Ingenius. Revista de Ciencia y Tecnología
ISSN: 1390-650X
ISSN: 1390-860X
revistaingenius@ups.edu.ec
Universidad Politécnica Salesiana
Ecuador

Análisis de sentimiento de los datos de twitter de COVID-19 utilizando modelos de aprendizaje profundo y aprendizaje máquina

Darad, Simran; Krishnan, Sridhar

Análisis de sentimiento de los datos de twitter de COVID-19 utilizando modelos de aprendizaje profundo y aprendizaje máquina
Ingenius. Revista de Ciencia y Tecnología, núm. 29, 2023
Universidad Politécnica Salesiana, Ecuador
Disponible en: https://www.redalyc.org/articulo.oa?id=505573889011
DOI: https://doi.org/10.17163/ings.n29.2023.10
2023.Universidad Politécnica Salesiana

Análisis de sentimiento de los datos de twitter de COVID-19 utilizando modelos de aprendizaje profundo y aprendizaje máquina

Sentimental analysis of COVID-19 twitter data using deep learning and machine learning models

Simran Darad
Toronto Metropolitan University, Canadá
sdarad@ryerson.ca
https://orcid.org/0000-0003-4629-3980

Sridhar Krishnan
Toronto Metropolitan University, Canadá
https://orcid.org/0000-0002-4659-564X

DOI: https://doi.org/10.17163/ings.n29.2023.10
Redalyc: https://www.redalyc.org/articulo.oa?id=505573889011

Recepción: 15 Octubre 2021
Recibido del documento revisado: 01 Diciembre 2022
Aprobación: 16 Diciembre 2022
Publicación: 01 Enero 2023

Resumen:

En este artículo, aplicamos técnicas de aprendizaje automático para predecir el sentimiento de las personas que usan las redes sociales como Twitter durante el pico de COVID-19 en abril de 2021. Los datos contienen tweets recopilados en las fechas entre el 16 de abril de 2021 y el 26 de abril de 2021, donde el texto de los tweets se ha etiquetado mediante la formación de los modelos con un conjunto de datos ya etiquetado de tweets de virus de corona como positivo, negativo y neutro. El análisis del sentimiento se llevó a cabo mediante un modelo de aprendizaje profundo conocido como Representaciones de Codificadores Bidireccionales de Transformers (BERT) y varios modelos de aprendizaje automático para el análisis de texto y el rendimiento, que luego se compararon entre sí. Los modelos ML utilizados son Bayes ingenuas, regresión logística, bosque aleatorio, máquinas vectoriales de soporte, descenso de gradiente estocástico y aumento de gradiente extremo. La precisión de cada sentimiento se calculó por separado. La precisión de clasificación de todos los modelos de ML producidos fue de 66.4 %, 77.7 %, 74.5 %, 74.7 %, 78.6 % y 75.5 %, respectivamente y el modelo BERT produjo 84.2%. Cada modelo clasificado de sentimiento tiene una precisión de alrededor o superior al 75 %, que es un valor bastante significativo en los algoritmos de minería de texto. Vemos que la mayoría de las personas que tuitean están adoptando un enfoque positivo y neutral.

Palabras clave: COVID-19, coronavirus, twitter, tweets, análisis de los sentimientos, tweepy, clasificación de texto.

Abstract:

The novel coronavirus disease (COVID-19) is an ongoing pandemic with large global attention. However, spreading fake news on social media sites like Twitter is creating unnecessary anxiety and panic among people towards this disease. In this paper, we applied machine learning (ML) techniques to predict the sentiment of the people using social media such as Twitter during the COVID-19 peak in April 2021. The data contains tweets collected on the dates between 16 April 2021 and 26 April 2021 where the text of the tweets has been labelled by training the models with an already labelled dataset of corona virus tweets as positive, negative, and neutral. Sentiment analysis was conducted by a deep learning model known as Bidirectional Encoder Representations from Transformers (BERT) and various ML models for text analysis and performance which were then compared among each other. ML models used were Naïve Bayes, Logistic Regression, Random Forest, Support Vector Machines, Stochastic Gradient Descent and Extreme Gradient Boosting. Accuracy for every sentiment was separately calculated. The classification accuracies of all the ML models produced were 66.4%, 77.7%, 74.5%, 74.7%, 78.6%, and 75.5%, respectively and BERT model produced 84.2%. Each sentiment-classified model has accuracy around or above 75%, which is a quite significant value in text mining algorithms. We could infer that most people tweeting are taking positive and neutral approaches.

Keywords: COVID-19, coronavirus, Twitter, tweets, sentiment analysis, tweepy, text classification.

Forma sugerida de citación:

Darad, S. y Krishnan, S. "Análisis de sentimiento de los datos de twitter de COVID-19 utilizando modelos de aprendizaje profundo y aprendizaje máquina," *Ingenius, Revista de Ciencia y Tecnología*, N.# 29, pp. 108-117, 2023. DOI: https://doi.org/10.17163/ings.n29.2023.10

1. Introducción

Existen varias plataformas de redes sociales que son utilizadas por los usuarios por muchas razones. Recientemente, las plataformas de redes sociales más utilizadas para comunicaciones informales han sido Facebook, Twitter, Reddit, etc. Entre estas, Twitter, la plataforma de microblogging, tiene una Interfaz de Programación de Aplicaciones (API, Application Programming Interface en inglés) bien documentada para acceder a los datos (tweets) disponibles en la plataforma. Por lo tanto, Twitter se ha convertido en una fuente de información primaria para investigadores que trabajan en el área de Computación Social [1].

Las plataformas de Redes Sociales como Twitter son un gran recurso para capturar emociones y pensamientos humanos. Durante estos tiempos difíciles, la gente ha adoptado las redes sociales para discutir sus miedos, opiniones y conocimientos acerca de la pandemia mundial [2]. Esta investigación se enfocó en un conjunto de datos que contiene tweets de Twitter y tweets a los que se accedió relacionados con la "Pandemia de COVID-19".

La enfermedad por Coronavirus 2019 (COVID-19) fue inicialmente detectada en Wuhan, China, en diciembre de 2019 y se ha propagado mundialmente a más de 198 países [3]. El brote de COVID-19 tiene un impacto socio-económico. La Organización Mundial de la Salud lo declaró una epidemia el 30 de enero de 2020. Desde entonces se ha propagado exponencialmente causando serios problemas de salud, incluyendo muertes dolorosas [4]. Se requieren conjuntos de datos de gran tamaño para entrenar modelos de aprendizaje automático o para llevar a cabo cualquier tipo de análisis.

El conocimiento extraído de conjuntos de datos pequeños y conjuntos de datos de regiones específicas no puede ser generalizado debido a las limitaciones en el número de tweets y la cobertura geográfica. Por lo tanto, este artículo introduce un conjunto de datos de gran tamaño constituido por tweets en inglés acerca del COVID-19 [5].

El objetivo principal de este trabajo es predecir el sentimiento de la gente durante el pico de la pandemia en abril de 2021. ¿Cómo podemos clasificar los tweets acerca del coronavirus como positivos, negativos y neutrales, lo cual nos indica lo que está sintiendo la gente? Por lo tanto, hay dos formas de etiquetar los tweets extraídos utilizando la API de Twitter con tweepy. La primera forma es entrenar un modelo BERT y varios modelos de aprendizaje automático con datos ya etiquetados, evaluando qué modelo clasificador podría etiquetar correctamente los tweets, y luego usar ese modelo para etiquetar el texto de los tweets extraídos.

La segunda forma de encontrar el sentimiento es usar VADER, una librería predesarrollada en código abierto para análisis de sentimiento. Esta librería pronostica de manera automática la puntuación del sentimiento de los tweets, usando la capacidad del aprendizaje automático para clasificarlos y hacer inferencias acerca de los tweets extraídos. A partir de la clasificación de diferentes tweets, el esfuerzo era ser capaz de proporcionar más conocimientos acerca de cómo la pandemia afecta la salud mental y la reacción de la gente sobre qué tan bien están manejando esta situación.

1.1. Revisión de la literatura

El objetivo principal de este trabajo es analizar las reacciones de la gente por Twitter sobre la pandemia mundial de COVID-19, y clasificarlas como positivas, negativas o neutrales. Esto se hace mediante el análisis

de sentimiento de la data obtenida en Twitter. Se utilizaron varias técnicas de aprendizaje automático para obtener los resultados. En esta sección se ofrece un resumen de los artículos que fueron usados como referencia para este trabajo.

Se han realizado muchos estudios sobre este tópico en un período de tiempo corto. Para empezar, este artículo captura y presenta las tendencias de tweets positivos, negativos y neutrales por estado y por mes en la India. Primero se hace un análisis por estado y luego se calcula la frecuencia de tweets Positivos, Negativos y Neutrales. A partir del análisis realizado en este artículo, se observa que la gente en la India estaba expresando sus opiniones mayoritariamente con sentimientos positivos [1]. En otro artículo se consideró un conjunto de datos muy grande, de más de 310 millones de tweets. Este estudio especifica la puntuación del sentimiento de tweets únicamente en inglés. Se observó un hashtag común que fue utilizado en la mayoría de los tweets [5].

En otro trabajo de investigación se realizó un análisis de sentimiento de los tweets por país. Este trabajo tomó en consideración los tweets de doce países entre el 11 y el 31 de marzo de 2020. Los tweets fueron recolectados, preprocesados y luego usados para minado de textos y análisis de sentimiento. El resultado del estudio concluye que mientras la mayoría de la gente alrededor del mundo tuvo un enfoque positivo y lleno de esperanza, también hay casos alrededor del mundo en los que se mostró miedo, tristeza y disgusto [6]. En otro artículo de investigación se utilizó un modelo BERT para analizar los sentimientos detrás de los tweets emitidos por los internautas de la India. Varias palabras comunes surgieron en el análisis en base a las cuales los tweets fueron clasificados en cuatro sentimientos tales como miedo, tristeza, rabia y alegría.

Este modelo tuvo una precisión de 89 % en comparación con otros modelos como Regresión Logística (LR, Logistic Regression en inglés), Máquinas de Soporte Vectorial (SVM, Support Vector Machines en inglés), y Redes de Memoria Larga de Corto Plazo (LSTM, Long Short Term Memory) [7]. Por otra parte, se realizó una investigación corta centrada en analizar los sentimientos y emociones de la gente durante el COVID-19 en base a los tweets emitidos entre el 11 y el 31 de marzo de 2020, cuyos resultados nos indican que la forma de pensar de la gente estuvo casi al mismo nivel alrededor del mundo [8].

Existen algunos artículos en los que se realizó un análisis exploratorio de los datos para obtener los resultados. Por ejemplo, en un artículo de investigación se realizó un análisis exploratorio para un conjunto de datos, suministrando información acerca del número de casos confirmados por día en algunos de los países más golpeados, para hacer una comparación entre el cambio en el sentimiento y el cambio en los casos desde el inicio de esta pandemia hasta junio de 2020 [2]. En este artículo, los autores han tratado de entender y analizar los tweets acerca del COVID-19 en la India, usando procesadores NVIVO y nubes de palabras.

El estudio involucra las palabras y hashtags utilizados y los sentimientos entrañados por estas palabras. La conclusión ofrece un entendimiento acerca de palabras de alto impacto y bajo impacto [9]. En este artículo de investigación se recolectan datos de los usuarios que compartieron su ubicación como 'Nepal' entre el 21 y el 31 de mayo de 2020. El resultado del estudio concluyó que mientras la mayoría de la gente de Nepal adoptó un enfoque positivo y lleno de esperanza, también se mostraron casos de miedo, tristeza y disgusto [10].

Considerando que Twitter es un lugar donde las personas pueden expresar sus opiniones sin revelar su identidad, muchas de esas personas utilizan esto como una ventaja para presentar opiniones positivas o negativas en base a sus sentimientos. Un análisis de sentimiento que se realizó en data de Twitter sobre COVID utilizando varias técnicas de aprendizaje automático y conocimiento de las redes sociales, nos dio resultados positivos y negativos. El algoritmo de regresión logística se utilizó para realizar el análisis, y se obtuvo una precisión de 78.5% [11].

Por otra parte, se realizó minería de datos en Twitter para recolectar un total de 107990 tweets relacionados con COVID-19, entre el 13 de diciembre de 2019 y el 9 de marzo de 2020. Se utilizó un enfoque de procesamiento de Lenguaje Natural (NLP, Natural Language Processing en inglés) y el algoritmo de asignación latente de Dirichlet para identificar los tópicos más comunes de los tweets, así como también categorizar clústeres e identificar temas en base al análisis de palabras clave. Los resultados indican los aspectos

principales de la conciencia y preocupación pública con respecto a la pandemia de COVID-19. Primero, la tendencia de la propagación y los síntomas del COVID-19 pueden dividirse en tres etapas.

Segundo, los resultados del análisis de sentimiento mostraron que la gente tiene una perspectiva negativa hacia el COVID-19 [12]. En este artículo, nuestro objetivo es realizar un análisis de sentimiento de los tweets durante la pandemia de COVID-19 y clasificarlos como positivos, negativos o neutrales.

Luego de aprender acerca del conjunto de datos, el próximo paso fue resolver el problema de clasificación, que en este artículo es el análisis de sentimiento. Muchas de los artículos mencionados previamente [1,5] realizaron análisis de sentimiento sobre tweets para clasificarlos en tres categorías diferentes. Estos artículos de investigación suministraron información vital acerca de cómo puede utilizarse el análisis de sentimiento para clasificar los tweets del conjunto de datos. El siguiente paso fue crear un clasificador. "The impact of preprocessing on text classification" es un artículo ingenioso que proporcionó detalles y pistas sobre cómo realizar preprocesamiento de los datos y cuál clasificador sería el óptimo.

Menciona que la SVM es un clasificador de patrones de última generación, y se recomienda su uso como el algoritmo de clasificación [13]. Los artículos utilizan Random Forest, Bayes ingenuo y SVM para la clasificación, e indican que las SVM lineales alcanzaron los mejores resultados, con una precisión cercana al 95%. En base a esta investigación, hemos decidido utilizar Bayes ingenuo, Regresión Logística, Random Forest, SVM, descenso por gradiente estocástico (SGD, Stochastic Gradient Descent en inglés), refuerzo de gradiente extremo (eXtreme Gradient Boosting, en inglés) y Representaciones de Codificadores Bidireccionales de Transformers (BERT, Bidirectional Encoder Representations from Transformers en inglés).

Antes de continuar con el conjunto de datos, es importante saber sobre el conjunto de datos y aprender tanto como sea posible. Se realizó un análisis exploratorio detallado del conjunto de datos utilizando referencias de varios documentos.

2. Materiales y métodos

2.1. Materiales

La data para este trabajo fue adquirida de Twitter, utilizando su API tweepy. Tweepy es un paquete de Python de código abierto de fácil uso, para acceder a las funcionalidades proporcionadas por la API de Twitter. Tweepy incluye un conjunto de clases y métodos que representan modelos de Twitter y puntos finales de APIs, y maneja de forma transparente varios detalles de implementación, tales como codificación y decodificación de datos. La extracción de un total de 200000 tweets de la API de Twitter se realizó entre el 16 y el 26 de abril de 2021, lo que representa un conjunto de datos más grande y permite obtener mejores resultados.

El otro conjunto de datos es de código abierto y fue recolectado de un blog [14] que contiene tweets sobre coronavirus con sentimientos etiquetados. El conjunto de tweets que fue recolectado por el blog fue un conjunto de datos de análisis de sentimiento etiquetados. Este conjunto de datos fue dividido en dos subconjuntos para entrenamiento y prueba de los diferentes clasificadores. El conjunto de datos que fue buscado y recolectado de Twitter no está etiquetado.

2.1.1. Analítica Descriptiva

El conjunto de datos contiene campos de texto, por lo que el análisis de texto de los tweets se realizó como se describe a continuación. Sin embargo, antes de realizar el análisis fue necesario aprender más acerca del conjunto de datos. Primero, aún antes de llevar a cabo el proceso de limpieza, es necesario familiarizarse con

el tipo de datos que se estará manejando. Esto ayuda a proporcionar más contexto y antecedentes al científico de datos. Por lo tanto, después de cargar el archivo csv, se ejecutaron algunas funciones sobre los datos para familiarizarse con ellos.

Es necesario conocer el tamaño del conjunto de datos, los tipos de datos de cada columna, el número de registros nulos, la distribución de las diferentes clases, etc. Luego se eliminan las filas duplicadas, en caso de que existan. Luego se detectó que algunas columnas no serían necesarias en el análisis posterior, por lo que fueron eliminadas.

Luego, se aplicaron a los datos esas técnicas de preprocesamiento para limpiar los tweets. Esto incluye convertir el texto a letras minúsculas, tokenization y eliminar etiquetas de nombres de usuario, símbolos de retweet, hashtags, espacios en blanco, signos de puntuación, números, emojis y URLs para limpiar el texto. Posteriormente se realizó el análisis de texto sobre este texto limpio, según se describe a continuación. El análisis se realizó al conjunto de datos recolectado de la API de Twitter, con 200000 tweets (Figura 1).

```
print('There are {} rows and {} columns in the dataset.'.format(df.shape[0],df.shape[1]))
```

```
There are 200000 rows and 13 columns in the dataset.
```

FIGURA 1.
Tamaño del conjunto de datos

Luego se mira a la información del conjunto de datos, lo que indica el tipo de campo y cuántos valores no nulos están presentes, lo cual ayuda a entender mejor el conjunto de datos (Figura 2).

Con las redes sociales nunca se puede recuperar toda la data. Siempre hay algunos valores faltantes en el conjunto de datos. A la gente le gusta mantener algunas cosas discretas, tales como su ubicación y descripción en el caso de Twitter. Asimismo, puede verse que algunas personas no están cómodas con el uso de hashtags se muestra en la Figura 3.

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200000 entries, 0 to 199999
Data columns (total 13 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   user_name         199990 non-null  object
 1   user_location     142121 non-null  object
 2   user_description  180498 non-null  object
 3   user_created      200000 non-null  object
 4   user_followers    200000 non-null  int64
 5   user_friends      200000 non-null  int64
 6   user_favourites   200000 non-null  int64
 7   user_verified     200000 non-null  bool
 8   date              200000 non-null  object
 9   text              200000 non-null  object
 10  hashtags          55136 non-null   object
 11  source            200000 non-null  object
 12  is_retweet        200000 non-null  bool
dtypes: bool(2), int64(3), object(8)
memory usage: 17.2+ MB
```

FIGURA 2.
Información del conjunto de datos

```
df.isna().sum()

user_name              10
user_location       57879
user_description    19502
user_created            0
user_followers          0
user_friends            0
user_favourites         0
user_verified           0
date                    0
text                    0
hashtags           144864
source                  0
is_retweet              0
dtype: int64
```

FIGURA 3.
Total de valores nulos

Entonces debe encontrarse la frecuencia de las palabras, mostrando las palabras de uso más frecuente de acuerdo con su conteo. Se observó que "COVID-19" es la palabra más utilizada (Figura 4).

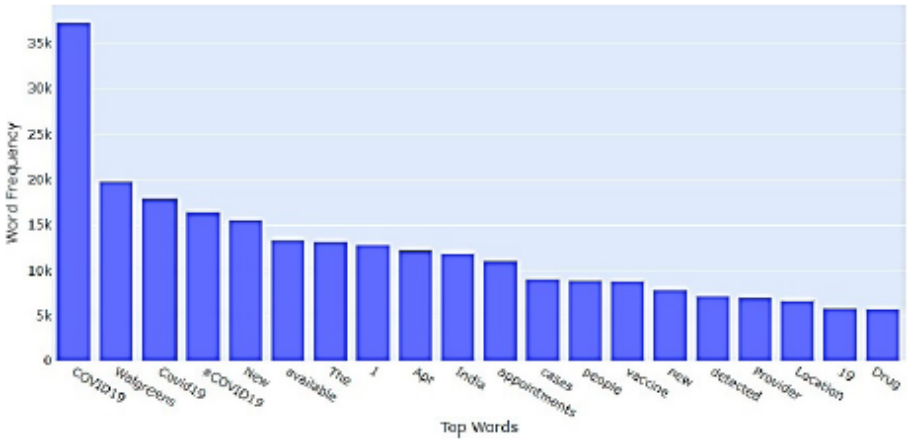

FIGURA 4.
Palabras más usadas en los tweets

Para tener una mirada más cercana al texto contenido en el conjunto de datos, se creó una visualización de la nube de palabras (Figura 5).

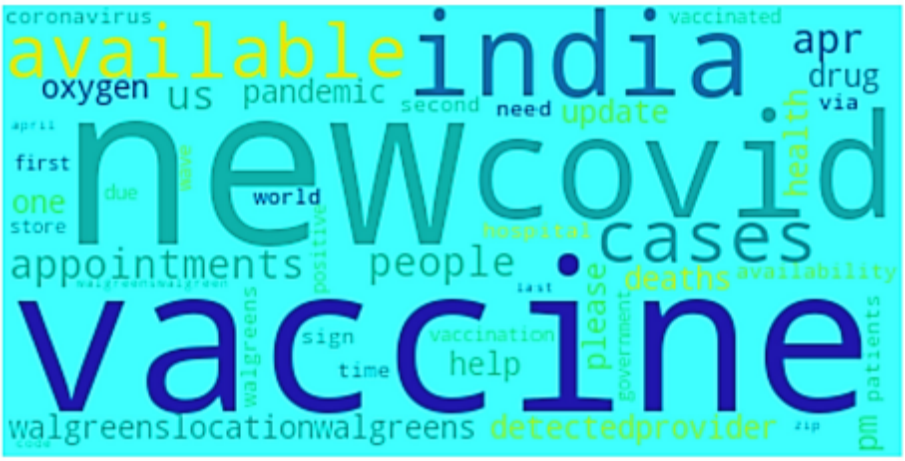

FIGURA 5.
Nube de las 50 palabras más utilizadas

La nube de palabras que se muestra en la Figura 5 lista las 50 palabras más utilizadas. Las nubes de palabras son útiles para entender lo que los usuarios están publicando. La mayoría de las palabras están relacionadas con el COVID y los nuevos casos, y parece que la mayoría de las personas también hizo publicaciones acerca de las vacunas (Figura 6).

Después de mirar un resumen del texto de los tweets en el corpus, se consideran los hashtags mirando a los que exhiben mayor tendencia.

FIGURA 6.
Nube de palabras para los Hashtags

La nube de palabras que se muestra en la Figura 6 lista todas las palabras con los hashtags más utilizados. Las nubes de palabras son útiles para entender lo que los usuarios están publicando. La mayoría de las palabras están relacionadas con el COVID y los nuevos casos, y parece que la mayoría de las personas también hizo publicaciones acerca de las vacunas.

La Figura 7 muestra el lugar desde donde la mayoría de la gente está tuiteando. Puede verse que un mayor número de personas está tuiteando desde la India y desde los Estados Unidos de América, ya que el período de tiempo seleccionado para extraer los tweets fue durante la tercera ola y el número de casos fue más alto en esos países.

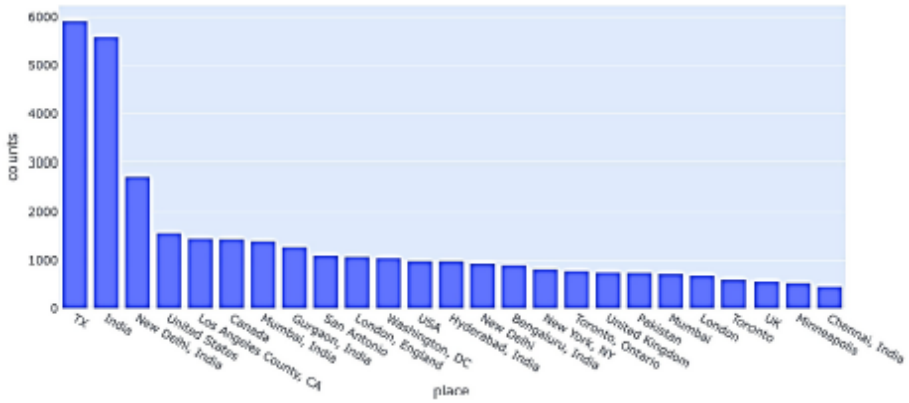

FIGURA 7.
Los 25 lugares desde donde más se originaronlos tweets

La Figura 8 muestra los usuarios verificados quetuitearon más acerca del COVID. Puede verse quecasi todos son canales de noticias tuiteando acerca delas actualizaciones más recientes sobre el COVID y elnúmero de casos en sus respectivos países.

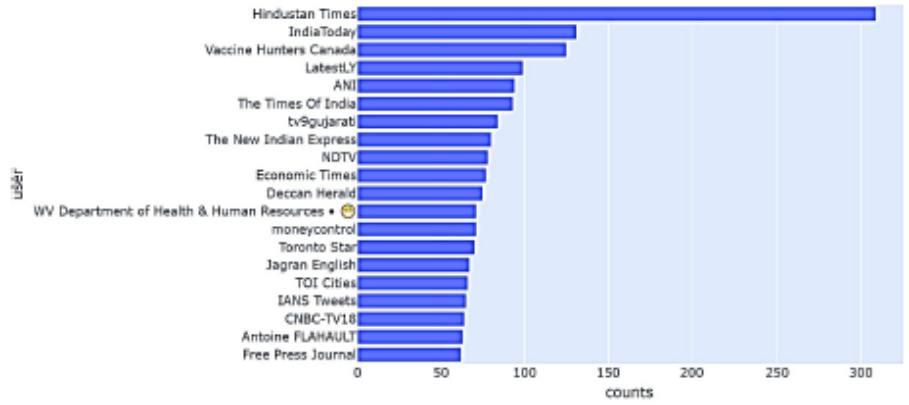

FIGURA 8.
Los 20 usuarios verificados que más tuitearon

Tras hacer un resumen de la data, se limpió y preprocesó el texto de los tweets en el corpus, procediendo a hacer un análisis de n-gramas. Los n-gramas ofrecen un mejor contexto de los temas acerca los cuales los usuarios están publicando, ya que a medida que se consideran bigramas o trigramas se tiene información acerca de las frases más frecuentes en vez de palabras individuales. La Figura 9 muestra que los unigramas más frecuentes están basados en nuevos casos, vacunas,salud, pandemia, gente, disponibilidad y citas.

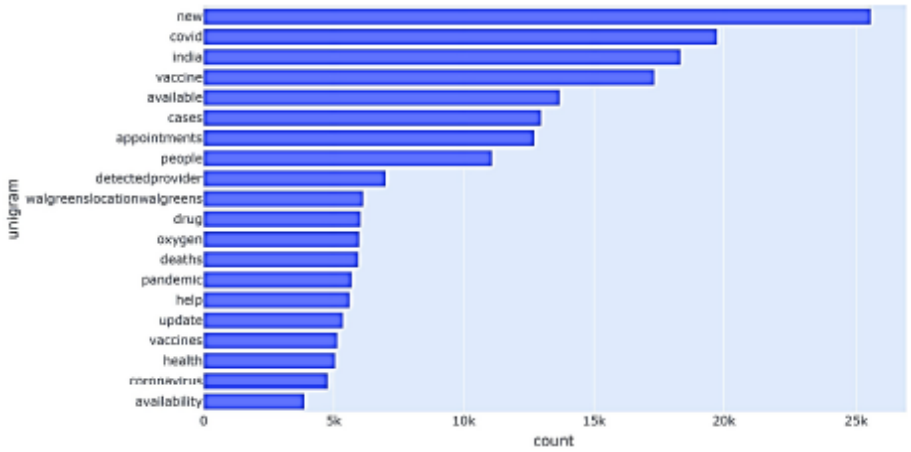

FIGURA 9.
Los 20 unigramas más frecuentes

Un análisis de bigramas (Figura 10) suministra más detalles acerca de las tendencias durante ese tiempo,de la disponibilidad de citas de vacunas, nuevos casos y de la segunda ola.

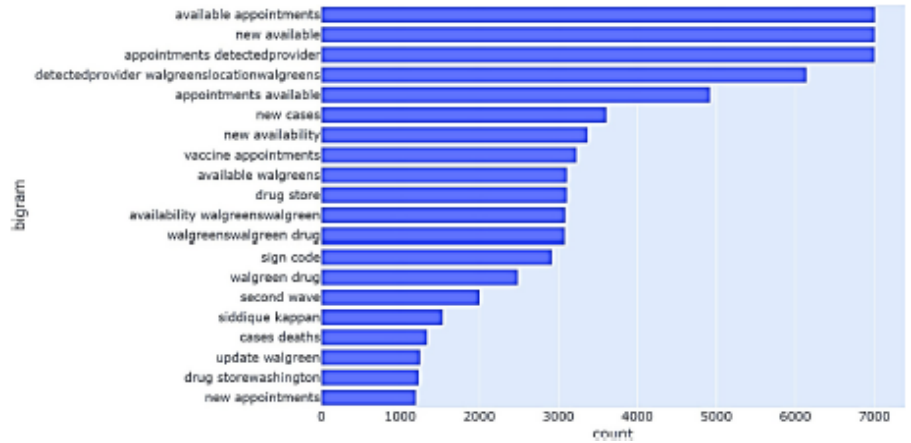

FIGURA 10.
Los 20 bigramas más frecuentes

Un análisis de trigramas (Figura 11) suministra más detalles sobre dónde hay disponibilidad de nuevas citas de vacunación contra el COVID. Parece que la mayoría de ellas están en Walgreens, que es una compañía estadounidense que opera como la segunda cadena más grande de tiendas de farmacia, detrás de CVS Health.

FIGURA 11.
Los 20 trigramas más frecuentes

2.2. Métodos

El objetivo de este estudio es entrenar utilizando texto de los tweets etiquetados, para evaluar automáticamente si el tweet no etiquetado recolectado es positivo, negativo o neutral. Después de entrenar los modelos sobre datos etiquetados de Twitter, los modelos fueron aplicados a data extraída para etiquetar los sentimientos y comparar los resultados de los diferentes algoritmos. El segundo método de etiquetado de tweets se hizo utilizando el paquete de Python NLKT VADER basado en léxicos.

En este trabajo la respuesta es etiquetar los tweets como positivos, negativos o neutrales. El conjunto de datos recolectado contiene una gran cantidad de información del usuario como nombre, descripción, seguidores, amigos y mucha más, pero sólo el texto del tweet fue utilizado para etiquetar la data a partir de la data de entrenamiento etiquetada disponible.

2.2.1. Diseño Experimental 1

Es muy complicado etiquetar los sentimientos entrañados en la data de COVID-19, debido a las palabras utilizadas para representar la situación. Por ejemplo, si hay nuevos casos existe un tweet que dice "Fui testeado para Corona positivo", que puede ser técnicamente etiquetado como positivo por la técnica ML.

Entonces, existe una gran incertidumbre al momento de predecir los sentimientos de la pandemia. Por lo tanto, se aplicaron dos técnicas diferentes para conocer los sentimientos.

a) Procesamiento de Texto

El conjunto de datos llamado "coronavirustweets" contenía data etiquetada que muestra el sentimiento como extremadamente positivo, positivo, neutral, negativo y extremadamente negativo. Al restringir las etiquetas de categorías a una clasificación de solo tres clases, se tiene neutral, negativo combinado con extremadamente negativo, y positivo combinado con extremadamente positivo para tener una mayor precisión. Antes de generar la data de entrenamiento y prueba, es necesario preprocesar el texto original del tweet para remover signos de puntuación, palabras de parada, espacios, emoticones y el origen de los datos.

El preprocesamiento de los datos de texto es un paso esencial, ya que prepara el texto para el minado.

El objetivo de este paso es limpiar el texto que es irrelevante para encontrar el sentimiento en los tweets, tal como los signos de puntuación (.,?," etc.), caracteres especiales (,%,&,$, etc.), números (1,2,3, etc.), manejo de Twitter, enlaces (HTTPS: / HTTP:) y palabras de parada que no significan nada en el contexto del texto.

Las palabras de parada son aquellas palabras en lenguaje natural que tienen muy poco significado, tales como "is", "an", "the", etc. Para removerlas de una oración el texto se divide en palabras, y luego se eliminan aquellas palabras que están en una lista de palabras de parada suministrada por NLTK.

b) Aleatorización

El conjunto de datos fue dividido aleatoriamente en dos conjuntos estratificados de acuerdo a los valores de sentimiento, el de entrenamiento con el 80 % de la data y el de prueba con el 20 % de la data.

c) Vectorización de los tweets

Antes de implementar los diferentes clasificadores de texto basados en ML, es necesario convertir los datos de texto en vectores. Esto es crucial ya que los algoritmos esperan la data en alguna forma matemática y no en forma de texto. El contador de vectorización cuenta el número de veces que una palabra aparece en el documento (en cada tweet). Este proceso ayuda convirtiendo la data como la entendemos en data numérica, que es más fácil de entender por parte del computador.

d) Clasificadores

Luego de vectorizar los tweets, estamos listos para implementar los algoritmos de clasificación. Existen tres tipos de sentimientos, por lo que deben entrenarse los modelos para que den la etiqueta correcta para el conjunto de datos de prueba. Se construyeron diferentes modelos de aprendizaje automático tales como Bayes ingenuo, Regresión Logística, Random Forest, Máquina de Soporte Vectorial, Descenso por Gradiente Estocástico y Refuerzo de Gradiente Extremo junto con BERT, un modelo de aprendizaje profundo. Se aplica una combinación de clasificadores, tales como bagging and boosting, sobre el conjunto de datos para minimizar cualquier sobreentrenamiento de los clasificadores.

Se utiliza la exactitud como puntuación para medir el desempeño del modelo (también se calculan la precisión, recall y matriz de confusión). La precisión, recall y matriz de confusión permiten saber qué tan correctamente están etiquetados los valores reales. El BERT es una técnica desarrollada por Google basada en el mecanismo de Transformers. En nuestra aplicación de análisis de sentimiento, el modelo se entrena sobre un modelo BERT preentrenado. Los modelos BERT han reemplazado las redes LSTM basadas en redes neuronales recurrentes (RNN, recurrent neural networks en inglés) convencionales, que sufren de pérdida de información en secuencias de texto largas [15]. Los resultados del artículo explicaron que un modelo de lenguaje preparado bidireccionalmente puede tener un sentido más profundo sobre la configuración y el flujo del idioma que los modelos de una sola dirección. En contraste con los modelos direccionales que permiten una lectura secuencial de la entrada de texto (de derecha a izquierda o de izquierda a derecha), el transformer codificador reconoce de una vez la secuencia total de palabras. Por lo tanto, se considera bidireccional, pero es un modelo no direccional con una precisión más alta que otros modelos establecidos [7].

e) Etiquetado de nuevos tweets

Dado que la data recolectada no está etiquetada, los modelos entrenados se almacenan y se cargan con pickle. Esto permite almacenar el modelo en un archivo y cargarlo posteriormente para hacer predicciones. Luego puede aplicarse el modelo para etiquetar la data extraída y preprocesada.

f) Comparación de los algoritmos

Una vez obtenidos los sentimientos de los tweets de los diferentes modelos y almacenados los archivos csv de dichos modelos, se comparan los resultados sobre data etiquetada.

2.2.2. Diseño Experimental 2

VADER significa Valence Aware Dictionary and sEntiment Reasoner. VADER pertenece a un tipo de análisis de sentimiento basado en léxicos de palabras relacionadas con sentimientos. Es un modelo basado en reglas para análisis general de sentimiento, y su efectividad fue comparada con 11 benchmarks típicos, incluyendo Word Count (LIWC), Affective Norms for English Words (ANEW), the General Inquirer, Linguistic Inquiry, Senti WordNet, y técnicas de aprendizaje automático que se basan en Máquinas de

Soporte Vectorial (SVM), Bayes ingenuo, y Entropía Máxima. En este enfoque, cada una de las palabras en el léxico se clasifica como positiva o negativa, y en michos casos qué tan positiva o negativa.

VADER tiene un buen desempeño en el análisis de sentimientos expresados en las redes sociales. Estos sentimientos deben estar presentes en forma de comentarios, tweets, retweets o descripciones posteriores, y también trabaja bien sobre textos de otros dominios. Se diseñó el modelo de sentimiento VADER, el cual extrae características de data de Twitter, formula las puntuaciones de sentimiento y los clasifica en las clases positivo, negativo o neutral.

a) Limpieza de la Data

El conjunto de datos extraído de Twitter requiere que el texto sea preprocesado, removiendo signos de puntuación, palabras de parada, espacios, emoticones y el origen de los datos.

b) Encontrar Polaridad

La puntuación compuesta (polaridad) se calcula sumando la valencia de cada palabra en el léxico, ajustada de acuerdo a las reglas, y luego normalizadas a valores entre -1 (extremo más negativo) y +1 (extremo más positivo).

c) Encontrar Sentimientos

Luego de obtener la puntuación compuesta, la polaridad de los tweets se utiliza para categorizarlos en 3 clases: Positivo, Negativo o Neutral. Los Sentimientos Positivos son aquellos con puntuación por encima de 0, los Sentimientos Negativos son aquellos con puntuación por debajo de 0, y los Sentimientos Neutrales son aquellos que tienen una polaridad de 0.0. Estas 3 clases fueron almacenadas junto con los tweets en el conjunto de datos denominados “Sentiments”.

3. Resultados y discusión

3.1. Resultados

3.1.1. Resultado Experimental 1

Se aplicó la clasificación multiclase de los diferentes modelos a la data de entrenamiento para encontrar la exactitud de la etiqueta correcta en el conjunto de prueba. Se construyeron diferentes modelos ML tales como Bayes ingenuo, Regresión Logística, Random Forest, Máquina de Soporte Vectorial, Descenso por Gradiente Estocástico y Refuerzo de Gradiente Extremo (Figura 12).

Se observó que el clasificador basado en Descenso por Gradiente Estocástico da el mejor resultado, alcanzando una precisión de 78.64 %. Esta exactitud es muy cercana a la exactitud de la Regresión Logística, y ambos modelos pueden ser utilizados para predecir el sentimiento de la data no etiquetada. La menor exactitud fue la del clasificador Bayes ingenuo, el cual trabaja bien con datas grandes. El clasificador de Bayes ingenuo trabaja sobre n-gramas y se probó sobre diferentes n-gramas, pero la exactitud se mantiene alrededor del 65 %.

	Model	Test accuracy
4	Stochastic Gradient Decent	0.786460
1	Logistic Regression	0.777095
5	XGBoost	0.755143
3	Support Vector Machines	0.747467
2	Random Forest	0.745011
0	Naive Bayes	0.664569

FIGURA 12.
Comparación de la exactitud de los modelos

El modelo BERT se desempeña extremadamente bien en comparación con otros modelos ML. Su exactitud alcanza 84.2 %, que es la exactitud más alta alcanzada en el entrenamiento y prueba de los modelos. BERT es una técnica excelente y diferente, que ofrece la mejor precisión porque fue diseñada para leer en ambas direcciones de una vez. Esta capacidad, habilitada por la introducción de los Transformers, se conoce como bidireccionalidad. Sin embargo, BERT fue preentrenado usando solamente un corpus de texto normal no etiquetado (específicamente la totalidad de Wikipedia en inglés, y el Corpus Brown). Este modelo continúa aprendiendo de forma no supervisada a partir del texto no etiquetado, y mejora aun cuando se utilice en aplicaciones prácticas (es decir, búsqueda en Google). Su preentrenamiento sirve como capa base de "conocimiento" sobre la cual se construye. Desde ahí, BERT puede adaptarse al cuerpo en constante crecimiento de contenido que puede buscarse y de consultas, y ser ajustado de acuerdo a las especificaciones del usuario. Este proceso se conoce como transferencia de conocimiento (transfer learning) [16].

Luego, este modelo entrenado se usa sobre nuestro conjunto de datos, obteniéndose los siguientes resultados en base a la exactitud en prueba (Figura 13).

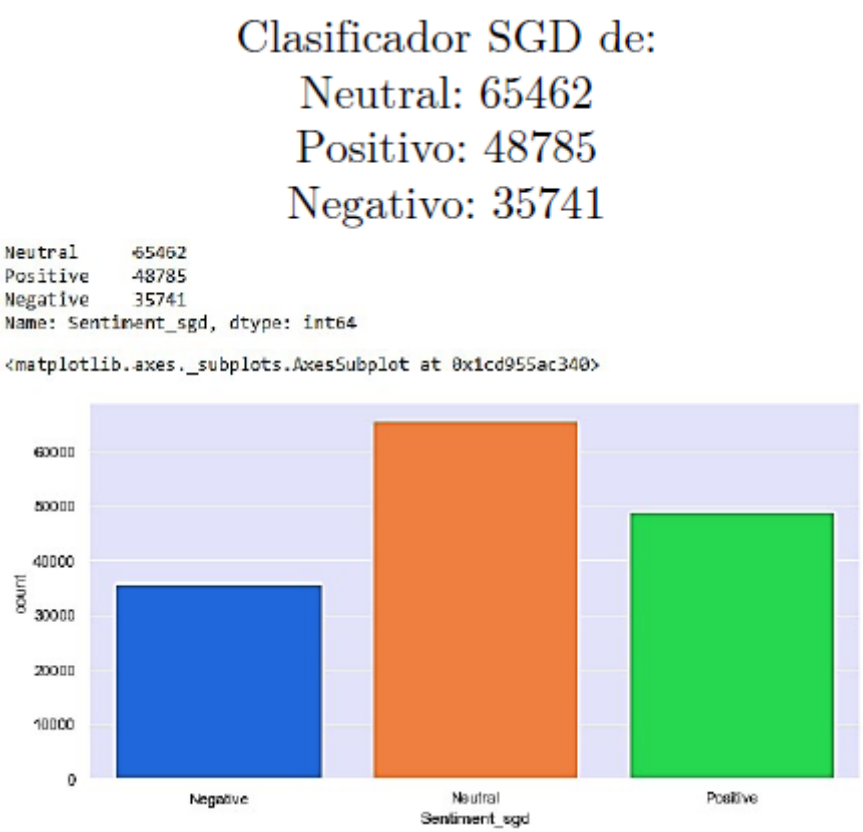

FIGURA 13.
Resultados de Clasificador SGD

El Descenso por Gradiente Estocástico es un enfoque simple pero muy eficiente para ajustar clasificadores y regresores lineales, bajo funciones convexas de pérdida. El SGD ha sido aplicado exitosamente a los problemas grandes y dispersos de aprendizaje automático encontrados frecuentemente en clasificación de texto y procesamiento de lenguaje natural, que es por lo que se desempeña mejor que todos los demás modelos (Figura 14).

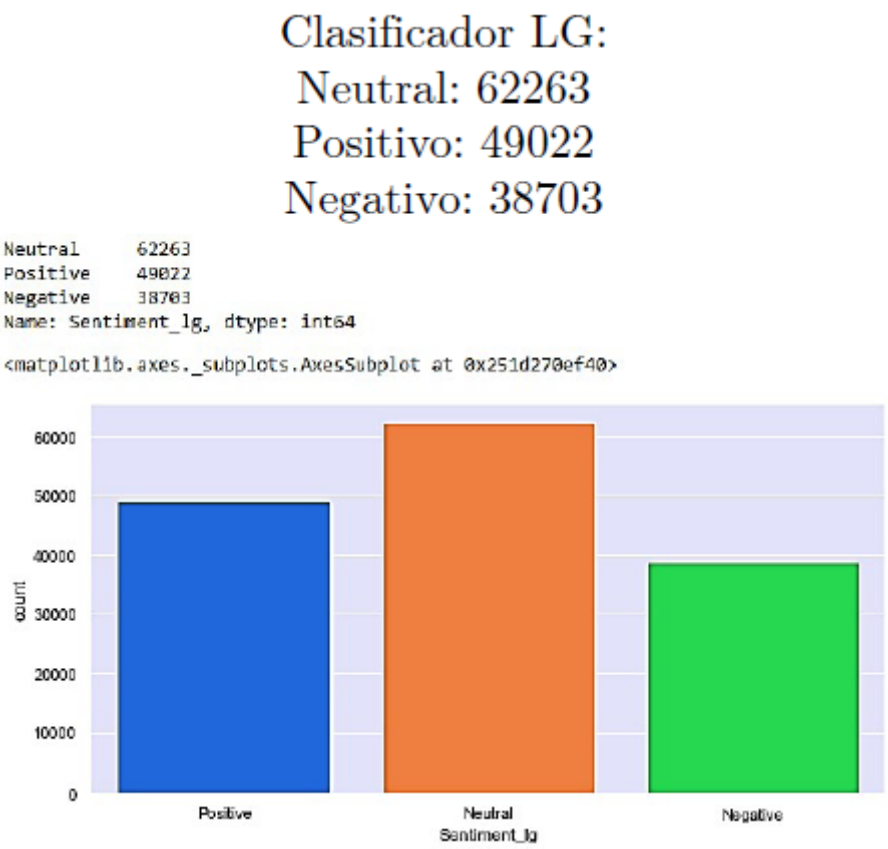

FIGURA 14.
Resultados del clasificador LR

En los resultados de observa que la Regresión Logística da más tweets etiquetados como Positivos y Negativos, mientras que el Descenso por Gradiente Estocástico predice algunos de ellos como Neutrales. A pesar de que la exactitud de ambos es casi la misma, existe una diferencia de aproximadamente 3000 tweets etiquetados como Neutrales. La regresión logística multinomial es una extensión de la regresión logística que agrega soporte nativo para problemas de clasificación multiclase. Por defecto, la regresión logística está limitada a dos clases, que es la razón por la cual SGD tiene una mejor exactitud para predecir sentimientos.

Se aplicó la combinación de clasificadores como bagging y boosting sobre el conjunto de datos, para minimizar cualquier sobreentrenamiento de los clasificadores. Sin embargo, no hay ningún sobreentrenamiento de la data porque la exactitud obtenida con bagging es 72.1 %, que es aproximadamente igual, mientras que la exactitud con boosting es 51.4.

En [17] se presenta un análisis similar para conocer la ansiedad entre usuarios de Twitter debido a la pandemia, en base a palabras clave. Cerca de 900000 tweets fueron extraídos de la API de Twitter y analizados usando modelos de Bayes ingenuo y de regresión logística. La exactitud de los modelos en tweets cortos es de 91 % y 74 %, respectivamente. Sin embargo, la mayor limitación de este estudio es que todos los sentimientos dependen de la palabra "miedo" de ciudadanos de Estados Unidos únicamente, y son tweets cortos [7].

3.1.2. Resultado Experimental 2

El modelo de sentimiento VADER es una técnica de etiquetado automático, en el que se formuló la puntuación del sentimiento clasificando los tweets como positivos, negativos o neutrales. La diferencia principal que se observa aquí es que da menos (alrededor de 5000 menos) tweets neutrales, y los clasifica

como positivos o negativos. Puede verse que casi empareja la exactitud de nuestros modelos entrenados con data etiquetada, al mostrar los siguientes resultados (Figura 15).

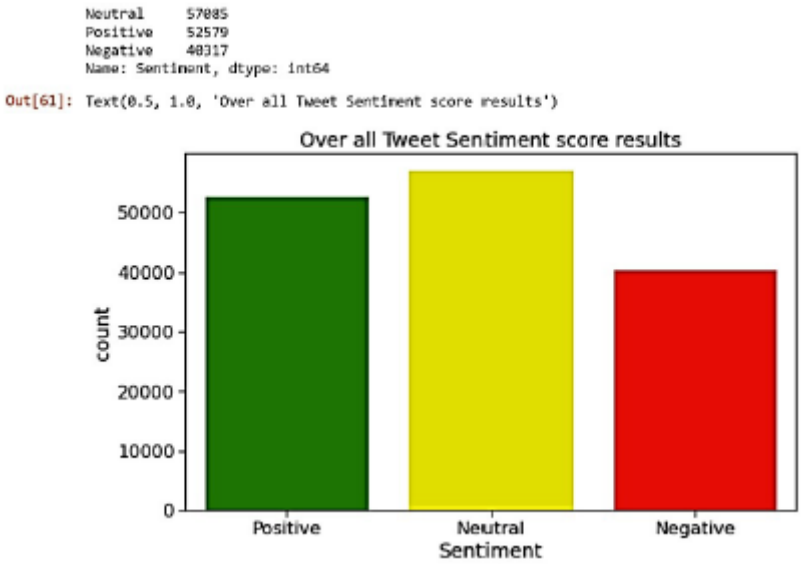

FIGURA 15.
Resultados de VADER

3.2. Discusión

Este estudio puede ser utilizado para analizar los sentimientos cambiantes de la gente alrededor del mundo, y verificar si hay variaciones importantes en ellos en el período de tiempo junto con el aumento en el suministro de vacunas. Se espera que a medida que la propagación de la pandemia aumente en la gente no vacunada, la mayoría de los sentimientos en los tweets serán positivos a medida que las cosas vayan retornando a la normalidad.

Se realizó un análisis similar utilizando TextBlob, como se hizo en el Experimento 2 pero usando VADER. Sin embargo, de acuerdo a la documentación de TextBlob, este utiliza un modelo Bayes ingenuo (NB, Naïve Bayes en inglés) para la clasificación. El clasificador NB fue entrenado sobre el NLTK (Natural Language Tool Kit) para detectar la valencia de los tweets agregados [10]. Como se observó, el modelo de Bayes ingenuo alcanza la menor exactitud, por lo que TextBlob no es preciso para etiquetar sentimientos. Los métodos de ML clásicos alcanzaron una precisión de poco más de 70 %, mientras que el modelo de aprendizaje profundo que utiliza BERT alcanzó una exactitud impresionante de 84.2 %.

4. Conclusiones

Los resultados del estudio concluyen que la mayoría de las personas alrededor del mundo adoptó un enfoque positivo y lleno de esperanza. Sin embargo, países como la India y los Estados Unidos de América mostraron señales de un tuiteo a mayor escala debido a la tercera ola, en comparación con los países restantes.

Se utilizaron dos técnicas sobre nuestro conjunto de datos, pero tal como se muestra, siempre existe un margen de error en la clasificación de texto. También se muestra que BERT requiere un gran poder computacional, GPU y un tiempo largo para entrenar el modelo. En la predicción de texto en cualquier red social es casi imposible alcanzar una exactitud perfecta. A través de esto, se puede aprender la cuestión principal para ayudar a los proveedores de salud a identificar algún tipo de enfermedad mental antes que sea demasiado tarde.

Agradecimientos

Los autores agradecen a la Ryerson University por el apoyo brindado a este proyecto

Referencias

[1] T. Vijay, A. Chawla, B. Dhanka, and P. Karmakar, "Sentiment analysis on covid-19 twitter data," in 2020 *5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE),* 2020, pp. 1–7. [Online]. Available: https: //doi.org/10.1109/ICRAIE51050.2020.9358301

[2] M. Mansoor, K. Gurumurthy, A. R. U, and V. R. B. Prasad, "Global sentiment analysis of COVID-19 tweets over time," *CoRR*, vol. abs/2010.14234, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2010.14234

[3] H. Drias and Y. Drias, "Mining twitter data on covid-19 for sentiment analysis and frequent patterns discovery," *medRxiv,* 2020. [Online]. Available: https://doi.org/10.1101/2020. 05.08.20090464

[4] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood, and G. S. Choi, "A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis," *PLOS ONE*, vol. 16, no. 2, pp. 1–23, 02 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0245909

[5] R. Lamsal, "Design and analysis of a large-scale COVID-19 tweets dataset," *Applied Intelligence*, vol. 51, no. 5, pp. 2790–2804, May 2021. [Online]. Available: https://doi.org/10.1007/s10489-020-02029-z

[6] A. D. Dubey, "Twitter sentiment analysis during covid-19 outbreak," *SSRN*, 2021. [Online]. Available: https://dx.doi.org/10.2139/ssrn.357202

[7] N. Chintalapudi, G. Battineni, and F. Amenta, "Sentimental analysis of COVID-19 tweets using deep learning models," *Infect Dis Rep*, vol. 13, no. 2, pp. 329–339, Apr. 2021. [Online]. Available: https://doi.org/10.3390/idr13020032

[8] M. A. Kausar, A. Soosaimanickam, and M. Nasar, "Public sentiment analysis on twitter data during covid-19 outbreak," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021. [Online]. Available: http: //dx.doi.org/10.14569/IJACSA.2021.0120252

[9] A. Mitra and S. Bose, "Decoding Twitter-verse: An analytical sentiment analysis on Twitter on COVID-19 in india," *Impact of Covid 19 on Media and Entertainment*, 2020. [Online]. Available: https://bit.ly/3YMj1c3

[10] B. P. Pokharel, "Twitter sentiment analysis during covid-19 outbreak in nepal," *SSRN*, 2020. [Online]. Available: https: //dx.doi.org/10.2139/ssrn.3624719

[11] C. R. Machuca, C. Gallardo, and R. M. Toasa, "Twitter sentiment analysis on coronavirus: Machine learning approach," J*ournal of Physics: Conference Series*, vol. 1828, no. 1, p. 012104, feb 2021. [Online]. Available: https: //dx.doi.org/10.1088/1742-6596/1828/1/012104

[12] S. Boon-Itt and Y. Skunkan, "Public perception of the COVID-19 pandemic on twitter: Sentiment analysis and topic modeling study," *JMIR Public Health Surveill*, vol. 6, no. 4, p. e21978, Nov. 2020. [Online]. Available: h ttps://doi.org/10.2196/21978

[13] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Information Processing & Management*, vol. 50, no. 1, pp. 104–112, 2014. [Online]. Available: https://doi.org/10.1016/j.ipm.2013.08.006

[14] S. Gujral, "Sentiment analysis: Predicting sentiment of COVID-19 tweets," *Analytics Vidhya*, 2021. [Online]. Available: https://bit.ly/3j9tMVj

[15] S. Gujral, "Amazon product review sentiment analysis using bert," *Analytics Vidhya*, 2021. [Online]. Available: https://bit.ly/3Vad9WE

[16] B. Lutkevich. (2022) Bert language model. TechTarget Enterprise Al. [Online]. Available: https://bit.ly/3Wo5Pb4

[17] J. Samuel, G. G. M. N. Ali, M. M. Rahman, E. Esawi, and Y. Samuel, "Covid-19 public sentiment insights and machine learning for tweets classification," *Information*, vol. 11, no. 6, 2020. [Online]. Available: https://doi.org/10.3390/info11060314

Enlace alternativo

https://revistas.ups.edu.ec/index.php/ingenius/article/view/5512 (html)