COVID-19 VACCINES TWEETS SENTIMENT ANALYSIS

Daniel Martínez
Emiliano Etienne
Josué Maldonado
Xavier Rocabado

TABLA DE CONTENIDOS

FORMULACIÓN

Tras un largo periodo de lucha contra el Covid-19, el desarrollo de vacunas despertó una amplia gama de sentimientos y abrió las puertas a discusiones sobre el impacto de las mismas.

¿Qué opinan los ciudadanos de Estados Unidos sobre las vacunas?

SENTIMIENTOS

Positivos

Mediante el análisis de tweets podremos conocer la proporción de reacciones positivas o negativas en relación a las vacunas.

Negativos

Al conocer la evolución de esta proporción en un periodo determinado, podremos observar el impacto de noticias en los sentimientos de la población.

DATOS

Los datos se obtuvieron del API de Twitter y del dataset proporcionado y cuentan con las siguientes características seleccionadas para el análisis:

- Dataset:
  - User_name: Twitter handle [string]
  - User_followers: Número de seguidores [int]
  - User_friends: Número de amigos [int]
  - Date: Fecha y hora del tweet [datetime]
  - Text: Contenido del tweet [string]
  - Hashtags: Hashtags utilizados en el tweet [string]
  - Retweets: Retweets recibidos hasta la fecha de extracción de la información [int]
  - Favourites: Número de likes obtenidos hasta la fecha de extracción de la información [int]
  - Verified: Si el usuario está verificado o no [boolean]
- API:
  - Country_code: País de donde provienen los tweets [string]
  - Place_full_name: Nombre de donde se publicó el tweet (Ciudad, Estado) [string]
  - Lat: Latitud [float]
  - Lng: Longitud [float]

Data set de entrenamiento

Obtenidos del API de twitter y la información brindada.

53,188 obs

LIMPIEZA

Para realizar el análisis de sentimientos encontrados en los tweets, se realizó una limpieza del dataset. Se removieron:

```
data.text = data.text.str.lower()
data.text = data.text.apply(lambda x:re.sub('@[^\s]+','',x)) #Remove twitter handlers
data.text = data.text.apply(lambda x:re.sub(r'\B#\S+','',x)) #Remove hashtags
data.text = data.text.apply(lambda x:re.sub(r"http\S+", "", x)) # Remove URLS
data.text = data.text.apply(lambda x:' '.join(re.findall(r'\w+', x))) # Remove all the special characters
data.text = data.text.apply(lambda x:re.sub(r'\s+[a-zA-Z]\s+', ' ', x)) #remove all single characters
data.text = data.text.apply(lambda x:re.sub(r'\s+', ' ', x, flags=re.I)) # Substituting multiple spaces with single space
```

Caracteres innecesarios

Incluyendo espacios dobles e hipervínculos

Información adicional

Twitter handlers y hashtags

PREPARACIÓN

Adicionalmente, se crearon variables que nos permitan conocer información más detallada de la locación de los usuarios al momento de crear el tweet, utilizando el dataset *uscities.csv*:

```
[ ]  # Data preparation
     f_data = f_data[f_data.place_type=='city']
     f_data.drop(columns=['sentiments'],inplace=True)
     us_cities=us_cities[['city','state_id','state_name','county_name','lat','lng','id']]
     us_cities['place_full_name']=us_cities['city']+', '+us_cities['state_id']
     final_df = f_data.merge(us_cities,on='place_full_name',how='left')
     final_df['date']=pd.to_datetime(final_df['date'])
     final_df['round_date']=final_df['date'].dt.floor('h') #Fecha por hora utilizada en el dashboard
     final_df.drop(columns=['id','Positive Sentiment','Neutral Sentiment','Negative Sentiment'],inplace=True)
```

Obtener Latitud y Longitud

En base de los nombres de las ciudades

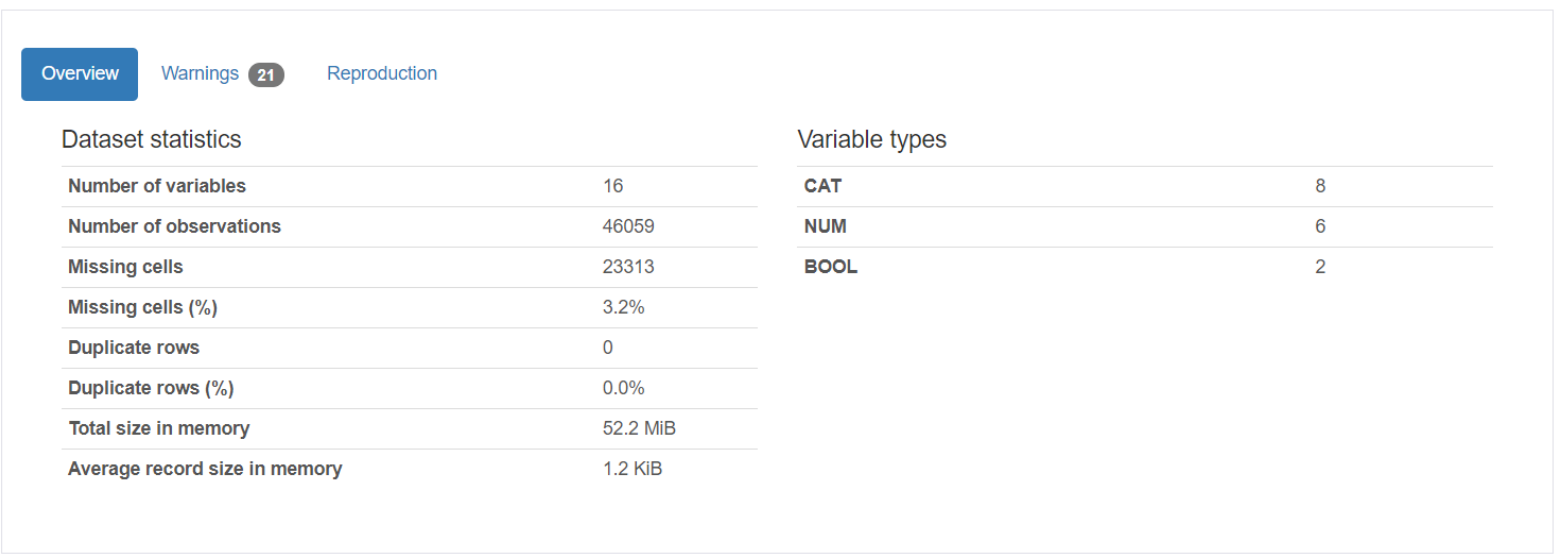
Merge

Entre la información de los tweets y las ciudades

ANÁLISIS EXPLORATORIO

El análisis exploratorio completo puede encontrarse aquí: https://drive.google.com/file/d/19pQ9B4xyscxF3E0WY5J9EWpAOCa34kEa/view?usp=sharing

Información general del dataset.



Overview | Warnings 21 | Reproduction

Dataset statistics

Number of variables	16
Number of observations	46059
Missing cells	23313
Missing cells (%)	3.2%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	52.2 MiB
Average record size in memory	1.2 KiB

Variable types

CAT	8
NUM	6
BOOL	2

ANÁLISIS EXPLORATORIO

Observamos información general de las variables más importantes del dataset.
Sobre el número de followers:

user_followers
Real number ($\mathbb{R}_{\geq 0}$)

Distinct	9752	**Mean**	103497.3731
Distinct (%)	21.2%	**Minimum**	0
Missing	0	**Maximum**	14919786
Missing (%)	0.0%	**Zeros**	335
Infinite	0	**Zeros (%)**	0.7%
Infinite (%)	0.0%	**Memory size**	360.0 KiB

Sobre los hashtags:

hashtags
Categorical

HIGH CARDINALITY
MISSING

Distinct	16835
Distinct (%)	46.5%
Missing	9816
Missing (%)	21.3%
Memory size	360.0 KiB

ANÁLISIS EXPLORATORIO

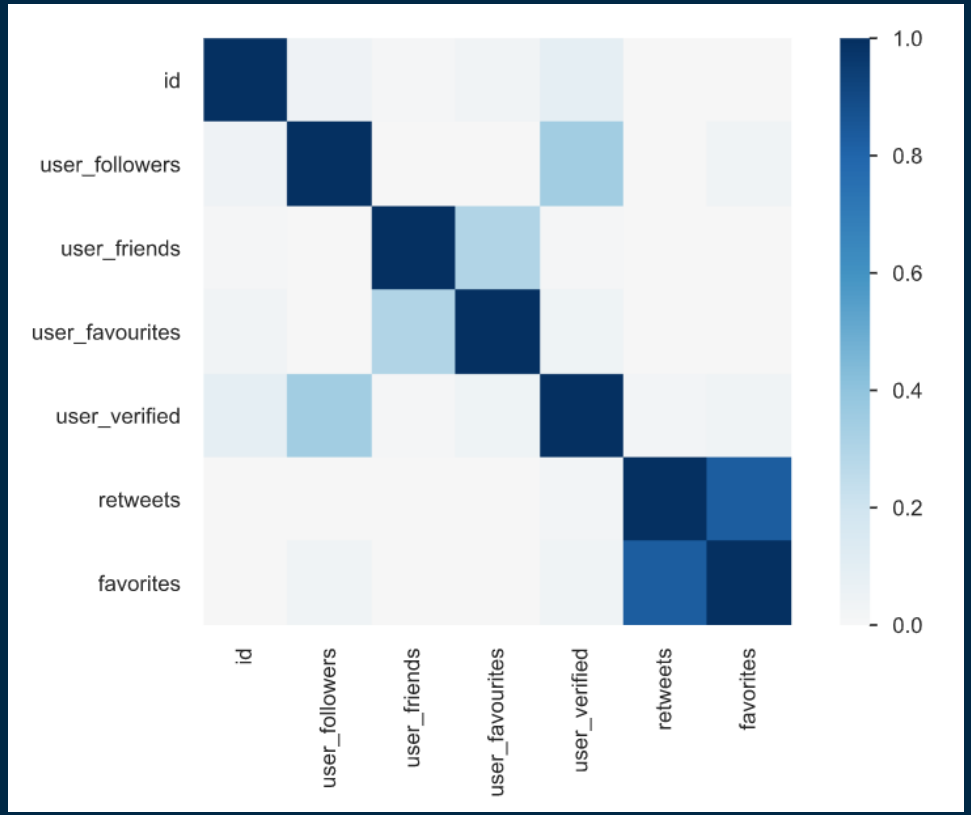
Usuarios verificados:

user_verified
Boolean

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	45.1 KiB

Correlaciones

Correlación Phik:



Análisis de sentimiento

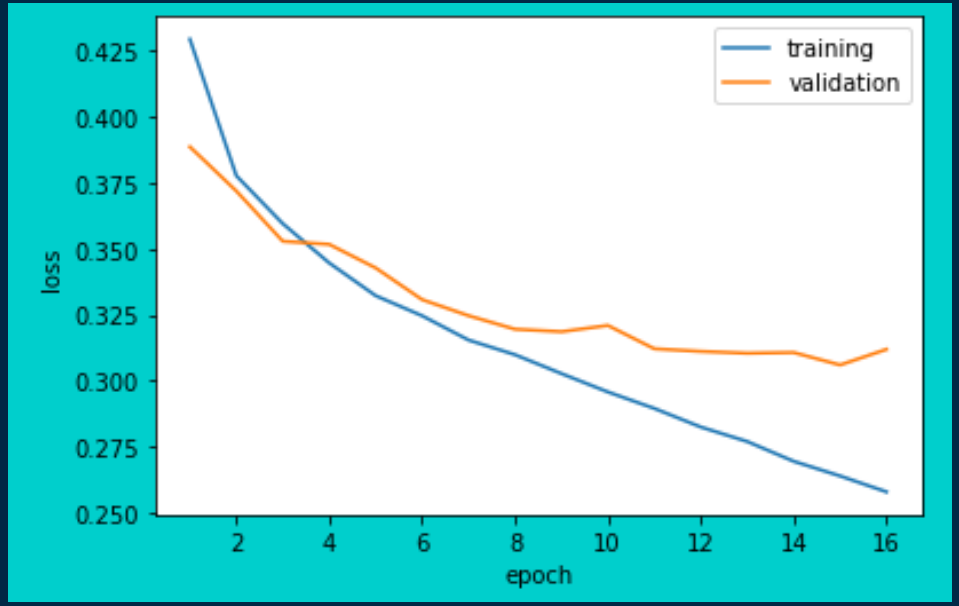
- Se utilizó el clasificador Vader de nltk para dar las labels al data set dado.
- Utilizamos datos de la Api de twitter para nivelar el dataset.
- Utilizamos un word2vect pre-entrenado de keras para convertir las palabras a vectores.
- Se entrenó una Long Short Term Memory Neural network.

Resultados.

	ACCURACY	Loss function
Red neuronal con neutral	70%	Categorical cross entropy
Red neuronal sin neutral	86%	Binary cross entropy

RESULTADO DEL MODELO

Gráfica de pérdida por validación y entrenamiento del modelo durante 16 epochs.



88% accuracy (training-set)

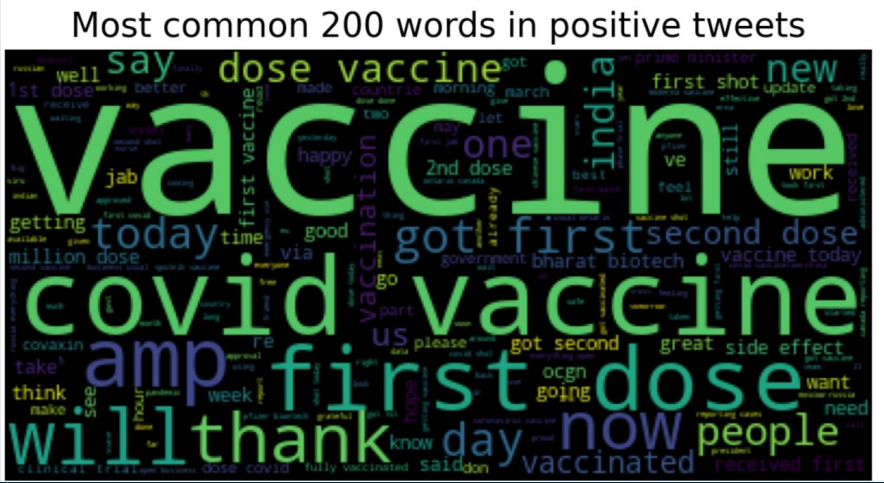
86% validation accuracy (test-set)

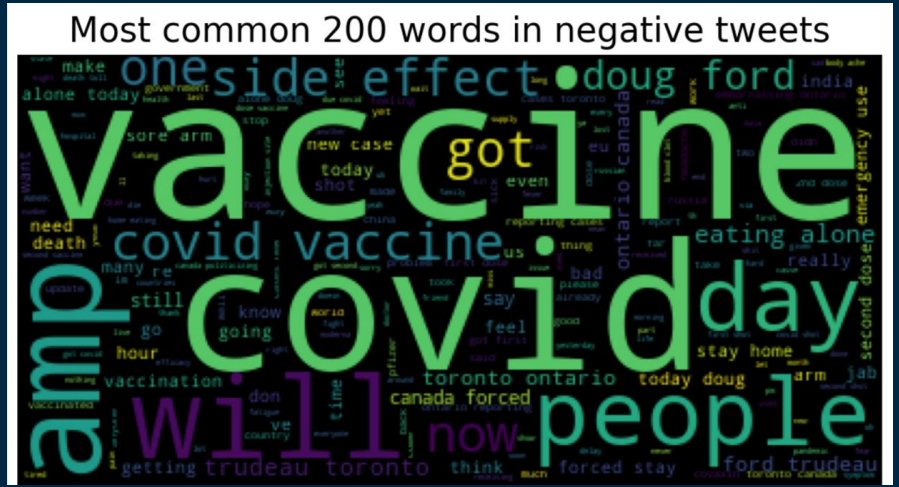
```
Epoch 16/16
1164/1164 [==============================] - 7s 6ms/step - loss: 0.2542 - accuracy: 0.8886 - val_loss: 0.3119 - val_accuracy: 0.8688
```

INSIGHTS DEL MODELO

- Utilizar 16 EPOCs nos brinda el mejor resultado.
- El modelo funciona mejor sin valores neutrales [positivos, negativos].

PALABRAS MÁS USADAS POR SENTIMIENTO





STREAMLINE

De visualización en un dashboard diseñado en Power BI, utilizando múltiples herramientas de Azure.

Next steps.

Una app con azure que nos de el sentimiento en tiempo real.

Twitter API

Azure Databricks

Event Hub

Analytics stream

Blob storage

Gracias