UNIVERSIDAD NACIONAL DE SAN MARTIN

RCSI
https://revistas.unsm.edu.pe/index.php/rcsi
e-ISSN: 2709-992X

Artículo original
Original article
Jul-Dic, 2023

Midiendo la carga emocional: Análisis de las emociones presentes en contenido de tweets sobre COVID-19 en Lima

Measuring the emotional charge: Analysis of the emotions present in the content of tweets about COVID-19 in Lima

Holgado-Apaza, Luis Alberto[1]*
Ancco-Calloapaza, Coren Luhana[2]
Bedregal-Flores, Octavio[2]
Quispe-Layme, Marleny[1]
Miranda-Castillo, Ralph[1]

[1]Universidad Nacional Amazónica de Madre de Dios, Puerto Maldonado, Perú
[2]Universidad Nacional de San Agustín, Arequipa, Perú

Recibido: 28 Mar. 2023 | **Aceptado:** 13 Jun. 2023 | **Publicado:** 10 Jul. 2023

Autor de correspondencia*: lholgado@unamad.edu.pe

Como citar este artículo: Holgado-Apaza, L, A., Ancco-Calloapaza, C. L., Bedregal-Flores, O., Quispe-Layme, M. & Miranda-Castillo, R. (2023). Midiendo la carga emocional: Análisis de las emociones presentes en contenido de tweets sobre COVID-19 en Lima. *Revista Científica de Sistemas e Informática, 3*(2), e587. https://doi.org/10.51252/rcsi.v3i2.587

RESUMEN

Durante el estado de emergencia y las cuarentenas implementadas por los líderes mundiales, se ha observado un aumento significativo en la actividad de las personas en las redes sociales, como Twitter, donde comparten opiniones y noticias cargadas de emociones. En este estudio, presentamos una herramienta de visualización para el análisis de sentimientos en tweets relacionados con COVID-19 en la ciudad de Lima, Perú, durante el año 2020. Para ello, entrenamos un modelo BERT llamado BETO, diseñado específicamente para el procesamiento de lenguaje natural en español. Utilizamos el conjunto de datos SenWave, que comprende 11 emociones, para entrenar el modelo. Posteriormente, validamos el modelo utilizando un conjunto de datos compuesto por 33,770 tweets recolectados en la ciudad de Lima, Perú. El resultado de nuestro estudio es un panel de control interactivo que muestra el flujo de sentimientos expresados en los tweets analizados. Nuestros hallazgos revelan que las tres emociones más frecuentes durante el año 2020 fueron: humor, aburrimiento y optimismo. Además, identificamos las cinco palabras más populares utilizadas en los tweets: contagio, salud, distanciamiento, aislamiento y Martín Vizcarra, en referencia al expresidente del Perú.

Palabras clave: BERT; BETO; COVID-19; emociones; NLP

ABSTRACT

During the state of emergency and quarantines implemented by world leaders, there has been a significant increase in people's activity on social networks, such as Twitter, where they share opinions and emotionally charged news. In this study, we present a visualization tool for sentiment analysis in tweets related to COVID-19 in the city of Lima, Peru, during the year 2020. For this purpose, we train a BERT model called BETO, specifically designed for natural language processing in Spanish. We used the SenWave dataset, comprising 11 emotions, to train the model. Subsequently, we validate the model using a dataset composed of 33,770 tweets collected in the city of Lima, Peru. The result of our study is an interactive dashboard showing the flow of sentiments expressed in the analyzed tweets. Our findings reveal that the three most frequent emotions during 2020 were: humor, boredom and optimism. In addition, we identified the five most popular words used in the tweets: contagion, health, distancing, isolation and Martín Vizcarra, referring to the former president of Peru.

Keywords: BERT; BETO; emotions; covid-19; NLP

1. INTRODUCCIÓN

Comprender el sentimiento de la población con respecto a la COVID-19 es de suma importancia para las autoridades, ya que les permite evaluar cómo sus acciones impactan en las emociones de los ciudadanos. La pandemia de la COVID-19 que tuvo lugar en el año 2020, junto con las medidas de cuarentena implementadas por los líderes de distintos países, resultó en un incremento significativo de la actividad en las redes sociales, como Twitter (IPSOS, 2020).

Durante este período, las personas se vieron obligadas a aislarse, lo cual generó una mayor participación en dichas plataformas, incluyendo la publicación de opiniones y noticias con una carga emocional considerable (Mendoza Castillo, 1970). Por consiguiente, los tweets generados durante el año 2020 se convirtieron en una valiosa fuente de datos para analizar el comportamiento y el sentimiento público con relación a la pandemia.

En la mayoría de los estudios relacionados con el análisis de sentimientos, se han utilizado tres etiquetas para la clasificación de emociones: positivo, neutro y negativo (Yang et al., 2020). Sin embargo, durante la pandemia, el análisis de los sentimientos de las personas se vuelve mucho más complejo que simplemente toca categorizarlos en estas tres dimensiones.

En un texto, se pueden expresar una amplia gama de emociones, por lo que se requieren etiquetas más detalladas para comprender de manera más precisa los sentimientos y emociones de las personas durante la crisis de COVID-19 (Alturayeif & Luqman, 2021).

A la fecha, se han llevado a cabo diversos estudios que abordan el análisis de esta información. Por ejemplo, Topbas et al. (2021) proponen modelos de aprendizaje profundo para el análisis de sentimientos en tweets relacionados con COVID-19. Emplean técnicas como redes neuronales recurrentes (RNN) y el modelo de representaciones de codificador bidireccional (BERT) para clasificar los tweets en las tres categorías de positivo, neutro y negativo.

Por otro lado, Sitaula et al. (2021) proponen el uso de tres métodos de extracción de características diferentes, a saber, basado en texto rápido (ft), específico de dominio (ds) y agnóstico de dominio (da), para representar los tweets. Luego, utilizan redes neuronales convolucionales (CNN) para clasificar los tweets en las tres clases de positivo, neutral y negativo.

En un estudio realizado por Blanco & Lourenço (2022), se propone la clasificación del sentimiento en optimista y pesimista. Utilizan algoritmos de aprendizaje automático comunes, como Support Vector Machines, Random Forest y Naïve Bayes, en combinación con la técnica TF-IDF, y logran obtener buenos resultados.

Además, Imvimol & Chongstitvatana, 2021 consideran la clasificación de emociones en seis categorías: ira, disgusto, miedo, tristeza, alegría y sorpresa. Para ello, emplean diversas técnicas, como el perceptrón multicapa, RNN, LSTM, LSTM bidireccional y GRU.

En el estudio presentado por Alturayeif & Luqman (2021), se propone un modelo multietiqueta para clasificar 11 emociones en tweets de idioma árabe. Utilizan transformadores bidireccionales específicamente entrenados con el conjunto de datos SenWave (Yang et al., 2020). Además, los autores consideran la clasificación de emojis, previa conversión a su equivalente en texto.

Los estudios anteriores proporcionan una visión de los enfoques y técnicas empleados para el análisis de sentimientos y emociones en tweets relacionados con el COVID-19. Sin embargo, hay una falta de investigación específica enfocada en el contexto de Lima, Perú. Por lo tanto, en este estudio, nos proponemos abordar esta brecha y analizar las emociones presentes en los tweets sobre COVID-19 en la ciudad de Lima durante el año 2020.

2. MATERIALES Y MÉTODOS

La Figura 1 muestra el flujo de trabajo general del modelo, que abarca desde la extracción de tweets hasta la visualización de datos.

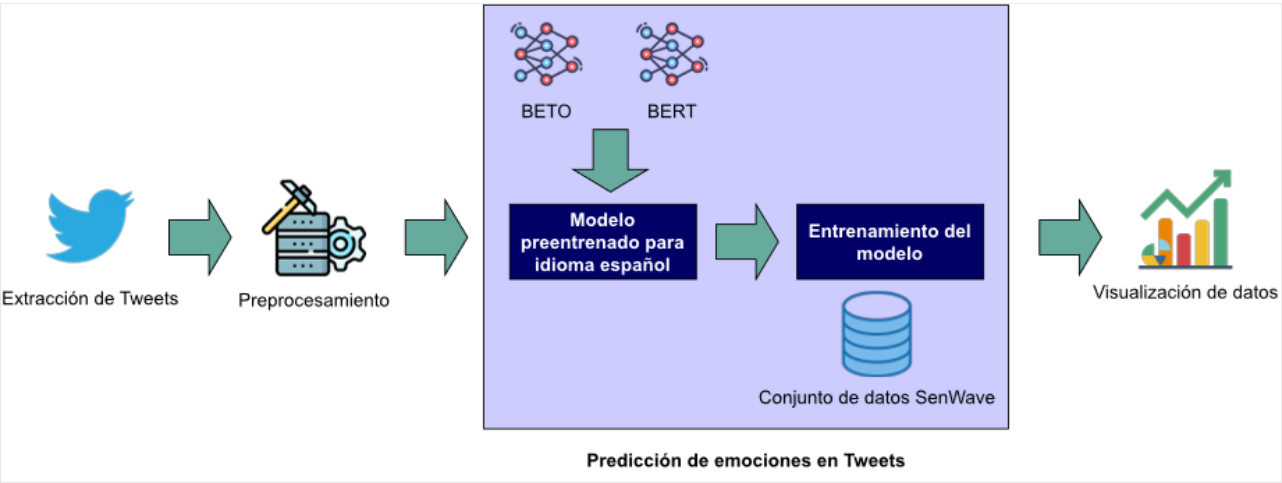

Figura 1. *Flujo de trabajo principal para el análisis de emociones presentes en contenidos de Tweets*

2.1. Extracción de Tweets

La Figura 2 muestra a detalle el proceso de extracción de Tweets.

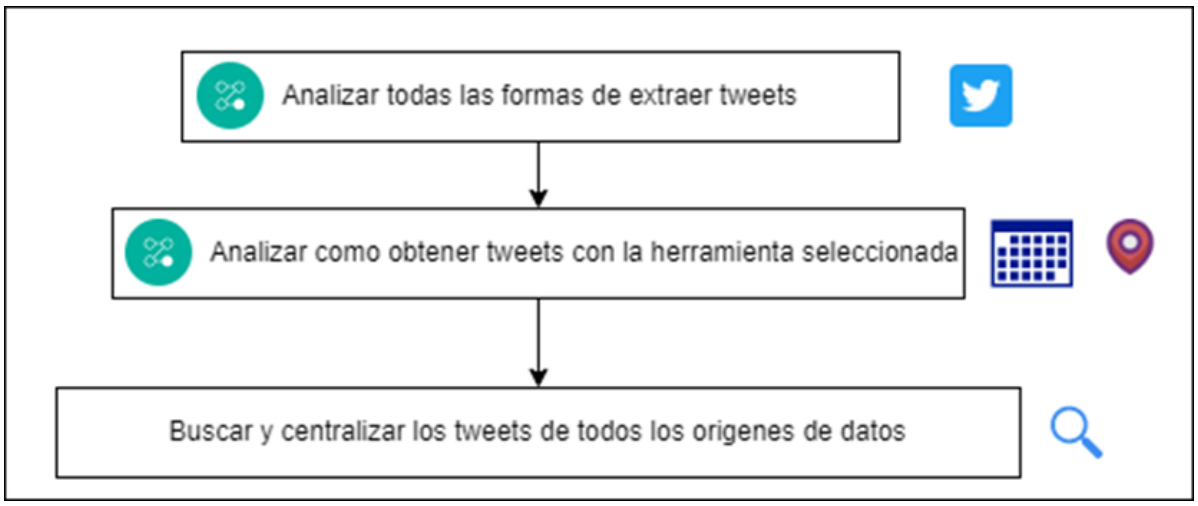

Figura 2. *Extracción de Tweets para el análisis de emociones*

Como primera tarea analizamos las formas más comunes para extraer Tweets, tras ello elegimos las técnicas de web scraping (Snscrape y Twint) que demostraron las mejores condiciones para obtener Tweets.

Después de seleccionar el método de extracción de datos, procedimos a analizar cómo obtener tweets utilizando la herramienta seleccionada, Twint. Para ello, establecimos una serie de términos relevantes como "Coronavirus", "Covid", "Cuarentena", "Contagio", "Epidemia", "Pandemia", "SARS", "Aislamiento" y "distanciamiento".

Configuramos la ubicación geográfica en la ciudad de Lima, con un radio de 200 kilómetros. Con el fin de recopilar una amplia variedad de datos, decidimos realizar búsquedas diarias con una ventana de tiempo que abarcara dos días previos y dos días posteriores. Es importante mencionar que estos ajustes fueron

aplicados con la finalidad de obtener datos específicos y relevantes para nuestro análisis sobre la situación del Coronavirus en la ciudad de Lima y sus alrededores.

Con relación a la librería snscrape, realizamos la búsqueda considerando los siguientes parámetros:

- Texto de búsqueda: “covid”

- near: "Lima"

- within: 100km

- since: 2019-12-01 until:2021-12-31

- Cantidad de registros: 50000

Luego de ello obtuvimos 22 968 tweets con la herramienta Snscrape y 22 000 tweets con Twint. Unimos ambos conjuntos de datos, considerando solo las columnas idTweet, fecha y tweet, obteniendo un total de 44 968 registros.

2.2. Preprocesamiento

Realizamos la limpieza de tweets, mediante el empleo de expresiones regulares para el reemplazo de ciertos caracteres como “xq” por su equivalente “por qué”, “d” por “de”, “sr” por “señor” y “q” por “que”. Adicional a ello se procede a eliminar los valores de localizador de recursos uniforme (URL), hashtag, arroba, caracteres especiales, caracteres individuales, rt, varios espacios en blanco y números.

Con relación a esta tarea Aygun et al., 2022; Caraballo Ayala et al., 2021; Wankhade & Rao, 2022 recomiendan que es necesario la eliminación de dichos datos para disminuir el ruido de la información textual extraída.

Luego de la limpieza, se procede con el proceso de tokenización. Utilizamos función word_tokenize () de la librería NLTK de Python, esta función considera una palabra de una oración como un token, estos tokens son almacenados en una estructura de datos de tipo lista. Una vez separado en tokens se procede a recorrer la lista para excluir todas las palabras que forman parte de la lista de Stopwords.

Finalmente procedemos a realizar la lematización que consiste en convertir una palabra en su forma base, luego de esta tarea se obtiene un total de 33,370 tweets para el proceso de predicción de sentimientos.

2.3. Predicción de emociones en Tweets

En esta fase utilizamos un modelo pre-entrenado de Representación de Codificador Bidireccional de Transformadores (BERT) para idioma español denominado BETO (Cañete et al., 2020). Entrenamos y validamos nuestro modelo con el conjunto de datos denominado Senwave que posee 11 categorías de emociones (Yang et al., 2020). Tras ello, utilizamos el modelo construido para etiquetar el conjunto preprocesado de tweets. La Figura 3, muestra el proceso de predicción de emociones en tweets.

En la etapa inicial del proceso de investigación, se llevó a cabo una exhaustiva búsqueda de artículos científicos utilizando fuentes de información académicas reconocidas. Priorizando bases de datos en línea de prestigio como Scielo, DOAJ, MIAR, Web of Science Group y Redalyc, poniendo énfasis en fuentes indexadas en Scopus.

Para optimizar y refinar nuestras búsquedas, utilizamos operadores booleanos tales como "AND", "OR", "NOT", " " y (). Estos operadores nos permitieron combinar y filtrar términos de búsqueda de manera precisa, obteniendo así un conjunto de resultados más relevantes y pertinentes a nuestra temática de investigación, además del uso de palabras claves como: Data Warehouse, implementación, soluciones DwH, DwH y la toma de decisiones, gracias a esto como resultado logramos identificar un total de 60 artículos.

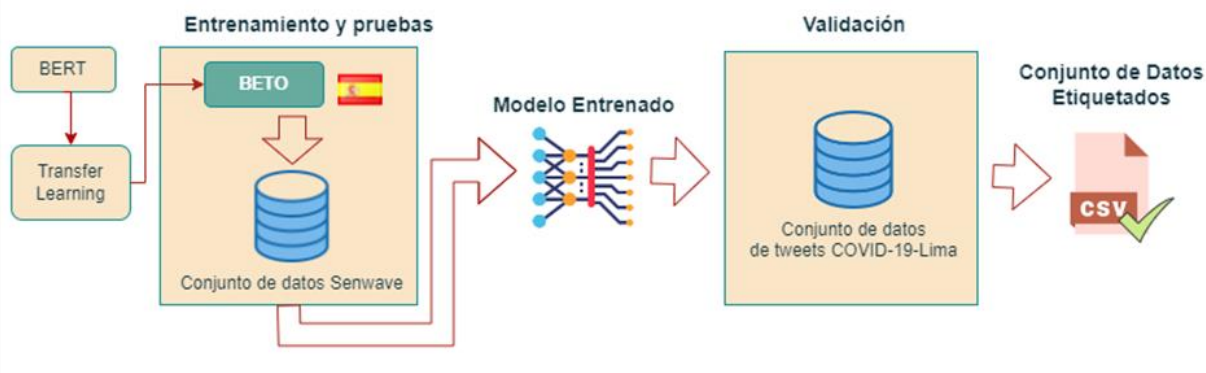

Figura 3. *Predicción de emociones y obtención del conjunto de datos de Tweets COVID-19-Lima etiquetados*

2.4. Visualización de datos

Realizamos esta tarea a partir del conjunto de datos de Tweets COVID-19-Lima etiquetados, para ello utilizamos la herramienta Power BI de Microsoft en su versión trial para la construcción de un tablero para visualizar el flujo del sentimiento público durante el año 2020, palabras más usadas por emoción, además de tweets según una palabra, emoción y/o mes.

3. RESULTADOS Y DISCUSIÓN

En la Figura 4, observamos la herramienta de visualización desarrollada en este estudio para presentar los resultados. Este gráfico ofrece una representación visual de los resultados del estudio, incluyendo el flujo de emociones a lo largo del tiempo, el ranking de emociones más frecuentes, una nube de palabras con términos clave y la visualización de los propios tweets. Estas visualizaciones ayudan a comprender mejor las emociones y temas predominantes en los tweets relacionados con COVID-19 en la ciudad de Lima durante el año 2020.

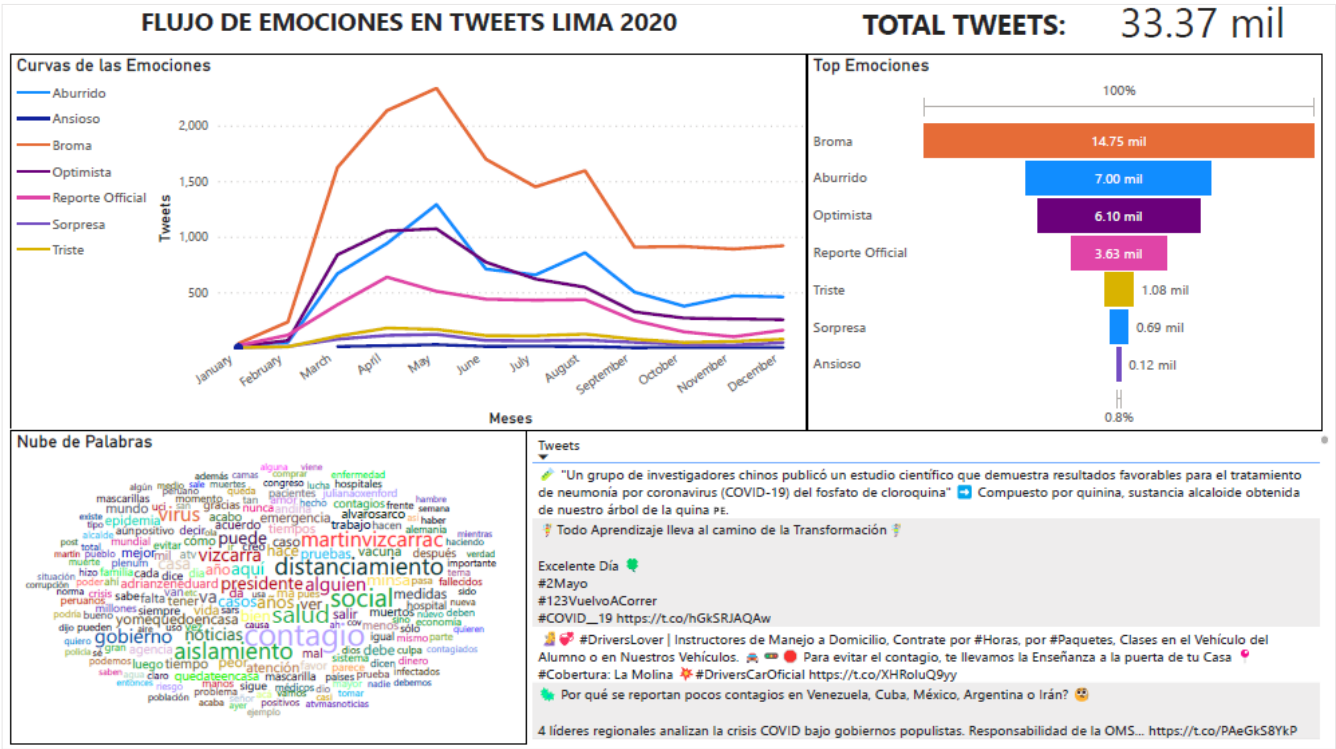

Figura 4. *Herramienta de visualización de Tweets*

La Figura 5 revela el ranking de las emociones más frecuentes durante la pandemia de COVID-19 en el año 2020. El análisis muestra una variedad de emociones presentes en los tweets, desde emociones humorísticas hasta emociones negativas como el aburrimiento, la tristeza y la ansiedad. Estos hallazgos proporcionan una visión más completa de las respuestas emocionales de los usuarios de Twitter frente a la pandemia.

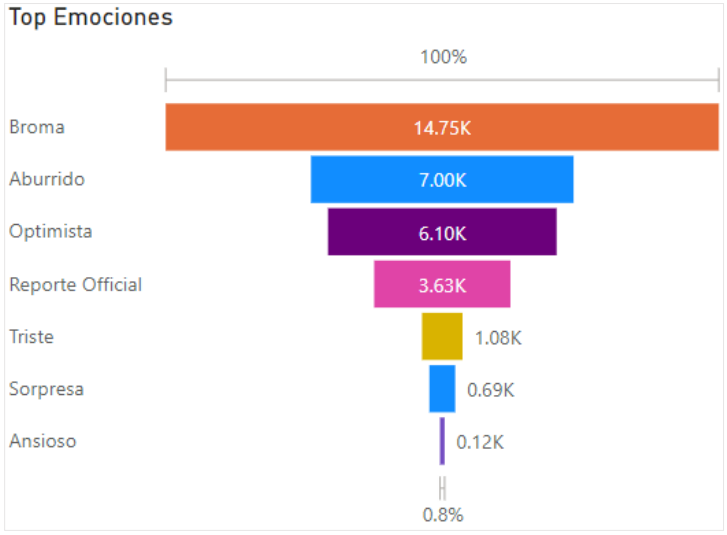

Figura 5. *Ranking de emociones más frecuentes*

En la Figura 6, se presenta la nube de palabras que refleja las palabras más populares encontradas en los tweets analizados. Esta nube de palabras revela las cinco palabras más prominentes, las cuales son: "contagio", "social", "salud", "distanciamiento" y "aislamiento".

Estos resultados resaltan las principales preocupaciones y temas discutidos en los tweets relacionados con la pandemia de la COVID-19 en la ciudad de Lima durante el período analizado. La Figura 6, destaca que la palabra "contagio" indica que esta fue una preocupación central para los usuarios de Twitter, reflejando su interés en comprender y abordar la propagación del virus.

Además, las palabras "social", "salud", "distanciamiento" y "aislamiento" sugieren la importancia dada a las medidas preventivas y de protección en la comunidad. Estas palabras reflejan la conciencia sobre la necesidad de mantener el distanciamiento social y el aislamiento para evitar la propagación del virus y proteger la salud tanto individual como colectiva.

Figura 6. *Nube de palabras de tweets durante el año 2020 en Lima*

La Figura 7 presenta el flujo de emociones encontradas en los tweets durante la pandemia del COVID-19 en la ciudad de Lima durante el año 2020. En esta gráfica, se destaca el sentimiento predominante de "broma". Es interesante observar que, de enero a abril, las personas mostraron un sentimiento de optimismo frente a la pandemia. Sin embargo, a partir de abril, la gráfica muestra un aumento en el sentimiento de aburrimiento. Esto puede ser atribuido a las cuarentenas obligatorias decretadas por el gobierno peruano, como las implementadas el 16 de marzo. A partir de ese momento, las cuarentenas se prolongaron, lo que puede explicar el aumento en el sentimiento de aburrimiento. No obstante, el sentimiento de optimismo se mantiene como el tercer sentimiento más prevalente a lo largo del año 2020.

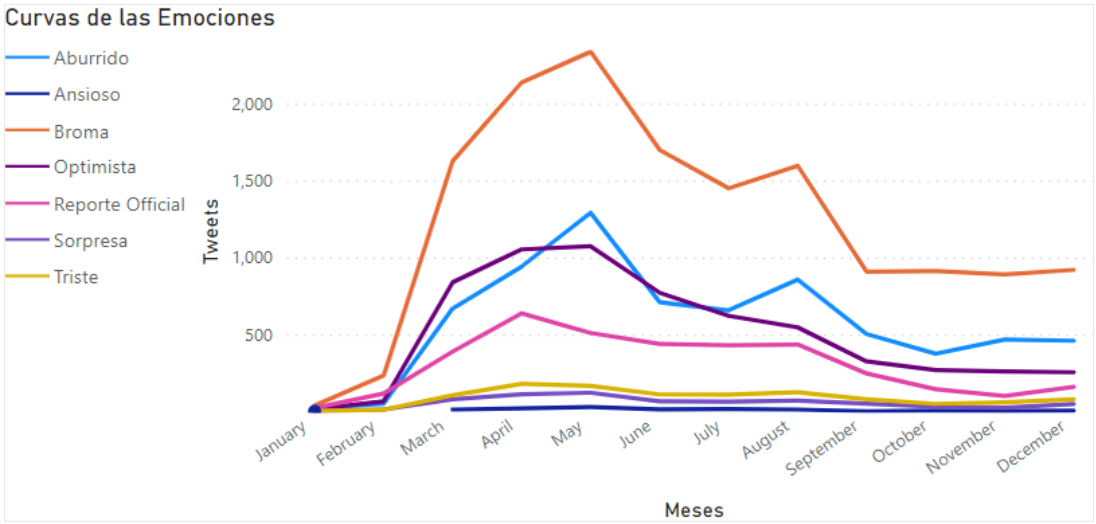

Figura 7. *Flujo de emociones durante la pandemia del COVID-19 en la ciudad de Lima durante el 2020*

El sentimiento de optimismo inicial puede reflejar la esperanza y la confianza en que la situación mejoraría. Sin embargo, el aumento en el sentimiento de aburrimiento después de la implementación de las cuarentenas prolongadas sugiere la fatiga y el desgaste emocional experimentado por la población debido a las restricciones y el distanciamiento social. Esta técnica de visualización también fue utilizada por Cañete et al., 2020; Garcia & Berton, 2021; Mohamed Ridhwan & Hargreaves, 2021; Yang et al., 2020.

Estos resultados resaltan la importancia de comprender las respuestas emocionales de las personas durante una crisis sanitaria y cómo estos sentimientos pueden evolucionar en diferentes momentos. Además, proporcionan información relevante para comprender el impacto psicológico de las medidas de contención y el desarrollo de estrategias de apoyo emocional adecuadas para la población afectada.

CONCLUSIONES

En este estudio, logramos el análisis de las emociones presentes en los tweets relacionados con COVID-19 en la ciudad de Lima durante el año 2020. Además, identificamos las palabras que están relacionadas con los sentimientos de la población de Lima. Para lograr este objetivo, se recopiló un total de 44,968 tweets utilizando la técnica de web scraping. Estos tweets pertenecen a la ciudad de Lima, Perú, durante el año 2020. El preprocesamiento de los datos involucró la limpieza de los tweets, lo cual consistió en eliminar las URLs, hashtags, menciones, caracteres especiales, caracteres individuales, retweets, múltiples espacios en blanco y números. A continuación, se realizó la tokenización utilizando la librería NLTK de Python. Como resultado de este proceso, se obtuvo un conjunto de datos que consta de 33,370 tweets etiquetados utilizando el modelo BETO, los cuales se utilizaron para la visualización de los resultados. Los resultados obtenidos revelaron que la palabra "contagio" mantuvo su popularidad a lo largo de todo el año 2020. Por otro lado, la palabra "salud" se hizo más prominente a partir de febrero, mientras que "distanciamiento" se volvió relevante desde marzo. La palabra "aislamiento" solo fue popular entre marzo y septiembre, y la

mención de "martinvizcarrac" estuvo presente desde febrero hasta septiembre. Estos hallazgos proporcionan una visión más clara de las preocupaciones y emociones de la población limeña durante el año 2020 en relación con la pandemia de COVID-19. Además, demuestran la utilidad de la metodología utilizada, que combina técnicas de procesamiento de lenguaje natural y análisis de datos para extraer información valiosa de los tweets relacionados con eventos de interés público. Se espera que estos resultados contribuyan a una mejor comprensión de las dinámicas emocionales y preocupaciones de la población en situaciones de crisis, lo que podría ser útil para la toma de decisiones en el ámbito de la salud pública y la comunicación de riesgos durante futuras emergencias sanitarias.

FINANCIAMIENTO

Ninguno

CONFLICTO DE INTERESES

No existe ningún tipo de conflicto de interés relacionado con la materia del trabajo.

CONTRIBUCIÓN DE LOS AUTORES

Conceptualización: Holgado-Apaza, L. A, Ancco-Calloapaza, C. L. y Vedregal-Flores, O.
Curación de datos: Holgado-Apaza, L. A.
Análisis formal: Ancco-Calloapaza, C. L.
Metodología: Holgado-Apaza, L. A, Ancco-Calloapaza, C. L.
Software: Holgado-Apaza, L. A, Ancco-Calloapaza, C. L. y Vedregal-Flores, O.
Visualización: Vedregal-Flores, O.
Redacción - borrador original: Holgado-Apaza, L. A.
Redacción - revisión y edición: Holgado-Apaza, L. A.

REFERENCIAS BIBLIOGRÁFICAS

Alturayeif, N., & Luqman, H. (2021). Fine-Grained Sentiment Analysis of Arabic COVID-19 Tweets Using BERT-Based Transformers and Dynamically Weighted Loss Function. *Applied Sciences*, *11*(22), 10694. https://doi.org/10.3390/app112210694

Aygun, I., Kaya, B., & Kaya, M. (2022). Aspect Based Twitter Sentiment Analysis on Vaccination and Vaccine Types in COVID-19 Pandemic With Deep Learning. *IEEE Journal of Biomedical and Health Informatics*, *26*(5), 2360–2369. https://doi.org/10.1109/JBHI.2021.3133103

Blanco, G., & Lourenço, A. (2022). Optimism and pessimism analysis using deep learning on COVID-19 related twitter conversations. *Information Processing & Management*, *59*(3), 102918. https://doi.org/10.1016/j.ipm.2022.102918

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., &, & Pérez, J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. *PML4DC at ICLR*. https://github.com/dccuchile/beto

Caraballo Ayala, N. E., Carreño Miranda, R., & Paternina Salgado, V. A. (2021). Análisis de sentimientos en Twitter: Opiniones en Colombia de los Juegos Olímpicos 2021. *Uniwersytet Śląski*. http://hdl.handle.net/10584/9874

Garcia, K., & Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing*, *101*, 107057. https://doi.org/10.1016/j.asoc.2020.107057

Imvimol, C., & Chongstitvatana, P. (2021). Sentiment analysis of messages on Twitter related to COVID-19 using deep learning approach. *2021 25th International Computer Science and Engineering Conference (ICSEC)*, 363–367. https://doi.org/10.1109/ICSEC53205.2021.9684587

IPSOS. (2020). *Uso de Redes Sociales entre peruanos conectados 2020*. Institut de Publique Sondage d'Opinion Secteur. https://www.ipsos.com/es-pe/uso-de-redes-sociales-entre-peruanos-conectados-2020

Mendoza Castillo, L. (2020). Lo que la pandemia nos enseñó sobre la educación a distancia. *Revista Latinoamericana de Estudios Educativos*, *50*(ESPECIAL), 343–352. https://doi.org/10.48102/rlee.2020.50.ESPECIAL.119

Mohamed Ridhwan, K., & Hargreaves, C. A. (2021). Leveraging Twitter data to understand public sentiment for the COVID-19 outbreak in Singapore. *International Journal of Information Management Data Insights*, *1*(2), 100021. https://doi.org/10.1016/j.jjimei.2021.100021

Sitaula, C., Basnet, A., Mainali, A., & Shahi, T. B. (2021). Deep Learning-Based Methods for Sentiment Analysis on Nepali COVID-19-Related Tweets. *Computational Intelligence and Neuroscience*, *2021*, 1–11. https://doi.org/10.1155/2021/2158184

Topbas, A., Jamil, A., Hameed, A. A., Ali, S. M., Bazai, S., & Shah, S. A. (2021). Sentiment Analysis for COVID-19 Tweets Using Recurrent Neural Network (RNN) and Bidirectional Encoder Representations (BERT) Models. *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, 1–6. https://doi.org/10.1109/ICECube53880.2021.9628315

Wankhade, M., & Rao, A. C. S. (2022). Opinion analysis and aspect understanding during covid-19 pandemic using BERT-Bi-LSTM ensemble method. *Scientific Reports*, *12*(1), 17095. https://doi.org/10.1038/s41598-022-21604-7

Yang, Q., Alamro, H., Albaradei, S., Salhi, A., Lv, X., Ma, C., Alshehri, M., Jaber, I., Tifratene, F., Wang, W., Gojobori, T., Duarte, C. M., Gao, X., & Zhang, X. (2020). *SenWave: Monitoring the Global Sentiments under the COVID-19 Pandemic*. http://arxiv.org/abs/2006.10842