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

Simran Darad, Sridhar Krishnan

Recibido: 15-10-2021, Recibido tras revisión: 01-12-2022,
Aceptado: 16-12-2022, Publicado: 01-01-2023

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 precisión de alrededor o superior al 75 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.

1 Introdución

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. Por lo tanto, Twitter se ha convertido en una fuente de información primaria para investigadores que trabajan en el área de Computación Social. 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. 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. 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. Se requieren conjuntos de datos de gran tamaño para entrenar modelos de aprendizaje automático o para llevar a cabo cualquier tipo de análisis. 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. 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. 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. 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. El estudio involucra las palabras y hashtags utilizados y los sentimientos entrañados por estas palabras. 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 78.5Por 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. 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.

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. 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

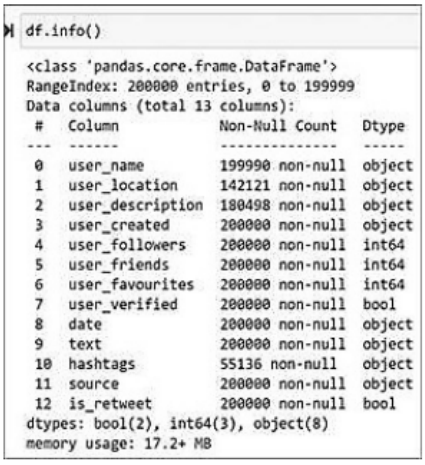
El conjunto de datos contiene campos de texto, por lo que el análisis de texto de los tweets se realizó como se describe a continuación. Sin embargo, antes de realizar el análisis fue necesario aprender más acerca del conjunto de datos. Primero, aún antes de llevar a cabo el proceso de limpieza, es necesario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

```
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.
```

Figure 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.



```
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
```

Figure 2: Información del conjunto de datos

2.2 Métodos

El objetivo de este estudio es entrenar utilizando textode 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, losmodelos 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 tweetscomo 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.

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.

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 (alrededorde 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 3).

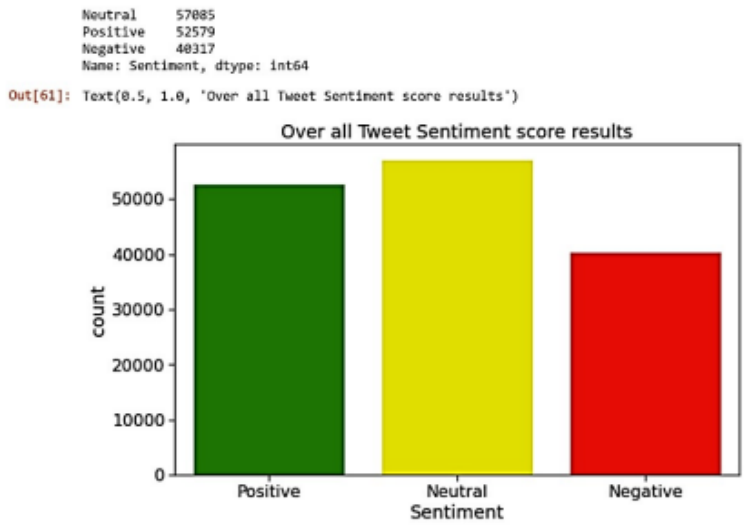


Figure 3: Tamaño del conjunto de datos

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.

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 covid19 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.3572023

[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," Journal 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: https://doi.org/10.2196/21978

[13] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," Information Processing y 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] ——, "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