-9159-7581

ARK: ark:/42411/s15/a123
DOI: 10.48168/innosoft.s15.a123
PURL: 42411/s15/a123

RECIBIDO 05/10/2023 • ACEPTADO 22/12/2023 • PUBLICADO 30/03/2024

RESUMEN

El proyecto se enfoca en el desarrollo de un algoritmo de Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés) diseñado para detectar comentarios suicidas en la plataforma Reddit y posteriormente realizar un análisis de sentimientos negativos con el propósito de brindar apoyo a los usuarios que puedan encontrarse en riesgo de suicidio. Para lograr este objetivo, el proyecto combina conceptos y técnicas de inteligencia artificial, procesamiento de lenguaje natural y psicología/psiquiatría.

Para evaluar la eficiencia del proyecto aplicamos la métrica F1 obteniendo un resultado bastante aceptable respecto a una clasificación textual.

Palabras claves: BERT, clasificación, NLP, depresión, suicidio.

ABSTRACT

The project focuses on the development of a Natural Language Processing (NLP) algorithm designed to detect suicidal comments on the Reddit platform and subsequently perform a negative sentiment analysis for the purpose of providing support to users who may be at risk of suicide. To achieve this goal, the project combines concepts and techniques from artificial intelligence, natural language processing and psychology/psychiatry.

To evaluate the efficiency of the project we applied the F1 metric obtaining a fairly acceptable result with respect to a textual classification

Keywords: *BERT, classification, depression, NLP, suicide.*

INTRODUCCIÓN

El Procesamiento de Lenguaje Natural es esencial para analizar el contenido de los comentarios, identificando signos de advertencia relacionados con la ideación suicida. Además, los aportes de la psicología y la experiencia clínica contribuyen significativamente a una comprensión matizada de estas señales de advertencia, asegurando una identificación más precisa de los usuarios en riesgo.

En este contexto, es relevante señalar que investigaciones anteriores han contribuido significativamente al desarrollo de algoritmos que abordan la detección de la ideación suicida. Yeskuatov et al. (2022) aprovecharon Reddit para la detección de la ideación suicida, proporcionando una revisión exhaustiva de las técnicas de aprendizaje automático y procesamiento de lenguaje natural [1]. Además, estudios realizados por Aldhyani et al. (2022) [2], Tadesse et al. (2019) [3], Awatramani et al. (2021) [4], Pal et al. (2018) [5] y Rahat et al. (2019) [6] han explorado diversas metodologías, desde el aprendizaje profundo hasta el análisis de sentimientos, enriqueciendo el panorama de herramientas disponibles para tareas críticas como esta. Estas referencias constituyen una base sólida para el desarrollo y la evaluación de nuestro algoritmo.

Trabajos relacionados

En esta sección, revisaremos algunos enfoques y estudios relacionados que se han centrado en problemas similares al propuesto en este proyecto.
Leveraging Reddit for Suicidal Ideation Detection: A Review of Machine Learning and Natural Language Processing Techniques.

Este artículo proporciona una revisión de la literatura de los artículos recientes que detallan los métodos de aprendizaje automático y procesamiento del lenguaje natural (NLP) utilizados para detectar la ideación suicida en Reddit. Los autores identifican tres enfoques principales:

- **Enfoques basados en reglas:** Estos enfoques utilizan un conjunto de reglas predefinidas, basadas en palabras y frases clave asociadas a la ideación suicida, para identificar publicaciones y comentarios que pueden indicar un riesgo de suicidio.

- **Enfoques basados en el aprendizaje automático: Estos** enfoques utilizan algoritmos de aprendizaje automático para entrenar un modelo que pueda predecir la probabilidad de que una publicación o comentario indique ideación suicida. Los modelos de aprendizaje automático se entrenan en un conjunto de datos de publicaciones y comentarios etiquetados como suicidas o no suicidas.

- **Enfoques híbridos:** Estos enfoques combinan elementos de los enfoques basados en reglas y en el aprendizaje automático.

Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models

Propone un enfoque para detectar y analizar la ideación suicida en las redes sociales utilizando modelos de aprendizaje profundo y aprendizaje automático.

Los autores del artículo utilizaron un conjunto de datos de publicaciones de Reddit que habían sido etiquetadas como suicidas o no suicidas. El conjunto de datos también incluía información sobre el historial de publicaciones de los usuarios.

Los autores utilizaron dos modelos para predecir la probabilidad de que una publicación fuera suicida:

- Un modelo de aprendizaje profundo basado en una red neuronal recurrente (RNN).
- Un modelo de aprendizaje automático basado en el algoritmo XGBoost.

Los resultados de la evaluación mostraron que el modelo de aprendizaje profundo pudo predecir la probabilidad de que una publicación fuera suicida con una precisión del 95 por ciento.

Detection of Depression-Related Posts in Reddit Social Media Forum

Este Artículo investiga el uso del lenguaje en Reddit para identificar mensajes que podrían indicar que un usuario está deprimido. Para ello utilizaron una combinación de técnicas de procesamiento del lenguaje natural (NLP) y aprendizaje automático para entrenar un modelo que pudiera clasificar los mensajes como relacionados o no relacionados con la depresión.

Los investigadores encontraron que los mensajes relacionados con la depresión tendían a utilizar más palabras relacionadas con la tristeza, la ansiedad, la culpa y la desesperanza. También encontraron que los mensajes relacionados con la depresión tendían a tener una mayor proporción de palabras negativas y una menor proporción de palabras positivas.

El modelo entrenado por los investigadores pudo clasificar correctamente los mensajes relacionados con la depresión con una precisión del 80 por ciento.

Detection of Suicidality among Opioid Users on Reddit: A Machine Learning Based Approach

El modelo entrenado pudo clasificar correctamente los mensajes suicidas con una precisión del 82 por ciento. Los investigadores identificaron varias características lingüísticas que estaban asociadas con el suicidio, como el uso de palabras relacionadas con la muerte, la desesperanza y la ideación suicida. También encontraron que los mensajes suicidas tendían a ser más cortos y menos coherentes que los mensajes no suicidas.

Sentiment Analysis of Mixed-Case Language using Natural Language Processing

El artículo propone un enfoque basado en el aprendizaje automático para el análisis de sentimientos de textos en lenguaje mixto. El enfoque utiliza una combinación de técnicas de NLP, como la detección de idiomas y la traducción automática, para preprocesar los textos en lenguaje mixto antes de aplicar un modelo de aprendizaje automático para clasificar los textos en categorías de sentimiento (positivo, negativo o neutral).

Materiales y métodos o Metodología computacional

En esta sección, revisaremos algunos enfoques y estudios relacionados que se han centrado en problemas similares al propuesto en este proyecto.

Leveraging Reddit for Suicidal Ideation Detection: A Review of Machine Learning and Natural Language Processing Techniques.

Este artículo proporciona una revisión de la literatura de los artículos recientes que detallan los métodos de aprendizaje automático y procesamiento del lenguaje natural (NLP) utilizados para detectar la ideación suicida en Reddit. Los autores identifican tres enfoques principales:

Recopilación de datos

Para llevar a cabo el proyecto, se utilizaron dos conjuntos de datos preexistentes:

- Conjunto de datos de detección de suicidio y depresión en Reddit: Este conjunto de datos se obtuvo de la plataforma Kaggle y contiene comentarios etiquetados como suicidas y no suicidas.

Los comentarios fueron recopilados de la plataforma Reddit y se proporcionan con etiquetas que indican la presencia de ideación suicida.

- Conjunto de datos de sentimientos negativos: Este conjunto de datos se obtuvo de Hugging Face a través de su API. Contiene comentarios etiquetados con sentimientos negativos, como tristeza, miedo y enojo. Este conjunto de datos se utilizó para realizar un análisis de sentimientos en combinación con la detección de ideación suicida

Ambos conjuntos de datos se utilizaron en la fase de entrenamiento del modelo y en la evaluación de su rendimiento.

Preprocesamiento de datos

Antes de utilizar los datos en la fase de entrenamiento, se aplicaron diversas técnicas de preprocesamiento para limpiar y preparar los textos. Estas técnicas incluyeron:

- Tokenización: División de los comentarios en palabras o tokens.
- Eliminación de stopwords: Eliminación de palabras comunes que no aportan significado.
- Lematización: Reducción de palabras a sus formas base.
- Eliminación de caracteres especiales: Eliminación de caracteres no alfabéticos y símbolos.

El preprocesamiento aseguró que los textos fueran representativos y estuvieran listos para la fase de entrenamiento.

Diseño del Modelo

1. Pipeline

El diseño de la propuesta puede verse reflejado en el siguiente pipeline:

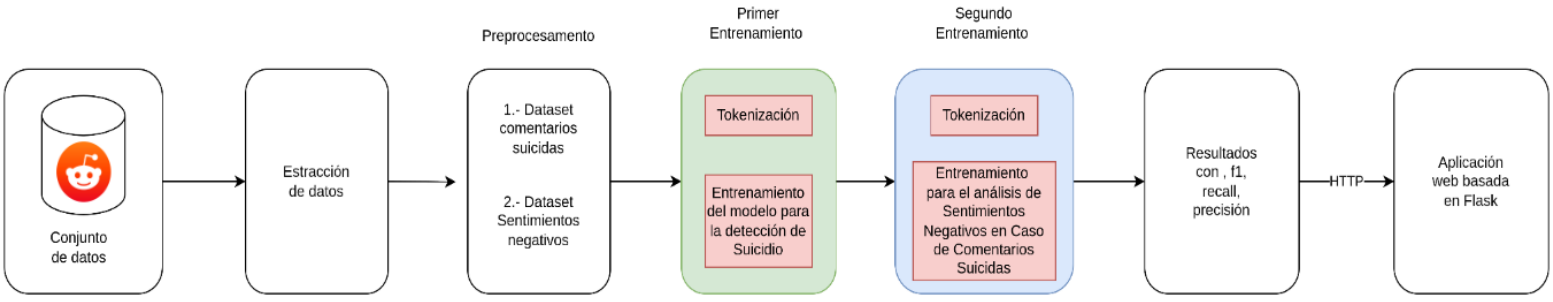

Figura 1. Pipeline descriptivo de la propuesta.

2. Entrenamiento del Modelo

Algorithm 1: Entrenamiento del Modelo

Data: Datos de entrenamiento y prueba
Result: Modelo entrenado y resultados del entrenamiento

```
while épocas do
    Inicializar métricas de entrenamiento;
    Inicializar métricas de validación;
    for cada época do
        Iniciar temporizador de época;
        // Entrenamiento
        for cada lote en datos de entrenamiento do
            Procesar lote;
            Calcular pérdida y actualizar pesos;
            Actualizar métricas de entrenamiento;
        end
        Calcular métricas promedio de entrenamiento;
        // Validación
        for cada lote en datos de validación do
            Procesar lote;
            Calcular pérdida y métricas de validación;
        end
        Calcular métricas promedio de validación;
        Registrar resultados y métricas en archivo;
    end
    Guardar modelo y tokenizer;
end
```

Figura 2. Entrenamiento del Modelo

3. Ventajas

- Contexto Bidireccional: BERT es capaz de capturar el contexto bidireccional en el que se encuentra una palabra en una oración, lo que le permite comprender mejor el significado de las palabras en relación con su contexto.

- Transferencia de Conocimiento: BERT se entrena en grandes cantidades de datos no etiquetados, lo que le permite capturar patrones lingüísticos generales. Esto hace que sea efectivo para tareas específicas con conjuntos de datos más pequeños, ya que el modelo ya tiene un conocimiento lingüístico general.

- Rendimiento de Estado del Arte: BERT ha demostrado superar a muchos modelos previos en una variedad de tareas de procesamiento del lenguaje natural, incluyendo traducción automática, respuesta a preguntas y análisis de sentimientos.

- Facilidad de Uso: BERT está disponible preentrenado y se puede ajustar fácilmente para tareas específicas mediante el entrenamiento con datos adicionales.

- Modelo de Atención: BERT utiliza un mecanismo de atención que le permite asignar pesos diferentes a diferentes partes de la entrada, lo que ayuda a capturar relaciones complejas.

4. Desventajas

- Requisitos Computacionales: BERT es un modelo grande y complejo que requiere recursos computacionales significativos. Su implementación puede ser costosa en términos de tiempo y potencia de cálculo.

- Memoria Necesaria: Debido a su tamaño, BERT puede requerir una cantidad considerable de memoria, lo que puede limitar su uso en dispositivos con recursos limitados.

- Interpretabilidad: Dada la complejidad del modelo, entender y explicar las

decisiones de BERT puede ser un desafío. La interpretabilidad del modelo es un área en la que se está trabajando activamente.

- No Considera el Orden de las Palabras: Aunque BERT es capaz de capturar el contexto bidireccional, no tiene en cuenta el orden exacto de las palabras en una oración, ya que su estructura de atención no es posicional.

Resultados y discusión

Comparativa

Tabla 1. Detección de comentarios suicidas

Modelo	SVM	LSTM	Naive Bayes(Bag of words)	Naive Bayes (TF-IDF)	Propuesto 10° Época
Precisión	87.82	88.63	96.70	98.40	94.00
Accuracy	89.51	90.12	91.12	90.12	94.00
Recall	91.41	91.71	85.16	81.60	94.00
F1 score	89.58	90.14	90.56	89.20	94.00

Tabla 2. Análisis de sentimientos negativos

Modelo	Propuesto
Precisión	97.00
Accuracy	98.00
Recall	97.00
F1 score	98.00

- SVM: Es un algoritmo de aprendizaje automático supervisado que se utiliza para clasificación binaria y multiclase. Se basa en la idea de encontrar un hiperplano que separe los datos de los dos o más grupos.
- LSTM: Es una arquitectura de red neuronal recurrente que se utiliza para tareas de PLN, como la traducción automática, el reconocimiento de voz y la generación de texto. Se basa en la idea de utilizar una celda de memoria para almacenar información a lo largo del tiempo.
- Naive Bayes: Se basa en la idea de que la probabilidad de que una instancia pertenezca a un cierto grupo es proporcional a la probabilidad de que se produzcan los atributos de la instancia dado que pertenece a ese grupo.

Resultados

Para todos estos resultados se usaron data sets equilibrados de 30K de datos, entre los que más resaltan es el entrenamiento con BERT y Naive Bayes (TF-IDF), aunque el método de NB es más rápido al sacar resultados BERT puede superar estos resultados con el entrenamiento exacto sin sobre entrenarlo.

En cuanto a este trabajo que se realizó en BERT, se usó el modelo de "bert-base-uncase" ya que nuestros recursos para trabajar este proyecto fueron muy limitados, al no contar con una tarjeta de video usamos google colab (T4 GPU) el cual nos prestaba su GPU por unas 6 horas promedio, limitándonos a usar data sets de 30K de datos por 10 epocas, cuando quisimos trabajar con otros modelos como "bert-large-uncase" o "bert-large-uncase-enmascared" que son más eficientes no pudimos ya que el T4 GPU de colab nos limitaba a solo usar "bert-base-uncased" y así no pudimos obtener los resultados deseados.

Conclusiones

- Potencial Impacto en la Salud Mental: El proyecto presenta una aplicación práctica de la tecnología para mejorar la salud mental al prevenir posibles tragedias. La detección temprana de usuarios en riesgo, incluso si no expresan abiertamente sus pensamientos suicidas, abre la puerta para la intervención oportuna de profesionales de la salud mental.
- Desafíos Computacionales: Aunque BERT ha demostrado ser una herramienta poderosa, su implementación puede ser computacionalmente costosa y requerir recursos significativos de hardware. Este desafío se debe tener en cuenta al considerar la escalabilidad y la implementación práctica del modelo.
- Resultados Aceptables: Los resultados obtenidos, especialmente en términos de precisión, recall y F1 score, indican que el modelo propuesto tiene un rendimiento aceptable. La aplicación de métricas como F1 score es crucial en tareas de clasificación de texto para evaluar el equilibrio entre precisión y recall.

Contribución de Autoría

Aron Josue Hurtado Cruz: Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Isabel Karina Ttito Campos:** Conceptualización, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos.

Referencias

[1] Yeskuatov E, Chua SL, Foo LK. Leveraging Reddit for Suicidal Ideation Detection: A Review of Machine Learning and Natural Language Processing Techniques. Int J Environ Res Public Health. 2022 Aug 19;19(16):10347. doi: 10.3390/ijerph191610347. PMID: 36011981; PMCID: PMC9407719.

[2] Aldhyani, Theyazn H. H., Saleh Nagi Alsubari, Ali Saleh Alshebami, Hasan Alkahtani, and Zeyad A. T. Ahmed. 2022. "Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models" International Journal of Environmental Research and Public Health 19, no. 19: 12635. https://doi.org/10.3390/ijerph191912635.

[3] M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Detection of Depression-Related Posts in Reddit Social Media Forum," in IEEE Access, vol. 7, pp. 44883-44893, 2019, doi: 10.1109/ACCESS.2019.2909180.

[4] P. Awatramani, R. Daware, H. Chouhan, A. Vaswani and S. Khedkar, "Sentiment Analysis of Mixed-Case Language using Natural Language Processing," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 651-658, doi: 10.1109/ICIRCA51532.2021.9544554.

[5] S. Pal, S. Ghosh, and A. Nag, "Sentiment Analysis in the Light of LSTM Recurrent Neural Networks," Int. J. Synth. Emot., vol. 9, pp. 33–39, 2018.

[6] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," in 8th Int. Conf. Syst. Model. Adv. Res. Trends, 2019, pp. 266–270.

Clasificador de Reseñas de Videojuegos de la Plataforma Steam

Steam Video Game Reviews Classifier

Luis Alberto Gonzáles Usca

Universidad La Salle. Arequipa, Perú.
lgonzalesu@ulasalle.edu.pe

Kevin Joel Linares Salinas

Universidad La Salle. Arequipa, Perú.
klinaress@ulasalle.edu.pe

Jose Alfredo Pinto Villamar

Universidad La Salle. Arequipa, Perú.
jpintov@ulasalle.edu.pe
https://orcid.org/0009-0000-5074-0187

ARK: ark:/42411/s15/a117

DOI: 10.48168/innosoft.s15.a117

PURL: 42411/s15/a117

RECIBIDO 18/10/2023 • ACEPTADO 11/01/2024 • PUBLICADO 30/03/2024

RESUMEN

Este documento utiliza un dataset ofrecido por la comunidad de Steam, el cual recopila más de 37 millones de recomendaciones de usuarios de distintos videojuegos, estos datos datos están cuidadosamente limpiados y preprocesados, y todos ellos son provenientes de la Steam Store, la cual es una plataforma de contenido descargable de videojuegos en línea. Lo que se hará con este dataset será un análisis de comentarios de cada usuario de la Steam Store con la finalidad de clasificar emociones, tanto negativas como positivas. El dataset está constituido por tres conjuntos de datos, donde utilizaremos solo las recomendaciones para realizar este trabajo.

Palabras claves: Videojuegos, Steam Store, Recomendaciones, Emociones

ABSTRACT

This paper leverages a dataset generously provided by the Steam community, encompassing over 37 million user recommendations for various video games. These meticulously cleaned and preprocessed data originate exclusively from the Steam Store, a platform for online downloadable content in the realm of video games. The primary objective of this study is to conduct a sentiment analysis of user comments within the Steam Store, discerning both negative and positive

emotions. The dataset comprises three distinct subsets, and this study focuses exclusively on the recommendations dataset for its analysis.

Keywords: *Video Games, Steam Store, Recommendations, Emotions*

INTRODUCCIÓN

Este documento usa una recopilación de datos de los comentarios de la plataforma Steam. Steam fue creado por la empresa de desarrollo Valve Corporation, y es creada con la finalidad de ofrecer al público actualizaciones automáticas de sus juegos, pero finalmente se optó por albergar juegos de terceros, haciendo que Steam se comporte como una unidad de comercio de videojuegos. Además, esta plataforma ofrece protección contra la piratería, garantizando a los usuarios la legitimidad de los títulos y liberándolo de los virus. Steam ha logrado tener bastante acogida por los usuarios, haciendo que estos formen comunidades y foros para discutir temas de los títulos de videojuegos, tanto que las empresas que desarrollan los mismos, tengan en consideración estos comentarios, con el fin de mejorar los videojuegos. Por ello se pretende con esta propuesta analizar las emociones de cada usuario frente a un conjunto de comentarios relacionados a los videojuegos. La interacción entre los usuarios y la plataforma ha generado un vasto y diverso conjunto de comentarios y reseñas. Estos comentarios reflejan la pasión y el compromiso de la comunidad de jugadores, y muchas veces son considerados por las empresas desarrolladoras como una valiosa fuente de retroalimentación para mejorar sus juegos. En este contexto, la propuesta de analizar las emociones de cada usuario frente a un conjunto de comentarios relacionados con videojuegos se vuelve aún más relevante. Este análisis puede arrojar luz sobre las preferencias, expectativas y experiencias de los jugadores en Steam, y proporcionar información valiosa para la industria del desarrollo de videojuegos. Al comprender las emociones de los usuarios, las empresas pueden ajustar sus estrategias y prioridades para ofrecer experiencias de juego más satisfactorias y alineadas con las necesidades de su audiencia.

Propuesta del Proyect

Motivación

La motivación de esta propuesta es saber y clasificar las emociones de los usuarios, frente a los títulos de los videojuegos, teniendo en cuenta cómo es que estos comentarios, influyen en la comunidad de la Steam Stor. A continuación, vamos a resolver algunas interrogantes, para concretar más los puntos de motivación de esta propuesta.

Q1: ¿Cómo varía la percepción de los usuarios hacia los videojuegos a lo largo del tiempo?

A medida que ha habido avances tecnológicos, los videojuegos han ido evolucionando a la par, tanto que han modificado la experiencia de juego, los gráficos y la historia de los mismos. Esto hace que los usuarios de hoy en día tengan una buena disposición a la hora de elegir un título para jugar. Ya sea que estos videojuegos están orientados a diferentes públicos, los videojuegos ofrecen una gama de experiencias a los usuarios, para que éstos disfruten largas horas de entretenimiento. Sin embargo, no todo en los videojuegos es positivo, existen usuarios que toman actitudes de violencia y problemas éticos, debido a la influencia de videojuegos de carácter más violento, haciendo que algunos, tomen una percepción de la realidad completamente distinta. Esto es cierta forma puede conllevar a problemas verdaderamente serios como convertirse en un individuo antisocial, por ello es necesario tomar todos los posibles escenarios de cómo los videojuegos alteran la percepción de algunos usuarios, y que mejor es a través de los comentarios pertinentes, frente a esta clase de sucesos, habrá comentarios en contra sobre los videojuegos de carácter violento, otros a favor y otros de forma neutral, tomando en cuenta, que solo se trata de entretenimiento.

Q2: ¿Qué características específicas de un juego influyen más en la opinión de los usuarios?

La comunidad siempre hace distinción entre calidad y precio, otros entre jugabilidad y reseña histórica, lo cierto es que actualmente, los usuarios, prefieren que los videojuegos tengan un mayor grado de visualización, obteniendo la mejor calidad de ellos, tanto para consolas específicas, como para computadores con tarjetas gráficas de gama media. Esto en su mayoría es muy difícil de compensar ya que hay títulos que realmente consumen bastantes recursos visuales, haciendo que la experiencia sea casi real, no obstante esto limita bastante a los usuarios, ya que no todos cuentan con las posibilidades de recrear esa misma experiencia en sus host de videojuegos, esto por el costo de tecnología que conlleva y otra por el precio del título del videojuego, ya que por supuesto, demanda más horas de dedicación a un videojuego que ofrece ese nivel de experiencia. Por otro lado, las características de reseña, jugabilidad, música, soporte, multijugador, son importantes también, pero no lo es tanto como la parte visual, así que si pudiéramos obtener un conjunto de usuarios que "gocen" de buenas prestaciones tecnológicas, los comentarios, serían más enriquecedores, no solo enfocándose a lo visual sino a la reseña y de que trata el videojuego por ejemplo.

Q3: ¿Cuáles son las frases o comentarios más recurrentes asociados con reseñas positivas o negativas?

Dentro de la comunidad de Steam, se encuentran un sin fin de títulos de videojuegos que ofrecen horas de entretenimiento, pero se hace notorio los comentarios de las personas que realmente no disfrutan de ciertos títulos y otros si. Comentarios como por ejemplo en el título DOTA 2 :

- "buen juego si quieres perder tu vida"
- "una !"#$ % lleno de peruanos"

- "buen juego mucho flamer pero buen juego"

Deduciendo, de forma simple y a percepción de usuario claro está, estos tres simples comentarios la mayoría de jugadores de este título frecuentemente están en desacuerdo a con qué otros jugadores se les calibre o empareje con servidores de Perú, otros que a pesar de que no les importa el emparejamiento, disfrutan el juego, pero en su mayoría todos se refieren no directamente al videojuego sino a la gente que juega en él.

Problema

En la industria de los videojuegos y en las decisiones relacionadas con el desarrollo y distribución de juegos en la plataforma Steam, se enfrenta un desafío similar: la falta de herramientas efectivas para evaluar de manera precisa y automatizada las emociones y opiniones expresadas en las reseñas escritas por los jugadores. La carencia de esta información dificulta la capacidad de los desarrolladores, editores y distribuidores para comprender y responder a las reacciones de la comunidad de jugadores de manera oportuna y precisa, lo que puede resultar en la creación de juegos que no cumplen con las expectativas de los jugadores o que no alcanzan su máximo potencial en términos de ´éxito comercial y crítico en la plataforma Steam. Por lo tanto, el problema central radica en la necesidad de desarrollar un enfoque efectivo y preciso para calificar emocionalmente las reseñas de juegos a través de oraciones [1], lo que permitirá una mejor comprensión de las opiniones y emociones de los jugadores y, en última instancia, mejorará la toma de decisiones en la industria de los videojuegos en Steam

Objetivo

Habiendo explorado un poco de este universo de los videojuegos y como es que los usuarios influyen en gran parte al desarrollo de los mismos, con la retroalimentación mediante los comentarios, hacia los desarrolladores. El objetivo principal de esta investigación es desarrollar un clasificador de texto de emociones de reseñas de los videojuegos en la plataforma Steam, mediante el modelo de procesamiento de lenguaje natural (NLP) que pueda analizar de manera precisa y automatizada las oraciones en las reseñas de videojuegos de Steam, asignando calificaciones emocionales a cada oración. Este modelo permitirá una comprensión más profunda de las emociones expresadas por los jugadores en sus reseñas, lo que facilitará la identificación de tendencias y patrones en la percepción de los videojuegos de la plataforma Steam

Datos

Para esta propuesta, hemos obtenido un dataset de la comunidad de Steam, basado en las recomendaciones de los usuarios, frente a distintos títulos de videojuegos, esto con la finalidad de que el clasificador obtenga un porcentaje significativo a la hora de analizar nuevos comentarios ingresados. A continuación, se resolverán las siguientes interrogantes:

Q1: ¿Qué datos necesitara?

Se necesitará el dataset de la plataforma de distribución de videojuegos Steam, sobre todo de la Steam Store, donde los usuarios antiguos, hacen una pequeña reseña del título ya jugado por ellos, y así obtener una mejor perspectiva de que es lo que busca el usuario a través de sus comentarios.

Q2: ¿Cómo recolectarán los datos?

La recopilación de datos se llevará a cabo buscando conjuntos de datos disponibles en páginas web o fuentes públicas en línea que contengan reseñas de películas junto con calificaciones emocionales. Esto puede implicar la descarga de conjuntos de datos existentes.

Q3: ¿Dónde planea obtenerlos?

Se planea obtener los datos de un conjunto de datos disponible en la página web Kaggle. Kaggle es una plataforma conocida por proporcionar conjuntos de datos para una variedad de tareas de análisis de datos y aprendizaje automático.

Q4: ¿Cómo planea almacenarlos?

Para almacenar los datos, se considera que los datos se ingresarán en un archivo con formato .xls de Excel, lo que posibilitará su fácil visualización y facilitará su posterior modificación y análisis.

Q5: ¿Cómo accederá a ellos para utilizarlos en su proyecto?

Los datos se encuentran en formato CSV, lo que facilita su acceso y la utilización en este proyecto ya que puedes cargar los datos desde el archivo CSV en el software que utilizaremos para el análisis de sentimientos en las reseñas de videojuegos. El formato CSV es ampliamente compatible y permitirá manipular y analizar los datos de manera efectiva en este proyecto.

Diseño

Q1: ¿Cómo sugiere usar la herramienta para ayudar a los usuarios a realizar las tareas que respondan a las preguntas que enumeró en la motivación?

Para ayudar a los usuarios a responder las preguntas que se enumeraron en la motivación y lograr una comprensión efectiva de las opiniones en comentarios de la plataforma Steam basadas en sentimientos, proponemos el diseño de una herramienta basada en una interfaz amigable y un sistema robusto en la parte del backend: En la parte de las representaciones específicas:

- Búsqueda Avanzada: Una herramienta de búsqueda que permite a los usuarios filtrar reseñas basadas en determinados criterios, como fecha, género de la película, entre otros.
- Resumen de Sentimientos: Una sección que muestra frases o oraciones más comunes asociadas a reseñas positivas o negativas
- Sobre el diseño de interfaz: Sección de Filtrado: Herramientas de filtrado y clasificación para que los usuarios puedan segmentar las reseñas según sus necesidades. métricas y representaciones gráficas.
- Modelo NLP Robusto: Implementar un modelo de NLP que haya sido entrenado y validado en el conjunto de datos, capaz de clasificar oraciones y reseñas según el sentimiento.
- Optimización de Consultas: Dado que se trabajará con un volumen significativo de reseñas, el código debe estar optimizado para realizar consultas rápidas y eficientes, posiblemente utilizando técnicas como indexación y caché.
- Modularidad: El código debe estar organizado en módulos o clases específicas, facilitando la posibilidad de ampliaciones futuras o mejoras en la herramienta.
- Seguridad: Implementar medidas de seguridad para proteger la integridad de los datos y garantizar que la herramienta no sea vulnerable a ataques.

Revisión Literaria

En el mundo contemporáneo, la adquisición de videojuegos ha experimentado una transición fundamental, pasando de las tiendas locales a las plataformas digitales, como Steam. Esta evolución ha transformado la forma en que los consumidores obtienen información sobre los videojuegos que desean comprar. Anteriormente, la adquisición de conocimiento sobre lanzamientos de videojuegos se basaba en la retroalimentación de entusiastas del juego, revistas especializadas y recomendaciones de amigos [2]. Hoy en día, las reseñas de usuarios en línea se han convertido en una fuente esencial de información para los jugadores, desarrolladores y las propias plataformas de venta. En el primer estudio, los autores analizan diferencias entre las revisiones de videojuegos etiquetadas como útiles y no útiles en la plataforma Steam. Recopilan una gran cantidad de revisiones, extraen diversas características, realizan pruebas de hipótesis estadísticas y llevan a cabo experimentos predictivos. Sus hallazgos indican que existen

diferencias significativas entre ambos grupos, destacando la influencia de la longitud de la reseña y el tiempo dedicado a jugar en la percepción de utilidad. Estos resultados proporcionan valiosas ideas para los desarrolladores sobre como apoyar a la comunidad, como brindar retroalimentación inmediata a los autores de las reseñas [3].

En el segundo estudio, se aborda la importancia del análisis de sentimiento en las revisiones de videojuegos en plataformas de comercio electrónico, cómo Steam. Los resultados demuestran que tanto la magnitud del texto como el sentimiento pueden predecir la recomendación final de un jugador para un videojuego. Esto destaca la relevancia del análisis de sentimiento en el contexto de la toma de decisiones de compra en el ámbito de los videojuegos [4].

En el tercer estudio, se investiga el rendimiento de las técnicas de análisis de sentimiento en revisiones de videojuegos. Se evalúan tres clasificadores ampliamente utilizados y se identifican las causas subyacentes de las clasificaciones incorrectas. El estudio sugiere que la mayoría de los clasificadores no funcionan de manera ´optima en revisiones de videojuegos, y señala cuatro causas principales de clasificaciones erróneas, como las revisiones que mencionan ventajas y desventajas del juego. Los autores hacen un llamado a la comunidad de investigadores y desarrolladores para abordar estos desafíos y mejorar el análisis de sentimiento en revisiones de videojuegos [5].

Diseño

La propuesta que nosotros presentamos, pretende brindar una ayuda o soporte, a los usuarios, empresas desarrolladoras de videojuegos, en cuanto a cómo los jugadores, mediante sus comentarios de títulos comprados, influyen en la compra, desarrollo o jugabilidad de los videojuegos, ya que se quiere clasificar las emociones mediante estos comentarios, y así poder "medir" las emociones de cada usuario. Para ello, hemos generado una lista de etapas donde son necesarias para lograr este objetivo y el diseño del clasificador de emociones:

1. Método de análisis Utilizamos un enfoque cuantitativo, para analizar las emociones de la Steam Store, y para ello hemos recopilado una muestra de 80000 registros de reseñas o comentarios de jugadores, que han calificado mediante los mismos, diferentes títulos de videojuegos. Después de esta recopilación, lo que necesitamos es usar NLP, constituido por diferentes tareas de procesamiento, lo cual ayudara clasificar los comentarios en tres categorías: emociones positivas, emociones negativas y emociones neutras.
2. Decisiones de diseño Las decisiones del diseño fueron tomadas de acuerdo a: fecha de publicación, divertido, útil, hora de reproducción, revisión de acceso temprano, recomendación, revisión, título. Ya habiendo clasificado estos criterios, nuestro paso será tomar las muestras necesarias que cumplan estos requisitos.
3. Público Objetivo El público objetivo de este clasificador de emociones, son los jugadores, desarrolladores de videojuegos y cualquier otro individuo que pueda interesar esta información. Además, lo que se quiere es mostrar una diversidad de emociones que nacen

frente a un título de videojuego.

4. Pipeline Finalmente el pipeline refleja el proceso que nos llevará al resultado final que es el clasificador de emociones de reseñas de videojuegos de la Steam Store. A continuación, describiremos de forma gráfica el flujo de trabajo de esta propuesta:

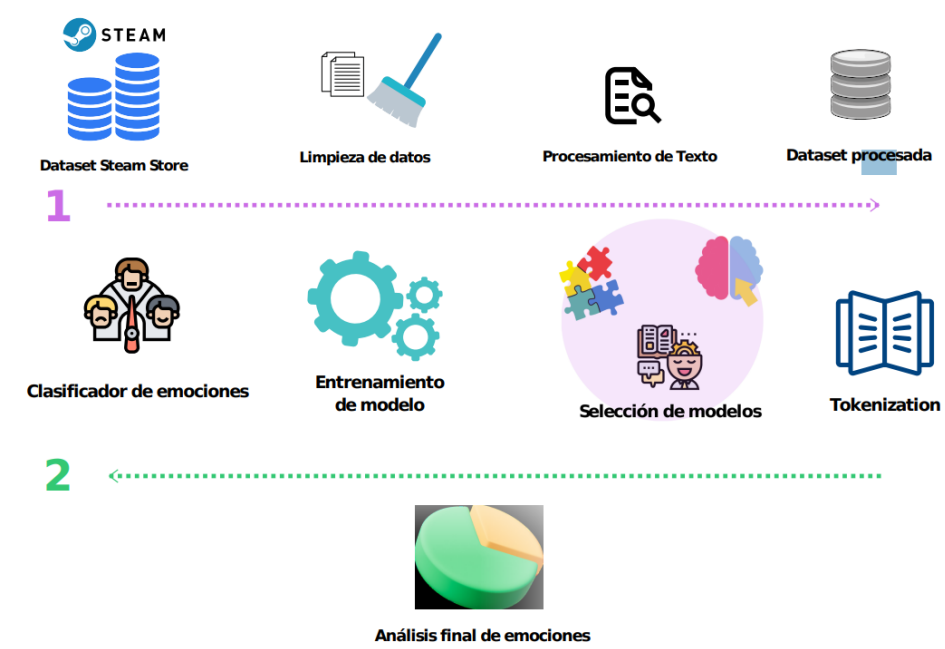

Figura 1. Pipeline del clasificador de emociones basado en reseñas.

- Dataset Steam Store, es la base de datos obtenida de las reseñas de todos los jugadores, el cual para este caso concreto solo tomaremos 80000 registros.
- Limpieza de datos, ya que existen comentarios donde algunos usuarios, realizan inserciones no deseadas en los textos, como por ejemplo caracteres numéricos, símbolos, es necesario hacer una limpieza o filtración de los mismos.
- Procesamiento de texto, una vez ya limpio, tenemos que agrupar los comentarios, de manera que se puedan clasificar por título del videojuego, esto con la finalidad de obtener un mejor resultado.
- Dataset procesado, ya la data está limpia y ordenada, y es momento de realizar tareas propias de NLP.
- Tokenization, lemmatization, post-tagging, NLG, entre otras tareas de NLP.
- Selección de modelos, aquí debemos elegir metodologías acordes a realizar nuestro clasificador de emociones, que incluyen NLP, Machine Learning.
- Entrenamiento del modelo, una vez decidido el modelo, tenemos que entrenar el mismo con un 80 % de los registros, con la finalidad de que tenga una precisión optima y un 20 % con fines de pruebas.

- Clasificador de emociones, ya habiendo pasado por cada etapa, el clasificador de emociones, estará listo para ser probado de manera abierta y posteriormente le agregaremos una interfaz al usuario.
- Análisis final de emociones, al final de esta propuesta, obtendremos datos significativos, que serán relevantes para los usuarios, como para los desarrolladores de videojuegos y otros, que quieran hacer uso de estos datos. Figura 1.

Propuesta de Implementación de Algoritmo

Algorithm 1: Train the Model

Data: model: Initialized BERT model, train_dataloader: Prepared training data, num_epochs: Number of training epochs

Result: model: Trained BERT model

```
for epoch in range(num_epochs): do
    model.train();
    total_loss ← 0;
    all_predictions ← [];
    all_labels ← [];
    for batch in train_dataloader: do
        inputs, labels ← batch;
        optimizer.zero_grad();
        inputs ← inputs.to(device);
        labels ← labels.to(device);
        outputs ← model(inputs, labels=labels);
        loss ← outputs.loss;
        total_loss + = loss.item();
        predictions ← torch.argmax(outputs.logits, dim=1);
        all_predictions.extend(predictions.cpu().numpy());
        all_labels.extend( labels.cpu().numpy()) ;
        loss.backward();
        optimizer.step();
    accuracy ← accuracy_score(all_labels, all_predictions);
    precision ← precision_score(all_labels, all_predictions, average='weighted');
    recall ← recall_score(all_labels, all_predictions, average='weighted');
    f1 ← f1_score(all_labels, all_predictions, average='weighted');
return model
```

Comparación

Esta tabla compara el rendimiento de tres modelos diferentes (BERT, Naive Bayes y RoBERTa) en términos de métricas de evaluación comunes para la clasificación de texto, como accuracy, precisión, recall y F1-score.

Modelo	Acuaracy	Precisión	Recall	F1-score
BERT	0.96	0.96	0.96	0.96
Naive Bayes	0.78	0.81	0.74	0.77
RoBERTa	0.85	0.85	0.85	0.85

Tabla 1. Comparación de modelos

- BERT muestra un rendimiento consistente y fuerte en todas las métricas con valores de 0.96 en accuracy, precisión, recall y F1-score.
- Naive Bayes [6], aunque presenta un rendimiento decente, exhibe resultados inferiores en comparación con BERT y RoBERTa, con un accuracy de 0.78 y valores más bajos en precisión, recall y F1-score.
- RoBERTa también muestra un rendimiento sólido y equilibrado con valores de 0.85 en todas las métricas.

En esta comparativa, BERT sigue siendo el líder en términos de rendimiento general, seguido por RoBERTa, mientras que Naive Bayes muestra resultados más modestos en todas las métricas evaluadas.

Resultados

Resultados de la implementación del modelo de BERT Se obtuvieron los siguientes resultados de efectividad durante el entrenamiento de BERT. El entrenamiento duró 2 horas con 30 minutos con una cantidad de 13 épocas las cuales desde la época 11 siguió con un accuracy y precisión de 0.96 por lo cual no el entrenamiento no iba a mejorar por más épocas que sigamos usando.

Estos valores representan la evolución de las métricas de evaluación durante el entrenamiento del modelo. Se observa una tendencia alentadora de mejora progresiva en todas las métricas a medida que el modelo avanza en las ´épocas de entrenamiento. La pérdida (Loss), una medida de la discrepancia entre las predicciones del modelo y los valores reales, disminuye notablemente desde 1.081 en la primera época hasta 0.114 en la época trece, indicando una mejor capacidad del modelo para ajustarse a los datos de entrenamiento. Las métricas de Accuracy, Precisión, Recall y F1-score también reflejan un aumento constante, indicando un mejor rendimiento del modelo en la clasificación correcta y balanceada de las clases a lo largo del tiempo.

Conclusiones

Este proyecto ha desarrollado un clasificador de comentarios de usuarios en la plataforma Steam mediante el uso de BERT, una arquitectura avanzada en el campo del Procesamiento del Lenguaje Natural (NLP).

Época	Pérdida	Accuracy	Precisión	Recall	F1
1	1.08133341	0.52	0.52	0.52	0.52
2	0.80510457	0.66	0.66	0.66	0.66
3	0.64701713	0.73	0.73	0.73	0.73
4	0.51307071	0.8	0.8	0.8	0.8
5	0.40484561	0.85	0.85	0.85	0.85
6	0.3164701	0.88	0.88	0.88	0.88
7	0.24869163	0.91	0.91	0.91	0.91
8	0.19934077	0.93	0.93	0.93	0.93
9	0.17175587	0.94	0.94	0.94	0.94
10	0.14212216	0.95	0.95	0.95	0.95
11	0.12243625	0.96	0.96	0.96	0.96
12	0.1069883	0.96	0.96	0.96	0.96
13	0.1144044	0.96	0.96	0.96	0.96

Tabla 2. Resultados de entrenamiento de BERT

El objetivo principal fue evaluar el rendimiento del modelo a medida que se entrenaba para clasificar comentarios de usuarios de Steam. A lo largo de trece épocas de entrenamiento, se observó un progreso significativo en las métricas clave de evaluación. La métrica de pérdida (Loss) disminuyó consistentemente de 1.081 en la primera época a 0.114 en la época trece, indicando una mejora sustancial en la capacidad del modelo para adaptarse a los datos de entrenamiento y predecir las etiquetas de los comentarios. Además, las métricas de accuracy, precisión, recall y F1-score mostraron un crecimiento constante, lo que sugiere la capacidad del modelo para clasificar con precisión los comentarios de los usuarios en Steam, manteniendo un equilibrio entre la capacidad de predecir las clases positivas y negativas. La clasificación efectiva de comentarios en Steam tendría diversas aplicaciones beneficiosas. Por ejemplo, podría ser un componente crucial en el desarrollo de sistemas de recomendación más precisos y personalizados para los usuarios, permitiendo la identificación y promoción de contenido relevante de manera más efectiva. Además, la capacidad de clasificar automáticamente comentarios según su tono o sentimiento podría proporcionar información valiosa para mejorar la satisfacción del usuario y la gestión de la comunidad en la plataforma.

Contribución de Autoría

Luis Alberto Gonzales Usca: Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Kevin Joel Linares Salinas:** Conceptualización, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos. **Jose**

Alfredo Pinto Villamar Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición.

Referencias

[1] U. Rajapakshe, "Development centric player feedback analysis for video games: A review," in 2019 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS). IEEE, 2019, pp. 190–194.

[2] H.-N. Kang, H.-R. Yong, and H.-S. Hwang, "A study of analyzing on online game reviews using a data mining approach: Steam community data," International Journal of Innovation, Management and Technology, vol. 8, no. 2, p. 90, 2017.

[3] L. Eberhard, P. Kasper, P. Koncar, and C. G¨utl, "Investigating helpfulness of video game reviews on the steam platform," in 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE, 2018, pp. 43–50.

[4] G. Andreev, D. Saxena, and J. K. Verma, "Impact of review sentiment and magnitude on customers'recommendations for video games," in 2021 International Conference on Computational Performance Evaluation (ComPE). IEEE, 2021, pp. 992–995.

[5] M. Viggiato, D. Lin, A. Hindle, and C.-P. Bezemer, "What causes wrong sentiment classifications of game reviews?" IEEE Transactions on Games, vol. 14, no. 3, pp. 350–363, 2021.

[6] Z. Zuo, "Sentiment analysis of steam review datasets using naive bayes and decision tree classifier," 2018.

Concientización sobre la obesidad en Latinoamérica en los centros de salud utilizando un árbol de decisión

Obesity awareness in Latin America in health centers using a decision tree

Diego Moises Chuctaya Ruiz
Universidad Nacional de San Agustín.
Arequipa, Perú.
dchuctayar@unsa.edu.pe
https://orcid.org/0000-0002-7296-4592

Luis Pablo Condori Villalba
Universidad Nacional de San Agustín.
Arequipa, Perú.
lcondorivill@unsa.edu.pe
https://orcid.org/0000-0002-7119-4856

Gilbert Wil Ramos Ticona
Universidad Nacional de San Agustín.
Arequipa, Perú.
gramosti@unsa.edu.pe
https://orcid.org/0000-0001-2200-3333

Esteba Cruz Santos Adilson
Universidad Nacional de San Agustín.
Arequipa, Perú.
sestebac@unsa.edu.pe
https://orcid.org/0000-0002-2094-5760

ARK: ark:/42411/s15/a156
DOI: 10.48168/innosoft.s15.a156
PURL: 42411/s15/a156

RECIBIDO 29/10/2023 • ACEPTADO 04/01/2024 • PUBLICADO 30/03/2024

RESUMEN

El presente documento busca concientizar acerca de la obesidad localizada en Latinoamérica y sus centros de salud, incitando a reducir la obesidad en la población, tomando como herramienta un software que a partir de los datos tomados en la investigación pueda determinar los niveles de obesidad en un paciente utilizando técnicas relacionadas a la Inteligencia Artificial donde el modelo aplicado a los datos se puede observar cuales son las personas con un mayor grado de obesidad y analizar sus respectivas causas

Palabras claves: Inteligencia Artificial, obesidad, árbol de decisión, entropía

ABSTRACT

This paper seeks to raise awareness about localized obesity in Latin America and its health centers, inciting to reduce obesity in the population, taking as a tool a software that from the

data taken in the research can determine the levels of obesity in a patient and recommend some ways to improve health using techniques related to Artificial Intelligence where the model applied to the data can be observed which are the people with a higher degree of obesity and analyze their respective causes.

Keywords: *Artificial Intelligence, obesity, decision tree, entropy.*

INTRODUCCIÓN

En la actualidad la obesidad se le puede denominar "la Epidemia del Siglo 21" ya que abarca el problema que se presenta en forma reciente mundialmente [7]. El país que puede representar esta epidemia a grandes rasgos son los Estados Unidos donde el porcentaje de adultos que son obesos (IMC >30) subió de 15, 3% en 1995 a 23,9% en 2005 de los cuales el 4,8 % posee un IMC >40. La obesidad viene acompañada con riesgo aumentado de hipertensión, diabetes, hiperlipidemia y enfermedad coronaria; por lo cual se estima que la obesidad podría llevar a una disminución en la esperanza de vida en USA en el Siglo 21, siendo el indicador de cómo probablemente se muestren el los países latinoamericanos que son influenciados por este mismo [5].

En América tiene la prevalencia más alta de todas las regiones de la OMS, con 62,5% de los adultos con sobrepeso u obesidad (64.1% de los hombres y 60.9% de las mujeres) [8]. Donde la obesidad afecta tanto a adultos como niños; y en América Latina los centros de salud buscan alternativas para afrontar y detener el aumento de las tasas de obesidad, promoviendo y apoyando ideas políticas que permitan a las personas mejorar la alimentación, la actividad física y la salud en la Región de las Américas [6].

La justificación por la cual se realiza esta investigación es mejorar la forma de vida del paciente. Además que la obesidad no es un problema singular, sino que también puede traer otros problemas de salud como enfermedades del corazón, diabetes, enfermedades pulmonares, entre otros [1].

Como otro punto, se sabe que la obesidad afecta al ámbito psicosocial, ya que está asociada con la depresión, la baja autoestima, el bullying, entre otros. Por otro lado, en el ámbito laboral, aparece el absentismo laboral, mayor riesgo de accidentes, dificultad de movilidad, peor desempeño académico, somnolencia, falta de concentración, entre otros, y a la vez no está apto para trabajos que requieran mucha movilidad como podría ser el trabajo de un bombero o militar [2].

En el año 2012, el trabajo "Prevention of Obesity using Artificial Intelligence Techniques" realizado por Bouharati S. et al. nos habla que los desórdenes de obesidad varían de acuerdo a cada

persona, hábitos e incluso el ambiente que lo rodea, por lo que una forma para prevenir la obesidad, tomando en cuenta todos estos factores ambiguos, es utilizar el modelo de lógica difusa y con ello poder reconocer patrones causantes de la obesidad [3].

En el año 2017, el trabajo "Artificial Intelligence technologies to manage obesity" realizado por Bruna Marmett et. al. nos habla que la enorme creciente de casos de obesidad los niveles de mortalidad va en aumento y un método eficaz para manejarlo es la Inteligencia Artificial. Estos sistemas son complejos por lo que para realizar uno se requiere de pruebas no tradicionales de métodos de verificación y validación [4].

Además se recalca que estos sistemas no buscan reemplazar al personal médico experto, sino que este sistema debe ser tratado como un apoyo en la decisión de los médicos para tener un diagnóstico más preciso[3]. El objetivo trazado para la investigación es incitar a los centros de salud de Latinoamérica a reducir la obesidad en la población, tomando como herramienta un software que a partir de los datos tomados en la investigación pueda determinar los niveles de obesidad en un paciente y recomendar algunas formas para mejorar la salud. En esta ocasión los niveles de obesidad se calculan en función de la condición física y hábitos alimenticios

Marco referencial

Trabajos Relacionados

1. Herrera, D. en su trabajo "Hábitos Alimentarios y su Relación con el Sobrepeso y Obesidad en Adolescentes en la Unidad Educativa Julio María Matovelle en el año 2016" sostiene que en la Unidad Educativa Matovelle en la ciudad de Quito, la población se realizó un estudio a adolescentes con edades comprendidas entre 12 y 18 años. El tipo de estudio empleado fue observacional y el enfoque cualicuantitativo analítico y de corte transversal. Para determinar el estado nutricional, se empleó el programa AnthroPlus OMS, en 722 estudiantes, correspondiente a la totalidad de alumnos, posteriormente se seleccionó mediante muestreo aleatorio sistemático a individuos con normopeso (n=40), sobrepeso (n=35) y obesidad (n=32) para evaluar los hábitos alimentarios. Los principales resultados indican que la prevalencia de sobrepeso y obesidad es de 28% (sobrepeso y obesidad), sin diferencias por sexo. Mientras al relacionar hábitos alimentarios de los normopeso con los de sobrepeso muestran diferencia estadísticamente significativa, lo que no demostró con el grupo de obesidad, donde se llegó a la conclusión de que la relación entre la existencia de sobrepeso y obesidad y los hábitos alimenticios en los estudiantes de la Unidad Educativa Matovelle, determina, que la mejor calidad de alimentación la tienen los adolescentes con estado nutricional normal, ya que su índice de calidad promedio es 5,12 que es mayor a los adolescentes con sobrepeso que tienen un índice de calidad promedio de 3,67 y los adolescentes con obesidad con un índice de calidad promedio de 4,40.[12]
2. Gardi, P., Gonzalo, L. y Medina, J. en su trabajo "Hábitos alimentarios y su relación con la obesidad en adolescentes" sostienen que los hábitos alimentarios no saludables

constituyen factores de riesgo modificables, por lo que la identificación es importante para tomar medidas por lo que le objetivo del estudio fue estimar la prevalencia de obesidad en adolescentes de la Institución Educativa de tipo Experimental de la Universidad Nacional de Educación Enrique Guzmán y Valle e identificar su asociación con los hábitos alimentarios no saludables. El estudio se realizó en adolescentes de 14 a 16 años, en el periodo de septiembre 2016 a diciembre de 2016 tomando en cuenta los hábitos alimentarios y la obesidad. Se llegó a la conclusión de que el consumo de alimentos es una práctica que contribuye significativamente en los adolescentes de la Institución Educativa[9].

3. Álvarez, N. en su trabajo "Alimentación Y Salud: La Obesidad Como Factor De Riesgo" tuvo como objetivo realizar una revisión y puesta al día del problema de la obesidad, desde un punto de vista global, por su frecuencia e importancia tanto a nivel mundial como en nuestro país, así como de los diversos factores que contribuyen a su desarrollo y mantenimiento y las consecuencias que produce en la salud y enfermedades asociadas. Este trabajo se realizó mediante el análisis de trabajos, reportes, informes y datos estadísticos relacionados a este tema. Se llegó a la conclusión de que la obesidad es una situación muy frecuente en los países desarrollados, que va aumentando progresivamente en las últimas décadas. Además, constituye un problema de Salud Pública a nivel mundial, pues influye negativamente en la calidad de vida de las personas que lo presentan[10].
4. Chew, H., Ang, W., & Lau, Y. en su trabajo "The potential of artificial intelligence in enhancing adult weight loss: a scoping review" tiene como objetivo en su trabajo presentar una visión general de cómo podría utilizarse la inteligencia artificial (IA) para regular los comportamientos alimentarios y dietéticos, las conductas de ejercicio y la pérdida de peso. Este trabajo se realizó mediante búsqueda literaria en ocho bases de datos (CINAHL, Cochrane-Central, Embase, IEEE Xplore, PsycINFO, PubMed, Scopus y Web of Science). Se llegó a la conclusión de que el uso de la IA para la pérdida de peso aún no se ha desarrollado. Un punto importante es que se propone un marco sobre la aplicabilidad de la IA para la pérdida de peso, pero se advierte de su dependencia del compromiso y la contextualización[11].

Marco teórico

1. Metodología ágil: Las metodologías ágiles son un formato más contemporáneo para el desarrollo de software donde funciona el compromiso activo del cliente, la comunicación recurrente entre el equipo de desarrollo y la entrega rápida del valor del cliente, de acuerdo con el Manifiesto Ágil [5]. Al tener un modelo de comunicación interna y externa muy fácil, ya que las reuniones de lanzamiento y planificación se llevan a cabo al comienzo de cada sprint, el cliente siempre tiene fácil acceso al prototipo desarrollado y le permite iterar entre lo que se está haciendo. desarrollado y el usuario final.

Materiales y métodos

1. **Google Colab:**

   También conocido como "Collaboratory", te permite programar y ejecutar Python en tu navegador con las siguientes ventajas:

   - No requiere configuración
   - Da acceso gratuito a GPUs
   - Permite compartir contenido fácilmente

   Esta herramienta fue utilizada por todos los integrantes de manera síncrona y asíncrona para la realización de nuestro árbol de decisión para el dataset respectivo a la obesidad.

2. **Lenguaje de programación Python:**

   La principal implementación de árboles de decisión en Python está disponible en la librería scikit-learn a través de la clase DecisionTreeClassifier. Una característica importante para aquellos que han utilizado otras implementaciones es que, en scikit-learn, es necesario convertir las variables categóricas en variables dummy (one-hot-encoding).

3. **Árbol de decisión:**

   Estos proveen de una herramienta de clasificación muy potente. Su uso en el manejo de datos la hace ganar en popularidad dadas las posibilidades que brinda y la facilidad con que son comprendidos sus resultados por cualquier usuario. El árbol en sí mismo, al ser obtenido, determina una regla de decisión. Esta técnica permite:

   - Segmentación: establecer qué grupos son importantes para clasificar un cierto ítem.
   - Clasificación: asignar ítems a uno de los grupos en que está particionada una población.
   - Predicción: establecer reglas para hacer predicciones de ciertos eventos.
   - Reducción de la dimensión de los datos: Identificar qué datos son los importantes para hacer modelos de un fenómeno.
   - Identificación-interrelación: identificar qué variables y relaciones son importantes para ciertos grupos identificados a partir de analizar los datos.
   - Recodificación: discretizar variables o establecer criterios cualitativos perdiendo la menor cantidad posible de información relevante.

4. **Dataset:**

   Es un conjunto de datos tabulados en cualquier sistema de almacenamiento de datos estructurados. Nuestro conjunto de datos está distribuido de la siguiente manera:

   a. Cantidad de variables de entrada: 16
   b. Cantidad de instancias: 2111

Tabla 1. Variables de entrada del dataset

Variable	**Definición**	**Tipo de variable**	**Dominio**
Gender	Género	nominal	Female, Male
Age	Edad	discreta	numérico
Height	Tamaño	continua	numérico
Weight	Peso	continua	numérico
FHWO	antecedentes familiares con sobrepeso	nominal	si,no
FAVC	Consumo frecuente de alimentos con alto contenido calórico	nominal	si, no
FCVC	Frecuencia de consumo de verduras	continua	numérico
NCP	Número de comidas principales	discreta	numérico
CAEC	Consumo de alimentos entre horas	nominal	no, a veces, con frecuencia, siempre
SMOKE	fuma	nominal	si, no
CH2O	Consumo de agua diaria	continua	numérico
SCC	Monitoreo del consumo de calorías	nominal	si, no
FAF	Frecuencia de la actividad física	continua	numérico
TUE	Tiempo de uso de dispositivos tecnológicos	continua	numérico
CALC	Consumo de alcohol	nominal	no, a veces, con frecuencia, siempre
MTRANS	Transporte utilizado	nominal	Automóvil, Motocicleta, Bicicleta, Transporte Público, Caminata

Tabla 2. Variables de salida del dataset

Variable	Definición	Tipo de variable	Dominio
NObeyesdad	Nivel de obesidad	ordinal	Insufficient_Weight,Normal_Weight,Overweight_Level_I,Overweight_Level_II,Obesity_Type_I,Obesity_Type_II,Obesity_Type_III

Resultados y discusión

Dentro de los resultados obtenidos se muestra la matriz de confusión el cual muestra que del 20% de datos de prueba, que son 423 casos, 402 generaron resultados correctos mientras que 21 generaron resultados incorrectos

```
Matriz de Confusión:
[[55  1  0  0  0  0  0]
 [ 2 36  0  0  0 10  0]
 [ 0  0 79  1  0  0  3]
 [ 0  0  1 51  0  0  0]
 [ 0  0  0  0 65  0  0]
 [ 0  2  0  0  0 63  1]
 [ 0  0  0  0  0  0 53]]
```

Figura 1: Matriz de confusión

Por lo tanto al momento de comparar los resultados bien clasificados con los que no fueron bien clasificados, nos damos cuenta que la exactitud del modelo es del 95.03%, tal como nos muestra los resultados del código al calcular dicha exactitud.

```
Exactitud del modelo:
0.950354609929078
```

Figura 2: Exactitud del modelo

Para asegurarnos de que estos datos se pueden considerar fiables y ver que tan cerca está el resultado de una predicción del valor verdadero se realizó la medición de la precisión del modelo, el cual nos arroja una precisión del 95.16%

```
Precisión del modelo:
0.9516427562434174
```

Figura 3: Precisión del modelo

En cuanto a la tasa de verdaderos positivos del modelo y por lo tanto para determinar la habilidad del modelo de detectar los casos relevantes se calculó la exhaustividad del modelo el cual nos da un valor del 95.03%. Este es un valor alto por lo que podemos decir que nuestro modelo es sensible, es decir, que no se le escapan muchos positivos.

```
Exhaustividad del modelo:
0.950354609929078
```

Figura 4: Exhaustividad del modelo

Para combinar las medidas de precisión y exhaustividad y poder comparar este rendimiento combinado se calcula el valor F1 del modelo el cual es de 94.93%.

```
F1 del modelo:
0.9493888070912593
```

Figura 5: Valor F1 del modelo

En la generación del árbol de decisión se tuvo en cuenta como nodos todas las variables predictoras ya que así conseguimos un árbol de decisión completo y que sea más preciso en el cual podemos ver que el atributo que se seleccionó como nodo raíz es el peso ya que este según su entropía es el que mayor ganancia de información tiene por lo que es más adecuado para ser el nodo inicial.

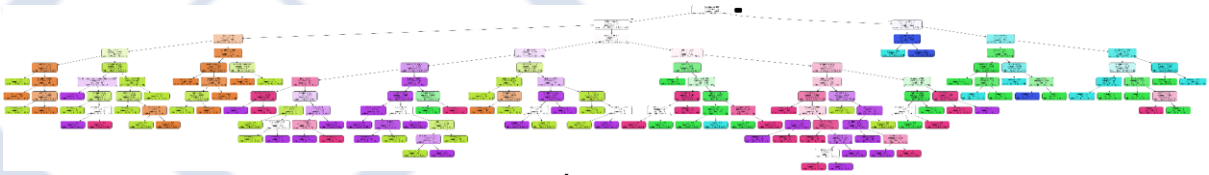

Figura 6: Árbol de decisión

Ya con el árbol de decisión generado podemos predecir el nivel de obesidad de una personas tomando en función de sus hábitos alimenticios y su condición física. Para interpretar este modelo se vio por conveniente generar un par de histogramas. Primeramente se está tomando en cuenta el nivel de obesidad de las personas en función de su edad. Podemos ver como el nivel de obesidad tipo 3, el más alto, es más predominante en las personas con edades entre los 20 y 30 años.

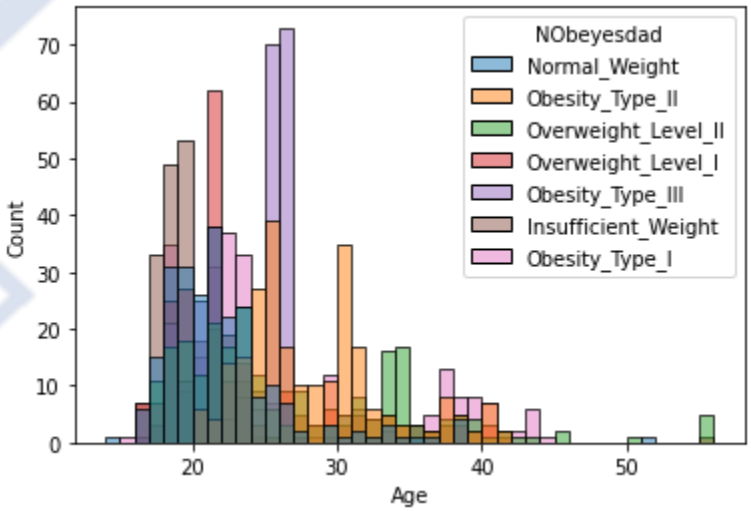

Figura 7: Histograma del nivel de obesidad en función de la edad

En el histograma generado con el nivel de obesidad respecto a la frecuencia de la actividad física de una persona, se visualiza que las personas de obesidad tipo 3, son las que menos se ejercitan y las que más se ejercitan son las personas de peso normal.

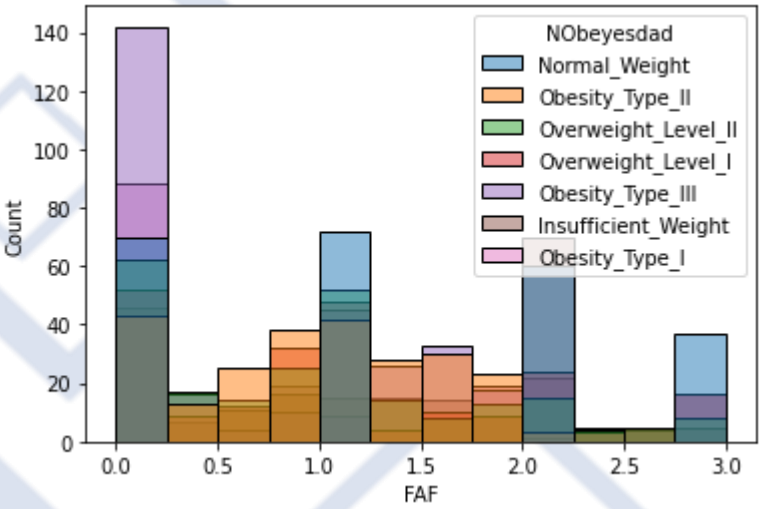

Figura 8: Histograma del nivel de obesidad en función de la Frecuencia de la actividad física(FAF)

En el histograma generado con el nivel de obesidad respecto al tiempo del uso de dispositivos de una persona, se visualiza que la gran parte de las personas que participaron en las encuestas no utilizan mucho los dispositivos.

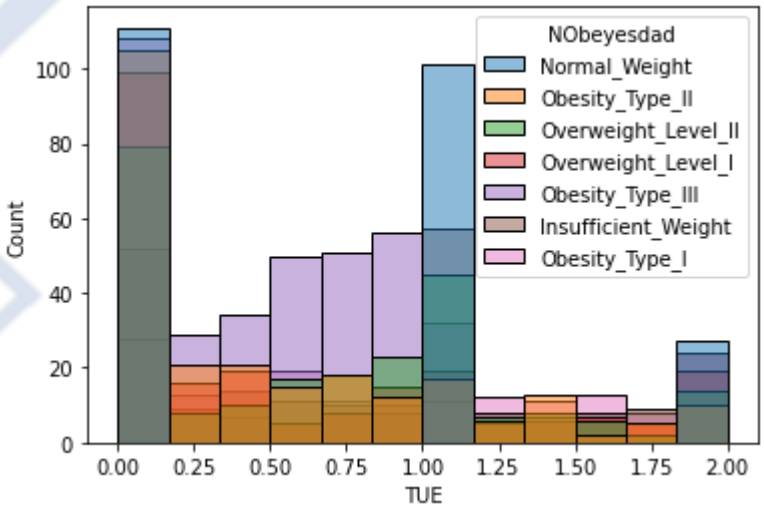

Figura 9: Histograma del nivel de obesidad en función del Tiempo de uso de dispositivos tecnológicos(TUE)

En el histograma generado con el nivel de obesidad respecto al consumo de agua diaria de una persona, se visualiza que la mayoría de las personas encuestadas consumen 2 litros de agua diaria. Las personas que toman de uno a dos litros de agua diaria son mayormente las personas de peso normal. Los que beben más agua son las personas de obesidad tipo 1 y 3.

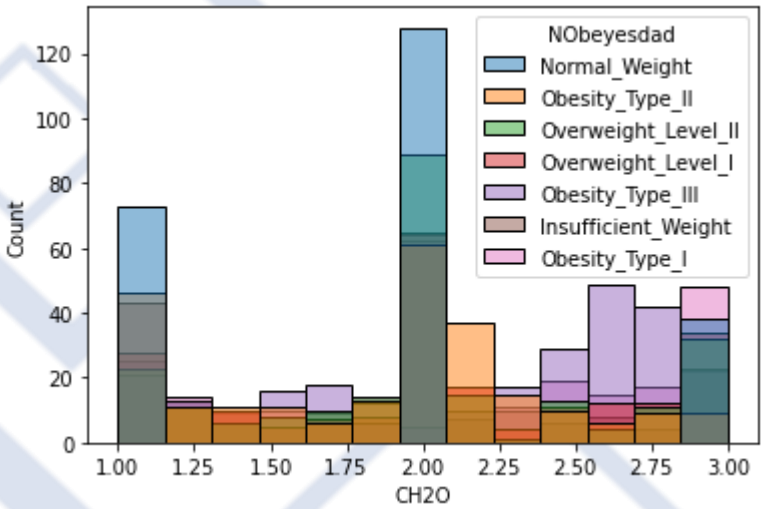

Figura 10: Histograma del nivel de obesidad en función del Consumo de agua diaria(CH2O)

Conclusiones

En el modelo aplicado a los datos se puede observar que las personas con un grado mayor de obesidad se encuentran entre los 20 a 30 años. También se pudo ver que una de las causas por el cual el grado de obesidad sube es porque las personas no se ejercitan, siendo una de las razones el sedentarismo causado en los últimos años por la pandemia y la evolución de la tecnología. Por esta razón es que es importante que los centros de salud deben de actuar ante estos sucesos, para que las personas que residen en Latinoamérica tengan un mejor estilo de vida.

El proyecto contribuye en el campo de la salud, específicamente en el ámbito de una vida saludable tratando de concientizar sobre un problema como lo es la obesidad. Por otro lado, el modelo de predicción aplicado puede ser muy útil para calcular el grado de obesidad de una persona respondiendo un par de preguntas simples.

Como trabajo futuro sería empezar a visitar diversas instituciones para concientizar a las personas de Latinoamérica sobre la obesidad, para esto se puede comenzar por los centros educativos, ya que los niños suelen ser los más vulnerables a sufrir bullying y otros tipos de discriminación.

Contribución de Autoría

Diego Moises Chuctaya Ruiz: Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Luis Pablo Condori Villalba:** Conceptualización, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos. **Gilbert Wil Ramos Ticona:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Esteba Cruz Santos Adilson:** Visualización, Software, Validación, Redacción - borrador original, Curación de datos.

Referencias

[1] Miguel Soca, P. y Niño Peña, A., 2009. Consecuencias de la obesidad . [en línea] Scielo.sld.cu. Disponible en: <http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352009001000006> [Consultado el 25 de junio de 2022].

[2] Martínez Gárate, I., Valdés del Olmo, L., Bayona González, A. y Martínez Castellanos, J., 2021. Relación entre la obesidad y el estrés laboral: una revisión sistemática . [en línea] scielo.isciii. Disponible en: <https://scielo.isciii.es/scielo.php?pid=S0465-546X2021000200112&script=sci_arttext&tlng=pt> [Consultado el 25 de junio de 2022].

[3] Bouharati S. et. al.,2012, Prevention of Obesity using Artificial Intelligence Techniques. International Journal of Science and Engineering Investigations. vol. 1, num. 9.

[4] Marmett, B., Böek Carvalhoa, R., Santos Fortesb, M., Cazellab, S., 2017. Artificial Intelligence technologies to manage obesity. vol. 30, num. 2. (2018). pag. 73-79. DOI: https://doi.org/10.14295/vittalle.v30i2.7654

[5] Parvez Hossain, Bisher Kawar, Meguid El Nahas. Obesity and Diabetes in the Developing World. A Growing Challenge. New Engl J Med 2007; 356: 213- 215

[6] Global Strategy on Diet, Physical Activity and Health [Internet]. OMS Organización Mundial de la Salud. 2017. [Citado 12 noviembre 2017]. Recuperado a partir de: http://www.who.int/dietphysicalactivity/goals/en/

[7] Juan A Rivera, Simón Barquera, Fabricio Campirano, Ismael Campos, Margarita Safdie, Víctor Tovar. Epidemiological and nutritional transition in Mexico: rapid increase of non-communicable chronic diseases and obesity. Public Health Nutrition 2002; 5(1A): 113-122

[8] Aguirre B. H., García T. J. F., Vázquez H. M. C., Alvarado A. M., Romero Z. H. Panorama general y programas de protección de seguridad alimentaria en

México. Rev. Méd Electrón [Internet]. 2017 [citado:12-noviembre-2017]; 39 Supl 1: S741-749. Recuperado a partir de: http://www.revmedicaelectronica.sld.cu/index.php/rme/article/view/2124/3525

[9] Gardi, P., Gonzalo, L. y Medina, J. 2019. Hábitos alimentarios y su relación con la obesidad en adolescentes. [Tesis de Licenciado en Nutrición Humana]. Universidad Nacional de Educación Enrique Guzmán y Valle

[10] Álvarez, N. 2019. ALIMENTACIÓN Y SALUD: LA OBESIDAD COMO FACTOR DE RIESGO, Volumen II. Número 17. Recuperado a partir de: https://www.npunto.es/revista/17/alimentacion-y-salud-la-obesidad-como-factor-de-riesgo

[11] Chew, H., Ang, W., & Lau, Y. 2021. The potential of artificial intelligence in enhancing adult weight loss: a scoping review. Public health nutrition, 24(8), 1993–2020. https://doi.org/10.1017/S1368980021000598

[12] Herrera, D. 2016. Hábitos Alimentarios y su Relación con el Sobrepeso y Obesidad en Adolescentes en la Unidad Educativa Julio María Matovelle en el año 2016. Pontificia Universidad Católica del Ecuador. Recuperado a partir de: https://core.ac.uk/download/pdf/143442581.pdf

Anexos

- Estimation of obesity levels based on eating habits and physical condition: https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+

Clasificación de comentarios de Android usando BERT

Android comment classification using BERT

Susana Rosa Elizabeth Mansilla Ancco
Universidad La Salle. Arequipa, Perú.
smansillaa@ulasalle.edu.pe

Marcelo Antony Pérez Treviños
Universidad La Salle. Arequipa, Perú.
mperezt@ulasalle.edu.pe

ARK: ark:/42411/s15/a120
DOI: 10.48168/innosoft.s15.a120
PURL: 42411/s15/a120

RECIBIDO 09/11/2023 • ACEPTADO 27/01/2024 • PUBLICADO 30/03/2024

RESUMEN

En este proyecto, se centra en el desarrollo de una herramienta de análisis de texto basada en NLP para evaluar comentarios de usuarios de aplicaciones Android, específicamente recopilados de F-Droid. La falta de una solución automatizada para analizar y entender estas opiniones, clasificándolas en tópicos específicos, motiva la investigación. El objetivo es proporcionar a desarrolladores, usuarios y analistas de datos una visión detallada de las preferencias y percepciones de los usuarios. Utilizando conjuntos de datos en inglés entre 2014 y 2017, la propuesta se implementa en Python con la librería Pandas. Se emplea el modelo BERT para la clasificación, con un enfoque específico en la comparación de diferentes modelos, dando como resultado un 97% de exactitud con el uso de BERT. La interfaz gráfica se construye en Visual Studio, permitiendo a los usuarios ingresar comentarios y obtener clasificaciones de tópicos, junto con visualizaciones de nubes de palabras.

Palabras claves: BERT, clasificación de tópicos, clasificación de texto, procesamiento de lenguaje natural.

ABSTRACT

In this project, he focuses on the development of an NLP-based text analysis tool to evaluate user feedback of Android applications, specifically collected from F-Droid. The lack of an automated solution to analyze and understand these opinions, classifying them into specific topics, motivates research. The goal is to provide developers, users, and data analysts with a detailed view of user preferences and perceptions. Using data sets in English between 2014 and 2017, the proposal is implemented in Python with the Pandas library. The BERT model is used for

classification, with a specific focus on the comparison of different models, resulting in 97% accuracy with the use of BERT. The graphical interface is built in Visual Studio, allowing users to enter comments and obtain topic rankings, along with word cloud visualizations.

Keywords: BERT, t*opic classification, text classification, natural language processing.*

INTRODUCCIÓN

En el contexto del creciente ecosistema de aplicaciones Android, la comprensión de las opiniones de los usuarios se ha vuelto esencial para mejorar la calidad y la experiencia general. Este proyecto se centra en la creación de una herramienta de análisis de texto basada en técnicas de Procesamiento del Lenguaje Natural (NLP) para examinar comentarios de usuarios de aplicaciones Android, específicamente recopilados de la plataforma F-Droid. La motivación subyacente radica en la necesidad de proporcionar a desarrolladores, usuarios y analistas de datos una visión integral de las opiniones expresadas en estos comentarios, clasificándolos en tópicos específicos para identificar patrones y tendencias.

La importancia de este proyecto se destaca en el contexto del constante desarrollo y la evolución del panorama de aplicaciones móviles, donde las opiniones de los usuarios desempeñan un papel crucial en la toma de decisiones para los desarrolladores y las empresas. El análisis de texto a través de técnicas avanzadas de NLP se presenta como una herramienta valiosa para extraer información significativa de grandes conjuntos de datos de comentarios como App, Feature/Functionality, Contents, GUI, Model, Update/Version y Other, permitiendo una comprensión más profunda de las preferencias y percepciones de los usuarios.

El trabajo se centra en abordar preguntas fundamentales relacionadas con las técnicas de clasificación de texto, la identificación de tipos de tópicos discutidos y las herramientas utilizadas en el análisis de texto. Se destaca la comparación de diferentes modelos de clasificación, lo que proporciona una evaluación crítica de sus fortalezas y limitaciones.

En cuanto al entrenamiento del modelo BERT, el trabajo detalla el proceso mediante un pipeline específico, describiendo los datos utilizados en este proceso. Además, se presenta el código de entrenamiento del modelo de manera comprensible, utilizando un pseudocódigo que facilita la comprensión del procedimiento.

Adicionalmente, se realiza una exploración de trabajos relacionados que han abordado problemas similares. Esta revisión literaria no solo brinda inspiración, sino que también establece un marco de referencia esencial para el diseño y desarrollo de la herramienta propuesta. La información recopilada en la revisión literaria se utilizará como base sólida para la propuesta detallada que sigue en el trabajo.

Motivación

A. Problema

El problema central radica en la ausencia de una herramienta automatizada que permita analizar y comprender las opiniones de los usuarios expresadas en los comentarios de las aplicaciones de Android. Este problema se agrava por la necesidad de realizar la clasificación en un tiempo reducido y con una adecuada categorización del contenido.

B. Objetivo

Este trabajo tiene como objetivo desarrollar una herramienta de análisis de texto para aplicaciones de Android a partir de datos recolectados de F-Droid, para que clasifique los comentarios de los usuarios y según los tópicos pueda tener una referencia para las actualizaciones o mantenimiento del aplicativo.

C. ¿En qué dominio del conocimiento está trabajando?

En cuanto al tema de clasificación de texto en comentarios de usuarios de aplicaciones Android, el dominio del conocimiento en el que se está trabajando es el campo de Procesamiento del Lenguaje Natural (NLP), Inteligencia Artificial (IA) y clasificación de textos.

D. ¿Quiénes son los usuarios objetivos?

Los usuarios objetivos abarcan desarrolladores de aplicaciones, usuarios de Android, empresas de desarrollo, investigadores en experiencia del usuario y analistas de datos. Estos actores desempeñan un papel clave en la mejora continua de los productos, contribuyendo así a una experiencia de usuario mejorada en el uso de aplicaciones móviles.

E. ¿Por qué es interesante el tema que proponen?

El análisis de texto aplicado a los comentarios de usuarios en F-Droid proporciona información valiosa sobre la percepción de los usuarios en diversos tópicos. Esta información se vuelve crucial para mejorar tanto la calidad de las aplicaciones como la experiencia del usuario en el ecosistema Android. El sistema desarrollado tiene la capacidad de gestionar y clasificar estos comentarios en áreas específicas, permitiéndonos enfocarnos en temas particulares y abordar de manera prioritaria aquellos que requieran mantenimiento.

F. ¿Cuáles son las preguntas que su proyecto de NLP intenta responder?

P1: ¿Cuáles son las técnicas utilizadas en la clasificación de texto?

P2: ¿Cuáles son los principales conjuntos de datos?

P3: ¿Qué tipos de tópicos se utilizan?

P4: ¿Se han implementado estrategias específicas en el código para mejorar la precisión del modelo BERT en comparación con otros modelos de clasificación?

P5: ¿Cuáles son las palabras de los comentarios más utilizadas en cada tópico?

Trabajos relacionados

El texto aborda la aplicación del Procesamiento del Lenguaje Natural (PLN) para optimizar la revisión de especificaciones de construcción, con un enfoque especial en la identificación de categorías de riesgo contractuales. Se introducen métodos fundamentales de preprocesamiento de texto, como normalización y tokenización, seguidos de una explicación detallada de técnicas de incorporación de texto, incluyendo el destacado modelo BERT [1].

Trata acerca del uso de Ecco que es una biblioteca de código abierto diseñada para potenciar la transparencia y explicabilidad de los modelos de lenguaje basados en la arquitectura Transformer, como BERT. Su principal enfoque es proporcionar herramientas avanzadas que permitan analizar y visualizar aspectos internos de estos modelos. La implementación con Ecco se realiza de manera eficiente a través de una biblioteca de Python que se integra fácilmente con modelos de lenguaje preentrenados, en particular con BERT. Esta integración permite aprovechar tecnologías web para generar visualizaciones interactivas, mejorando así la comprensión y la interpretabilidad de los modelos. Ecco se fundamenta en diversas bibliotecas de código abierto, incluyendo Scikit-Learn, Matplotlib, NumPy, PyTorch y Transformers, lo que garantiza una base sólida y confiable para sus funcionalidades [2].

En este trabajo de investigación, se analiza el aprendizaje automático supervisado. Admite la máquina de vectores y el algoritmo Naïve Bayes y compara su precisión general, precesión y valor de recuperación. El resultado muestra que, en el caso de las reseñas de aerolíneas, la máquina de vectores de soporte dio mejores resultados que el algoritmo Naïve Bayes [3].

Los modelos tradicionales generalmente necesitan obtener buenas características de muestra mediante métodos artificiales y luego clasificarlas con algoritmos clásicos de aprendizaje automático. Por lo tanto, la eficacia del método está restringida en gran medida por la extracción

de características. Sin embargo, a diferencia de los modelos tradicionales, el aprendizaje profundo integra la ingeniería de características en el proceso de ajuste del modelo mediante el aprendizaje de un conjunto de transformaciones no lineales que sirven para asignar características directamente a los resultados [4].

Trata acerca del sistema de clasificación de texto corto para descripciones de transacciones bancarias, que consta de tres etapas principales: preprocesamiento, análisis de aprendizaje automático (ML) y clasificación. En la etapa de preprocesamiento, se recopilan datos de transacciones bancarias, se tokenizan y eliminan palabras sin significado. En el análisis de ML, se extrae conocimiento lingüístico mediante la creación de léxicos basados en categorías de interés, utilizando diversas características como datos léxicos, cantidad de transacción, fecha y n-gramos de palabras y caracteres. La clasificación se realiza mediante un clasificador de Máquinas de Soporte Vectorial (SVM), abordando el desafío de clasificar texto breve. La evaluación del sistema se realiza en conjuntos de datos de descripciones de transacciones bancarias españolas, comparando resultados con enfoques competidores y utilizando métricas como precisión, recall y F-measure [5].

Aborda el desafío de la desigualdad de clases en conjuntos de datos, resaltando que muchos algoritmos funcionan mejor cuando las clases están representadas de manera equitativa. Para superar este problema, implementa la funcionalidad class_weight de sklearn.utils, lo que resulta en mejoras significativas para su conjunto de datos desequilibrado. El código proporcionado muestra la implementación práctica, desde el preprocesamiento de datos hasta el manejo del desequilibrio de clases mediante pesos calculados, el entrenamiento del modelo y las fases de validación y prueba de rendimiento. Demuestra la eficacia de abordar la desigualdad de clases y destaca la rapidez en la que se puede lograr el ciclo completo de desarrollo del modelo [6].

Propuesta

A. Datos

Se eligió utilizar un conjunto de datos en inglés que se centra en comentarios de aplicaciones de Android, seleccionándolo debido a su disponibilidad y a que es el conjunto de datos más extenso identificado, abarcando el periodo entre 2014 al 2017. Estos conjuntos de datos se obtuvieron del repositorio de bases de datos en GitHub y se almacenaron en formato estructurado CSV.

El análisis y procesamiento de los datos se llevarán a cabo en el lenguaje de programación Python, haciendo uso de la librería Pandas para la lectura eficiente de los archivos CSV.

En el proyecto, se emplearán dos conjuntos de datos relacionados con comentarios de aplicaciones de Android obtenidos a través de F-Droid. El primer conjunto, presentado en la Tabla 1, contiene información detallada sobre los usuarios. Por otro lado, el segundo

conjunto, presentado en la Tabla 2, abordará la clasificación de los tópicos según el contenido de los comentarios.

Tabla 1. Descripción de atributos de la primera base de datos

Atributos	**Descripción**
Id	Identificador del usuario
Package	Nombre del paquete
Review	Comentario del usuario
Date	Fecha
Star	Clasificación por estrella
Versión_id	Versión

Tabla 2. Descripción de atributos de la segunda base de datos

Atributos	Descripción
Id	Identificador del usuario
Review	Titulo del comentario
Intention	Intención del comentario
Topic	Clasificación del comentario

Con el objetivo de entrenar el modelo propuesto, se generará un nuevo archivo en formato CSV a partir de los dos conjuntos de datos mencionados. En este archivo, presentado en la Tabla 3, se utilizará el comentario del usuario del primer conjunto de datos, y se incorporará el tópico del segundo conjunto de datos asociado a dicho comentario. Este nuevo archivo servirá como base para el desarrollo del modelo propuesto en el proyecto donde se trabajara con 7 tópicos como App, Feature/Functionality, Contents, GUI, Model, Update/Version y Other.

Tabla 3. Descripción de atributos de la tercera base de datos

Atributos	Descripción
Review	Comentario del usuario
Topic	Clasificación del comentario

El tercer dataset se genera en formato CSV, con un total de 90801 registros, y con un tamaño aproximado de 6.04 MB en el disco duro.

B. Diseño

Para este proyecto, se emplearon dos bases de datos con el objetivo de mejorar la información disponible. Para lograr esto, se creó una tercera base de datos que contiene comentarios de la primera base, vinculando mediante un identificador compartido con la segunda base para clasificar al tópico. Este procedimiento permitió establecer la correspondencia correcta entre los comentarios y los respectivos tópicos.

Una vez obtenido el tercer conjunto de datos, se procedió con el preprocesamiento, que incluyó la eliminación de duplicados, espacios en blanco, signos, emojis y caracteres no ASCII. Después de esta fase de limpieza, se llevó a cabo un balanceo de datos para garantizar que los siete tópicos tuvieran una cantidad equitativa de muestras. Dado que algunos tópicos presentaban una falta de datos, se calcularon los pesos de tópico inversamente proporcionales a la frecuencia de cada uno, abordando así el desbalance.

Posteriormente, se dividió el conjunto de datos en conjuntos de entrenamiento y prueba. Se empleó el tokenizador BERT para procesar y codificar los comentarios en ambos conjuntos. A continuación, se definió un modelo BERT específico para la clasificación en función de los siete tópicos.

El entrenamiento del modelo se llevó a cabo a lo largo de varias épocas, utilizando una función de pérdida ponderada por los pesos de tópico para contrarrestar el desbalance existente. Tras completar el entrenamiento, se evaluó el modelo en el conjunto de prueba y se calcularon métricas de precisión, recall y F1.

Finalmente, el modelo entrenado y el codificador de etiquetas se guardaron en archivos para su posterior utilización. Estos procesos que se explicó con anterioridad se ve reflejado en la Figura 1.

Los usuarios de la herramienta pueden interactuar con la visualización generada para llevar a cabo un análisis de los tópicos. La interfaz proporciona un cuadro de texto que permite a los usuarios ingresar un comentario, y el sistema entrenado por el modelo responderá mostrando el tópico al que pertenece dicho comentario. Además, la herramienta presenta una Nube de Palabras para cada tópico que incluye App, Feature/Functionality, Contents, GUI, Model, Update/Version y Other.

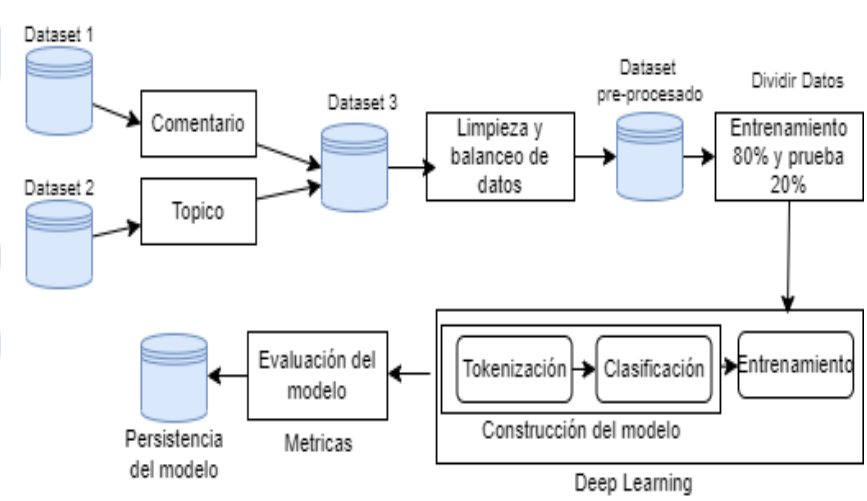

Figura 1. Diagrama para la clasificación de comentarios según tópico con Deep Learning.

C. Pseudocódigo

En la realización de este proyecto, seguimos meticulosamente los pasos delineados en el algoritmo de la Tabla 4, implementado en Python. A continuación, se detallan las fases clave del proceso:

- Obtención del Conjunto de Datos.
- Procedemos a identificar y obtener las etiquetas únicas de los diversos tópicos presentes en el conjunto de datos.
- Inicializamos el codificador de etiquetas para convertir estas en valores numéricos con label_encoder,y se almacena en nueva columna label.
- Se determinaron los pesos de clase, siendo inversamente proporcionales a la frecuencia de cada tópico de los valores almacenados en label.
- Comenzamos a dividir los datos en conjuntos de entrenamiento y prueba.
- Se procedió a la inicialización del tokenizador, configurando el procesamiento en bert base uncased.
- Se tokeniza y codifica los datos de entrenamiento y prueba, se llevó a cabo la conversión de los textos de los comentarios en vectores numéricos.
- Iniciamos la configuración de un modelo de clasificación de secuencias utilizando bert base uncased.
- Configuración del Optimizador en base al clasificador.
- En la función evaluar_modelo se evalúa el modelo en el conjunto de prueba y calcular métricas como la precisión, la precisión por clase, recall y la puntuación F1 para luego retornar estas métricas.
- Se define el número de épocas a evaluar en este caso es 9.

- Se realiza el entrenamiento del modelo 9 veces, y en cada entrenamiento donde se utiliza una función de pérdida ponderada para mejorar el aprendizaje del modelo, se retorna el total de pérdidas.
- Evaluación Final y Métricas, evaluamos el modelo en el conjunto de prueba y obtenemos la precisión del modelo. También imprimimos las métricas obtenidas durante el proceso en la función de evaluación.

Tabla 4. Algoritmo de entrenamiento BERT

Algoritmo 1: Proceso Tópicos Android
Entrada: CD => conjunto de datos label=>Datos de los tópicos en valores numéricos entrenamiento_df =>Datos de Entrenamiento prueba_df =>Datos de Prueba num_labels =>Cantidad de tópicos test_dataloader =>Carga de lotes de datos durante la prueba. train_dataloader => Carga de lotes de datos durante la entrenamiento. **Salida**: Metrica_Calculado 1: df= cargar datos(CD) 2: labels = obtener_etiquetas_unicas (df[topico]) 3:label= label encoder.codificador_etiquetas (df[topico]) 4: calcular_pesos_topico(df['label']) 5: dividir_datos(df,prueba size=02,entrenamiento size=0.8) 6: tokenizador=tokenizador.preenternado(bert base uncased) 7: tokenizar_codificar(entrenamiento_df, tokenizador) 8: tokenizar_codificar(prueba_ df, tokenizador) 9:clasificador= clasificador_preentrenado(bert base uncased) 10: optimizador(clasificador) 11: funcion evaluar_modelo(clasificador, test_dataloader): 12: retorna accuracy, precision, recall, f1 13: num_epocas = 9 14: para epocas =1 .. num_epocas(train_dataloader) 15: retorna total_perdidas 16: Imprimir Metrica_Calculado

Resultado

A. Entrenamiento del modelo BERT

El entrenamiento del modelo BERT se llevó a cabo en Colab debido a la necesidad de un proceso más eficiente utilizando la GPU. Se obtuvieron los siguientes resultados de efectividad durante el entrenamiento de BERT. El entrenamiento duró 3 horas con 15 minutos, con un total de 90,801 comentarios clasificados.

Los resultados proporcionados por el entrenamiento del modelo BERT se muestran en las métricas de precisión, recall, F1 y para calcular la exactitud.

Se tiene dos modelos adicionales como Máquinas de Vectores de Soporte (SVM) y Naive Bayes(NB) que también sirven para la clasificación de texto, y con el modelo que BERT se quiere medir el nivel de precisión y eficiencia que tiene en comparación con los demás.

Primero, se aplica la fórmula de precisión para cada tópico (1), midiendo la exactitud de las predicciones positivas, esto se muestra en la Tabla 5.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (1)$$

Tabla 5. Comparación de modelos métrica – precision

	Precision		
	SVM	**NB**	**BERT**
App	0.93	0.60	0.98
Feature/Functionality	0.92	0.73	0.97
Other	0.96	0.79	0.99
Contents	0.87	0.99	0.84
GUI	0.84	0.99	0.86
Model	0.85	0.99	0.84
Update/Version	0.87	0.99	0.84

Segundo, se aplica la fórmula de recall para cada tópico (2), midiendo la capacidad del modelo para capturar todas las instancias positivas, esto se muestra en la Tabla 6.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

Tabla 6. Comparación de modelos métrica – recall

	Recall		
	SVM	**NB**	**BERT**
App	0.94	0.79	0.98
Feature/Functionality	0.93	0.89	0.97
Other	0.97	0.56	0.96
Contents	0.76	0.03	0.97
GUI	0.67	0.06	0.95
Model	0.79	0.04	0.99
Update/Version	0.88	0.08	0.98

Tercero, se aplica la fórmula de F1 para cada tópico (3), que es la media ponderada de precisión y recall. Esto se muestra en la Tabla 7.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Tabla 7. Comparación de modelos métrica – F1

	F1		
	SVM	**NB**	**BERT**
App	0.93	0.68	0.98
Feature/Functionality	0.92	0.80	0.97
Other	0.97	0.65	0.97
Contents	0.81	0.05	0.90
GUI	0.74	0.11	0.90
Model	0.82	0.08	0.91
Update/Version	0.88	0.16	0.91

Finalmente, se aplica la fórmula de exactitud (4), que mide la proporción general de predicciones correctas, que se visualiza en la Tabla 8.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Examples} \quad (4)$$

Tabla 8. Comparación de modelos métrica – accuracy

	SVM	NB	BERT
Accuracy	0.93	0.69	0.97

Como se visualiza en la exactitud, el modelo propuesto utilizando BERT tiene un mejor rendimiento en la clasificación de los tópicos, alcanzando un 97% de instancias correctas.
NV muestra un rendimiento más bajo, especialmente en recall y F1-score. SVM tiene un rendimiento general sólido pero ligeramente inferior a BERT, y BERT muestra un rendimiento superior en todas las métricas y categorías en comparación con SVM y NV. BERT es especialmente fuerte en las categorías con más datos (App, Feature/Functionality, Other), donde supera significativamente a SVM y NV.
Según las comparaciones entre los 3 modelos, BERT ofrece el rendimiento más consistente y superior en todas las métricas evaluadas.

B. Word Cloud

Se utilizó la librería Word Cloud (nube de palabras) en Python para identificar qué palabras son más recurrentes en cada tópico. Esto se visualiza en palabras, cuanto más presente esté una palabra en el texto, más grande aparecerá en la nube de palabras. Esto se presentó desde las Figura 2 hasta la Figura 8

Figura 2. Nube de palabras del tópico App.

Figura 3. Nube de palabras del tópico Feature/Functionality.

Figura 4. Nube de palabras del tópico Contents.

Figura 5. Nube de palabras del tópico GUI.

Figura 6. Nube de palabras del tópico Model.

Figura 7. Nube de palabras del tópico Update/Version.

Figura 8. Nube de palabras del tópico Other.

C. Interfaz gráfica

Este modelo de entrenamiento se utilizó para clasificar comentarios en una página web con respecto a los temas para los cuales fue entrenado, se visualiza la interfaz gráfica para el uso de este modelo en la Figura 9.

Figura 9. Interfaz gráfica de la página web.

Conclusiones

Este proyecto ha logrado desarrollar con éxito una herramienta de análisis de texto basada en NLP para evaluar comentarios de usuarios de aplicaciones Android. La implementación del modelo BERT, respaldada por una cuidadosa metodología, ha demostrado un rendimiento superior, alcanzando una precisión del 97% en la clasificación de tópicos. La interfaz gráfica proporciona una experiencia interactiva para los usuarios, permitiéndoles ingresar comentarios y obtener clasificaciones junto con visualizaciones de nubes de palabras. Este proyecto no solo aborda la falta de soluciones automatizadas en la comprensión de opiniones de usuarios, sino que también destaca la importancia de las técnicas avanzadas de NLP en la extracción de información significativa de grandes conjuntos de datos.

Contribución de Autoría

Susana Rosa Elizabeth Mansilla Ancco: Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Marcelo Antony Pérez Treviños:** Conceptualización, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos.

Referencias

[1] S. Moon, S. Chi, and S.-B. Im, "Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from Transformers (Bert)," Automation in Construction, October, 2022. [Online]. Available:https://www.sciencedirect.com/science/article/pii/S0926580522003387.

[2] Alammar, J. "Ecco: An open source library for the explainability of transformer language models", Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing: System demonstrations, pp.249-257,2021.

[3] Rahat, A. Mohaimin, A. Kahir, and A. Mohammad. "Comparison of Naive Bayes and SVM Algorithm based on sentiment analysis using review dataset", 8th International Conference System Modeling and Advancement in Research Trends (SMART). IEEE, 2019.

[4] Li, Q., et al., "A survey on text classification: From traditional to deep learning", ACM Transactions on Intelligent Systems and Technology (TIST), vol. 13, no.2, pp. 1-41,2022.

[5] G.Mendez, S., et al. "Identifying banking transaction descriptions via support vector machine short-text classification based on a specialized labelled corpus", IEEE Access, vol. 8, pp. 61642-61655,2020.

[6] Nikhil, "Bert: Handling class imbalance in text classification," Medium, December, 2023. [Online]. Available: https://medium.com/@nikviz/bert-handling-class-imbalance-in-language-models-7fe9ccc62cb6.

Poker Hand Valuator, IA evaluadora de manos de poker

Poker Hand Valuator, poker hand evaluator AI

Estith Bryan Vargas Quispe
Universidad Nacional de San Agustin.
Arequipa, Perú.
evargasq@unsa.edu.pe

Eybert Macedo Pillco
Universidad Nacional de San Agustin.
Arequipa, Perú.
emacedop@unsa.edu.pe

Quispe Ttito Juan Carlos
Universidad Nacional de San Agustin.
Arequipa, Perú.
jquispett@unsa.edu.pe

Jose Miguel Cano Vilcapaza
Universidad Nacional de San Agustin.
Arequipa, Perú.
jcanov@unsa.edu.pe

ARK: ark:/42411/s15/a157
DOI: 10.48168/innosoft.s15.a157
PURL: 42411/s15/a157

RECIBIDO 22/11/2023 • ACEPTADO 14/02/2024 • PUBLICADO 30/03/2024

RESUMEN

Nuestro objetivo es la predicción de manos de poker, la probabilidad de que pueda sacar una buena mano cuando se hace 5 robos de una baraja de poker de 52 cartas, aplicamos la redes neuronales para realizar dicha predicción conjunto con diferentes librerías que ayudan a que el proceso sea más simplificado y los resultados sean más fiables, por lo tanto en el uso de esta metodología logramos obtener un average de 97% en la mayoría de los casos con una desviación de 2.5% lo cual consideramos aceptable debido a la cantidad muy desbalanceada de los datos de este dataset, por lo tanto este método de inteligencia artificial nos sirve para predecir nuevas manos y tomar mejores decisiones conforme te encuentres en una situación del juego.

Palabras claves: dataset, poker hand, poker texas, predicción, red neuronal.

ABSTRACT

Our objective is the prediction of poker hands, the probability that you can draw a good hand when you make 5 steals from a 52-card poker deck. We apply neural networks to make this prediction together with different libraries that help the process is more simplified and the results are more reliable, therefore in the use of this methodology we managed to obtain an average of 97% in most cases with a deviation of 2.5% which we consider acceptable due to the very

unbalanced amount of the data from this dataset, therefore this artificial intelligence method helps us predict new hands and make better decisions as you find yourself in a game situation.

Keywords: *Dataset, poker hand, poker texas, prediction, neural network.*

INTRODUCCIÓN

En este trabajo se realizará un algoritmo que se encargue de tomar la decisión según la mayor probabilidad de formar una mejor mano de poker cuando ya se recibieron las cinco cartas iniciales, cabe destacar que existen proyectos que ya realizan trabajos similares como cuando intentan estimar probabilidades en el juego de poker texas que proporcionar a diferentes usuarios un soporte de simulación que permita estimar la probabilidad de ganar una mano[1].

La mayoría de los proyectos de inteligencia artificial referentes al poker son para crear un bot que se encargue de ganar revisando unos que otros parámetros para conseguir su objetivo, entre estos tenemos el bot de [2] que analiza los parámetros de nivel de riesgo, el valor de las jugadas y la apuesta correspondiente, por otro lado otro bot revisa los parámetros [9], adaptar su modo de juego de juegos anteriores.

Uno de los mayores problemas en el poker es la toma de decisiones en la primera mano para cambiar las cartas necesarias según las cartas que te vas a quedar en la mano, esta decisión es fundamental al momento de continuar o no con la partida, por lo que este algoritmo ayuda bastante a un novato e incluso a un profesional en cual seria la mejor decisión cuando se le presente una de las situaciones ya producidas por el algoritmo.

Entonces nos planteamos el método de heurísticas que hacen inferencias deductivas e inductivas, forma y optimizan reglas heurísticas para hacer que las máquinas alcancen resultados similares a una decisión humana[7], otro método de uso también son los algoritmos genéticos[6] los cuales son adaptativos los cuales se emplean para la resolución de problemas de búsqueda y optimización, permiten optimizar funciones numéricas como explica el artículo [5] que utilizar este método para obtener distintos resultados de optimización que sean necesarios para resolver la mejor mano en cierto momento del juego.

Estas decisiones pueden determinar la victoria en una partida importante por lo que aun cuando una de las partes más importantes del poker es el blofeo de lo que uno podría tener en la mano, independientemente de lo que tengan los demás, tomar una buena decisión implica la mayor probabilidad de que salgas librado de un blofeo del oponente.
Entonces nuestro objetivo es que el algoritmo analice la primera mano y escoja las cartas que debe devolver al mazo para que pueda conseguir una mejor combinación, dicho algoritmo tendrá

que volver a analizar la nueva mano y guardar la información para futuros análisis de las siguientes manos.

Materiales y métodos o Metodología computacional

Redes neuronales: Las redes neuronales tratan de emular el comportamiento del cerebro humano, caracterizado por el aprendizaje genérico a partir de un conjunto de datos. Estos sistemas imitan esquemáticamente la estructura neuronal del cerebro.[17]

Machine learning: Es una forma de la IA el cual permite a los sistemas aprender de los datos en lugar de aprender de la programación explícita, pero esto no es un proceso sencillo. Un modelo de machine learning es la salida de información que se genera cuando se entrena un algoritmo de machine learning con datos, después del entrenamiento al proporcionar un tipo de modelo, se nos da una salida con el mismo modelo.

Aprendizaje iterativo: Nos permite entrenar modelos con conjuntos de datos antes de ser implementados, algunos de los modelos de Machine Learning están online y son continuos, este proceso iterativo de modelos permite que se logre una optimización en la predicción de datos y que debido al tamaño y la complejidad de los datos, estos pueden ser pasados por alto muy fácilmente por un ser humano.

Aprendizaje Supervisado: comienza típicamente con un conjunto establecido de datos y una cierta comprensión de cómo se clasifican los datos, este aprendizaje pretende encontrar patrones para luego aplicarlos a un proceso de análisis. Un ejemplo claro es la aplicación con base en imágenes y descripciones escritas que distingan entre millones de animales

Aprendizaje no Supervisado: Se utiliza cuando el problema no requiere una cantidad masiva de datos sin etiquetar, claro ejemplo son las redes sociales las cuales tienen grandes cantidades de datos sin etiquetar, la comprensión detrás de estos datos requiere algoritmos que clasifican los datos con base a los patrones que encuentra. El aprendizaje no supervisado lleva a cabo un proceso iterativo, el cual analiza los datos sin intervención humana.

Poker Hand Data Set: El data set fue creado el 2007 por Robert Cattral y Franz Oppacher con el objetivo de predecir una buena mano de poker, cada fila de este dataset está compuesta por 5 cartas robadas de una baraja estándar de 52 piezas, cada carta es descrita usando 2 atributos (palo y número), para un total de 10 variables predictivas. Finalmente tenemos un atributo más el cual se encarga de describir el tipo de mano que se formó al momento del último robo de carta. El número de instancias o filas que tenemos en el dataset para el entrenamiento es de 25010 filas, mientras que para el testeo tenemos 1 000 000 filas.

Número de atributos predictivos tenemos 10 y un atributo objetivo el cual nos indica el tipo de mano que se formó con los otros atributos.

Tabla 1. Información de los atributos o variables que componen una fila del dataset de Poker Hand

	nombre variable	tipo de atributo	valor de atributo	representa
1	S1	Ordinal	1-4	Corazones, espadas, diamantes, trébol
2	C1	Numérico	1-13	As, 2, 3, 4, 5, 6, 7, 8, 9, 10, jota, reina, rey
3	S2	Ordinal	1-4	Corazones, espadas, diamantes, trébol
4	C2	Numérico	1-13	As, 2, 3, 4, 5, 6, 7, 8, 9, 10, jota, reina, rey
5	S3	Ordinal	1-4	Corazones, espadas, diamantes, trébol
6	C3	Numérico	1-13	As, 2, 3, 4, 5, 6, 7, 8, 9, 10, jota, reina, rey
7	S4	Ordinal	1-4	Corazones, espadas, diamantes, trébol
8	C4	Numérico	1-13	As, 2, 3, 4, 5, 6, 7, 8, 9, 10, jota, reina, rey
9	S5	Ordinal	1-4	Corazones, espadas, diamantes, trébol
10	C5	Numérico	1-13	As, 2, 3, 4, 5, 6, 7, 8, 9, 10, jota, reina, rey
11	Tipo	Ordinal	0-9	revisar tabla 2

A continuación, revisamos los valores que puede tomar el atributo Tipo los cuales dependen de los otros atributos que son de las cartas robadas de la baraja.

Tabla 2. Descripción de valores que puede tomar el atributo de tipo según las cartas robadas de la baraja de Poker Hand

valor del atributo	nombre de la mano	descripción
0	nada en mano	no es una mano de poker reconocida
1	un par	un par de rango iguales entre las 5 cartas
2	dos pares	dos pares de rango iguales entre las 5 cartas
3	tres de un tipo	tres cartas iguales de mismo rango entre las 5 cartas
4	escalera	5 cartas secuencialmente numeradas

5	Color	5 cartas de un mismo palo
6	Full	un par y un trío del mismo número respectivamente
7	poker	4 cartas de Ases en una mano de 5 cartas
8	escalera color	5 cartas numeradas secuencialmente de un mismo palo
9	escalera real de color	as, rey, reina, jota, 10 todas de un mismo palo

La distribución de las manos que se aprecian en el dataset según el tipo de mano se ve bastante desbalanceado, el la figura 1 observamos un gráfico en el cual la mano (nada en mano y un par), son bastante abundantes, por lo que ambos componen un aproximado de 92.33% de los datos en el dataset de entrenamiento, aun incluyendo casi todos los valores que pueden tomar las manos menos frecuentes la diferencia es muy notable.

El dataset original sobre el cual trabajaremos tiene la siguiente distribución y tamaño:

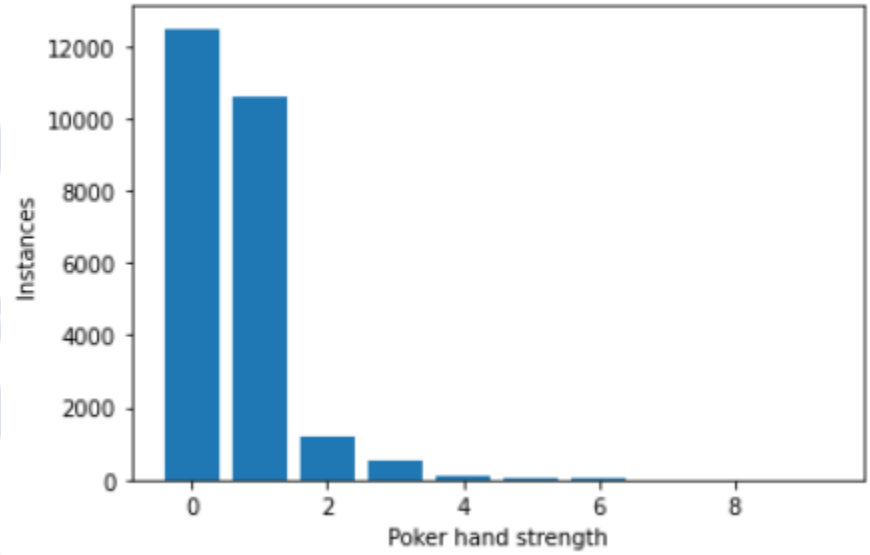

Figura 1. Gráfico de cantidad de manos encontradas en el entrenamiento de la red neuronal.

En la tabla 3 veremos claramente en porcentaje la cantidad de manos que tiene el dataset de entrenamiento con su porcentaje respectivo y la cantidad de manos que tiene el dataset total de manos que tiene el Poker Hand con sus porcentajes.

Tabla 3. Cantidad y porcentajes de combinaciones de tipos de mano en el dataset de entrenamiento y dataset global

valor de atributo	nombre de mano	cantidad instancias de entrenamiento (CIA)	porcentaje CIA	cantidad de instancias global (CIG)	porcentaje CIG
0	nada en mano	12 493	49.95202%	1 302 540	50.1177%
1	un par	10 599	42.37905%	1 098 240	42.2569%
2	dos pares	1 206	4.82207%	123 552	4.7935%
3	tres de un tipo	513	2.05118%	54 912	2.1128%
4	escalera	93	0.37185%	10 200	0.3925%
5	Color	54	0.21591%	5 108	0.1980%
6	Full	36	0.14394%	3 744	0.1441%
7	poker	6	0.02399%	624	0.0240%
8	escalera color	5	0.01999%	36	0.0014%
9	escalera real de color	5	0.01999%	4	0.0002%
total		25 010	99.9999%	2 598 960	100%

Herramientas:

Google Colaboratory: Según el sitio oficial Google Colab [11] es un producto que permite escribir y ejecutar el lenguaje Python de manera online sin costo alguno. En producto es donde ejecuta la mayoría del código para el presente trabajo, principalmente para el entretenimiento del modelo de redes neuronales aplicadas a Poker Hand.

Python: Challenger, Díaz, Becerra. [15] Es un lenguaje de programación que fue desarrollado bajo el concepto de ser libre de uso, ejecución, distribución y modificación.

Tensor Flow: Según el sitio oficial Tensor Flow[16] es una plataforma de extremo a extremo de código abierto para el aprendizaje de desarrollo y entrenamiento de modelos de AI.

Librería Pandas: Librería especializada en el manejo y análisis de estructuras de datos. [13]

Librería Numpy: Según el manual de Python[12] es una librería de python que se encarga de hacer cálculos numéricos y analizar los datos a gran volumen.

Librería Sklearn: Según la UOC[14] Librería que se encarga del procesamiento de un conjunto de datos.Conjunto de rutinas escritas en Python para hacer el análisis predictivo de algoritmos. Esta librería estaba en Numpys, SciPy y Matplotlib.

Procedimientos

Escogimos realizar una red neuronal para detectar la fortaleza de las manos de poker, realizamos varios modelos cada uno mejor que el anterior implementando nuevas técnicas que nos ayuden a mejorar las métricas accuracy, precision y recall, las cuales son indicadores muy utilizados para medir el rendimiento de las redes neuronales según Juba B. [18] sobre todo para aquellas muy grandes.

Data cleaning

Como se observa en la tabla 3 anteriormente vista, el dataset esta muy desbalanceado y vemos que existe filas repetidas, esto se nota en la mano de escalera real de color, segun el dataset global solo existen 4 manos posibles pero en el dataset de entrenamiento observamos que existen 5 manos, por lo cual podemos decir que existe una fila repetida, segun esta observacion podriamos inducir que las demas manos tambien podrian tener filas repetidas por lo cual antes de trabajar con este dataset de entrenamiento debemos hacer tratamiento previo a todo el dataset de entrenamiento.

Para el primer modelo (y los siguientes) se realizó el tratamiento de datos eliminación de valores “missing” con la función dropna() de pandas para eliminar las filas que no tengan datos si es que los hubiere, esto lo hacemos como un tratamiento estándar de los datos, igualmente, se eliminan los valores duplicados con la función df.drop_duplicates(keep='first') el cual se encarga de eliminar las filas duplicadas que se encuentran en el dataset, manteniendo la primera instancia encontrada y eliminando las otras repetidas.

Ahora nos encontramos con otro problema, antes que nada explicare que una mano de poker por ejemplo escalera real de color que es “As, rey, reina, jota” de mismo palo, 10, es lo mismo en todas sus formas desordenadas en que se robaron de la baraja, por ejemplo su homonimos “rey, As, reina, 10, jota”, “10, rey, As, reina, jota” y sus demas combinaciones del mismo palo, dicho esto la funciones anteriormente mencionadas para eliminacion de filas vacias y eliminacion de duplicados no elimina estos homonimos que a fin de cuentas son la misma mano, por lo tanto desarrollamos una funcion que se encargue de resolver este problema.

Función de mapeo: Esta función se encarga de mapear todas las manos que tengamos en el dataset de entrenamiento, descomponiendo en un array adicional y ordenando los atributos, según encuentre el mapeo en el array de una fila la compara con el mapeo de otras filas, y si dos array son iguales elimina el segundo array encontrado (fila) del dataset véase la figura 2.

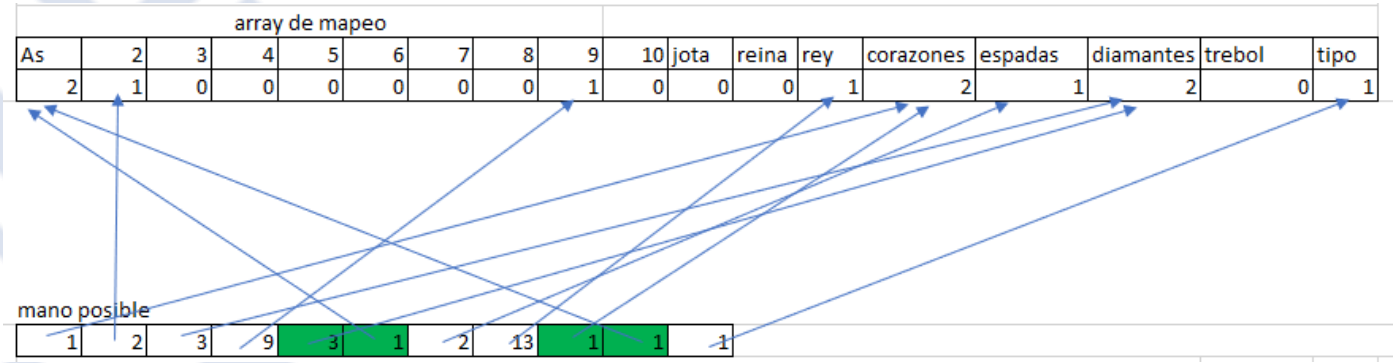

Figura 2. mapeo de manos utilizando un arreglo y contando la cantidad de apariciones.

Observamos que usando el método de transform(df) a realizar un ejemplo para una fila no hace el conteo de cuantas apariciones tenemos de cada palo y cada número en la mano de 5 cartas, por lo que ahora que está ordenado y mapeado tomamos como primer elemento y el las siguiente lecturas y mapeos de otras filas si encuentra alguna fila que sea igual al primer arreglo encontrado podemos deducir que es la misma mano, por lo tanto procedemos a eliminar esa segunda fila del dataset, asi eliminamos todas las filas que sean repetidas aun cuando las variables están en desorden.

Utilización de SMOTE:

También se utilizó SMOTE (Synthetic Minority Over-sampling Technique) para la creación de clases sintética, sin embargo no otorgó buenos resultados pues existían clases con cantidad de muestras menores a 5 vecinos, lo cual imposibilitaba la utilización correcta de SMOTE ya que cambia todas las clases a la menor terminando con 124930
muestras.

```
array([9, 9, 9, ..., 9, 9, 9])
```

Figura 3. Clases sintéticas al utilizar (k_neighbors<=4)

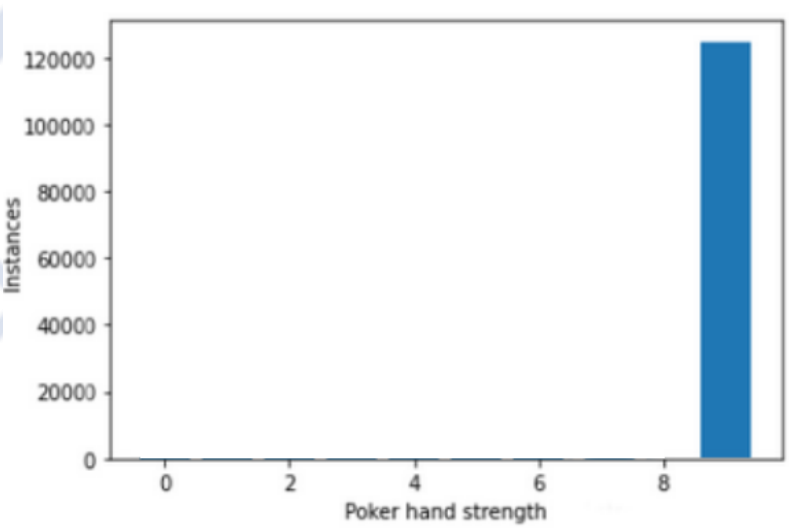

Figura 4. Clases al utilizar SMOTE

Utilización de class weights:

Se utilizó el paquete compute_class_weight de la librería sci-kit learn para asignar pesos dependiendo de las instancias existentes para cada clase, de esta forma asignando un peso mayor a las clases más raras.

```
[2.00176099e-01 2.35991318e-01 2.07363184e+00 4.87485380e+00
 2.68903226e+01 4.63111111e+01 6.94666667e+01 4.16800000e+02
 5.00160000e+02 5.00160000e+02]
```

Figura 5. Pesos por clase

Resultados y discusión

Primero debemos resaltar la importancia de la conversión o mapeamiento de los datos de entrada, pues esto nos significó un gran aumento en la precisión de todos los modelos que teníamos hasta el momento (20%).
Luego de la utilización de los pesos de clase que también ayudó en buena medida a mejorar los resultados obtenidos.

Y finalmente de la utilización de SCCE en vez de CCE, los resultados son los siguientes:
Al utilizar SparseCategoricalCrossEntropy se obtuvo:

Sparse categorical accuracy : 98.26%

```
313/313 [==============================] - 1s 3ms/step - loss: 30.0896 - sparse_categorical_accuracy: 0.9826
[30.08958625793457, 0.9826052188873291]
```

Figura 6. Métricas SCCE

Al utilizar CategoricalCrossEntropy se obtuvo:
Accuracy:97.73%
Recall: 97.73%
Precision: 97.72%

```
accuracy: 0.9773 - precision_2: 0.9773 - recall_2: 0.9772
```

Figura 7. Métricas CCE

Con lo cual podemos hallar el f1-score :

F1 score

$$F_1 = 2 \cdot \frac{p \cdot r}{p + r}$$

F1: 97.7249997442 %

Finalmente se implementó un método para la visualización de las predicciones junto con los valores esperados en tiempo real, para poder corroborar que el modelo en efecto predice las fortalezas de cada mano particular:

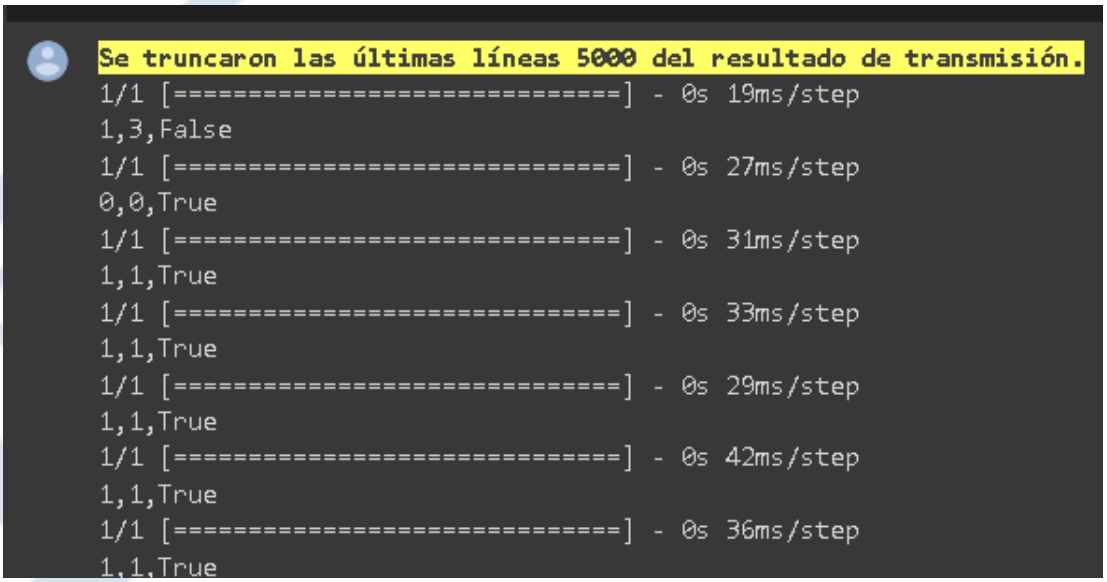

Figura 9. Predicción en tiempo real

Conclusiones

Como conclusiones tenemos que:
Es importante escoger la función de pérdida correcta, pues algunos mostraron mejores resultados que otros al no ser compatibles con el tipo de clase. Un cambio en los parámetros de entrada simple como fue el mapeo de las manos originales puede afectar en gran medida a las métricas obtenidas, obteniendo un 20% de precisión extra (de 60% a 80%), teniendo en cuenta que este dataset tiene una característica especial, que por más que se intente balancear aún conserva un poco del dataset original desbalanceado, eso se puede equilibrar mapeando las manos y equilibrando los pesos según la importancia de las manos que nos encontramos en el dataset.

Finalmente, con la aplicación de técnicas de balanceo se puede obtener el 20%, faltante así llegando a altísimos puntajes de precisión, accuracy y recall.

Contribución de Autoría

Estith Bryan Vargas Quispe: Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Eybert Macedo Pillco:** Conceptualización, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos. **Quispe Ttito Juan Carlos:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Redacción - borrador original, Escritura, revisión y edición. **Jose Miguel Cano Vilcapaza:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación.

Referencias

[1] González, J. (2013, 07). Inteligencia artificial aplicada al poker texas hold'em. UCrea. Retrieved June 22, 2022, from https://repositorio.unican.es/xmlui/handle/10902/3107

[2] Gascón, H., Bordallo, M., & Torres, E. (n.d.). Jugador Ganador de Poker Basado en Inteligencia Artificial. Retrieved June 23, 2022, from http://www.it.uc3m.es/jvillena/irc/practicas/07-08/IAPoker.pdf

[3] Zamora Díez, Fernando. (2021). PokerRun: desarrollo de una aplicación web de póker online. Core.ac.uk. https://core.ac.uk/display/459228542?source=2

[4] Superior, E., Detección, Y., Reconocimiento De Elementos En, M., De, P., Online, M., Jorge, G., Campo, Á., García, M., Ponente, María, J., & Sánchez, M. (2018). UNIVERSIDAD AUTONOMA DE MADRID TRABAJO FIN DE GRADO. https://repositorio.uam.es/bitstream/handle/10486/688137/g%C3%B3mez_campo_manuel_jorge_tfg.pdf?sequence=1

[5] Profesor, I., Berlanga, A., Jesús, D., Luís, J., & Madrid, G. (n.d.). TRABAJO FIN DE GRADO Título: Diseño y evaluación de una heurística de juego de póker Autor: Jacobo Conrado Pérez-Fajardo Titulación: Grado en Ingeniería. https://e-archivo.uc3m.es/bitstream/handle/10016/16334/TFG_Jacobo_Conrado_Perez_Fajardo.pdf?sequence=1&isAllowed=y

[6] Marco, G., Miguel, J., & Murillo, L. (2012). Diseño de Estrategias Óptimas en el Póker mediante Algoritmos Genéticos. https://e-archivo.uc3m.es/bitstream/handle/10016/16936/TFG_Gabriel_Marco_Angeles.pdf?sequence=4&isAllowed=y

[7] Findler, N. V. (1977). Studies in machine cognition using the game of poker. Communications of the ACM, 20(4), 230–245. https://doi.org/10.1145/359461.363617

[8] Cattral, R., & Oppacher, F. (2007). Discovering rules in the poker hand dataset. Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation - GECCO '07. https://doi.org/10.1145/1276958.1277329

[9] Şirin, V., & Polat, A. (n.d.). A MACHINE LEARNING APPROACH TO THE POKER PLAYING PROBLEM. Retrieved June 23, 2022, from https://users.metu.edu.tr/e163109/Term_Paper_CENG_562.pdf

[10] Da, B. (2018). Approximating Poker Probabilities with Deep Learning.https://arxiv.org/pdf/1808.07220.pdf

[11] "Colaboración de Google." https://research.google.com/colaboratory/faq.html (accessed Aug. 12, 2022)

[12] "La librería Numpy | Aprende con Alf." https://aprendeconalf.es/docencia/python/manual/numpy/ (accessed Aug. 12, 2022).

[13] "La librería Pandas | Aprende con Alf." https://aprendeconalf.es/docencia/python/manual/pandas/ (accessed Aug. 12, 2022).

[14] U. O. de Catalunya, "Espacio de recursos de ciencia de datos." http://datascience.recursos.uoc.edu/es/preprocesamiento-de-datos-con-sklearn/ (accessed Aug. 12, 2022).

[15] I. Challenger Pérez, Y. Díaz Ricardo, and R. Becerra García, "El lenguaje de programación Python/The programming language Python," Rev. Ciencias Holguín, vol. 20, pp. 1–13, 2014.

[16] "TensorFlow." https://www.tensorflow.org/ (accessed Aug. 12, 2022).

[17] "Las Redes Neuronales Artificiales - Raquel Flórez López, José Miguel Fernández Fernández - Google Libros." https://books.google.com.pe/books?hl=es&lr=&id=X0uLwi1Ap4QC&oi=fnd&pg=PA11&dq=redes+neuronales+&ots=gONwmsjqZl&sig=AMCEIhaM4AzG4aZctmXAybBjgkE&redir_esc=y#v=onepage&q=redes neuronales&f=false (accessed Aug. 17, 2022).

[18] Juba, B., & Le, H. S. (2019, July). Precision-recall versus accuracy and the role of large data sets. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 4039-4048).

Sistema Web para mejorar la gestión comercial y de talento humano utilizando la metodología Scrum

Web system to improve commercial and human talent management using the Scrum methodology

Maricielo Caciano Arroyo
Universidad Nacional de Trujillo.
Trujillo, Perú.
t513300120@unitru.edu.pe
https://orcid.org/0009-0007-3444-7985

Juan Pedro Santos Fernández
Universidad Nacional de Trujillo.
Trujillo, Perú.
jsantos@unitru.edu.pe
https://orcid.org/0000-0002-8882-9256

Antony Vasquez Cabrera
Universidad Nacional de Trujillo.
Trujillo, Perú.
t013300420@unitru.edu.pe
https://orcid.org/0009-0001-3151-2936

Luis Enrrique Boy Chavil
Universidad Nacional de Trujillo.
Trujillo, Perú..
lboy@unitru.edu.pe
https://orcid.org/0000-0002-3488-2668

Juan Luis Córdova Otero
Universidad Nacional de Trujillo.
Trujillo, Perú.
jcordovao@unitru.edu.pe
https://orcid.org/0000-0003-4159-7037

ARK: ark:/42411/s15/a147
DOI: 10.48168/innosoft.s15.a147
PURL: 42411/s15/a147

RECIBIDO 04/12/2023 • ACEPTADO 18/02/2024 • PUBLICADO 30/03/2024

RESUMEN

Este estudio se enfocó en implementar un Sistema Web para desarrollar la administración de talento humano y comercial en un supermercado, empleando la metodología Scrum. Se integraron sistemas de ventas y recursos humanos para optimizar la percepción del cliente y los resultados económicos, destacándose la eficacia de Scrum en casos prácticos, como la implementación de un sistema de ventas en línea para MIPYMES ocasionado por la crisis de salud originada por la propagación del COVID-19. Los resultados económicos, respaldados por indicadores clave como VAN, B/C y TIR, fortalecieron la viabilidad del proyecto. El sistema se estructuró en tres Sprints, cada uno enfocado en etapas específicas, utilizando el framework

Laravel y pruebas de rendimiento con JMeter, además la fase de desarrollo incluyó la codificación de todas las actividades planificadas. En conclusión, este sistema promete ser altamente beneficioso para mejorar la gestión de talento humano y de ventas. La eficacia de la implementación se respalda mediante pruebas de rendimiento, evidenciando su capacidad para gestionar un considerable volumen de solicitudes. La factibilidad financiera, respaldada por indicadores positivos, junto con un enfoque ágil, promete mejorar significativamente la operación y competitividad de un supermercado.

Palabras claves: Gestión comercial, Gestión de talento humano, Scrum, Sistema web.

ABSTRACT

This study focused on implementing a Web System to develop the management of human and commercial talent in a supermarket, using the Scrum methodology. Sales and human resources systems were integrated to optimize customer perception and economic results, highlighting the effectiveness of Scrum in practical cases, such as the implementation of an online sales system for MSMEs caused by the health crisis caused by the spread of COVID-19. The economic results, supported by key indicators such as NPV, B/C and IRR, strengthened the viability of the project. The system was structured into three Sprints, each focused on specific stages, using the Laravel framework and performance testing with JMeter. Additionally, the development phase included the coding of all planned activities. In conclusion, this system promises to be highly beneficial to improve human and sales talent management. The effectiveness of the implementation is supported by performance testing, demonstrating its ability to handle a considerable volume of requests. Financial feasibility, supported by positive indicators, together with an agile approach, promises to significantly improve the operation and competitiveness of a supermarket.

Keywords: *Commercial management, Human talent management, Scrum, Web system*

INTRODUCCIÓN

Este estudio se centró en implementar una plataforma en línea innovadora para desarrollar la gestión comercial y de talento humano en el sector de supermercados. Este sistema integró dos aspectos fundamentales: el sistema de la gestión de recursos de talento humano y el sistema dedicado a las operaciones de ventas. De acuerdo con [1], un sistema web es accesible a través de un navegador web, la aplicación facilita transacciones diarias, seguimiento de inventario, gestión de clientes y otros aspectos cruciales para la eficiencia operativa y sostenibilidad en el ámbito empresarial.

En primer lugar, se debe mencionar al sistema de ventas que según [2], se ocupa de las transacciones diarias, el seguimiento de inventario, la gestión de clientes y demás aspectos que

impactan directamente en la sensación del cliente y en los efectos financieros de la organización. Este sistema está evolucionando, ya que puede realizar transacciones de ventas, registrarlas, generar facturas, realizar cálculos de dividendos y reparar los servicios ofrecidos a empresarios y emprendedores.

De acuerdo con el estudio [3], cuyo objetivo fue abordar los desafíos enfrentados por las empresas de tamaño Micro, Pequeño y Mediano (MIPYMES) a raíz de los efectos económicas originarias de la pandemia de COVID-19, incluyó el desarrollo de un sistema de ventas en línea mediante un sitio web, utilizando la metodología Scrum. Obtuvo como resultados la elaboración de un sitio web para vender sus productos, lo que les permitió ajustarse y tener éxito durante la pandemia.

Adicionalmente [4], resalta la importancia de abordar las tensiones y desafíos inherentes a los proyectos de control de gestión en el ámbito empresarial. En este contexto, esta investigación se enfocó en la idea y creación de un sistema web eficaz mediante la implementación de la metodología Scrum. Esta metodología, se aplicó de manera específica en el caso de estudio, donde se delinearon y ejecutaron cinco Sprints, cada uno de ellos fundamentado en etapas particulares del desarrollo de la aplicación.

Por otro lado [5], define que el sistema de recursos humanos aborda la gestión integral del talento, abarcando desde la contratación y capacitación hasta la valoración del rendimiento laboral y el crecimiento profesional de los colaboradores, además menciona que, la interacción entre empleados y clientes desempeña un papel vital, la agilización de la gestión de talento se traduce directamente en un ambiente laboral más productivo y satisfactorio. Según [6], recursos humanos, originalmente llamados Relaciones Industriales, surgieron como intermediarios para gestionar conflictos entre objetivos organizacionales e individuales considerados incompatibles.

El estudio [7], tuvo el propósito de crear un sistema de control de nómina para Chalicen SAC usando Scrum. Los resultados incluyeron prototipos alineados con las funciones propuestas por el equipo Scrum, promoviendo eficiencia, satisfacción del cliente y comunicación constante. Además, el estudio [8], realizaron una serie de medidas para evaluar el producto de tecnología de los recursos humanos lo cual mostró problemas, como falta de sprints e insatisfacción del cliente, lo cual propusieron una propuesta de evaluar y mejorar las prácticas de Scrum para el futuro.

El propósito principal de este estudio fue implementar un Sistema Web que mejore la gestión de talento humano y comercial, empleando la metodología Scrum como componente crucial. La implementación de esta metodología permitió una estructura ágil y adaptativa que facilitó la optimización de los procesos comerciales, la gestión del talento en la organización, la optimización de la toma de decisiones estratégicas además de un entorno colaborativo que impulsó la innovación y la competitividad en el mercado..

Materiales y métodos o Metodología computacional

Metodología Scrum

Conforme [9], scrum es un marco eficaz para proyectos cambiantes, demostrando su capacidad en la resolución de problemas complejos y fomentando la creatividad e innovación. Según [10], los métodos ágiles, como Scrum, son ampliamente aceptados en diversos entornos, desde startups hasta instituciones gubernamentales. Así mismo, menciona que scrum busca aumentar la eficiencia del equipo y la excelencia del producto final al centrarse en las preocupaciones individuales de los miembros, promoviendo un entorno laboral cómodo y elevando la satisfacción. De acuerdo con [11], Scrum requiere roles bien definidos: el Scrum Master facilita y apoya al equipo, el Product Owner representa al cliente y prioriza el backlog, mientras que el Equipo de Desarrollo se autogestiona para realizar eficientemente el trabajo del sprint. Por otro lado, el estudio [12], menciona que scrum se basa en sprints para crear incrementos funcionales de software. Estos sprints aseguran un ritmo constante de desarrollo, con la transición fluida de uno a otro para mantener la dinámica continua. La metodología Scrum, aplicada en este caso, se compone de varias fases, cada una de las cuales aporta valor al proceso de desarrollo. Esta consta de 4 etapas como se muestra en la Figura 1.

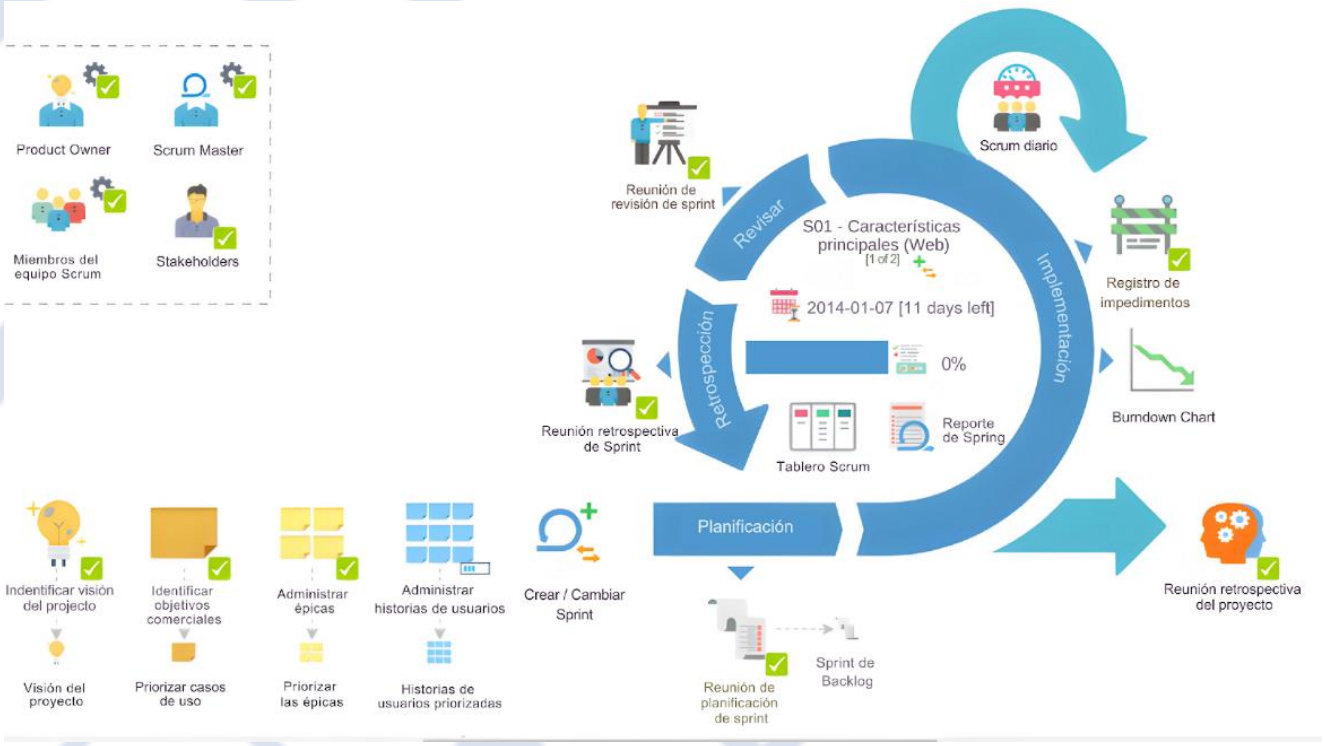

Figura 1. Etapas de la metodología Scrum. Adaptado de Visual Paradigm

En la Figura 1 se muestra un procedimiento laboral de Scrum que se fragmenta en períodos de iteración breves denominados sprints, cada uno con una duración de dos semanas o menos. Durante este intervalo, el equipo se dedica a cumplir un conjunto específico de tareas.

Framework Laravel

Según [13], Laravel, la plataforma PHP más ampliamente empleada, es ideal para aquellos nuevos en la programación y para aquellos con más experiencia en el campo. Reduce el período de creación de aplicaciones para la web y acelera el proceso de introducción al mercado gracias a sus enfoques dirigidos a PHP y módulos integrados. Laravel versión 10 con PHP 8.3 es especialmente robusto y adecuado para sistemas comerciales. Se implementó funcionalidades como gestión de inventario, procesamiento de pedidos y seguimiento de ventas utilizando Laravel, que permitió una construcción sólida y confiable del sistema del estudio.

Pruebas con Jmeter

Para evaluar el rendimiento en el estudio, se utilizó JMeter, una herramienta basada en Apache y desarrollada en Java. De acuerdo con [14], JMeter es diseñada para evaluar la carga en páginas web y diversas fuentes, JMeter calcula métricas como TPS (Transacciones por segundo). Se implentó JMeter versión 5.6.3 y Java 19.0.1 para llevar a cabo evaluaciones de desempeño en sistemas de bases de datos y solicitudes web. Esto permitió evaluar la administración de los datos guardados en el sistema de almacenamiento de información y la eficacia de la página web para manejar cargas intensas, generando resultados clave de rendimiento como tiempo en términos de reacción y habilidad para procesar.

Resultados

Fase de Planificación

Durante esta etapa, se dio inicio con la evaluación detallada de la problemática en cuestión. Se llevó a cabo una valoración exhaustiva del entorno y contexto de un Supermercado. Se observó de cerca diversos factores que influyen en la operación del negocio y que son esenciales para el éxito del proyecto como la gestión comercial y de talento humano.

Se configuró el Product Backlog que se muestra en la Tabla 1 donde se evaluó la duración y el esfuerzo necesario para las historias de usuario mediante el método de puntos de historia, y se utilizó el planning poker para asignar valores a cada historia. Durante esta fase [15], afirma que, en todas las características y funcionalidades necesarias, se estable la prioridad y se estima el tiempo requerido para la implementación. En la Tabla 1 se muestra las historias de

usuario del proyecto, indicando su descripción, prioridad, tamaño, puntos de historia y tiempo estimado en los sprints.

Tabla 1. Product Baclkog

Sprint	HU	Descripción	Prioridad	Size	Puntos de historia	Tiempo [Días]
Primer sprint	HU1	Ingresar al sistema	1	S	2	1
	HU2	Gestionar postulante	1	S	2	1
	HU3	Inscribir postulante	1	XS	1	½
	HU4	Gestionar proceso de selección	1	M	3	2
	HU5	Emitir resultados	1	S	2	1
	HU6	Gestionar tipo de clientes	1	S	2	1
	HU7	Gestionar clientes	1	S	2	1
	HU8	Gestionar tipo de comprobantes	1	XS	1	½
Segundo sprint	HU9	Gestionar empleado	1	S	2	1
	HU10	Gestionar área	2	XS	1	½
	HU11	Gestionar puesto	2	XS	1	½
	HU12	Gestionar criterios de evaluación	2	S	2	1
	HU13	Gestionar evaluación de desempeño	2	S	2	1
	HU14	Gestionar beneficios	2	XS	1	½
	HU15	Gestionar descuentos	2	XS	1	½
	HU16	Emitir comprobante de pago	2	L	5	3
	HU17	Gestionar roles	2	S	2	1
	HU18	Gestionar permisos	2	M	3	2
	HU19	Modificar contraseña	2	S	2	1
	HU20	Consultar historial de ventas	3	M	3	2
Tercer sprint	HU21	Gestionar planilla	3	M	3	2
	HU22	Gestionar capacitaciones	3	S	2	1
	HU23	Gestionar pagos	3	M	3	2
	HU24	Gestionar usuarios	3	XS	1	½
	HU25	Generar reportes de gestión	3	M	3	2
Puntos de historia / Tiempo estimado (Time Boxing)					52	26

Se realizó un análisis económico del proyecto, con costos de inversión (S/ 6 692,35) y desarrollo (S/ 2 507,20). Los beneficios tangibles sumaron S/ 12,36, mientras que los costos operativos, incluyendo recursos humanos y otros, fueron de S/ 7 098,09. El costo total de desarrollo fue de S/ 28 657,64. El análisis financiero abarcó un flujo de caja a tres años y la evaluación se realizó utilizando indicadores clave, como se muestra en la Tabla 2.

Tabla 2. Indicadores económicos

Indicador	Valor obtenido	Condición	Estado
VAN	S/ 3 341,59	VAN > 0	Aprobado
B/C	S/ 1,13	B/C > 1	Aprobado
TIR	33%	TIR > 12%	Aprobado

El Valor Actual Neto (VAN) de S/ 3 341,59 fue positivo, lo que señaló su viabilidad. El Índice de Beneficio/Costo (B/C) fue 1,13, superior a 1 y la Tasa Interna de Retorno (TIR) del 33% que fue mayor que el porcentaje de interés del 12%. Por lo tanto, el proyecto se consideró económicamente viable según los criterios aplicados.

Fase de Desarrollo

Durante esta etapa, [16] define que se codifica las actividades planificadas para el sistema web, utilizando la estructura Modelo Vista-Controlador (MVC) para cumplir con los requisitos establecidos. Se desarrolló el diseño estructural como se muestra en la Figura 2, garantizando así una solución integral a los requerimientos especificados. Además, en la Figura 3 se muestra los componentes, que compone el sistema, cómo los componentes del software, hardware, y las redes que se distribuyen y se comunican entre sí en un entorno de ejecución real.

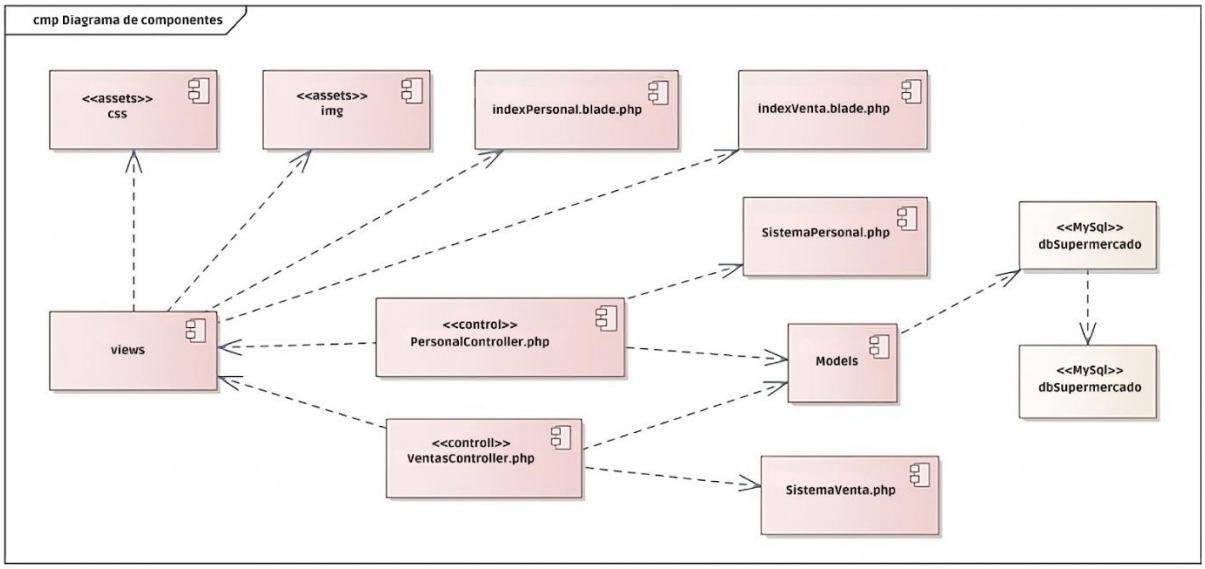

Figura 2. Diagrama de Componentes del Sistema web

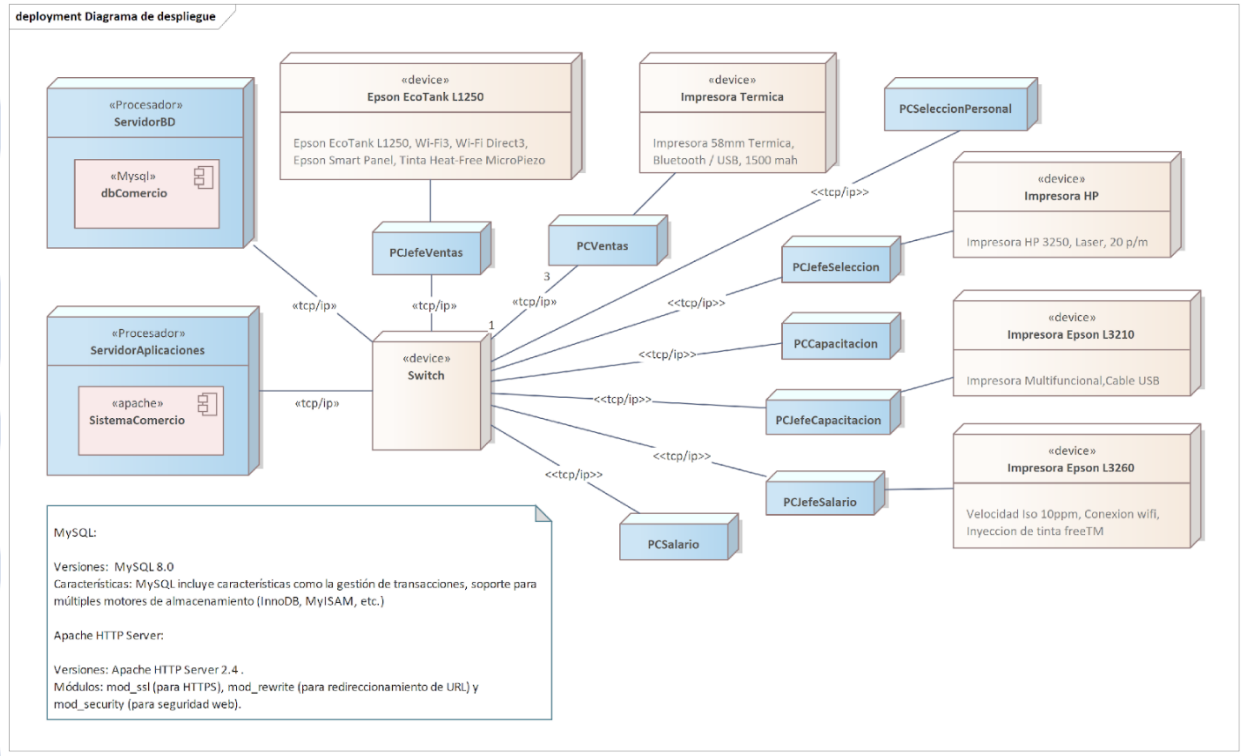

Figura 3. Diagrama de Despliegue

La gestión de la base de datos durante el transcurso de implementación de un sistema de software es un componente crítico que incide directamente en lograr los objetivos predeterminados para el proyecto. El proceso consta de varias etapas, siendo la primera el análisis de requisitos, donde se identificaron la información que se requiere ingresar en el almacenamiento de datos y se estableció la estructura de esta última conforme a dichos requisitos. Después, durante la etapa de diseño, se desarrolló detalladamente la configuración de la estructura, abarcando la definición de tablas, campos y relaciones entre ellas. En la Figura 4 se muestra una representación visual en conjunto con el gestor MySQL versión 8.0.

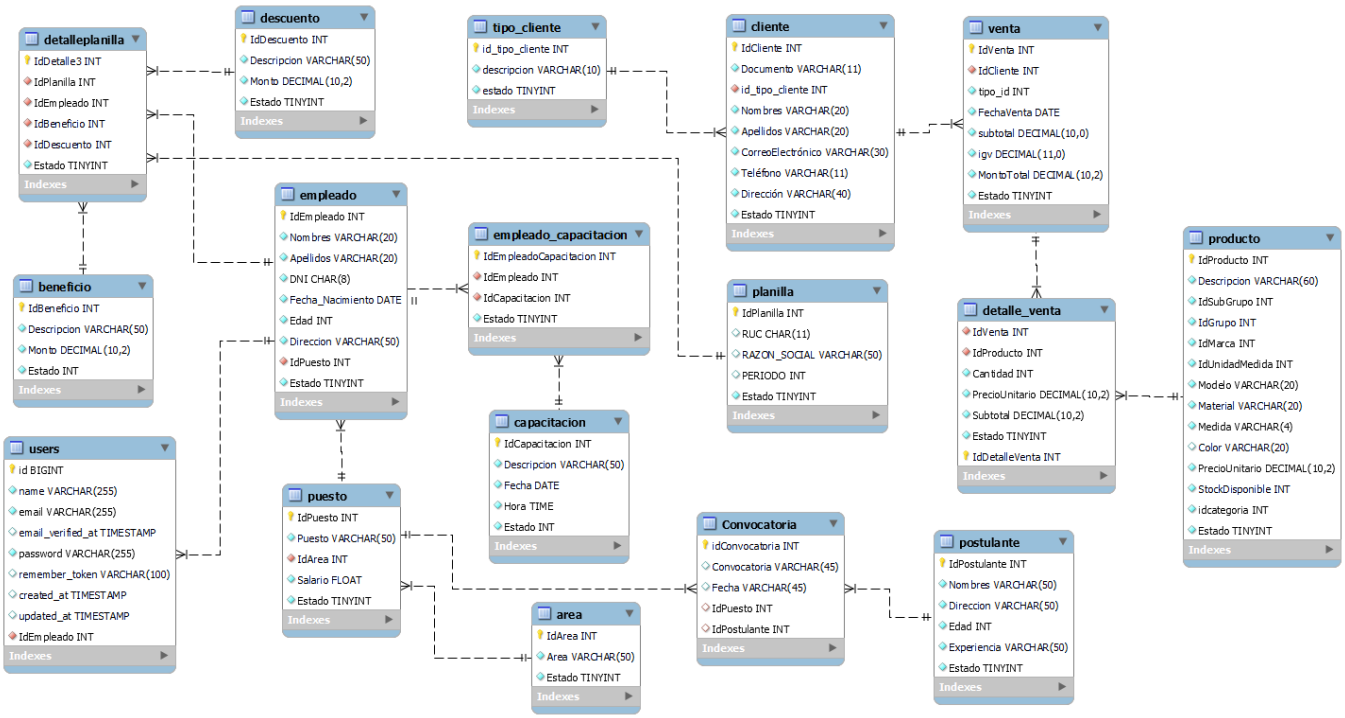

Figura 4. Base de datos del Sistema

Las pruebas de carga y estrés son componentes fundamentales en el proceso de desarrollo de software, diseñadas para evaluar el rendimiento y la resistencia de una aplicación frente a situaciones de demanda intensa. Mientras que las pruebas de carga determinan cómo responde el sistema bajo cargas normales o esperadas, las pruebas de estrés evalúan sus límites y capacidad para mantener la estabilidad operativa ante condiciones extremas.

Tabla 3. Reporte resumen de prueba de estrés

Etiqueta	# Muestras	Media	Mín	Máx	Desv. Estándar	% Error	Rendimiento	Kb/sec	Sent KB/sec	Media de Bytes
Login	1000	99	0	138	6,64	0,00%	910,75	265,04	124,52	298
Registrar postulante	1000	49	0	85	4,75	0,00%	956,02	278,22	130,71	298
Editar postulante	1000	50	0	90	5,23	0,00%	963,39	280,36	131,71	298
Eliminar postulante	1000	49	0	177	6,12	0,00%	969,93	282,27	132,61	298
Registrar empleados	1000	49	0	85	4,62	0,00%	969,93	282,27	132,61	298
Editar empleados	1000	49	0	179	6,16	0,00%	968,05	281,72	132,35	298
Eliminar empleados	1000	49	0	198	7,12	0,00%	968,99	281,99	132,48	298
Listar detalle planillas	1000	49	0	85	4,36	0,00%	964,32	280,63	131,84	298
Registrar detalle planillas	1000	49	0	84	4,54	0,00%	956,02	278,22	130,71	298
Registrar beneficios	1000	50	0	81	5,71	0,00%	956,93	278,48	130,83	298
Listar clientes	1000	50	0	86	5,46	0,00%	956,93	278,48	130,83	298
Registrar clientes	1000	49	0	179	7,01	0,00%	953,28	277,42	130,33	298
Listar tipo clientes	1000	50	0	180	7,49	0,00%	953,28	277,42	130,33	298
Registrar tipo comprobante	1000	49	0	85	5,17	0,00%	952,38	277,16	130,21	298
Actualizar tipo comprobante	1000	48	0	71	3,6	0,00%	967,11	281,45	132,22	298
Eliminar tipo de comprobante	1000	48	0	70	3,41	0,00%	967,11	281,45	132,22	298
Listar ventas	1000	49	0	72	3,85	0,00%	967,11	281,45	132,22	298
Crear venta	1000	49	0	71	3,87	0,00%	967,11	281,45	132,22	298
Registrar venta	1000	48	0	70	3,4	0,00%	967,11	281,45	132,22	298
Generar carrito de venta	1000	48	0	74	2,81	0,00%	967,11	281,45	132,22	298
Agregar productos	1000	48	0	70	2,02	0,00%	966,18	281,17	132,1	298
Total	22000	51	0	198	11,66	0,00%	10323,79	3004,38	1411,46	298

En la Tabla 3 se muestra la prueba de estrés realizada con JMeter a una aplicación web con diferentes funcionalidades, mostró que la aplicación tiene un buen rendimiento en general, con un tiempo de respuesta medio de 51 milisegundos, un rendimiento de 10323 peticiones por segundo y un ancho de banda de 3004 KB por segundo. Además, la aplicación no presentó ningún error, lo que indica una alta calidad. Sin embargo, la aplicación también mostró una alta variabilidad en los tiempos de respuesta, con una desviación estándar de 11,66 milisegundos y

una diferencia de 198 milisegundos entre el mínimo y el máximo. La funcionalidad con el mejor rendimiento fue la de agregar productos, mientras que la de login fue la que tuvo el peor rendimiento.

Fase de Finalización

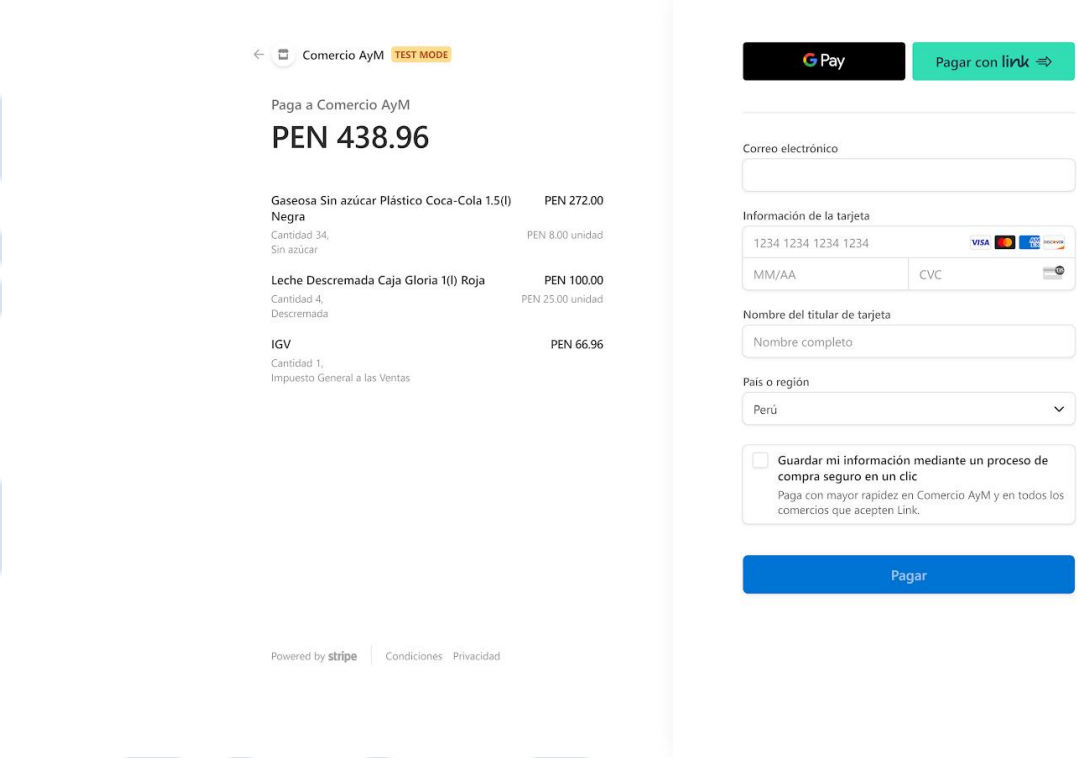

Figura 5. Pasarela de pagos del software

En la Figura 5 se muestra la pasarela de pago del software. La página permite al cliente pagar por sus compras utilizando diversas opciones de pago, como tarjetas de débito, crédito, y efectivo.

SUPERMERCADO TOTTUS
291 N 4th St, San Jose, CA 95112, La Libertad

COMPROBANTE DE PAGO 125
26-01-2024
Sede Trujillo

PRODUCTO	CANTIDAD	PRECIO	SUBTOTAL
Gaseosa Sin azúcar Plástico Coca-Cola 1.5(l) Negra	1	8.00	8.00
Mantequilla Clásico Taper Gloria gr. 200 azul	1	2.50	2.50
		SUBTOTAL	10.5
		IGV	1.89
		TOTAL	12.39

Figura 6. Boleta de venta

En la Figura 6 se muestra una boleta de venta que presenta con un código QR que permite verla fácilmente mediante el escaneo con un dispositivo móvil o escáner.

REPORTE DE DETALLE DE PLANILLA

Fecha del reporte: 21-01-2024

IDC	RUC DE PLANILLA	EMPLEADO	PUESTO	SALARIO INICIAL	BENEFICIO	DESCUENTO	SALARIO REAL
1	89343433102	José Paredes	Personal de Almacenamiento	1300	10.00	10.00	1300
2	89343433102	Antony Suárez	Vendedor	1260	40.00	30.00	1270
3	89343433102	Gloria Sánchez	Jefe de Almacén	1500	40.00	10.00	1530
4	89343433102	Paola Muñoz Suares	Encargado de Almacén	1036	20.00	0.00	1056
5	89343433102	Antony Vasquez Cabrera	Jefe de Ventas	1120	50.00	23.00	1147
6	89343433102	Sandra Jimenez Haro	Jefe de Área de Calidad	1025	30.00	0.00	1055
7	89343433102	Mauricio Arroyo Polo	Empaquetador	1400	50.00	23.00	1427
8	89343433102	Marley Casanova Lopez	Vendedor	1260	40.00	50.00	1250

Figura 7. Reporte de Planilla de Empleados

En la Figura 7 se muestra un informe detallado de la planilla de empleados. Esto permitió conocer salarios, beneficios y descuentos, simplificando la elección de opciones en la gestión de talento humano.

REPORTE DE VENTAS

Fecha del reporte: 21-01-2024

ID	NOMBRE CLIENTE	COMPROBANTE	FECHA	TOTAL
1	Juancito Miguel	BOLETA	2023-07-18	27.73
2	Carlos	BOLETA	2023-07-18	4.13
3	Roberto Carlos	BOLETA	2023-07-18	4.13
4	Ana	BOLETA	2023-07-18	47.20
5	Juan	BOLETA	2023-07-18	20.06
6	Carla	BOLETA	2023-07-18	23.60
7	Ana Sofia	BOLETA	2023-07-21	95.46
32	Juancito Miguel	BOLETA	2024-01-16	6454716.70

Figura 8. Reporte de ventas del software

En la Figura 8 se muestra los informes de los clientes que han realizado compras, que permitió así llevar un exhaustivo control sobre diversas facetas relacionadas con la administración de las ventas y la complacencia del cliente.

Discusiones

En primer lugar, el presente estudio incorporó un análisis económico para evaluar la viabilidad del proyecto, en donde se obtuvo resultados positivos en todos sus indicadores. A diferencia de la investigación desarrollada por [16], que se centró exclusivamente en la creación de un sistema en línea con el propósito de mejorar el proceso de ventas, sin llevar a cabo ningún tipo de evaluación económica.

En segundo lugar, se llevaron a cabo pruebas exhaustivas de carga y estrés para garantizar la disponibilidad del software frente a cargas excepcionales, logrando un rendimiento óptimo. Por otro lado, en el estudio [11], se establecieron roles definidos como el Product Owner, el Scrum Master y el equipo de desarrollo, sin embargo, careció de la implementación de estas pruebas, lo que podría haber comprometido la fiabilidad del software ante condiciones de alto estrés.

En tercer lugar, este sistema cuenta con una pasarela de pago, que permite realizar transacciones financieras de forma segura, diferenciándose del sistema propuesto por [15], que solo se centra en las reuniones diarias de Scrum para mantener el control y la comunicación abierta en el equipo, pero no aborda la implementación de una pasarela de pago.

Finalmente, se generó el recibo de venta que detalla la transacción efectuada por el cliente, junto con los productos adquiridos. En contraste, el software mencionado por [2], no incluyó la función de generar un recibo de venta, lo que lo deja incompleto en términos de sistema.

Conclusiones

El estudio logró crear con éxito un sistema en línea a través del enfoque de Scrum, enfocado en potenciar la gestión comercial y de talento humano en un supermercado. El análisis financiero validó la viabilidad económica del proyecto mediante indicadores positivos (VAN, B/C, TIR). Utilizando Laravel y pruebas con JMeter, se aseguró la solidez del sistema. Este enfoque ágil, respaldado por resultados financieros favorables, se presenta como una mejora significativa para la gestión y competitividad de un supermercado.

Se recomienda para futuros investigadores abordar áreas no exploradas como la integración de tecnologías emergentes en la gestión de personal y ventas. La adaptación del sistema a cambios en el entorno comercial requerirá atención continua. Se recomienda un enfoque constante en la investigación y desarrollo para mantener la relevancia y eficiencia en un entorno comercial dinámico.

Contribución de Autoría

Maricielo Estefany Caciano Arroyo: Conceptualización, Investigación, Metodología, Software, Validación, Redacción - borrador original. **Antony Fernando Vasquez Cabrera:** Conceptualización, Investigación, Metodología, Software, Validación, Redacción - borrador original. **Juan Pedro Santos Fernández:** Conceptualización, Investigación, Metodología, Software, Validación, Redacción - borrador original. **Luis Enrrique Boy Chavil:** Conceptualización, Investigación, Metodología, Software, Validación, Redacción - borrador original. **Juan Luis Córdova Otero:** Conceptualización, Investigación, Metodología, Software, Validación, Redacción - borrador original.

Referencias

[1] Y. C. Roca Avila y R. A. Revollo Linares, Artists, Desarrollo de un sistema web para mejorar la administración del Condominio Las Terrazas de Surco utilizando el marco SCRUM, 2021. [Art]. Universidad Tecnológica del Perú, 2021.

[2] W. A. M. Wan Dorishah y M. R. Annis Al Barakah, «A Point-of-Sale System for Measuring Sales Performance,» International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, nº 1.5, pp. 151-155, 2019.

[3] W. Fitri Hidayat, A. Purnamawat and F. Sarasati, "IMPLEMENTATION OF THE SCRUM MODEL IN THE DEVELOPMENT OF ONLINE SALES SYSTEMS OF MSMEs DURINGTHE COVID-19 PANDEMIC," Jurnal Techno Nusa Mandiri, vol. 18, no. 1, pp. 55-65, 2021.

[4] A. Delgado, E. Lee Huamaní and S. Samaniego Diego, "Design of web systems for inventory control in the E-commerce sector under the Agile methodologies approach," International Journal of Emerging Trends in Engineering Research, vol. 8, no. 7, p. 3129–3133, 2020.

[5] T. Javdani Gandomani, A. Mashmool, M. Dashti, S. Khosravi, M. Najafi Sarpiri, M. Radnejad, M. Afshari y S. Mansouri, «Talent management in agile software development: The state of the art,» de 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT), Surabaya, 2021.

[6] I. Chiavenato, ADMINISTRACIÓN DE RECURSOS HUMANOS, Santa Fe: McGraw-Hill Interamericana de España S.L., 2011.

[7] A. Delgado and C. P. Antunez-Maguiña, "Web System Design for Human Resources Management in an SME in the Textile Sector," International Journal of Emerging Trends in Engineering Research, vol. 8, no. 4, pp. 1471-1476, 2020.

[8] Y. Candra Kurniawan y T. Raharjo, «Scrum Effectiveness Measurement in Human Resources Technology Product at Telecommunication Company,» Journal of Informatics and Communications Technology (JICT), vol. 5, nº 1, pp. 149-158, 2023.

[9] SCRUM, «Scrum: A Framework To Reduce Risk And Deliver Value Sooner,» Marzo 2021. [En línea]. Available: https://scrumorg-website-prod.s3.amazonaws.com/drupal/2021-

03/Scrum-%20A%20Framework%20to%20Reduce%20Risk%20and%20Deliver%20Value%20Sooner.pdf. [Último acceso: 20 Enero 2024].

[10] D. Babik, "Scrum Boot Camp: Introducing Students to Agile System," Journal of Information Systems Education, vol. 3, no. 33, pp. 195-208, 2022.

[11] A. Johanes Fernandes, R. Rengga Eko, W. Rakkha Leonardi, P. Agustinus Adi y P. Tonny, «Development Point of Sales Using SCRUM Framework,» Journal of Systems Integration, vol. 10, nº 1, pp. 1804-2724, 2019.

[12] D. Oluwaseun Alexander y S. Ismaila Temitayo, «The adoption of Software Engineering practices in a Scrum environment,» AfricanJournal of Science, Technology, Innovation and Development, vol. 14, nº 6, pp. 1429-1446, 2021.

[13] Z. Subecz, «Web-development with Laravel framework,» Gradus, vol. 8, nº 1, p. 211–218, 2021.

[14] N. Husufa y I. Prihandi, «Optimizing JMeter on Performance Testing Using the Bulk Data Method,» Journal of Information Systems and Informatics, vol. 4, nº 2, pp. 205-215, 2022.

[15] M. S. Anggreainy, A. Sagiterry Setiawan, M. Subekti, K. Jingga, Noprianto and J. Hartanto, "Implementing Online Food Ordering System for Food Court Using Scrum Approach," in ICSEC 2021 - 25th International Computer Science and Engineering Conference, Chiang, 2021.

[16] A. Delgado y C. Cieza-Palma, «Design of a Web System for Sales Processes in a Microenterprise in Peru,» International Journal of Emerging Trends in Engineering Research, vol. 8, nº 4, pp. 1466-1470, 2020.

Aplicación de técnicas de Inteligencia Artificial para la diferenciación del nivel socioeconómico

Application of Artificial Intelligence techniques for the differentiation of the socioeconomic level

Crhistian Ziegler Pacori Paucar
Universidad Nacional de San Agustín.
Arequipa, Perú.
cpacori@unsa.edu.pe
https://orcid.org/0000-0003-4444-1273

Moises Enrique Mayta Condori
Universidad Nacional de San Agustín.
Arequipa, Perú.
mmaytac@unsa.edu.pe

Luis Fernando Quispe Sanomamani
Universidad Nacional de San Agustín.
Arequipa, Perú.
lquispesan@unsa.edu.pe

Diego Gustavo Montana Neyra
Universidad Nacional de San Agustín.
Arequipa, Perú.
dmontanan@unsa.edu.pe

ARK: ark:/42411/s15/a158
DOI: 10.48168/innosoft.s15.a158
PURL: 42411/s15/a158

RECIBIDO 19/12/2023 • ACEPTADO 01/03/2024 • PUBLICADO 30/03/2024

RESUMEN

En este proyecto se hace una diferenciación entre personas a travez de diferentes parametros como edad,sexo,nivel educativo entre otros,para tratar de calcular a cuanto podria asender su salario. Este problema es importante a resolver por que así una persona podría predecir su futuros ingresos a través de las decisiones que tomaría en el presente, como por ejemplo hasta qué grado de educación debe recibir y cuando ya comenzar a trabajar para obtener experiencia. Nuestro procedimiento para resolver este problema han sido dos análisis estadísticos ,el primero regresión lineal y un árbol de decisión para poder hacer una comparativa entre estos, las hemos probado usando herramientas como Colab (Python) y un dataset. Nuestra población de nuestro trabajo fue de 32000 registros (filas).Los resultados fueron que a través del árbol de decisión hubo una precisión de 0.879 y un accuracy de 0.817 .Y con respecto a la regresión logística obtuvimos una precisión de 0.80 cuando para el sueldo <=50K y 0.72 cuando el sueldo es >50K, el accuracy obtenido es de 0.7912. Dando por conclusión que entre estas dos herramientas nos quedamos con el Árbol de decisión.

Palabras claves: Inteligencia Artificial,árboles de decisión,regresión logística,dataset,nivel socioeconómico.

ABSTRACT

In this project, a differentiation is made between people through different parameters such as age, sex, educational level, among others, to try to calculate how much their salary could rise. This problem is important to solve because then a person could predict her future income through the decisions she would make in the present, such as how much education she should receive and when to start working to gain experience. Our procedure to solve this problem has been two statistical analyses, the first linear regression and a decision tree to be able to make a comparison between them, we have tested them using tools such as Colab (Python) and a dataset. Our population for our work was 32,000 records (rows). The results were that through the decision tree there was a precision of 0.88 and an accuracy of 0.82. And with respect to the logistic regression we obtained a precision of 0.80 when for the salary <=50K and 0.72 when the salary is >50K, the accuracy obtained is 0.7912. Concluding that between these two tools we are left with the Decision Tree.

Keywords: *Artificial Intelligence, decision trees, logistic regression, dataset, socioeconomic status.*

INTRODUCCIÓN

Los ingresos económicos de una persona vendrían a ser las entradas de dinero percibidos de manera regular en un periodo y magnitud constante. Entre ellos están los salarios, pensiones, subsidios, etc. Según [1] el ingreso promedio se calcula por el ingreso nacional bruto y la población. Al dividir todos los ingresos y ganancias anuales entre la cantidad de población del país, mostrará el ingreso promedio per cápita. Se incluyen en esta cantidad todos los sueldos y salarios, pero también otros ingresos no ganados en inversiones o ganancias de capital. El ingreso promedio más alto del mundo se obtiene en las Bermudas. El presupuesto per cápita más bajo existe en Afganistán. En la comparación sobre 67 países, Perú ocupa el 49° lugar con un ingreso anual promedio de 6030 USD y un ingreso mensual promedio de 503 USD.

El Perú,[2] considerado una de las estrellas de crecimiento económico internacional en las dos últimas décadas, se ha convertido ahora en el país con mayor caída del PBI en América Latina, esperándose una contracción de 13.9% hacia finales del año 2020, según el FMI. Este resultado, y la consiguiente destrucción de millones de empleos y el aumento de la pobreza generalizada, nos ha hecho perder en pocos meses todo lo alcanzado en una década de esforzado avance económico. Según [3] datos de la Encuesta Nacional de Hogares (ENAHO), en el segundo trimestre de 2020, la población ocupada disminuyó en más de 6 millones de personas en relación a similar periodo de 2019. Los mayores incrementos en la tasa de desocupación se registraron en hombres, personas entre 25 a 44 años de edad y personas con estudios superiores no universitarios. La disminución de la población ocupada fue mayor en el área urbana (-49,0%) que

rural (-6,5%), y en las actividades de construcción (-67,9%), manufactura (-58,2%), servicios (-56,6%) y comercio (-54,5%), principalmente.

Las autoridades políticas y sanitarias de un país limitaron temporalmente la cantidad de contagios o infecciones, restringiendo el funcionamiento de empresas y mercados, y obligando a las personas a permanecer en sus respectivos domicilios [4]. Obviamente, no fue posible mantener a la totalidad de la población confinada, ya que siempre algunas actividades esenciales tienen que seguir funcionando, como la producción de alimentos, el transporte de mercancías, los mercados de abastos, hospitales, farmacias, vigilancia policial, etc. Pero, aparte de este tipo de actividades, el gobierno se encontró ante un dilema. ¿Cuántas y cuáles de las restantes labores no esenciales deben de permanecer cerradas mientras dure la pandemia?

La cuarentena rígida, que obligaba a las personas a permanecer la mayor parte del tiempo en sus respectivos domicilios y obligaba también al cierre de la mayoría de empresas y actividades económicas, duró un poco más de 100 días, desde el 16 de marzo hasta los primeros días de julio [5]. Un mes antes, a inicios de junio hubo una primera apertura de la economía, que permitió la operación de algunos sectores de servicios públicos y otras operaciones de servicios técnicos privados y de distribución o reparto de mercancías y alimentos preparados a domicilio, con lo cual según cifras del MEF solo el 27.2% de la economía nacional permaneció cerrada.

La crisis de la COVID-19 [3] y la consiguiente interrupción masiva de la actividad económica afectó potencialmente a los más de 17,1 millones de trabajadores que conformaban la fuerza laboral peruana en 2019. En base a la ENAHO 2019 y a la metodología de la Organización Internacional del Trabajo (OIT) se estimó que un 40,8% del empleo de Perú se encuentra en sectores de riesgo alto y otro 8,4%, en sectores de riesgo medio-alto, lo cual se reflejó en que estos trabajadores perdieron su empleo y muchos vieron reducidas sus horas de trabajo, con recortes salariales.

El propósito de este artículo, es recolectar la información de la población vulnerable y saber a través de qué medios y cómo obtienen los recursos para mantener sus hogares y así poder brindar una ayuda más especializada y específica a ellos. Para tal efecto se utilizará la Inteligencia Artificial utilizando la Regresión Logística, junto a las librerías de aprendizaje que nos proporciona Python para realizar el análisis de los datos obtenidos.

Estado del Arte

Nuestra fundamentación teórica está basada en los siguientes artículos:

INVESTIGACIÓN INTERNACIONAL N-1

Ricardo Timarán-Pereira, R. ; Caicedo-Zambrano, J. ; Hidalgo-Troya, A. ;"Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11° "; Revista de Investigación Desarrollo e Innovación: RIDI; Vol.11, Nº.1; 2019, Barcelona, España, doi: 10.19053/20278306.v9.n2.2019.9184

OBJETIVO:

Se presentan los resultados obtenidos al aplicar el modelo de clasificación basado en árboles de decisión, con el fin de detectar factores asociados al desempeño académico de los estudiantes colombianos de grado undécimo de educación media,se seleccionó,la información socioeconómica, académica e institucional de estos estudiantes.Y se generaron árboles de decisión que permitieron identificar patrones asociados al buen o mal desempeño académico de los estudiantes en las pruebas para mejorar la calidad de la educación en Colombia.

MUESTRA:

Las encuestas realizadas en KDNuggets en 2002,2004, 2007 y 2014 se comprobó que CRISP-DM era la principal metodología utilizada.La esta metodología para proyectos de minería de datos no es la "más actual" o "la mejor", pero es muy útil para comprender esta tecnología o extraer ideas para diseñar o revisar métodos de trabajo para proyectos de similares características.CRISP-DM es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de minería de datos y contempla seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

CONCEPTOS CLAVES QUE SE ESTÁ ANALIZANDO:

Árboles de decisión:

Diagrama en forma de árbol que muestra la probabilidad estadística o determina un curso de acción. Muestra a los analistas y, a los que toman las decisiones, qué pasos deben tomar y cómo las diferentes elecciones podrían afectar todo el proceso.

Nivel Socioeconómico:

Descripción de la situación de una persona según la educación, los ingresos y el tipo de trabajo que tiene. El nivel socioeconómico por lo general se define como bajo, medio o alto.

Calidad de la educación:

La calidad del sistema educativo es la cualidad que resulta de la integración de las dimensiones de pertinencia, relevancia, eficacia interna, eficacia externa, impacto, suficiencia, eficiencia y equidad.

INVESTIGACIÓN INTERNACIONAL N-2

Rodríguez Garcés, C.Sandoval Muñoz, D.; "Consumo tecnológico: Análisis de los determinantes del equipamiento doméstico mediante Árboles de Decisión" Revista Internacional de Investigación en Ciencias Sociales, Vol. 11, Nº. 1, 2015 ,Chile,2015,págs. 70-85

OBJETIVO:

Se aprovecha la base de datos del Centro de aprendizaje de Chile, haciendo un análisis de tendencia a través de un árbol de decisión acerca de los niveles de penetración de tecnología doméstica y factores diferenciadores.
Los resultados muestran que, a pesar de cierta diferenciación por tipo de dispositivo y del perfil de usuario, la rápida y masiva integración de dispositivos tecnológicos se ha dado según el nivel socioeconómico y también que es el vector de mayor segmentación.
Lo observado ha revelado que los dispositivos bien integrados, como teléfonos celulares; denotan un mayor poder adquisitivo, como se hizo notar también en la tecnología del pasado, como la televisión por cable o satélite.

MUESTRA:

El universo de estudio está definido por la población de 18 años y más de zonas urbanas y rurales. El muestreo es probabilístico estratificado por conglomerados múltiples entrevistando a una muestra de alrededor de 1.500 personas en cada año, con un error de muestreo del +3% y un nivel de confianza del 95%, estableciéndose un margen de respuesta efectiva promedio para todos los años cercano al 85%.

CONCEPTOS CLAVES QUE SE ESTÁ ANALIZANDO:

Árboles de decisión:

Mapa de los posibles resultados de una serie de decisiones relacionadas. Permite que un individuo o una organización comparen posibles acciones entre sí según sus costos, probabilidades y beneficios. Se pueden usar para dirigir un intercambio de ideas informal o trazar un algoritmo que anticipe matemáticamente la mejor opción.

Nivel Socioeconómico:

Es una medida total económica y sociológica que combina la preparación laboral de una persona, de la posición económica y social individual o familiar en relación a otras personas, basada en sus ingresos, educación y empleo.

Calidad de la educación:

Implica una búsqueda de constante mejoramiento en todos sus elementos, en insumos (recursos disponibles en las escuelas), procesos de enseñanza (tiempo destinado a la enseñanza escolar, cantidad de tareas y estipulaciones curriculares) y en los productos (logros estudiantiles).

INVESTIGACIÓN INTERNACIONAL N-3

3-Blanca CUJI; Wilma GAVILANES; Rina SANCHEZ; ”Modelo predictivo de deserción estudiantil basado en árboles de decisión”;Revista Espacios; Vol. 38 (Nº 55) Año 2017. Pág. 17 Colombia

OBJETIVO:

Muestra la construcción de un modelo predictivo de deserción estudiantil, para pronosticar la probabilidad, que un estudiante abandone su programa académico, mediante técnicas de clasificación, basadas en árboles de decisión. Se construyó un árbol con cuatro niveles de profundidad y mismo número reglas, que evalúan a los posibles desertores. Llevando a concluir que las variables nivel y notas tienen mayor influencia en la deserción.

MUESTRA:

Se tomó datos, de 485 estudiantes, almacenados en hojas de cálculo y base de datos relacionales, de la DITIC. Los datos fueron transformados, a variables, según los tipos de atributos propuestos por el estadístico S. Stevens (1946), se clasificaron en tres tipos: nominal, ordinal y cuantitativo, estos fueron: Género,estado civil,etnia,edad,lugar de nacimiento, ciudad de residencia, nivel.

CONCEPTOS CLAVES QUE SE ESTÁ ANALIZANDO:

Árboles de decisión:

Permite evaluar mediante una representación gráfica los posibles resultados, costos y consecuencias de una decisión compleja. Este método es muy útil para analizar datos cuantitativos y tomar una decisión basada en números

Nivel Socioeconómico:

La condición socioeconómica, una medida de situación social que incluye típicamente ingresos, educación y ocupación, está ligada a una amplia gama de repercusiones de la vida, que abarcan desde capacidad cognitiva y logros académicos hasta salud física y mental.

Calidad de la educación:

Está determinada por los conocimientos y competencias por las que se adquieren el reconocimiento a los derechos humanos. Para avanzar en la mejora de la calidad educativa es necesario integrar las aptitudes, la innovación educativa, la eficiencia y la igualdad.

INVESTIGACIÓN INTERNACIONAL N-4

EMMANUEL VAZQUEZ "SEGREGACIÓN ESCOLAR POR NIVEL SOCIOECONÓMICO: MIDIENDO EL FENÓMENO Y EXPLORANDO SUS DETERMINANTES"

OBJETIVO:

Proveer una cuantificación de los niveles y la evolución de la segregación escolar por nivel socioeconómico en el mundo y contribuir a la discusión de sus determinantes.
Muestra: Este trabajo utiliza una base de datos producida por el Programa para la Evaluación Internacional de Estudiantes (PISA) como fuente de información. La primera prueba PISA se realizó en 2000 con la participación de 43 países. 7 La segunda (2003) se realizó en 41 países, la tercera (2006) en 57 países, la cuarta (2009) y quinta (2012) en 65,8 y la sexta edición (2015) en 72. De hecho, además de los países miembros de la OCDE, cada vez más países de diferentes partes del mundo se han sumado a la iniciativa, ampliando la cobertura del programa. En 2015, un total de 542 385 estudiantes en 18 602 escuelas completaron las evaluaciones, lo que representa casi 27 millones de estudiantes en todo el mundo.

CONCEPTOS CLAVES QUE SE ESTÁ ANALIZANDO:

Nivel Socioeconómico: Situación de una persona según la educación, los ingresos y el tipo de trabajo que tiene.
Calidad de la educación: Funcionamiento en los centros educativos que permite tener un control de todos los procesos llevados a cabo en los mismos, así como la correcta gestión de éstos.

Marco Teórico

Inteligencia Artificial:

Según Winston[6] la inteligencia artificial está definida como un estudio de computación que hace posible percibir, razonar y actuar. En el campo de la ingeniería artificial es resolver problemas del mundo real utilizándose como un arsenal de ideas sobre la representación del conocimiento, el uso del conocimiento y el montaje de sistemas

Árboles de decisión:

Según Fletcher et al. [7] Los árboles de decisión son un método de aprendizaje supervisado no paramétrico utilizado para clasificación y regresión. No hacen suposiciones sobre la distribución de los datos subyacentes y están capacitados en datos etiquetados para clasificar correctamente los datos no vistos anteriormente.

Regresión Logística:

Según Dominguez et al. [8] Los modelos de regresión logística son modelos estadísticos en los que se evalúa la relación entre una variable cualitativa dependiente, dicotómica (regresión logística binaria o binomial) o variable con más de dos valores (regresión logística multinomial). Una o más variables explicativas independientes, o co-variables, ya sean cualitativas o cuantitativas.

Proceso de la IA:

Describir el proceso de transformación de datos.

Limpieza de datos

Según Lomet [9] la limpieza de datos, también llamada limpieza o depuración de datos, se ocupa de detectar y eliminar errores e inconsistencias de los datos para mejorar la calidad de los datos. Los problemas de calidad de los datos están presentes en recopilaciones de datos individuales, como archivos y bases de datos, por ejemplo, debido a faltas de ortografía durante el ingreso de datos, falta de información u otras colecciones, como archivos y bases de datos, por ejemplo, debido a faltas de ortografía durante la entrada de datos, falta de información u otros

Transformación de datos

Según Astera [10] la transformación de datos es el proceso de convertir datos de un formato a otro formato que sea más utilizable por el sistema o la aplicación de destino. Incluye múltiples actividades: puede 'transformar' sus datos filtrándose según ciertas reglas y uniendo diferentes campos para obtener una vista consolidada. Las herramientas de transformación ayudan a lograr su resultado final con facilidad.

Cargar los datos

En esta etapa [11], los datos procedentes de la fase anterior (fase de transformación) son cargados en el sistema de destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes.

Eliminación de Outliers

En esta etapa[12] se eliminan los Outliers(valores atípicos) que son observaciones que se desvían tanto de otras observaciones como para despertar la sospecha de que fue generada por un mecanismo diferente.

Herramientas

Dataset:

Los datos usados tienen los siguientes campos:

Age: edad de la persona
Workclass : la clase de trabajo del individuo
fnlwgt: el peso del muestreo
Education: el grado de educación de la persona
Education-num: número de años de educación en total
Marital-status: estado civil del individuo.
Occupation: la ocupación/trabajo que desempeña el individuo
Relationship: el tipo de relación familiar
Race: la raza del individuo.
Sex: el género del individuo.
Capital-gain: ingresos ganados de fuentes de inversión que no son sueldos/salarios
Capital-loss: ingresos perdidos de fuentes de inversión que no son sueldos/salarios
Hours-per-week: horas de trabajo por semana
Native-country: Ciudad de nacimiento

El Dataset está compuesto por 32561 entradas sin valores nulos, de los cuales 24720 entradas son <= 50K y 7841 entradas son > 50K.

Colab:

Colab,[13] también conocido como "Colaboratory", permite programar y ejecutar Python en el navegador con las siguientes ventajas: No requiere configuración, Da acceso gratuito a GPUs, Permite compartir contenido fácilmente. Colab puede facilitar el trabajo de estudiantes, científicos de datos o investigadores de IA.

Librerías:

Pandas: ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales. Es un software libre distribuido bajo la licencia BSD versión tres cláusulas
Numpy: da soporte para crear vectores y matrices grandes multidimensionales, junto con una gran colección de funciones matemáticas de alto nivel para operar con ellas.
Sklearn: Cuenta con varios algoritmos de clasificación , regresión y agrupamiento , que incluyen máquinas de vectores de soporte , bosques aleatorios , aumento de gradiente , k -means y DBSCAN , y está diseñado para interactuar con las bibliotecas numéricas y científicas de Python NumPy y SciPy

Resultados Obtenidos:

INTERPRETACIÓN DE NUESTROS RESULTADOS

ARBOL DE DECISION

Metricas :

Confusión Matrix:

[[4346 614]

[560 993]]

Precision: 0.88; 0.61
Recall: 0.88; 0.63
F1 - Score: 0.88; 0.63
Accuracy: 0.82
True Positives(TP) = 4346
True Negatives(TN) = 993
False Positives(FP) = 560
False Negatives(FN) = 614

REGRESION LOGISTICA

1. **Precisión :** 0.80; 0.72
2. **Accuracy:** 0.792199815743679

3. Confusion matrix

```
[[7096  256]
[ 1774  643]]
```

True Positives(TP) = 7136
True Negatives(TN) = 594
False Positives(FP) = 1762
False Negatives(FN) = 277

La matriz de confusión muestra 7136 + 594 = 7730 predicciones correctas, y 1762 + 277 = 2039 predicciones incorrectas.

	precision	recall	f1-score	support
<=50K	0.80	0.97	0.87	7352
>50K	0.72	0.27	0.39	2417
accuracy			0.79	9769

Conclusiones

1- Podemos ver que la puntuación de precisión de nuestro modelo en Árbol de Decisión es 0.87968241 y R. Logística 0.80. Entonces, podemos concluir que nuestro modelo de Árbol de Decisión está haciendo un mejor trabajo al predecir.

2- Dado que la Precisión identifica la proporción de resultados positivos predichos correctamente Para nuestro caso la precisión obtenida con el Árbol de decisión es de 0.88; esto implica que de cada 100 personas, a 88 las clasifica con una predicción correcta y al resto no.

En Regresión Logística Precisión nuestra precisión es de 0.80 quiere decir que nuestra predicción es correcta en un 80% para personas que perciben ingresos menores e iguales a $50K y 0.68(68%) para los que ganan más de $50K.

3- A nivel de las recall y F1 el Árbol de Decisión tiene valores mayores con respecto a la Regresión Logística.

4- Dado que nuestro Recall en ambos casos son mayores al 0.8 , 87% en Árbol de Decisión y 97% R. Logística concluimos que de nuestros resultados predichos correctamente abarcan la mayoría de los resultados positivos reales.

5- Ya que ninguno de nuestros F1-score llega a 1.0 pero están por encima de 0.8 podríamos decir que nuestros valores solo son buenos.

Contribución de Autoría

Crhistian Ziegler Pacori Paucar: Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Moises Enrique Mayta Condori:** Conceptualización, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos. **Luis Fernando Quispe Sanomamani:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Diego Gustavo Montana Neyra:** Visualización, Software, Validación, Redacción - borrador original, Curación de datos.

Referencias

[1] "Ingresos promedio a nivel mundial." https://www.datosmundial.com/ingreso-promedio.php .

[2] J. Vega, "Departamento de economía," *Pontif. Univ. Católica del Perú*, p. 25, 2020, [Online]. Available: https://repositorio.pucp.edu.pe/index/handle/123456789/176236

[3] J. Gamero and J. Pérez, "Perú: Impacto de la COVID - 19 en el empleo y los ingresos laborales," *Organ. Int. de Trab. Panor. Labor. en tiempos la COVID- 19*, vol. I, no. I, p. 64, 2020, [Online]. Available: https://www.ilo.org/wcmsp5/groups/public/---americas/---ro-lima/documents/publication/wcms_756474.pdf

[4] "Decreto Supremo N° 051-2020-PCM" https://cdn.www.gob.pe/uploads/document/file/572157/DECRETO_SUPREMO_N%C2%BA_051-2020-PCM.pdf (accessed Jun. 27, 2022).

[5] "Decreto Supremo N° 116-2020-PCM" https://cdn.www.gob.pe/uploads/document/file/898487/DS_116-2020-PCM.pdf

[6] Patrick Henry Winston, *Artificial Intelligence*, 3rd ed., vol. 110, no. 5. Addison-Wesley Publishing Company, 1993.

[7] S. Fletcher and M. Z. Islam, "Decision tree classification with differential privacy: A survey," *ACM Comput. Surv.*, vol. 52, no. 4, 2019, doi: 10.1145/3337064.

[8] S. Domínguez-Almendros, N. Benítez-Parejo, and A. R. Gonzalez-Ramirez, "Logistic regression models," *Allergol. Immunopathol. (Madr).*, vol. 39, no. 5, pp. 295–305, 2011, doi: 10.1016/j.aller.2011.05.002.

[9] D. B. Lomet, "Bulletin of the Technical Committee on Data Engineering," *Bull. Tech. Comm. Data Eng.*, vol. 24, no. 4, pp. 1–56, 2001, [Online]. Available: papers2://publication/uuid/30073F7F-1B7C-4496-ADA4-94FF4E6EE8F7

[10] "Transformación de datos y por qué es importante para las empresas | Astera." https://www.astera.com/es/type/blog/data-transformation-tools/

[11] "ETL: Extracción, transformación y carga de datos - Evaluando Software." https://www.evaluandosoftware.com/etl-extraccion-transformacion-carga-datos/

[12] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "OPTICS-OF: Identifying local outliers," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 1704, pp. 262–270, 1999, doi: 10.1007/978-3-540-48247-5_28.

[13] "Te damos la bienvenida a Colaboratory - Colaboratory." https://colab.research.google.com/?hl=es#scrollTo=5fCEDCU_qrC0.

Explorando los Principales Atributos de Blockchain para la protección de Datos médicos: Una Revisión Sistemática

Exploring the Key Attributes of Blockchain for Medical Data Protection: A Systematic Review

Anderson Jhanyx Reyes Riveros
Universidad Nacional de Trujillo.
Dirección postal.
ajreyesr@unitru.edu.pe
https://orcid.org/0000-0002-7324-5055

Jean Marco Cárdenas Iglesias
Universidad Nacional de Trujillo.
Dirección postal.
jcardenasi@unitru.edu.pe
https://orcid.org/0000-0003-0315-3953

Alberto Mendoza de los Santos
Universidad Nacional de Trujillo.
Dirección postal.
amendozad@unitru.edu.pe
https://orcid.org/0000-0002-0469-915X

ARK: ark:/42411/s15/a130
DOI: 10.48168/innosoft.s15.a130
PURL: 42411/s15/a130

RECIBIDO 04/01/2023 • ACEPTADO 05/03/2024 • PUBLICADO 30/03/2024

RESUMEN

Este artículo aborda la protección de datos médicos en sistemas de información médica, centrándose en la creciente adopción de registros médicos electrónicos (EHR). Reconoce los desafíos de seguridad inherentes a los sistemas centralizados y aboga por un intercambio seguro de datos médicos. La metodología sigue los principios de la declaración PRISMA, utilizando motores de búsqueda como SCOPUS, PUBMED e IEEE XPLORE para identificar 20 documentos relevantes. Estos documentos se centran en atributos clave de la tecnología Blockchain: control de acceso, privacidad de datos, seguridad de datos y encriptación. Los resultados indican que el control de acceso es el atributo más recurrente, seguido por la privacidad de datos, seguridad de datos y encriptación. La discusión resalta la aplicabilidad práctica de estos atributos, mejorando la confianza del paciente y la eficiencia del flujo de trabajo médico. Las conclusiones afirman la relevancia de la Blockchain en la protección de datos médicos, señalando oportunidades para investigaciones futuras, especialmente en entornos de salud menos desarrollados. El estudio proporciona un marco integral para profesionales de la salud y desarrolladores, subrayando la necesidad de una mayor aplicación y exploración de estrategias de implementación mediante

casos de estudio específicos. En resumen, la revisión sistemática aporta de manera significativa al conocimiento y aplicación de blockchain en la gestión segura de la información médica a nivel global. Destaca la importancia de atributos clave de blockchain en la mejora de la seguridad, privacidad e integridad de los datos médicos, ofreciendo una perspectiva completa para profesionales y desarrolladores interesados en este ámbito.

Palabras claves: blockchain, control de acceso, datos médicos, Seguridad de información

ABSTRACT

This article addresses the protection of medical data in health information systems, focusing on the growing adoption of electronic health records (EHRs). It recognises the security challenges inherent in centralised systems and advocates for the secure exchange of medical data. The methodology follows the principles of the PRISMA statement, using search engines such as SCOPUS, PUBMED and IEEE XPLORE to identify 20 relevant documents. These papers focus on key attributes of blockchain technology: access control, data privacy, data security and encryption. The results indicate that access control is the most recurring attribute, followed by data privacy, data security and encryption. The discussion highlights the practical applicability of these attributes, improving patient confidence and medical workflow efficiency. The conclusions affirm the relevance of the blockchain in medical data protection, pointing to opportunities for future research, especially in less developed healthcare settings. The study provides a comprehensive framework for healthcare professionals and developers, highlighting the need for further application and exploration of implementation strategies through specific case studies. In summary, the systematic review makes a significant contribution to the understanding and application of blockchain in the secure management of medical information globally. It highlights the importance of key attributes of blockchain in improving the security, privacy and integrity of medical data, providing a comprehensive perspective for practitioners and developers interested in this area.

Keywords: *blockchain, access control, medical data, information security*

INTRODUCCIÓN

Los datos médicos son cruciales para el cuidado de los pacientes y es debido al grado de importancia que es requerido para su uso en toda institución médica. El almacenamiento no solo consta de datos médicos, sino también de datos de diagnósticos y hospitalizaciones. Según el estudio [1] indica que debido a la amplitud de estos datos, los sistemas de información médica se vuelven cada vez más complejos y estructuralmente extensos, consecuentemente esto conlleva a que se opte por sistemas de información electrónicos.

En el presente, debido al adelanto tecnológico se manejan los registros médicos electrónicos (EHR, por sus siglas en inglés), Abeywardena [2] la define como un sistema interorganizacional, que almacena datos médicos del paciente (datos poblacionales, registro de avance, fármacos, señales corporales, historial médico, inmunizaciones, datos de laboratorio e informes de radiografía, etc.), este sistema mejora la óptima atención médica debido a que permiten el intercambio y acceso de datos en tiempo real a todo el entorno médico (laboratorios, especialistas, farmacias, escuelas médicas, etc.), a su vez brindan automatización de actividades y soporte en la toma de decisiones. Sin duda los sistemas EHR gestionan eficientemente los datos médicos, pero esto también trae consigo los principales problemas de toda información expuesta electrónicamente.

La información médica que afecta directamente a la salud de un paciente debe ser íntegra y fiable. Además, la privacidad del paciente debe protegerse de la exposición a usuarios no autorizados. Por lo tanto, es necesario desarrollar un sistema seguro de intercambio de datos médicos que pueda proporcionar la integridad y fiabilidad de los datos médicos y proteger la privacidad del paciente abordando los problemas de los actuales sistemas centralizados de intercambio de datos médicos. Se ha propuesto la descentralización del sistema para complementar los problemas del actual sistema de intercambio de datos médicos [3].

Una de las tecnologías más innovadoras que se ha desarrollado en los últimos años ha sido el blockchain, no solo por la versatilidad que presenta sino también por la seguridad que garantiza. Dentro del presente artículo tenemos como objetivo el detallar los beneficios, retos y oportunidades que ofrecen los sistemas de seguridad y control de acceso a los datos médicos mediante el uso de esquemas basados en la tecnología blockchain [4].

Por tanto, el objetivo de la investigación es dar respuesta a la interrogante ¿Cuáles son los principales atributos de blockchain que se han identificado y estudiado en la literatura científica y técnica en relación con la protección de datos médicos?

Concepto del Blokchain

La definición de blockchain, es descrita como la base de datos que posee universalidad, no centralizada, posibilita el registro de un historial de operaciones cifradas (reemplaza la historia clínica del paciente convencional) haciéndola inalterable a futuros cambios, siendo así una plataforma de registro descentralizada que favorece a la no centralización, visibilidad y la integridad de los datos privados de cada paciente, en ese punto sus tres ventajas claves: monitorización, visibilidad e inmutabilidad [5].

La blockchain suministra de una base de datos distribuida inalterable apoyada en una serie de crecientes bloques. Estos, por ser públicos, integran a un sistema accesible fortaleciendo la

fiabilidad en base a la visibilidad y solidez del método de creación de la blockchain [6]. La plataforma, como es accesible, asimismo es seudoanonimo: los miembros registrados se verifican con claves públicas (alias), no con nombres [7].

Según esto, la blockchain puede brindar solidez, seguridad, visibilidad y capacidad de crecimiento a amplios sistemas de datos, facilitando así afrontar una diversidad de peligros, comprendiendo desde las filtraciones de datos. Mediante la blockchain, estos peligros pueden contrarrestar documentando uno a uno todos los movimientos hechas hacia los datos, generando conservar el reconocimiento y privilegios, logrando limitar al sujeto que tiene permiso de acceso de los datos mismos [8].

Sistemas de seguridad de datos médicos

En la actualidad, merece la pena prestar atención a como permitir que los pacientes sean propietarios de sus datos médicos y compartan sus datos médicos de forma segura y dinámica entre diferentes instituciones médicas, lo cual no es un tema nuevo para el intercambio de datos médicos. Muchas instituciones médicas pueden almacenar los datos de forma centralizada en servidores. A la hora de compartir datos médicos es necesario tener en cuenta cuestiones de seguridad y privacidad [9]. En [10], menciona que de igual modo posee la capacidad de incidir sobre las propias prácticas laborales así como obligaciones legitimas de los profesionales de salud. Los médicos pueden adoptar las excelentes decisiones en los cuidados del acceso a su historial clínico total, debido a la ausencia de acceso causa el retraso de decisiones relevantes e influir en el bienestar de salud [11].

Los registros de salud de información privada son la información más confidencial que hay, por lo tantos los profesionales médicos como las clínicas y servicios médicos están forzados por ley a velar por su secreto medico[12].

Aunque por otro lado los últimos dos poseen mayores recursos para resguardar la información confidencial, el personal médico con frecuencia dispone de medios limitados para garantizar la preservación del resguardo de los registros médicos[13].

Aunque con el tiempo es más habitual emplear la tecnología para digitalizar y modernizar los registros médicos o interactuar con el paciente, y de que también es más recurrente las consultas virtuales y la asistencia médica en línea para supervisar y gestionar a distancia el bienestar del paciente [14].

Teniendo en cuenta el panorama del sector salud y la necesidad de una mejora eficaz en la seguridad de datos médicos es que podemos aplicar técnicas como la blockchain que tiene por objetivo el gestionar y contar con un banco de datos integral de registros relacionado a la salud del paciente así como historias médicas, información de ADN, radiografías, registros biológicos,

datos de seguimiento médico, datos personales y demográficos e incluso redes sociales[15]. Cabe destacar que la Blockchain cumpla el rol de administrador de control de accesos y permisos para datos y registros médicos.

Control de acceso a datos médicos

El resguardo de información vinculados a la salud de pacientes ha sido una inquietud, para preservar la confidencialidad del paciente, cabe destacar la importancia de comprender como se almacenan, comparten, utilizan y gestionan sus datos [16].

Los sistemas de control de acceso en el campo de atención medica son especiales, esto debido a la capacidad de limitar el acceso a espacios importantes, impedir el esparcimiento de enfermedades, prevenir la sustracción de aparatos de cuidado médico y fármacos indispensables, así como velar por el bienestar de pacientes y colaboradores[17].

Aun cuando la urgencia de poseer mecanismos de protección seguras y el acatamiento de las legislaciones que aplican en el área, los centros médicos, e instituciones de salud todavía cuentan con algún margen para la manejabilidad en cuanto a la manera de la gestión del control de acceso. A fin de cuentas, diversas instalaciones de salud poseen múltiples espacios, secciones de personas bajo cuidado médico y estructuras físicas, así que necesitan invertir en un sistema de resguardo del centro médico concebido de modo particular.

Teniendo en cuenta lo anterior es importante tomar como referencia una de las principales ventajas en la Blockchain puesto que permite mejorar aspectos como el control de acceso al sistema de salud, dado que éstas no tienen únicamente un sitio de acceso a los registros, más que esto implica dentro de los registros distribuidos en variados nodos, dificultando así efectuar ataques como DDoS [18].

En contraste, en cuanto a la tecnología que no faculta suprimir o alterar obviando la autorización previa, la información una vez almacenada, conlleva a no poder modificar tal información en el sistema y asegura el no rechazo de algún individuo que haya efectuado el cambio a la información de la red [19].

Ventajas de la tecnología Blockchain

Las fundamentales ventajas de Blockchain son:

- Invariabilidad de datos: es casi improbable alterar la data de la red, y en instancias de ocurrir tendría la capacidad de anular la cadena de bloque.
- Red resiliente: Blockchain es resistente a problema en cierto modo, debido a que, si cierta parte genera un contratiempo, la totalidad de la red continuara operante con versión más reciente de información [20].
- Confiere garantía entre desconocidos: La tecnología Blockchain se desempeña conforme al contenido de registros, en consecuencia, no requiere de un tercero para otorgar fiabilidad sobre sí mismo.
- Diversidad de utilidad e implementaciones: Blockchain representa un sistema de información versátil al emplearla con una vasta gama de funciones y utilizaciones, como, por ejemplo, en procesos de votación, de certificación de propiedad en relación a bienes y derechos de autor, así como también al monitoreo de materia prima y productos terminados[21].

Desventajas de la tecnología Blockchain

Las fundamentales desventajas de Blockchain son:

- Invariabilidad de datos: así como es su ventaja, a su vez puede ser lo contrario, esto debido al momento en que las personas erran en registrar información, generando así problemas en la realidad, dado que los datos no podrán ser editada[22].
- Cambio de celeridad del manejo de información. La red puede mermar la celeridad de los intercambios al momento que se suscite un contratiempo con la red o mientras las tarifas de ese proceso suelen ser de menor costo, reduciendo así el estímulo a los mineros, es decir aquellos relacionados a dicha actividad[23].
- Desmesurada cantidad de recursos. Blockchain representa una red de unanimidad, en consecuencia, sirve de múltiples elementos en el momento de autenticar diferentes versiones de similar registro.
- Potencial crecimiento de falta de empleo. Blockchain posee la capacidad descartar intermediarios que protejan la información y suministren garantía de su contenido. Entonces, ciertas áreas y ocupaciones se verían afectadas. Pasando algo similar con la innovación que produce la tecnología [24].

Metodología

Las revisiones sistemáticas son beneficiosas en numerosas facetas esenciales, del cual son capaces de brindar una consolidación de la situación actual de la información dentro de un campo particular, con base en la cual puede llegar a encontrar venideras preferencias en relación al estudio, enfocar dudas que desde otro enfoque no tienen la posibilidad de ser resueltas por estudios únicos, reconocer desafíos en el estudio inicial que necesitan ser solucionados en próximas investigaciones y crear o valorar teorías acerca de qué forma acontecen eventos de interés [25].

Con el propósito de esta misma revisión sistemática se empleó la metodología PRISMA, la cual fue diseñado sobre todo para revisiones sistemáticas que analizan los impactos de las acciones médicas, sin importar la estructura de los estudios añadidos. A pesar de ello, los elementos del registro de control son utilizables para las publicaciones de revisiones sistemáticas la cual valoran diferentes enfoques no ligados a la salud (por ejemplo, programas sociales o educativos), así mismo múltiples puntos son válidos a revisiones sistemáticas con fines diversos al estudio de enfoques (un caso, análisis de origen, predominio o proyección).

La declaración PRISMA 2020 se sitúa orientada con el propósito de ser aplicada en revisiones sistemáticas el cual contienen resumen (por ejemplo, metaanálisis de contrastaciones por parejas u otros procedimientos de síntesis estadística) o que no involucren resumen (por ejemplo, puesto que hay solo un estudio valido).

Los ítems de la declaración PRISMA 2020 son significativos para las revisiones sistemáticas de enfoques híbridos (abarcan estudios cuantificables y descriptivos), aunque de igual manera ellos mismos necesitan analizar las normativas en relación con la demostración y resumen de información descriptiva [26],[27]. Igualmente, se brinda un formato PRISMA para el esquema de secuencia, que es capaz de adaptarse en propósito de si la revisión sistemática es original o actualizada (Figura 1).

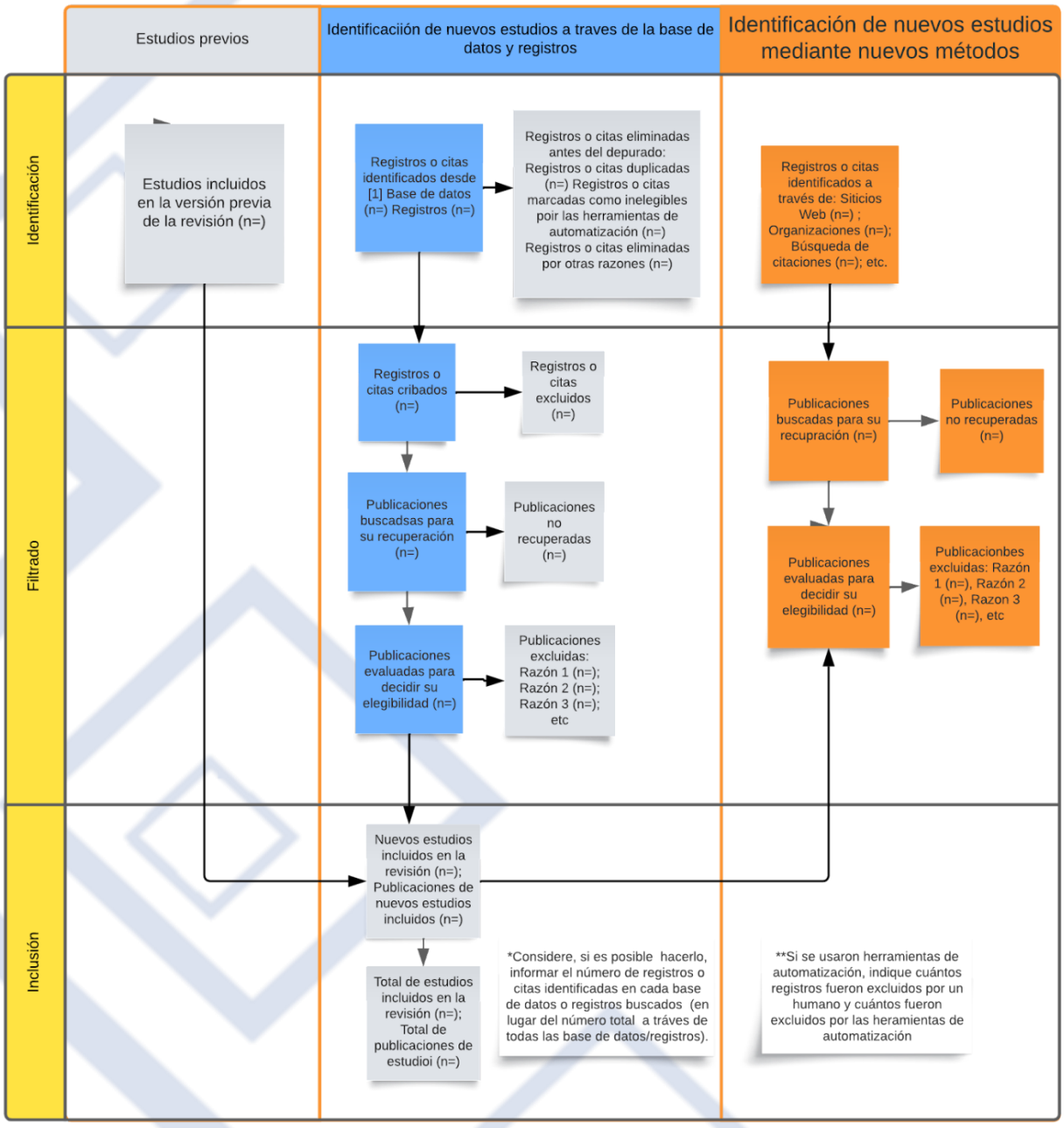

Figura 1. Diagrama de flujo PRISMA 2020

La versión más reciente ha sido ajustado por Yepes en [28]. Los cuadros en tono gris deben llenarse únicamente si son relevantes; en caso diferente, precisan ser eliminados del esquema de flujo. Es importante notar que un informe puede adoptar diversas formas, como un artículo en una publicación académica, borrador, resumen de ponencia, ficha de estudio, reporte de estudio médico, tesis o disertación, escrito no publicado, informe estatal u otro documento relevante.

Ecuaciones de búsqueda

Para poder comenzar el proceso de búsqueda, se implementó conectores booleanos de variables de estudio tomando en cuenta las diversas variables del tema. Con el fin de mejorar la precisión en la búsqueda de la literatura científica, se elaboró un protocolo que incluye la combinación de los términos predefinidos junto con los operadores booleanos detallados en la Tabla 1.

Tabla 1. Ecuación de búsqueda por fuente de búsqueda

Repositorio	Cadena de búsqueda
SCOPUS	TITLE-ABS-KEY (medical AND data AND security AND access AND contract AND blockchain)
PubMed	(medical data security and access control and blockchain)
IEEE xplore	("All Metadata": security) AND ("All Metadata": access control) AND ("All Metadata": medical data) AND ("All Metadata": blockchain)

Criterios de inclusión y exclusión

Los criterios de inclusión y exclusión son directrices particulares que se definen al llevar a cabo una revisión bibliográfica o un artículo de revisión. Su objetivo es determinar qué estudios o artículos serán considerados y cuáles serán descartados en la revisión. Estos criterios son esenciales con el fin de mantener la idoneidad y excelencia de los estudios incorporados en el análisis. Todos los criterios de inclusión planteados se visualizan en la Tabla 2 y todos los criterios de exclusión planteados se aprecia en la Table 3.

Tabla 2. Criterios de inclusión

N°	Criterios de inclusión
CI1	Los artículos deben abarcar la temática de los principales atributos de blockchain para la protección de datos médicos.
CI2	Seleccionar artículos redactados en inglés y español.
CI3	Los artículos hayan sido publicados entre los años 2018 y 2023.

Nota. CI= criterio de inclusión

Tabla 3. Criterios de exclusión

N°	Criterios de exclusión
CE1	Artículos duplicados.
CE2	Artículos que no hayan sido publicados entre los años 2018 y 2023.
CE3	Artículos que carecen de información relevante con nuestra temática.
CE4	Documentos que no sean artículos.

Nota. CE= criterio de exclusion

Proceso de recolección de información

Las primeras búsquedas se realizaron aplicando las directrices de inclusión mostrados en la Tabla 2, además de combinar los términos 'access control' y 'medical data security' en las bases de datos PubMed, SCOPUS usando combinaciones de términos booleanos AND y OR. Estas búsquedas arrojaron poca cantidad de resultados, algunos redundantes o poco valiosos para la revisión, pero nos brindaron una amplia comprensión del tema. Posteriormente se agregaron nuevos términos como 'blockchain', que ayudaron a que los resultados sean mucho mayores, sin embargo, no suficientes para nuestra revisión.

En la consulta de búsquedas en SCOPUS y PubMed se usaron las palabras anteriormente mencionadas haciendo diversas combinaciones: "access control"," medical security", "electronic health records" y "blockchain".

Al no tener mucha información acerca de la temática escogida se realizó una búsqueda manual mediante IEEE xplore con distintas combinaciones de los términos de búsqueda indicados como se logra visualizar en la Tabla 1.

Seguidamente, los documentos encontrados fueron depurados por los revisores aplicando los criterios de exclusión mostrados en la Tabla 3, con la finalidad de seleccionar aquellos que cumplieran con los estándares de relevancia y calidad para esta revisión sistemática. Se excluyeron los estudios que no estaban directamente relacionados con la temática, así como aquellos que no proporcionaban datos sustanciales para el análisis. Además, se descartaron aquellos documentos duplicados o de baja calidad metodológica. Este proceso de depuración aseguró la selección de fuentes confiables y pertinentes para el progreso de la actual revisión. Posteriormente, se procedió a la extracción de datos clave de los artículos seleccionados, lo cual constituye una fase crucial para la evaluación y resumen de la información.

En la Tabla 4, a continuación, se proporciona un desglose de los artículos correspondientes a cada base de datos y motores de búsqueda utilizados como punto de referencia.

Tabla 4. Artículos depurados empleando criterios de inclusión y exclusión

Base de datos	Artículos encontrados en total	Aplicando CE1	Aplicando CE2	Aplicando CE3	Aplicando CE4
SCOPUS	256	170	48	22	12
PubMed	57	43	21	8	5
IEEE xplore	16	14	8	4	3
TOTAL	329	227	77	34	20

Se aplicaron varios filtros a las publicaciones y revistas científicas seleccionadas (ver Figura 2), apegándonos a las directrices de inclusión y exclusión establecidos. Además, se evaluaron los principales atributos de blockchain que se han identificado y estudiado en la literatura científica y técnica en la relación con la protección de los datos médicos.

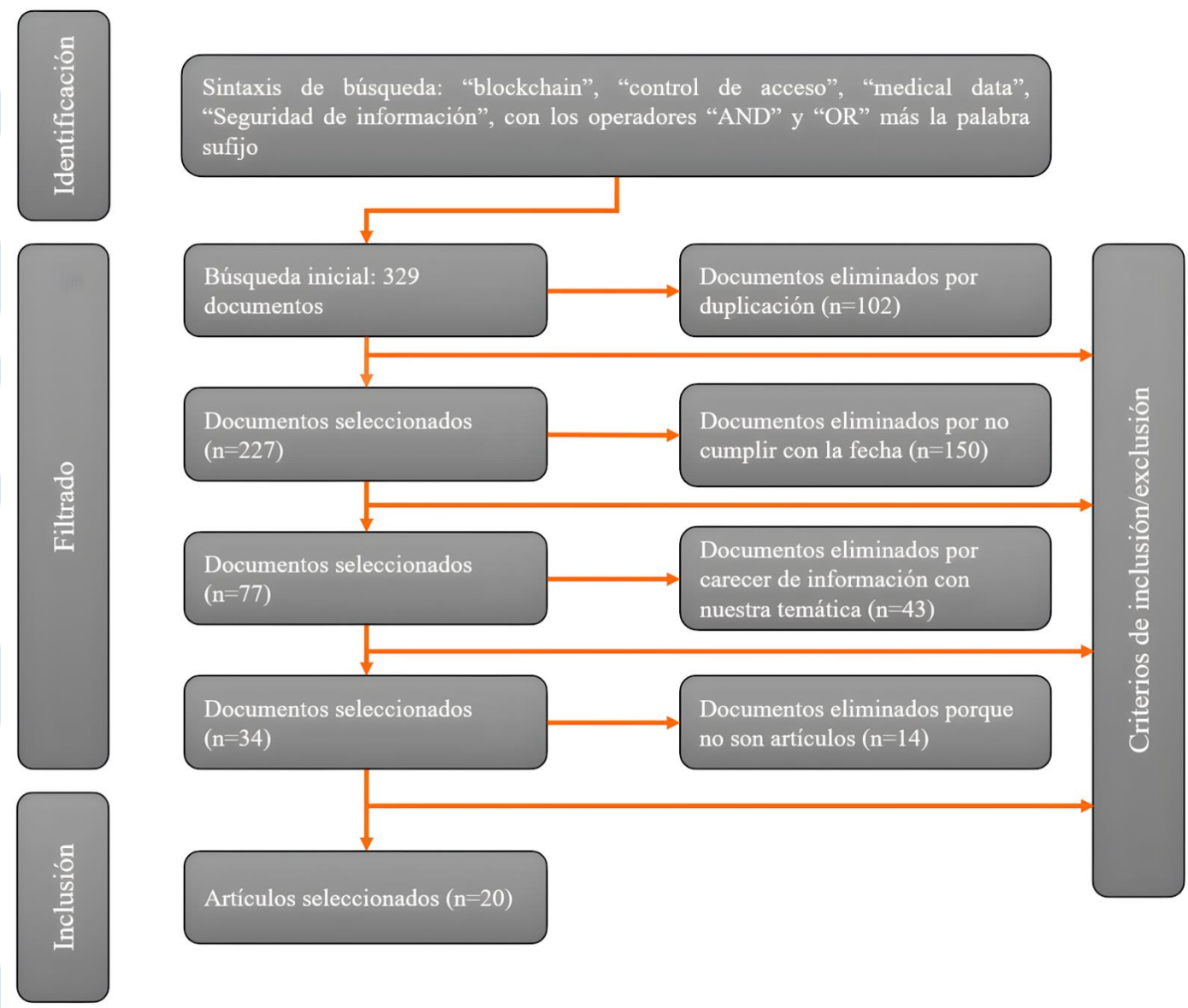

Figura 2. Diagrama de flujo PRISMA aplicado en este articulo

Resultados

Luego de ejecutar las directrices de inclusión y exclusión, se identificaron 20 artículos que cumplían con los requisitos establecidos. En la Tabla 5 se presenta un desglose detallado de estos

artículos, permitiendo así visualizar la evolución y enfoque de investigación orientada, tal como se refleja en las revistas donde fueron publicados.

Tabla 5. Resultados de búsqueda final y su respectivo enfoque de investigación orientada

N°	Autores	Atributos	Tipo	Tecnología
1	Liu, J., Li, X., Ye, L., Zhang, H., Du, X., Guizani, M	Privacidad de datos, control de acceso	Propuesta de sistema	Blockchain
2	Sharma, Balamurugan	Control de acceso, cifrado	Propuesta de sistema	Blockchain, Smart contract y cifrado proxy
3	Rupasinghe, T., Burstein, F., Rudolph, C.	Control de Acceso, privacidad de datos	Propuesta de arquitectura dinámica	Blockchain, smart contract,
4	Park, Y.-H., Kim, Y., Lee, S.-O., Ko, K.	Control de acceso	Propuesta de esquema	Blockchain, recifrado proxy, smart contract
5	Sun, Z., Han, D., Li, D., Wang, X., Chang, C.-C., Wu, Z.	Control de acceso	Propuesta de almacenamiento de seguridad de información	Blockchain, tejido Hyperledger, control de acceso basado en atributos, IPFS
6	Abeywardena, K.Y., Attanayaka, B., Periyasamy, K., Gunarathna, S., Prabhathi, U., Kudagoda, S	Control de acceso, privacidad de datos	Propuesta de sistema privado	Blockchain, criptografia, smart contract
7	Egala, B.S., Pradhan, A.K., Badarla, V., Mohanty, S. P	Control de acceso, privacidad de datos	Propuesta de esquema	Blockchain, smart contract
8	Younis, M., Lalouani, W., Lasla, N., Emokpae, L., Abdallah, M	Control de acceso, privacidad de datos	Propuesta de solución y nuevo protocolo	Blockchain, cloud, gestion de claves
9	Kumar, N.P.H., Prabhudeva, S.	Control de Acceso, privacidad de datos, encriptación	Propuesta de algoritmo	Blockchain Ethereum, sistema de archivos interplanetarios (IPFS), cifrado simétrico, smart contract

10	Majdoubi, D.E., Bakkali, H.E., Sadki, S	Control de acceso, privacidad de datos	Propuesta de sistema SmartMedChain	Blockchain en HyperLedger Fabric, sistema de archivos interplanetarios (IPFS), IoT
11	Hylock, R.H., Zeng, X	Privacidad de datos, control de acceso	Propuesta de sistema HealthChain	Blockchain, recifrado proxy, smart contract
12	Zhao, F., Yu, J., Yan, B.	Control de acceso	Propuesta de modelo de control	Blockchain, sistema de archivos interplanetarios (IPFS), smart contract
13	Devi Parameswari, C., Mandadi, V	Privacidad de datos, control de acceso	Propuesta de sistema	Blockchain, smart contract
14	Zhang, D., Wang, S., Zhang, Y., Zhang, Q., Zhang, Y	Control de acceso, privacidad de datos	Propuesta de esquema	Blockchain, smart contract
15	Chen, F., Huang, J., Wang, C., Tang, Y., Huang, C., Xie, D., Wang, T., Zhao, C	Control de acceso, seguridad de datos	Propuesta de diseño de control de acceso	Blockchain, smart contract, tejido Hyperledger
16	Hussien, H.M., Yasin, S.M., Udzir, N.I., Ninggal, M.I.H	Control de acceso, encriptación, privacidad de datos	Propuesta de esquema de control de acceso y autorización criptográfica	Smart contract, criptografía, cifrado, blockchain
17	Chen, Y., Meng, L., Zhou, H., Xue, G.	Control de acceso, encriptación, privacidad y seguridad de datos	Propuesta de esquema de preservación de la privacidad	Cifrado, Blockchain, tejido HyperLedger
18	Hafida Saidi; Nabila Labraoui; Ado Adamou Abba Ari; Leandros A. Maglaras; Joel Herve Mboussam Emati	Control de acceso, privacidad y seguridad de datos	Propuesta de sistema	Blockchain, smarts contracts
19	Eman-Yasser Daraghmi; Yousef-Awwad Daraghmi; Shyan-Ming Yuan	Control de acceso, encriptación, seguridad de datos	Propuesta de un diseño de sistema MedChain	Blockchain, smart contract, criptografía

20	Sabri Barbaria; Marco Casassa Mont; Essam Ghadafi; Halima Mahjoubi Machraoui; Hanene Boussi Rahmouni	Control de acceso, privacidad y seguridad de datos	Desarrollo de enfoque en intercambio de datos	Blockchain Hyperledger, smart contract

Tras un exhaustivo análisis de los 20 artículos seleccionados, se constató que diversos autores emplean una variedad de atributos en sus investigaciones sobre blockchain. Los atributos que fueron encontrados y seleccionado fueron (Control de acceso, privacidad de datos, seguridad de datos y encriptación). Con el fin de abordar la pregunta central de nuestra investigación, se hizo un escrutinio detallado y se procedió a la clasificación de estos atributos, cuyos resultados se presentan de manera detallada en la Tabla 6.

Tabla 6. Distribución de atributos de artículo seleccionados

N°	Control de acceso	Privacidad de datos	Seguridad de datos	Encriptación
1	1	1		
2	1			1
3	1	1		
4	1			
5	1			
6	1	1		
7	1	1		
8	1	1		
9	1	1		1
10	1	1		
11	1	1		
12	1			

13	1	1		
14	1	1		
15	1		1	
16	1	1		1
17	1	1	1	1
18	1	1	1	
19	1		1	1
20	1	1	1	
Total	**20**	**14**	**5**	**5**

A continuación, presentaremos el porcentaje de atributos aplicados en los artículos seleccionados (ver Figura 3). Se destaca que el Control de acceso es el atributo más prevalente, representando un 46% del total. Le sigue la Privacidad de datos con un 32%, mientras que tanto la Seguridad de datos como la Encriptación registran un 11% cada una.

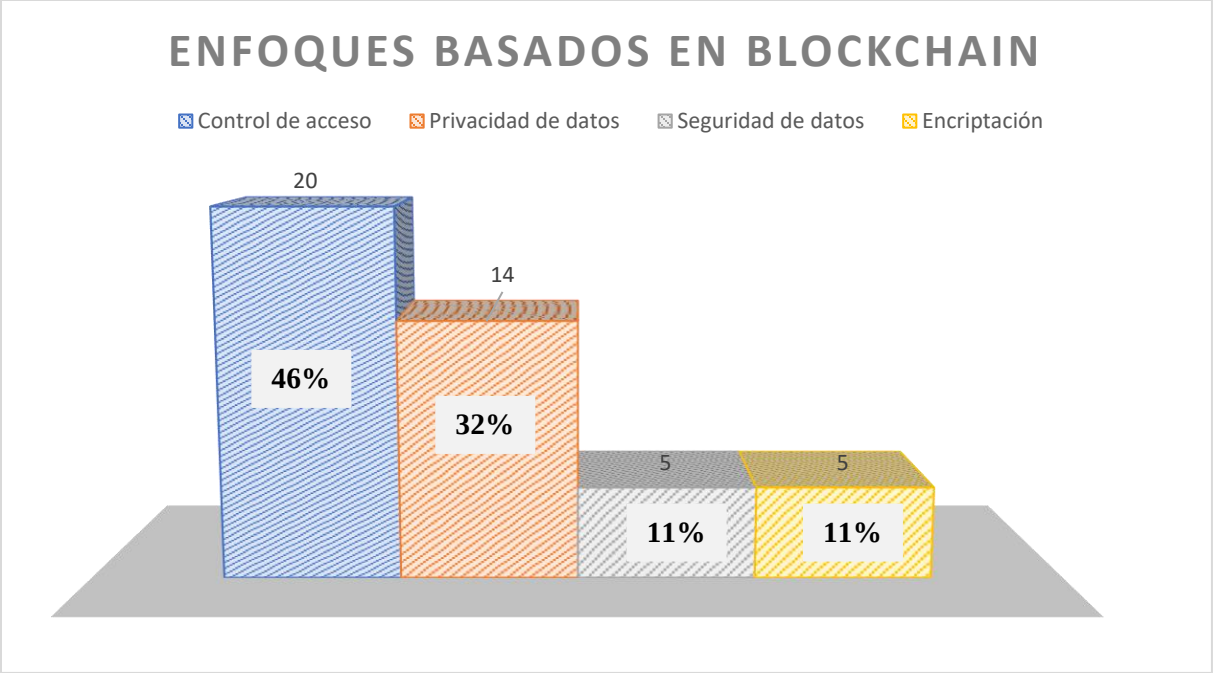

Figura 3. Distribución porcentual de los atributos de blockchain

Discusión

El empleo de Blockchain dentro de la atención médica y asistencia hacia el paciente, demanda el involucramiento de diversas entidades comprometidas, agregando proveedores, programas de bienestar y autoridades particulares. La integración de esta comunidad con el fin de conformar las directrices se volverá un desafío. Gracias a la índole de la información de la asistencia médica, poseer una garantía de datos siendo relevante. Como mencionamos en el presente artículo, hay diversas alternativas para tratar los desafíos de confidencialidad, tales como la Blockchain.

Los resultados encontrados con respecto a los principales atributos de blockchain para la protección de datos médicos, rige en cuatro atributos (control de acceso, privacidad de datos, seguridad de datos y encriptación) según los diferentes artículos, donde muestra de manera general que la implementación de dichos atributos aumenta la confianza, satisfacción del paciente y el flujo de trabajo eficiente. Estos resultados llegan a coincidir con Abubakar [29], lo cual señala que cuando se somete a análisis y evaluaciones de seguridad, el sistema de blockchain muestra mejoras de rendimiento en los niveles de privacidad de los datos, alta seguridad y diseño de control de acceso ligero en comparación con los modelos actuales de control de acceso centralizado, esto implica que los paciente tengan más confianza.

Una faceta fundamental de un sistema de asistencia médica equivale a la manera en la cual distribuyen la información mediante operaciones de secuencia de importancia. Esta tecnología posibilita la usabilidad compartida de la data sin prescindir de quitar las copias, equivocaciones y contradicciones de la cual puedan originarse mediante resguardo de información convencional, en otras palabras, suprime al mediador existente entre la solicitud de una prestación la cual generara un escenario más viable de intercambio de información médica.

Estos resultados coinciden con Tao [30], lo cual señala que a través de análisis de seguridad, rendimiento y comparación con otras soluciones, el esquema de este artículo puede satisfacer las necesidades de los escenarios de la vida real en términos de seguridad y viabilidad, y proporciona un nuevo modelo práctico para el intercambio de información médica.

Finalmente, es importante mencionar las limitaciones de este estudio. En su mayoría, los artículos analizados provienen de países desarrollados, lo que plantea retos con miras a la aplicación de Blockchain dentro del sector salud en países como Perú. Además, se observó una escasez de artículos sobre Blockchain publicados en países de Latinoamérica, y algunos de estos no están accesibles para su consulta.

Conclusiones

Esta revisión sistemática muestra un notable avance para el ámbito de la protección de datos médicos mediante tecnologías basadas en blockchain. La investigación ha proporcionado una visión detallada y exhaustiva de los atributos clave que destacan en esta área crucial. Específicamente, el control de acceso, la privacidad de datos, la seguridad de datos y la encriptación emergen como pilares fundamentales en la preservación y seguridad de la información médica en entornos blockchain.

Este estudio no solo consolida y sintetiza el estado actual del conocimiento sobre este tema, sino que también identifica áreas de oportunidad para futuras investigaciones. Uno de los hallazgos más relevantes es la necesidad de extender la aplicación de estas tecnologías a contextos menos desarrollados, como en el caso particular de Perú y países latinoamericanos, donde aún existe un amplio potencial por explorar.

En términos de contribución al campo, esta revisión sistemática proporciona un marco sólido y comprensivo para personal de salud, investigadores y desarrolladores de tecnología interesados en la protección de datos médicos. Ofrece una hoja de ruta valiosa al resaltar los atributos más prometedores y subraya la importancia de la aplicación de blockchain en la gestión segura de la información médica. Para futuras investigaciones, se sugiere un mayor énfasis en la adaptabilidad y viabilidad de estas soluciones en entornos de salud de países menos desarrollados. Además, sería beneficioso profundizar en la exploración de estrategias de implementación y casos de estudio específicos para analizar la influencia real de estos atributos en la práctica.

En resumen, esta revisión sistemática no solo enriquece el entendimiento sobre la protección de datos médicos con tecnologías blockchain, sino que también abre la puerta a una nueva era de investigación y desarrollo en este campo, con la aptitud de convertir el estilo en la cual se aborda la seguridad de la información médica a nivel global.

Contribución de Autoría

Anderson Jhanyx Reyes Riveros: Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Jean Marco Cárdenas Iglesias:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Alberto Carlos Mendoza de los Santos:** Visualización, Software, Validación, Redacción - borrador original, Curación de datos.

Referencias

[1] Z. Sun, D. Han, D. Li, X. Wang, C. C. Chang, and Z. Wu, "A blockchain-based secure storage scheme for medical information," Eurasip J. Wirel. Commun. Netw., vol. 2022, no. 1, 2022, doi: 10.1186/s13638-022-02122-6.

[2] K. Y. Abeywardena, B. Attanayaka, K. Periyasamy, S. Gunarathna, U. Prabhathi, and S. Kudagoda, "Blockchain based Patients' detail management System," ICAC 2020 - 2nd Int. Conf. Adv. Comput. Proc., pp. 458–463, 2020, doi: 10.1109/ICAC51239.2020.9357163.

[3] S. Lee, J. Kim, Y. Kwon, T. Kim, and S. Cho, "Privacy Preservation in Patient Information Exchange Systems Based on Blockchain: System Design Study," J. Med. Internet Res., vol. 24, no. 3, 2022, doi: 10.2196/29108.

[4] M. Abouali, K. Sharma, O. Ajayi, and T. Saadawi, "Performance Evaluation of Secured Blockchain-Based Patient Health Records Sharing Framework," 2022 IEEE Int. IOT, Electron. Mechatronics Conf. IEMTRONICS 2022, 2022, doi: 10.1109/IEMTRONICS55184.2022.9795759.

[5] S. Nandi, J. Sarkis, A. A. Hervani, and M. M. Helms, "Redesigning Supply Chains using Blockchain-Enabled Circular Economy and COVID-19 Experiences," Sustain. Prod. Consum., vol. 27, pp. 10–22, 2021, doi: 10.1016/j.spc.2020.10.019.

[6] F. Chen et al., "Data Access Control Based on Blockchain in Medical Cyber Physical Systems," Secur. Commun. Networks, vol. 2021, 2021, doi: 10.1155/2021/3395537.

[7] B. S. Egala, A. K. Pradhan, V. Badarla, and S. P. Mohanty, "Fortified-Chain: A Blockchain-Based Framework for Security and Privacy-Assured Internet of Medical Things with Effective Access Control," IEEE Internet Things J., vol. 8, no. 14, pp. 11717–11731, 2021, doi: 10.1109/JIOT.2021.3058946.

[8] S. Ghaffaripour and A. Miri, "Application of Blockchain to Patient-Centric Access Control in Medical Data Management Systems," 2019 IEEE 10th Annu. Inf. Technol.

Electron. Mob. Commun. Conf. IEMCON 2019, pp. 190–196, 2019, doi: 10.1109/IEMCON.2019.8936186.

[9] F. Zhao, J. Yu, and B. Yan, "Towards cross-chain access control model for medical data sharing," Procedia Comput. Sci., vol. 202, pp. 330–335, 2022, doi: 10.1016/j.procs.2022.04.045.

[10] C. D. Parameswari, "Healthcare data protection based on blockchain using solidity," pp. 577–580, 2020.

[11] J. Liu, X. Li, L. Ye, H. Zhang, X. Du, and M. Guizani, "BPDS: A Blockchain Based Privacy-Preserving Data Sharing for Electronic Medical Records," 2018 IEEE Glob. Commun. Conf. GLOBECOM 2018 - Proc., pp. 1–6, 2018, doi: 10.1109/GLOCOM.2018.8647713.

[12] H. Saidi, N. Labraoui, A. A. A. Ari, L. A. Maglaras, and J. H. M. Emati, "DSMAC: Privacy-Aware Decentralized Self-Management of Data Access Control Based on Blockchain for Health Data," IEEE Access, vol. 10, pp. 101011–101028, 2022, doi: 10.1109/ACCESS.2022.3207803.

[13] E. Y. Daraghmi, Y. A. Daraghmi, and S. M. Yuan, "MedChain: A design of blockchain-based system for medical records access and permissions management," IEEE Access, vol. 7, pp. 164595–164613, 2019, doi: 10.1109/ACCESS.2019.2952942.

[14] N. P. H. Kumar and S. Prabhudeva, "An Authorization Framework for Preserving Privacy of Big Medical Data via Blockchain in Cloud Server," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 3, pp. 140–150, 2022, doi: 10.14569/IJACSA.2022.0130319.

[15] T. Rupasinghe, F. Burstein, and C. Rudolph, "Blockchain based dynamic patient consent: A privacy-preserving data acquisition architecture for clinical data analytics," 40th Int. Conf. Inf. Syst. ICIS 2019, no. Bacchus 2017, pp. 1–9, 2019.

[16] D. El Majdoubi, H. El Bakkali, and S. Sadki, "SmartMedChain: A Blockchain-Based Privacy-Preserving Smart Healthcare Framework," J. Healthc. Eng., vol. 2021, 2021, doi: 10.1155/2021/4145512.

[17] H. M. Hussien, S. M. Yasin, N. I. Udzir, and M. I. H. Ninggal, "Blockchain-based access control scheme for secure shared personal health records over decentralised storage," Sensors, vol. 21, no. 7, pp. 1–36, 2021, doi: 10.3390/s21072462.

[18] D. Zhang, S. Wang, Y. Zhang, Q. Zhang, and Y. Zhang, "A Secure and Privacy-Preserving Medical Data Sharing via Consortium Blockchain," Secur. Commun. Networks, vol. 2022, 2022, doi: 10.1155/2022/2759787.

[19] R. H. Hylock and X. Zeng, "A blockchain framework for patient-centered health records and exchange (healthChain): Evaluation and proof-of-concept study," J. Med. Internet Res., vol. 21, no. 8, pp. 1–30, 2019, doi: 10.2196/13592.

[20] Y. Sharma and B. Balamurugan, "Preserving the Privacy of Electronic Health Records using Blockchain," Procedia Comput. Sci., vol. 173, no. 2019, pp. 171–180, 2020, doi: 10.1016/j.procs.2020.06.021.

[21] A. Ali et al., "Deep Learning Based Homomorphic Secure Search-Able Encryption for Keyword Search in Blockchain Healthcare System: A Novel Approach to Cryptography," Sensors, vol. 22, no. 2, 2022, doi: 10.3390/s22020528.

[22] Y. H. Park, Y. Kim, S. O. Lee, and K. Ko, "Secure outsourced blockchain-based medical data sharing system using proxy re-encryption," Appl. Sci., vol. 11, no. 20, 2021, doi: 10.3390/app11209422.

[23] S. Barbaria, M. C. Mont, E. Ghadafi, H. Mahjoubi Machraoui, and H. B. Rahmouni, "Leveraging Patient Information Sharing Using Blockchain-Based Distributed Networks," IEEE Access, vol. 10, pp. 106334–106351, 2022, doi: 10.1109/ACCESS.2022.3206046.


LaSalle
Universidad - PERÚ