Actividad NLP 2.0

Alan Ricardo Vilchis Arceo

En esta actividad se quiere analizar un corpus de oraciones para conocer qué sentimiento emite dicho texto, utilizando un modelo preentrenado.

```
!pip install -q transformers
```

```
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7.7/7.7 MB 49.4 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 302.0/302.0 kB 23.4 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 3.8/3.8 MB 69.0 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.3/1.3 MB 61.0 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 295.0/295.0 kB 17.4 MB/s eta 0:00:00
```

```
# Importar librerias
import pandas as pd
import csv
```

El corpus que elegí para realizar la actividad es un dataset de tweets, los cuales están etiquetados en tres categorías: positivos, negativos y neutros. Se obtuvieron de Kaggle (https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset/data). Estos tweets son comentarios sobre el primer ministro de la India, Narendra Modi.

```
# De la libreria de Hugging Face se importa el modulo pipeline y primero se utiliza el modelom default para el analisis de sentimientos
from transformers import pipeline
sentiment_pipeline = pipeline("sentiment-analysis")

csv_file = "Twitter_Data.csv"

# Leer csv
data = pd.read_csv(csv_file)

# Separa entre tweet y sentimiento del tweet
first_column = data.iloc[:, 0]
second_column = data.iloc[:, 1]

# Se seleccionan los primeros 10 registros del dataset para evaluarlos
first_10_elements = first_column[:10]
second_10_elements = second_column[:10]

# Mapeo para mostrar que tipo de sentimiento equivale cada valor numerico
sentiment_mapping = {-1: "negative", 0: "neutral", 1: "positive"}

# Se evalua y se muestra el modelo
for i, (text, sentiment) in enumerate(zip(first_10_elements, second_10_elements)):
    print(f"Row {i + 1}")
    print("First Column (Text):", text)
    # Convert numerical sentiment value to corresponding string
    sentiment_str = sentiment_mapping.get(sentiment, "Unknown")
    print("Second Column (Sentiment):", sentiment_str)
    results = sentiment_pipeline(text)
    print("Sentiment Analysis Results:", results)
    print()
```

```
No model was supplied, defaulted to distilbert-base-uncased-finetuned-sst-2-english and revision af0f99b (https://huggingface.co/distilb
Using a pipeline without specifying a model name and revision in production is not recommended.
Row 1
First Column (Text): when modi promised “minimum government maximum governance” expected him begin the difficult job reforming the state
Second Column (Sentiment): negative
Sentiment Analysis Results: [{'label': 'NEGATIVE', 'score': 0.9961429238319397}]

Row 2
First Column (Text): talk all the nonsense and continue all the drama will vote for modi
Second Column (Sentiment): neutral
Sentiment Analysis Results: [{'label': 'NEGATIVE', 'score': 0.9936436414718628}]

Row 3
First Column (Text): what did just say vote for modi  welcome bjp told you rahul the main campaigner for modi think modi should just rel
Second Column (Sentiment): positive
Sentiment Analysis Results: [{'label': 'NEGATIVE', 'score': 0.9632664918899536}]
```

```
Row 4
First Column (Text): asking his supporters prefix chowkidar their names modi did great service now there confusion what read what not nc
Second Column (Sentiment): positive
Sentiment Analysis Results: [{'label': 'NEGATIVE', 'score': 0.9983959794044495}]

Row 5
First Column (Text): answer who among these the most powerful world leader today trump putin modi may
Second Column (Sentiment): positive
Sentiment Analysis Results: [{'label': 'POSITIVE', 'score': 0.8206459283828735}]

Row 6
First Column (Text): kiya tho refresh maarkefir comment karo
Second Column (Sentiment): neutral
Sentiment Analysis Results: [{'label': 'POSITIVE', 'score': 0.9960861206054688}]

Row 7
First Column (Text): surat women perform yagna seeks divine grace for narendra modi become again

Second Column (Sentiment): neutral
Sentiment Analysis Results: [{'label': 'POSITIVE', 'score': 0.844713568687439}]

Row 8
First Column (Text): this comes from cabinet which has scholars like modi smriti and hema time introspect
Second Column (Sentiment): neutral
Sentiment Analysis Results: [{'label': 'NEGATIVE', 'score': 0.9508270621299744}]

Row 9
First Column (Text): with upcoming election india saga going important pair look current modi leads govt elected with deal brexit combir
Second Column (Sentiment): positive
Sentiment Analysis Results: [{'label': 'POSITIVE', 'score': 0.9950506091117859}]

Row 10
First Column (Text): gandhi was gay does modi
Second Column (Sentiment): positive
Sentiment Analysis Results: [{'label': 'NEGATIVE', 'score': 0.8542367815971375}]
```

Con el fin de profundizar más en la actividad, decidí buscar otro modelo más adecuado y realizar las mismas pruebas. Encontré el siguiente enlace: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest, el cual fue entrenado con 124 millones de tweets. Además, este modelo tiene las mismas categorías que el dataset de tweets de la India (positivo, neutral y negativo).

```
# Using a specific model for sentiment analysis
specific_model = pipeline(model="cardiffnlp/twitter-roberta-base-sentiment-latest")

# Perform sentiment analysis on each of the first 10 elements
for i, (text, sentiment) in enumerate(zip(first_10_elements, second_10_elements)):
    print(f"Row {i + 1}")
    print("First Column (Text):", text)
    # Convert numerical sentiment value to corresponding string
    sentiment_str = sentiment_mapping.get(sentiment, "Unknown")
    print("Second Column (Sentiment):", sentiment_str)
    results = specific_model(text)
    print("Sentiment Analysis Results:", results)
    print()
```

```
Some weights of the model checkpoint at cardiffnlp/twitter-roberta-base-sentiment-latest were not used when initializing RobertaForSeque
- This IS expected if you are initializing RobertaForSequenceClassification from the checkpoint of a model trained on another task or wi
- This IS NOT expected if you are initializing RobertaForSequenceClassification from the checkpoint of a model that you expect to be exa
Row 1
First Column (Text): when modi promised “minimum government maximum governance” expected him begin the difficult job reforming the state
Second Column (Sentiment): negative
Sentiment Analysis Results: [{'label': 'negative', 'score': 0.7419230341911316}]

Row 2
First Column (Text): talk all the nonsense and continue all the drama will vote for modi
Second Column (Sentiment): neutral
Sentiment Analysis Results: [{'label': 'negative', 'score': 0.5076758861541748}]

Row 3
First Column (Text): what did just say vote for modi  welcome bjp told you rahul the main campaigner for modi think modi should just rel
Second Column (Sentiment): positive
Sentiment Analysis Results: [{'label': 'neutral', 'score': 0.7030866742134094}]

Row 4
First Column (Text): asking his supporters prefix chowkidar their names modi did great service now there confusion what read what not nc
Second Column (Sentiment): positive
Sentiment Analysis Results: [{'label': 'negative', 'score': 0.518481969833374}]
```

```
Row 5
First Column (Text): answer who among these the most powerful world leader today trump putin modi may
Second Column (Sentiment): positive
Sentiment Analysis Results: [{'label': 'neutral', 'score': 0.6955121159553528}]

Row 6
First Column (Text): kiya tho refresh maarkefir comment karo
Second Column (Sentiment): neutral
Sentiment Analysis Results: [{'label': 'neutral', 'score': 0.8945183753967285}]

Row 7
First Column (Text): surat women perform yagna seeks divine grace for narendra modi become again

Second Column (Sentiment): neutral
Sentiment Analysis Results: [{'label': 'neutral', 'score': 0.6177760362625122}]

Row 8
First Column (Text): this comes from cabinet which has scholars like modi smriti and hema time introspect
Second Column (Sentiment): neutral
Sentiment Analysis Results: [{'label': 'neutral', 'score': 0.7386401295661926}]

Row 9
First Column (Text): with upcoming election india saga going important pair look current modi leads govt elected with deal brexit combir
Second Column (Sentiment): positive
Sentiment Analysis Results: [{'label': 'neutral', 'score': 0.5858953595161438}]

Row 10
First Column (Text): gandhi was gay does modi
Second Column (Sentiment): positive
Sentiment Analysis Results: [{'label': 'neutral', 'score': 0.71793133020401}]
```

Al realizar esta actividad, pude darme cuenta de que ambos modelos no son muy precisos. Lo que supongo que pudo haber ocurrido es justo lo que revisamos en clase sobre los diferentes lenguajes. Por ejemplo, el dataset que elegí consiste en tweets de la India que fueron traducidos al inglés. Probablemente, la traducción no fue del todo correcta, ya que al revisar los datos, algunos textos carecen de coherencia. Por otra parte, me resulta muy interesante cómo el segundo modelo, entrenado con una gran cantidad de tweets, no es muy preciso. Sin embargo, esto puede deberse a la calidad del dataset que estoy utilizando. Si hubiera elegido uno de mayor calidad, posiblemente habría obtenido mejores resultados. Asimismo, si la calidad de los tweets con los que se entrenaron los modelos hubiera sido mejor, los resultados habrían sido más precisos.