ISSN 1870-4069

Análisis preliminar del sentimiento sobre la vacunación del COVID-19 en México

Luis Norberto Zúñiga-Morales[3], Arturo Zúñiga-López[1], Juan Villegas-Cortez[2], Carlos Avilés-Cruz[1] Felipe Morales-Torres[3]

[1] Universidad Autónoma Metropolitana, Departamento de Electrónica, Unidad Azcapotzalco, México

[2] Universidad Autónoma Metropolitana, Departamento de Sistemas, Unidad Azcapotzalco, México

[3] Universidad Iberoamericana, Instituto de Investigación Aplicada y Tecnología, México

{azl, juanvc, caviles}@azc.uam.mx, {lzun.morales, felipe.mor.torres}@gmail.com

Resumen. El análisis de información en las redes sociales hoy día es un tema de interés general para muchas disciplinas del conocimiento, esto porque se han convertido en un instrumento de comunicación de información masivo. Actualmente es común que las personas determinen su criterio por lo que ven como información mayoritaria en la Internet y específicamente en redes sociales. En este trabajo presentamos un análisis de sentimientos, basados en textos compartidos en la red social Twitter, los llamados tuits, mismos que previamente han sido clasificados en tres estados. Los tuits tienen una complejidad lexicográfica que representan un nuevo reto para el reconocimiento de patrones, y es así que nosotros presentamos la aplicación de la máquina de soporte vectorial sobre una gran cantidad de tuits referidos al tema de la pandemia del COVID-19 y la vacunación. El resultado alcanzado en clasificación en esta etapa preliminar no es alta por la poca cantidad de tuist categorizados, sin embargo, podemos considerar que la continuación del trabajo a futuro puede ser de apoyo para criterios de políticas sociales y de salud, además de entender desde una nueva perspectiva este tipo de patrones.

Palabras clave: COVID-19, reconocimiento de patrones, aprendizaje profundo, análisis de sentimientos, redes sociales

Luis Norberto Zúñiga-Morales, Arturo Zúñiga-López, Juan Villegas-Cortez, et al.

Preliminary Sentiment Analysis of COVID-19 Vaccination in Mexico

Abstract. The analysis of information in social networks today is a topic of general interest for many disciplines of knowledge, because they have become an instrument of mass information communication. Nowadays it is common for people to determine their criteria by what they see as majority information on the Internet and specifically in social networks. In this paper we present a sentiment analysis, based on texts shared on the social network Twitter, the called tweets, which have been previously classified into three states. The tweets have a lexicographic complexity that represent a new challenge for pattern recognition, and so we present the application of the support vector machine on a large number of tweets referring to the topic of the COVID-19 pandemic and vaccination.The result achieved in classification at this preliminary stage is not high due to the small amount of categorized tweets, however, we can consider that the continuation of the work in the future can be of support for social and health policy criteria, in addition to understanding from a new perspective this type of patterns.

Keywords: COVID-19, pattern recognition, deep learning, sentiment analysis, social networks.

1. Introducción

Hoy en día los servicios de redes sociales han cambiado la forma en que las personas expresan sus opiniones y puntos de vista [22] e.g., Twitter es una red social muy popular de mensajes cortos, con 140 caracteres como máximo en su primera etapa, y desde el año 2020 permite 280 caracteres como máximo, que pretende ser un reflejo de lo que está pasando en un momento dado, como el brote de coronavirus que ha supuesto un problema grave para la economía mundial y ha afectado a la mayoría de las naciones.

Lo anterior ha provocado restricciones de viaje, cierre de negocios no esenciales y procedimientos de cuarentena. A la luz de estas medidas, la mayoría de la gente ha recurrido a las redes sociales para expresar su opinión sobre todo lo que está sucediendo en el mundo. El impacto de las plataformas de redes sociales se está volviendo más notable que nunca. Los sitios de redes sociales se consideran el gran centro de datos global porque las personas usan sus aplicaciones e invierten mucho tiempo en estos medios de comunicación [1].

Los recientes desarrollos en el campo de los sistemas de información y las plataformas de intercambio de opiniones han impulsado la investigación para analizar las opiniones expresadas en estas redes sociales, que se presenta en la literatura como "análisis de sentimientos" [1]. El límite de 140 caracteres por tuit hace que los tuits sean concisos y fáciles de entender, al tiempo que brindan una idea de las opiniones y sentimientos de las personas, de ahí que hay varios estudios que pretenden usar Twitter para analizar la situación del COVID-19 a nivel mundial.

Las vacunas son sin duda uno de los mayores logros de la medicina moderna, y hay esperanzas de que puedan constituir una solución para detener la pandemia de COVID-19 en curso [10], sin embargo, en la última década la oposición a las vacunas ha encontrado un lugar en los medios digitales y sociales como medio principal de organización y difusión de información, además aunado a la creciente preocupación por los derechos humanos y el escepticismo hacia la vacuna y sus efectos pueden hacer que el proceso de vacunación se convierta en una tarea complicada [16].

Desde el comienzo de la pandemia, las noticias recientes han puesto de relieve el aumento de la desinformación en línea y la oposición a las vacunas [4]. Estudios recientes han demostrado que los mensajes relacionados con la vacunación son uno de los vectores más activos para la propagación de información errónea y desinformación sobre salud. Aunque son una pequeña fracción del público en general, los oponentes a las vacunas tienen una presencia enorme en línea y especialmente en Twitter.

A pesar de los recientes esfuerzos de Twitter para limitar la propagación de afirmaciones de salud falsas y engañosas, muchas cuentas de usuario oponentes a las vacunas permanecen activas en la plataforma. Aunque algunas cuentas tuitean casi exclusivamente sobre vacunas, muchas otras también discuten otros tipos de contenido, lo que permite identificar subgrupos en función de intereses compartidos como la política, la salud pública o la actualidad [12].

Esta investigación tiene como objetivo identificar los sentimientos sobre la vacunación del COVID-19 en México a partir de tuits. El Análisis de Sentimientos (AS) o minería de opinión, es una tarea de clasificación automática de textos que utiliza diversas herramientas de disciplinas como Procesamiento de Lenguaje Natural, Lingüística Computacional y Minería de Textos [7, 11]. En este trabajo de tipo ingenieril nos apegamos a entender por "sentimiento" la primera acepción del diccionario VOX de la Lengua Española[4], que lo define como "Estado de ánimo o disposición emocional hacia una cosa, un hecho o una persona".

Los tuits se clasifican en positivos, neutrales o negativos. Nos permitimos aclarar esto para delimitar nuestro alcance de estudio desde la inteligencia artificial, respetando a los profesionales de la salud mental, psicólogos y psiquiatras, que consideramos les podemos aportar para ellos dar un análisis apoyado en nuestras conclusiones.

En la sección 2 presentamos el estado del arte de nuestra investigación, la metodología propuesta la mostramos en la sección 3, los experimentos y el análisis de sus resultados se exponen en la sección 4 y, finalmente compartimos nuestras conclusiones en la sección 5.

2. Estado del arte

La gran mayoría de los estudios de investigación que cubren el análisis del sentimiento de los tuits se inclinan más hacia los algoritmos de aprendizaje automático [18]. Los investigadores a menudo utilizan una metodología exploratoria y descriptiva, así como los datos visuales y textuales, para obtener información valiosa basada en el método de clasificación de aprendizaje automático [22].

[4] Diccionario General de la Lengua Española Vox. Copyright © 2012, 2020 Larousse Editorial, S.L., under licence to Oxford University Press. All rights reserved.

Sethi et al. [19], hizo predicciones de los sentimientos de las personas en Twitter mediante la construcción de un modelo para explorar el sentimiento real de las personas sobre COVID-19. Hicieron una comparación entre cinco clasificadores, que son regresión logística, Naïve Bayes multinomial, árboles de decisión, bosque aleatorio, XGBoost y Maquinas de soporte vectorial (SVM).

Los resultados mostraron que SVM y los árboles de decisión superan al otro clasificador. Sin embargo, el clasificador SVM es estable y confiable en todas las pruebas. Además, la precisión máxima del modelo propuesto fue del 93 %, lo que indica numéricamente que el modelo tiene la capacidad de analizar la emoción de las personas dentro de los tuits "COVID-19".

Chakraborty et al. [6], analizaron tuits retuiteados con aprendizaje profundo, el análisis revela que si bien las personas tuiteaban principalmente de manera positiva sobre COVID-19, los usuarios de Internet estaban ocupados re-tuiteando tuits negativos y que no se podían encontrar términos útiles. La precisión alcanzó hasta el 81 % cuando se usan clasificadores de aprendizaje profundo, mientras que el 79 % cuando se usa el modelo formulado basado en una regla difusa para identificar los sentimientos de los tuits.

Abdulaziz et al. [1], analizaron en un conjunto de datos de tuits en inglés sobre COVID-19. La implementación se realizó utilizando LDA (Latent Dirichlet Allocation) para encontrar los temas más importantes relacionados con el Coronavirus. Se entrenó con el 80 % del conjunto de datos y se probó con el 20 %. Además, se utilizó el sentimientos de los tuits recopilados utilizando un enfoques basados en el léxico para clasificar los sentimientos de las personas.

Sontayasara et al. [22], analizaron sentimientos utilizando el algoritmo de máquina de soporte vectorial. Los resultados mostraron una precisión de clasificación del 75.83 % basada en tres clasificaciones de sentimiento: positivo y negativo. Por tanto, este estudio podría proporcionar una idea de las opiniones y sentimientos de los viajeros relacionados con el negocio del turismo.

Rustam et al. [17], realizaron un análisis de opinión de los tuits de COVID-19 utilizando un enfoque de aprendizaje automático supervisado. Se utilizó la arquitectura Long Short-Term Memory (LSTM) del modelo de aprendizaje profundo, para la obtención de información. Singh et al. [20], estudian las opiniones de las personas para comprender su estado mental, para lo cual realizan un análisis de sentimientos utilizando el modelo BERT en los tuits.

Utilizan dos conjuntos de datos; un conjunto de datos que se recopila mediante tuits hechos por personas de todo el mundo, y el otro conjunto de datos que contiene los tuits hechos por personas de la India. Los resultados experimentales muestran que la precisión de la validación su modelo es de aproximadamente el 94 %.

3. Metodología

Nuestro texto de trabajo comprende diversas publicaciones de Twitter cuyo tema se encuentra relacionado con el COVID-19, para las cuales se busca determinar su sentimiento para analizar su comportamiento con el tiempo.

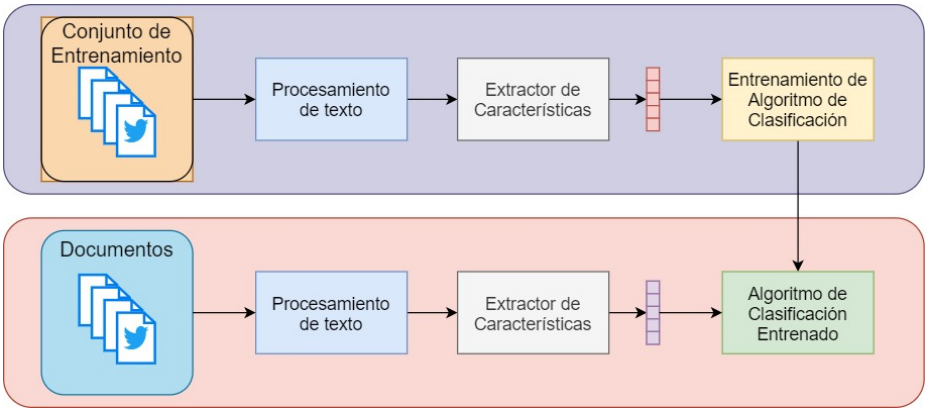


Fig. 1. Diagrama general de la metodología propuesta. El bloque superior representa el módulo de entrenamiento, mientras que el inferior representa el módulo de clasificación.

Para lograr esta clasificación, se propone un modelo compuesto de dos fases: 1) Fase de entrenamiento, y 2) Fase de clasificación. La Figura (1) muestra la idea general del modelo propuesto. En la fase de entrenamiento se busca entrenar un algoritmo de aprendizaje supervisado con información previamente anotada.

Durante la fase de clasificación, se utiliza el clasificador obtenido en el punto anterior para determinar la clase de cada documento facilitando su posterior análisis.

3.1. Construcción del conjunto de datos

La construcción del conjunto de datos consta de dos fases: la recopilación de información y el procesamiento de datos.

Recopilación de datos. Para la recopilación de datos se eligió Twitter como fuente de origen. Específicamente, se utilizó la API v2[5] de Twitter, con una cuenta con acceso a la modalidad de investigador académico.

El conjunto de datos incluye únicamente mensajes publicados entre el 1° de enero de 2020 y el 31 de marzo de 2021, relacionados con COVID-19, en idioma español y cuyo origen geográfico fuese México. Además, dentro de la consulta se especificó evitar publicaciones marcadas como retuit para evitar la repetición de publicaciones.

Para cada tuit se extrajeron los siguientes campos: texto, identificador del tuit, fecha de publicación y métricas públicas, la cual incluye conteos de retuits, respuestas, *likes* y *quotes*. La Figura 2 muestra la distribución de la información del conjunto de datos en un periodo mensual.

Anotación de datos. Para el modelo propuesto se consideran únicamente dos clases para los datos, "-1" y "1". La etiqueta -1 se utiliza para denotar aquellos mensajes que expresen un sentimiento negativo hacia el tema del COVID-19.

[5] Twitter API v2, URL: https://developer.twitter.com/en/docs/twitter-api/early-access

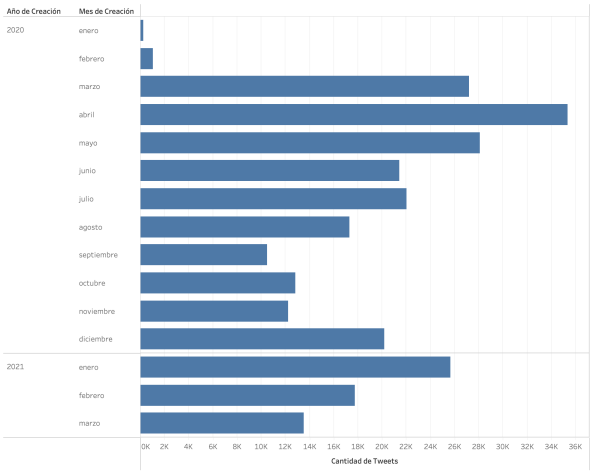


Fig. 2. Distribución temporal del año 2020 y parte del 2021, de los tuits recopiladas en el conjunto de datos.

Por otro lado, la etiqueta 1 se usa para identificar mensajes con una inclinación positiva y ajena hacia el COVID-19. Para su anotación, se recurrió a tres expertos en el campo de salud pública a los que se les facilitó una guía con las instrucciones de anotación. Al final, solamente se conservaron los tuits donde los tres expertos anotaron el mismo valor.

Procesamiento de datos. El preprocesamiento de datos se encarga de normalizar el texto para su posterior alimentación a los algoritmos de clasificación. El primer paso es cambiar el texto de cada tuit a letra minúscula. En un tuit se presentan diversos elementos que sirven para etiquetar diversos elementos de difusión, tales como el hashtag (#), cashtag ($) o menciones de usuario (@).

Estas etiquetas, aunque útiles para identificar información relevante, no suelen aportar información para el clasificador al no ser propios de una clase [13]. En nuestro caso, se eliminan cashtags debido a que no se habla de un tema financiero predominante en la investigación, y se eliminan las menciones de usuario por respeto a su privacidad.

Además, se elimina cualquier enlace a otro sitio presente en las publicaciones. Posteriormente, el texto filtrado se somete a un proceso de tokenización y lematización para su análisis previo, pero no para la extracción de características.

3.2. Análisis de sentimientos

Para realizar el AS sobre el conjunto de datos que se construyó, se utiliza la siguiente estrategia. Primero, se extraen características del texto de cada tuit por medio de BERT (Bidirectional Encoder Representations from Transformers). Después, las características extraídas se alimentan a un algoritmo de aprendizaje supervisado para su entrenamiento.

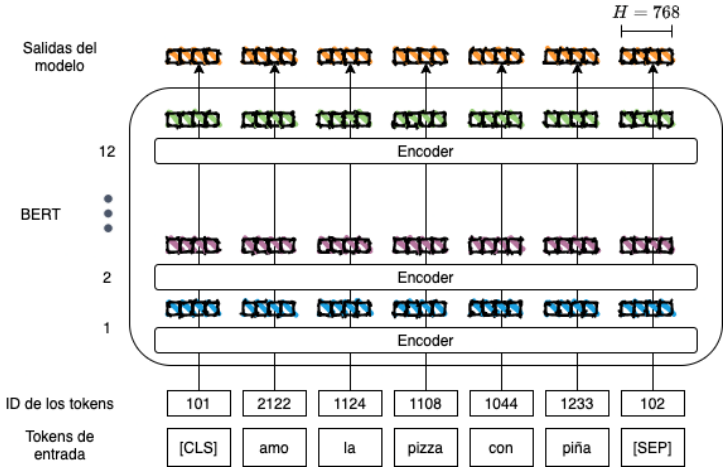


Fig. 3. Esquema general del modelo de lenguaje BERT en su modalidad base, la cual tiene 12 capas de encoders, 12 cabezales de auto atención y el tamaño del vector oculto es 768. Nótese los tokens adicionales [CLS] y [SEP] que se añaden al texto.

Extracción de características. BERT [8] es un modelo de representación de lenguaje basado en la idea de pre-entrenar modelos de lenguaje. La novedad de BERT es su entrenamiento bidireccional, ya que considera el contexto de las palabras de derecha a izquierda y viceversa, al mismo tiempo.

Para cumplir sus objetivos, BERT emplea Transformers [24], específicamente la arquitectura de sus codificadores. Inicialmente, BERT fue entrenado para trabajar con dos idiomas, chino e inglés, pero poco a poco se generaron modelos que abarcaron distintos idiomas.

Para el modelo del artículo se utiliza BETO [5], un modelo de BERT pre-entrenado con un gran corpus en español. El modelo consta de 12 capas de auto atención cada uno con 16 cabezales de atención, donde el tamaño vector oculto es 1024.

Algoritmo de clasificación. Para clasificar cada tuit según su polaridad, se utiliza un algoritmo de aprendizaje supervisado. En particular, se elige a la Máquina de Vectores de Soporte (MVS) [23], un método de clasificación que mapea datos de un conjunto de entrenamiento a diferentes espacios para construir un hiperplano que permita separar los miembros de cada clase.

La MVS ha demostrado ser un algoritmo de aprendizaje fuerte para la clasificación de texto [14] y puede considerarse como un método para establecer un marco de referencia al comparar distintos métodos de clasificación debido a su desempeño en tareas para clasificar la polaridad del sentimiento en tuits [9].

Validación del clasificador. La matriz de confusión [21] permite visualizar el desempeño de un algoritmo de clasificación. Cada columna indica la clase que que el clasificador predice y cada fila es la clase real a la que pertenece. La Tabla 1 muestra la matriz de confusión para el caso de clasificación binaria.

Tabla 1. Matriz de confusión para el caso de clasificación binaria, donde se elige una clase para que sea considerada la positiva y la otra la negativa.

		Valor Predicho	
		Clase 1	**Clase 2**
Valor	Clase 1	Positivo Verdadero(PV)	Falso Negativo(FN)
Real	Clase 2	Falso Positivo(FP)	Negativo Verdadero(NV)

Para evaluar el desempeño de los clasificadores se emplean cuatro medidas: precisión (precision), exhaustividad (recall), exactitud (accuracy) y la medida $F1$. En el caso de clasificación binaria se definen como:

– Exactitud: la medida más intuitiva, la razón de las instancias clasificadas correctamente y el total de los elementos clasificados:

$$\text{Exactitud} = \frac{PV + NV}{PV + NV + FP + FN}. \tag{1}$$

– Precisión: examina la razón de instancias clasificadas positivamente correctas:

$$\text{Precisión} = \frac{PV}{PV + FP}. \tag{2}$$

– Exhaustividad: efectividad del clasificador para identificar etiquetas positivas:

$$\text{Exhaustividad} = \frac{PV}{PV + FN}. \tag{3}$$

– Medida F1: es un promedio ponderado de la precisión y la exhaustividad:

$$\text{F1} = \frac{2 \cdot \text{Exhaustividad} \cdot \text{Precisión}}{\text{Exhaustividad} + \text{Precisión}}. \tag{4}$$

4. Experimentos y resultados

En esta sección se muestran los detalles de la implementación expuesta en la sección anterior, así como los resultados obtenidos y su análisis.

Para la implementación del modelo se utilizó un equipo de cómputo con sistema operativo MS-Windows ver 10, 64 bits, procesador Intel Core i7-7700HQ 2.80GHz, 16 GB RAM y una GPU GeForce GTX 1080. El modelo se programó en su totalidad con Python 3.8.

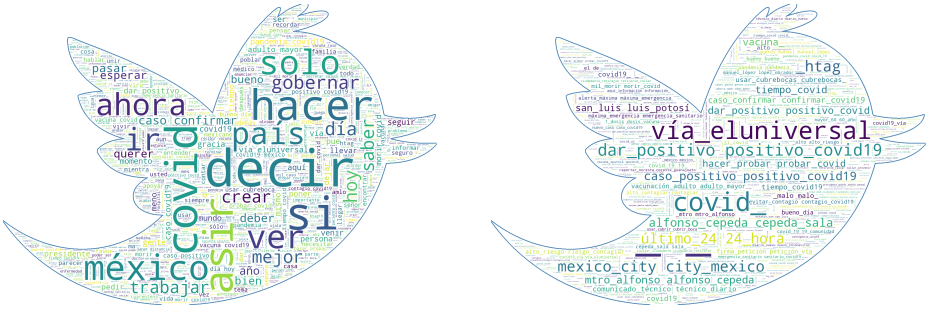
En la Tabla 2 se muestra un pequeño resumen del conjunto de datos.

A continuación, se presenta un análisis preliminar del conjunto de datos, el cual permite obtener una idea general del contenido del mismo. En particular, se realiza un análisis visual por medio de una diagrama de nubes de palabras y se explora los diferentes tópicos que conforman el conjunto de datos por medio de LDA.

En la Figura (4) se puede apreciar el diagrama de nube de palabras que se realiza al procesar cada tuit en el conjunto de datos.

Tabla 2. Resumen del conjunto de datos COVID-19 MX.

Periodo de Recolección	01/01/2020 - 31/03/2021
Cantidad de tuits recolectados	265,448
Query	(vacuna OR vacunación OR vacunar OR covid OR covid19 OR covid-19) lang:es, place_country:MX, -is:retweet
Campos solicitados	author_id, text, retweet_count, reply_count, like_count, quote_count, id, created_at



(a) (b)

Fig. 4. Diagrama de nube de palabras para (a) palabras lematizadas y (b) bigramas de palabras lematizadas que consideran todos los tuits del conjunto de datos COVID-19 MX. Entre más grande sea la palabra en la imagen, mayor es su frecuencia entre los tuits.

Cabe resaltar que se lematiza cada palabra, lo que ayuda a reducir la dimensión del conjunto conformado por las palabras únicas de los tuits. Como es de esperar, lo relacionado con COVID-19 resulta ser el tema principal de los documentos.

Sin embargo, también es posible apreciar aspectos relacionados como la política, la situación laboral y económica generada por la crisis, y la presencia de medios informativos digitales, entre otros.

El modelado de temas [2] es un tipo de modelo estadístico que permite inferir distintos temas que ocurren en una colección de documentos. La idea principal es que, entre más se relacionan ciertas palabras, se espera que aparezcan de forma conjunta en varios documentos de la colección.

El LDA [3] representa cada documento como una mezcla de temas por medio de palabras y su probabilidad. En este trabajo, se utiliza la implementación de LDA de Gensim[6], utilizando Term Frequency Inverse Document Frequency (TFIDF).

La Tabla (3) muestra los resultados obtenidos en el modelado de temas. Aunque el algoritmo permite modelar los documentos por medio de ciertas palabras, es responsabilidad de los autores interpretar los resultados.

[6] URL: https://radimrehurek.com/gensim/

Tabla 3. Resultado del modelado de temas usando TFIDF.

Tema	Composición
Vacunación y Vuelta a la Normalidad	Palabra: 0.012*covid + 0.008*si + 0.008*covid19 + 0.007*ir + 0.006*hacer + 0.005*casa + 0.005*poder + 0.004*vacuna + 0.004*ver + 0.004*dar
Casos confirmados y muertes	Palabra: 0.020*caso + 0.013*confirmar + 0.012*covid19 + 0.008*méxico + 0.007*2020 + 0.007*nuevo + 0.007*coronavirus + 0.006*defunción + 0.006*reportar + 0.006*mil
Vacunación	Palabra: 0.008*covid19 + 0.007*covid + 0.006*si + 0.006*decir + 0.006*hacer + 0.005*ir + 0.005*vacuna + 0.004*dar + 0.004*gobernar + 0.004*19
Consejos para la pandemia	Palabra: 0.008*covid19 + 0.008*mexico + 0.007*tiempo + 0.006*covid + 0.004*hacer + 0.004*día + 0.004*poder + 0.004*bueno + 0.004*probar + 0.004*si
Personal e instituciones de Salud	Palabra: 0.008*covid19 + 0.008*medida + 0.006*salud + 0.005*contingencia + 0.004*prevención + 0.004*sanitario + 0.004*hospital + 0.003*evitar + 0.003*pandemia + 0.003*médico

Tabla 4. Estadísticas del entrenamiento clasificador.

Precisión	Exhaustividad	Exactitud	F1
75.83 %	75.34 %	76.12 %	75.83 %

4.1. Extracción de características

Para la extracción de características se utiliza el modelo pre-entrenado BETO y su implementación en Huggingface[7] y Pytorch[8]. Para implementar BERT (y BETO), se utiliza un tipo especial de tokenización [8], la cual se encuentra incluida en el modelo, por lo que no se realiza el proceso de tokenización tradicional para extraer *word-embeddings*.

En cuanto al vector de características, Devlin et al. [8] ofrecen diversas opciones al momento de considerar los *word-embeddings* para las palabras. Sin embargo, para la tarea de clasificación, utilizan el último vector oculto del token especial [CLS], como se ve en la Figura (3). Este último es el vector de características que se alimenta al algoritmo de clasificación (MVS) para determinar la clase a la que pertenece.

4.2. Algoritmo de clasificación

Para la implementación de la MVS se utilizó la librería Scikit-learn [15], una librería de aprendizaje automático para Python. Utilizando los vectores de características extraídos con BETO, se utiliza una SVM con kernel de función de base radial utilizando como parámetros $\gamma = 2{,}3 \times 10^{-4}$ y $c = 1$.

[7] URL: https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

[8] URL: https://pytorch.org/

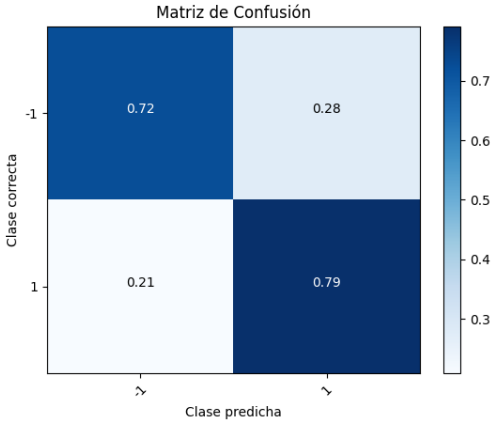


Fig. 5. Matriz de confusión para el clasificador entrenado.

Además, se realiza validación cruzada con 10 pliegues en una proporción $80 - 20$. En total, se usaron $2,113$ datos anotados previamente para el entrenamiento del clasificador.

La Figura (5) muestra la matriz de confusión del modelo entrenado, en la Figura (6) se observa la distribución de tuits positivos y negativos en un periodo mensual, y la Tabla (4) muestra las estadísticas del clasificador construido.

5. Conclusiones

El análisis en la redes sociales hoy en día es un tema de interés general para muchas disciplinas del conocimiento, esto porque se han convertido en un instrumento de comunicación de información masiva. Actualmente es común que las personas determinen su criterio por lo que ven como información mayoritaria en la Internet y específicamente en redes sociales.

Es este trabajo presentamos un análisis preliminar basados en textos compartidos en la red social Twitter, los llamados tuits, mismos que previamente han sido clasificados en dos estados.

Los tuits tienen una complejidad lexicográfica que representa un nuevo reto para el reconocimiento de patrones, y es así que nosotros presentamos la aplicación de una maquina de soporte vectorial sobre una cantidad tuits referidos al tema de la pandemia del COVID-19 y la vacunación para México, pero consideramos que se puede aplicar este estudio para otros países.

El resultado alcanzado en esta etapa preliminar no es alta si sólo contemplamos la poca cantidad de tuits categorizados (2,113), pero estamos por encima de otros trabajos que presentan metodologías numéricas más profundas, por todo esto podemos considerar que la continuación del trabajo a futuro puede ser de apoyo para criterios de políticas sociales y de salud, además de entender desde una nueva perspectiva este tipo de patrones.

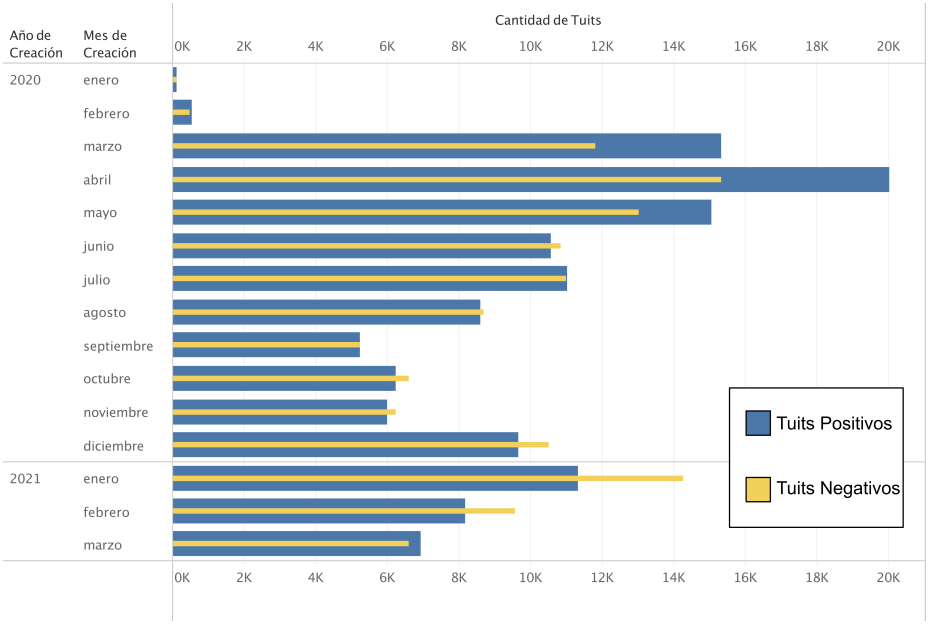


Fig. 6. Distribución temporal de tuits positivos (azul) y negativos (amarillo) del conjunto de datos completo. El etiquetado masivo se realiza mediante el clasificador entrenado (MVS) previamente.

La información en Twitter se genera de forma dinámica y creciente dependiendo los temas de tendencia en una región geográfica determinada, y esto nos ayuda a orientar nuestros esfuerzos hacia un análisis rápido con ventanas de tiempo-captura de la información, que pueda proporcionar a los especialistas de estudios sociales de los temas de interés, una herramienta en tiempo real de los hashtags de interés. Para trabajo a futuro se ha pensado en tener más datos etiquetados y probar con otras técnica de aprendizaje para compararlas y ver quien ofrece mejores resultados.

Referencias

1. Abdulaziz, M., Alsolamy, M., Alotaibi, A., Alabbas, A.: Topic based sentiment analysis for covid-19 tweets. International Journal of Advanced Computer Science and Applications, vol. 12, no. 1, pp. 626–636 (2021) doi: 10.14569/IJACSA.2021.0120172
2. Blei, D. M.: Probabilistic topic models. Communications of the ACM, vol. 55, no. 4, pp. 77–84 (2010) doi: 10.1145/2133806.2133826
3. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent dirichlet allocation. Journal of Machine Learning Research, vol. 3, pp. 993–1022 (2003)
4. Bonnevie, E., Gallegos-Jeffrey, A., Goldbarg, J., Byrd, B., Smyser, J.: Quantifying the rise of vaccine opposition on twitter during the COVID-19 pandemic. Journal of Communication in Healthcare, vol. 14, no. 1, pp. 12–19 (2021) doi: 10.1080/17538068.2020.1858222
5. Canete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., Pérez, J.: Spanish pre-trained BERT model and evaluation data. In: Practical Machine Learning for Developing Countries (2020)

6. Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., Hassanien, A. E.: Sentiment analysis of COVID-19 tweets by deep learning classifiers - a study to show how popularity is affecting accuracy in social media. Applied Soft Computing, vol. 97 (2020) doi: 10.1016/j.asoc.2020.106754
7. Daily, S. B., James, M. T., Cherry, D., Porter, J., Darnell, S. S., Isaac, J., Roy, T.: Affective computing: Historical foundations, current applications, and future trends. In: Emotions and Affect in Human Factors and Human-Computer Interaction, pp. 213–231 (2017)
8. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding, (2018) doi: 10.48550/arXiv.1810.04805
9. Elbagir, S., Yang, J.: Sentiment analysis of twitter data using machine learning techniques and scikit-learn. In: Proceedings of the International Conference on Algorithms, Computing and Artificial Intelligence, pp. 1–5, no. 57 (2018) doi: 10.1145/3302425.3302492
10. Germani, F., Biller-Andorno, N.: The anti-vaccination infodemic on social media: A behavioral analysis. Plos One, vol. 16, no. 3, pp. e0247642 (2021) doi: 10.1371/journal.pone.0247642
11. Guo, S., Zhang, G.: Using machine learning for analyzing sentiment orientations toward eight countries. Sage Open, vol. 10, no. 3, pp. 1–15 (2020) doi: 0.1177/2158244020951268
12. Jamison, A. M., Broniatowski, D. A., Dredze, M., Sangraula, A., Smith, M. C., Quinn, S. C.: Not just conspiracy theories: Vaccine opponents and proponents add to the covid-19 'infodemic' on twitter. Harvard Kennedy School Misinformation Review, vol. 1, no. 3 (2020) doi: 10.37016/mr-2020-38
13. Oliveira, N., Cortez, P., Areal, N.: Stock market sentiment lexicon acquisition using microblogging data and statistical measures. Decision Support Systems, vol. 85, pp. 62–73 (2016) doi: 10.1016/j.dss.2016.02.013
14. Patil, G., Galande, V., Kekan, V., Dange, K.: Sentiment analysis using support vector machine. International Journal of Innovative Research in Computerand Communication Engineering, vol. 2, no. 1, pp. 2607–2612 (2014)
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research, vol. 12, pp. 2825–2830 (2011)
16. Praveen, S., Ittamalla, R., Deepak, G.: Analyzing the attitude of indian citizens towards COVID-19 vaccine–a text analytics study. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 15, no. 2, pp. 595–599 (2021) doi: 10.1016/j.dsx.2021.02.031
17. Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., Choi, G. S.: A performance comparison of supervisedmachine learning models for covid-19 tweets sentiment analysis. PLoS ONE, vol. 16, no. 2, pp. 1–23 (2021) doi: 10.1371/journal.pone.0245909
18. Samuel, J., Ali, G. G. M. N., Rahman, M. M., Esawi, E., Samuel, Y.: COVID-19 public sentiment insights and machine learning for tweets classification. Information, vol. 11, no. 6, pp. 314 (2020) doi: 10.3390/info11060314
19. Sethi, M., Pandey, S., Trar, P., Soni, P.: Sentiment identification in covid-19 specific tweets. In: International Conference on Electronics and Sustainable Communication Systems, pp. 509–516 (2020) doi: 10.1109/ICESC48915.2020.9155674
20. Singh, M., Jakhar, A. K., Pandey, S.: Sentiment analysis on the impact of coronavirus in social life using the BERT model. Social Network Analysis and Mining, vol. 11, no. 1 (2021) doi: 10.1007/s13278-021-00737-z
21. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management, vol. 45, no. 4, pp. 427–437 (2009) doi: 10.1016/j.ipm.2009.03.002

22. Sontayasara, T., Jariyapongpaiboon, S., Promjun, A., Seelpipat, N., Saengtabtim, K., Tang, J., Leelawat, N.: Twitter sentiment analysis of bangkok tourism during covid-19 pandemic using support vector machine algorithm. Journal of Disaster Research, vol. 16, no. 1, pp. 24–30 (2021) doi: 10.20965/jdr.2021.p0024
23. Vapnik, V., Cortes, C.: Support-vector networks. Machine Learning, vol. 20, no. 3, pp. 273–297 (1995)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems, vol. 30 (2017)