LITCOIN NPL CHALLENGE

Rsultados Iniciales del LitCoin NPL Challenge

PROBLEMA CIENTIFICO

El reto consiste en analizar los títulos y resúmenes de artículos biomédicos para identificar entidades como genes, enfermedades o productos químicos, y predecir las relaciones entre esas entidades, como asociaciones, correlaciones o interacciones. Los participantes deben entrenar modelos de procesamiento de lenguaje natural (NLP) para realizar esta tarea de forma automática, utilizando los datos proporcionados en formato CSV para entrenar y evaluar sus modelos.

Resultados Iniciales

OBJETIVOS

- Desarrollar un modelo que prediga con precisión las relaciones entre entidades biomédicas en resúmenes de artículos científicos, como asociaciones, correlaciones o interacciones.
- Evaluar y optimizar el rendimiento del modelo para garantizar que sea eficiente y escalable, permitiendo analizar grandes volúmenes de datos biomédicos.

LOS MODELOS A UTILIZAR

LSTM

RNN

Transformers

Por que

LTSM

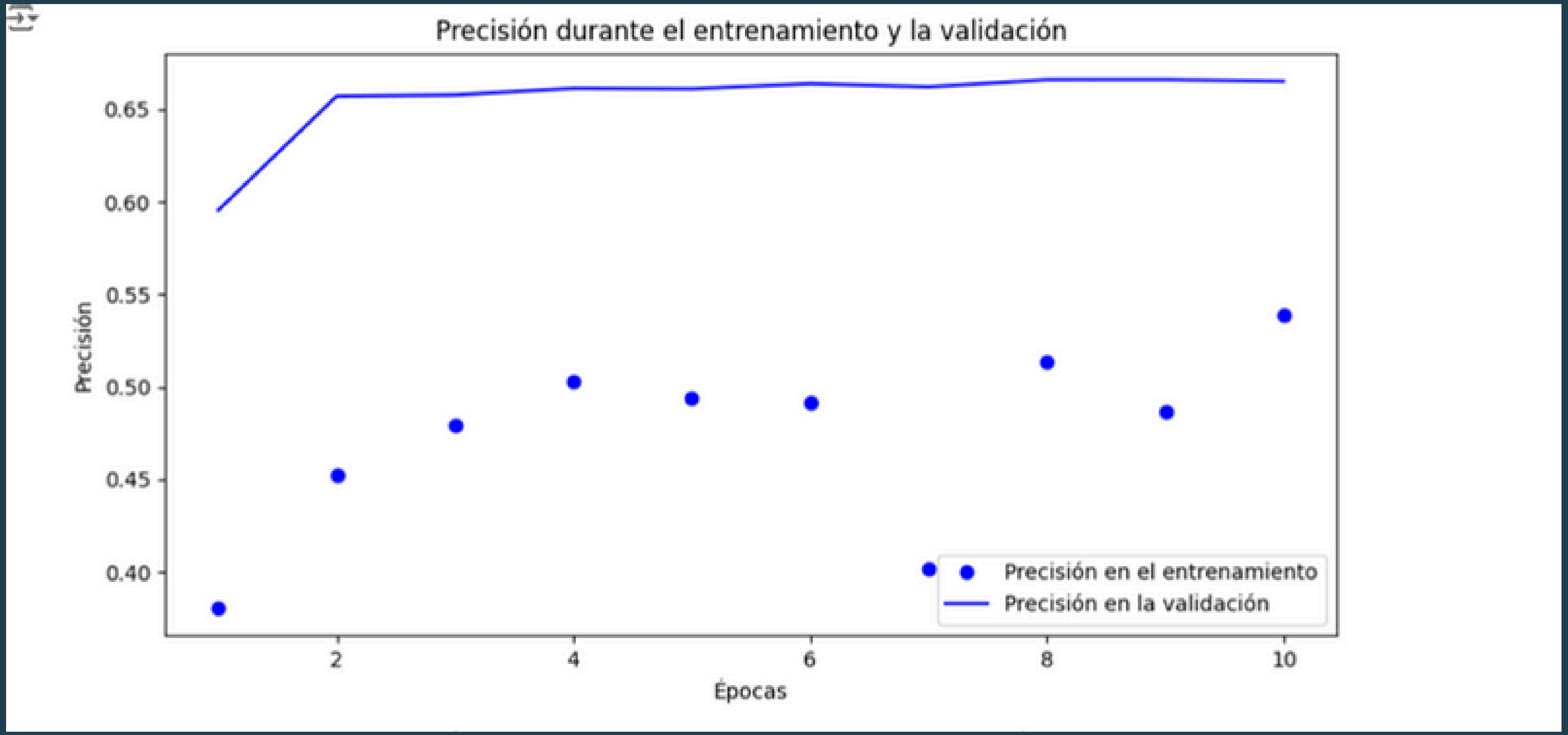
- Captura de dependencias a largo plazo
- Manejo efectivo de secuencias de texto
- Quitar valores nulos

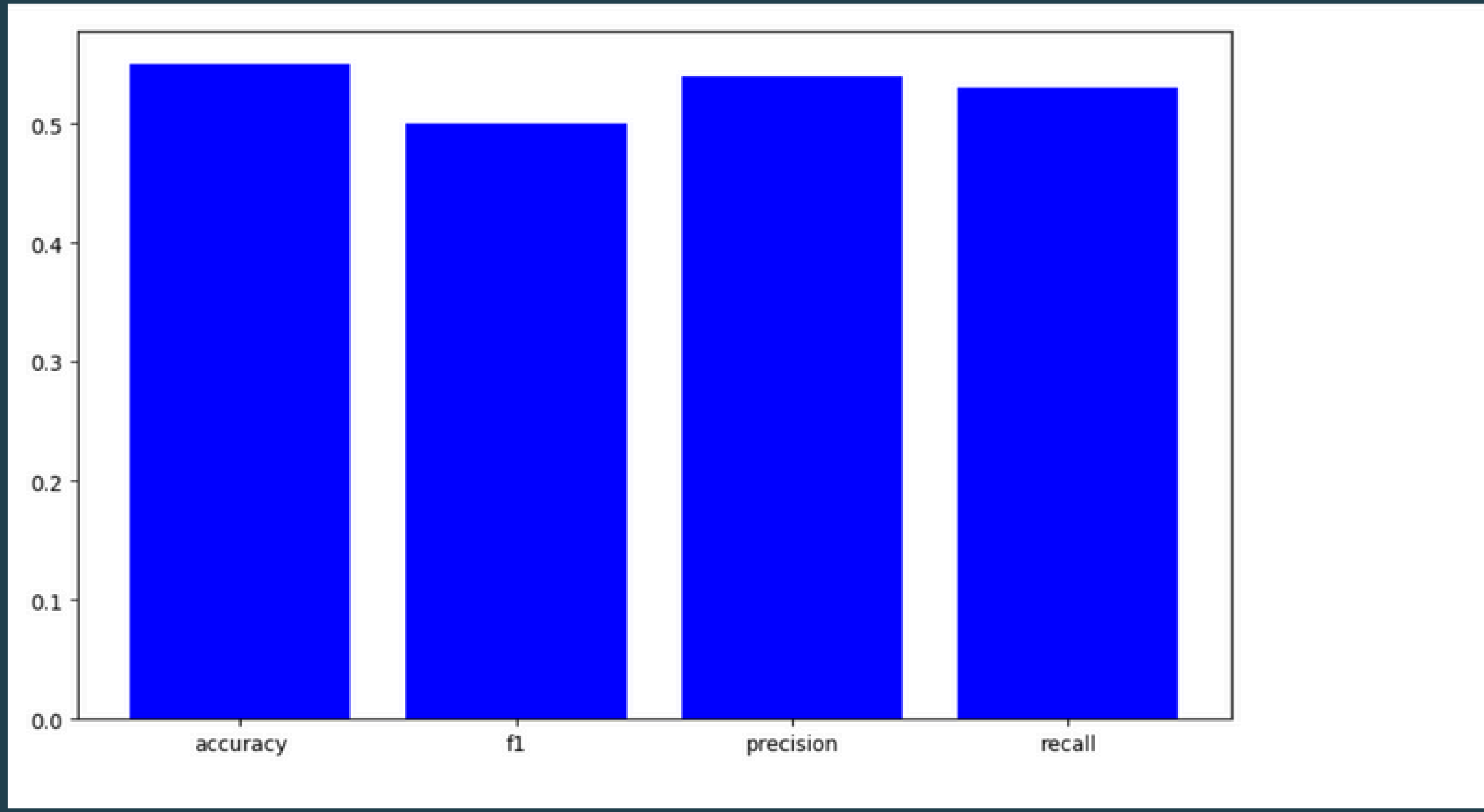
Resultados de

LTSM

Precisión Entrenamiento 66%

Precisión Validacion 55%





Parametros

LTSM

Model: "sequential_2"

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	?	0 (unbuilt)
bidirectional_1 (Bidirectional)	?	0 (unbuilt)
lstm_4 (LSTM)	?	0 (unbuilt)
dropout_2 (Dropout)	?	0 (unbuilt)
dense_2 (Dense)	?	0 (unbuilt)

```
 Total params: 0 (0.00 B)
 Trainable params: 0 (0.00 B)
 Non-trainable params: 0 (0.00 B)
Epoch 1/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 821s 439ms/step - accuracy: 0.5263 - loss: 1.3512 - val_accuracy: 0.3802 - val_loss: 1.4311
Epoch 2/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 855s 436ms/step - accuracy: 0.6571 - loss: 0.9120 - val_accuracy: 0.4521 - val_loss: 1.3434
Epoch 3/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 861s 435ms/step - accuracy: 0.6556 - loss: 0.8655 - val_accuracy: 0.4793 - val_loss: 1.4065
Epoch 4/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 891s 451ms/step - accuracy: 0.6593 - loss: 0.8360 - val_accuracy: 0.5026 - val_loss: 1.3387
Epoch 5/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 834s 436ms/step - accuracy: 0.6608 - loss: 0.8199 - val_accuracy: 0.4939 - val_loss: 1.3217
Epoch 6/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 865s 438ms/step - accuracy: 0.6648 - loss: 0.8059 - val_accuracy: 0.4916 - val_loss: 1.3124
Epoch 7/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 863s 438ms/step - accuracy: 0.6642 - loss: 0.7999 - val_accuracy: 0.4018 - val_loss: 1.2907
Epoch 8/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 860s 438ms/step - accuracy: 0.6653 - loss: 0.7979 - val_accuracy: 0.5137 - val_loss: 1.3434
Epoch 9/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 861s 437ms/step - accuracy: 0.6640 - loss: 0.7978 - val_accuracy: 0.4869 - val_loss: 1.3016
Epoch 10/10
```

Por qué

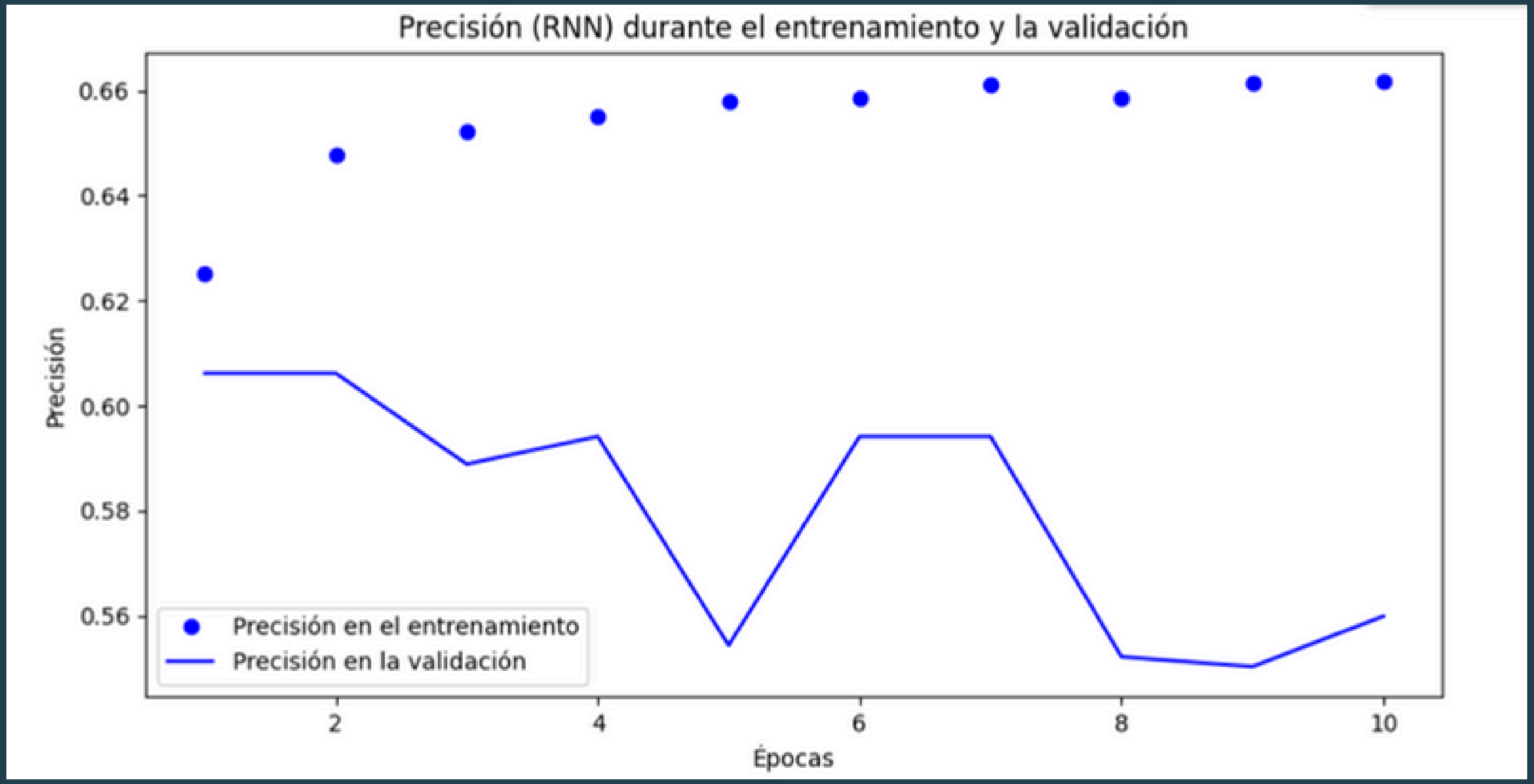
RNN

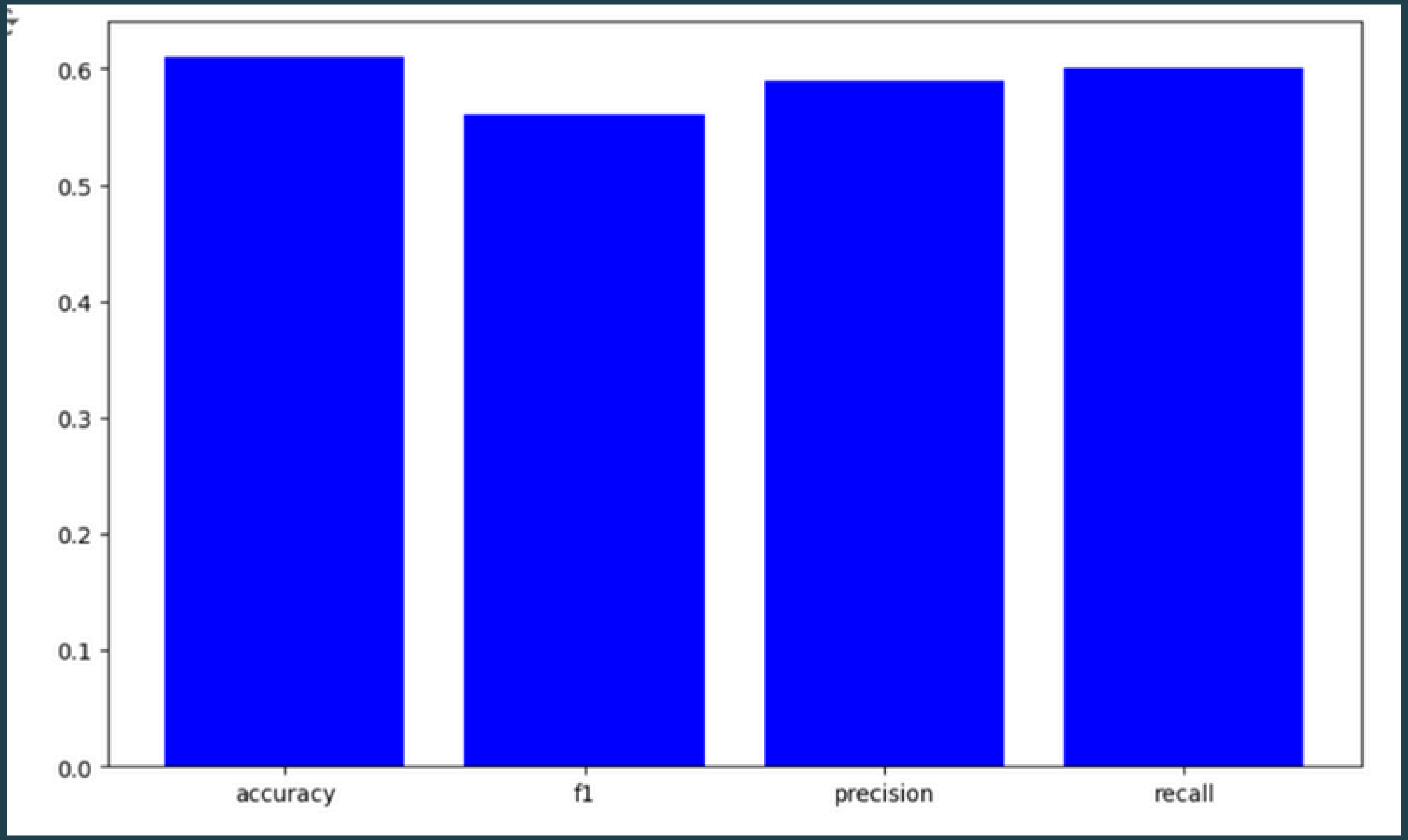
- Procesamiento secuencial de texto
- Captura de dependencias contextuales locales

Resultados de RNN

Precisión Entrenamiento: %66

Precisión Validacion 61%





RNN

bidirectional (Bidirectional)	?	0 (unbuilt)
simple_rnn_1 (SimpleRNN)	?	0 (unbuilt)
dropout (Dropout)	?	0 (unbuilt)
dense (Dense)	?	0 (unbuilt)

```
 Total params: 0 (0.00 B)
 Trainable params: 0 (0.00 B)
 Non-trainable params: 0 (0.00 B)
Epoch 1/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 298s 157ms/step - accuracy: 0.5731 - loss: 1.2182 - val_accuracy: 0.4835 - val_loss: 1.3446
Epoch 2/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 286s 155ms/step - accuracy: 0.6438 - loss: 0.8976 - val_accuracy: 0.4156 - val_loss: 1.2631
Epoch 3/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 324s 156ms/step - accuracy: 0.6541 - loss: 0.8499 - val_accuracy: 0.4578 - val_loss: 1.2461
Epoch 4/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 286s 155ms/step - accuracy: 0.6557 - loss: 0.8270 - val_accuracy: 0.4484 - val_loss: 1.2091
Epoch 5/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 287s 155ms/step - accuracy: 0.6588 - loss: 0.8194 - val_accuracy: 0.4277 - val_loss: 1.2207
Epoch 6/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 320s 155ms/step - accuracy: 0.6579 - loss: 0.8100 - val_accuracy: 0.4547 - val_loss: 1.2114
Epoch 7/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 321s 154ms/step - accuracy: 0.6598 - loss: 0.8058 - val_accuracy: 0.4330 - val_loss: 1.1945
Epoch 8/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 286s 155ms/step - accuracy: 0.6559 - loss: 0.8028 - val_accuracy: 0.4394 - val_loss: 1.2217
Epoch 9/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 286s 155ms/step - accuracy: 0.6637 - loss: 0.7938 - val_accuracy: 0.3866 - val_loss: 1.1973
Epoch 10/10
1848/1848 ━━━━━━━━━━━━━━━━━━━━ 323s 155ms/step - accuracy: 0.6595 - loss: 0.7940 - val_accuracy: 0.4263 - val_loss: 1.1788
```

Por qué

TRANSFORMERS

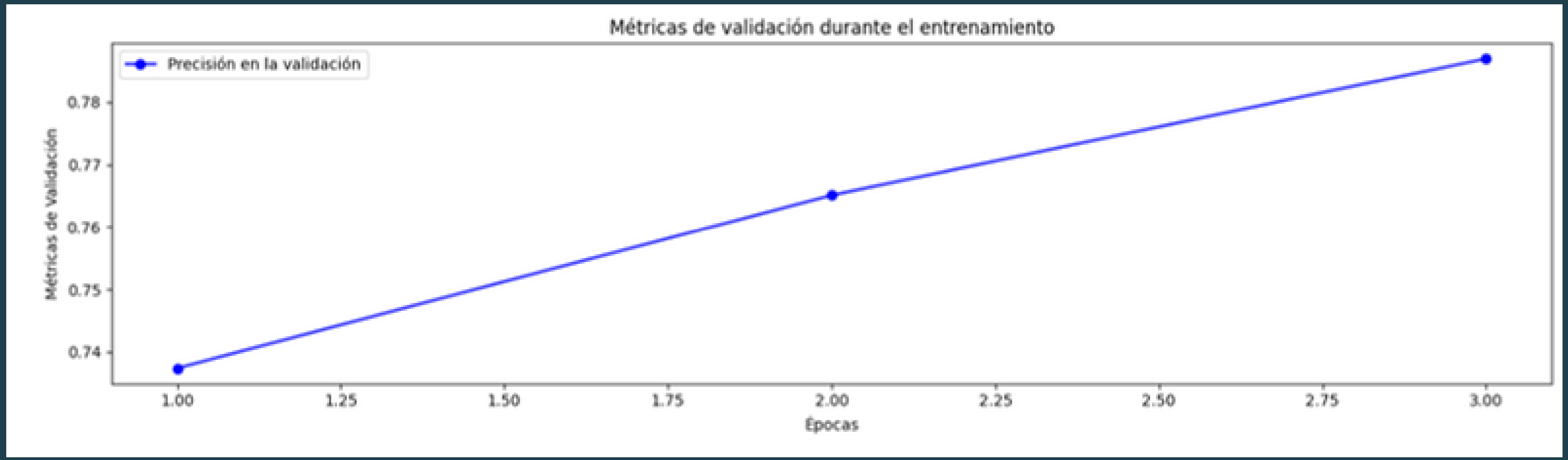
- Optimizado para datos secuenciales
- Mecanismos de atención para enfocar en las partes más relevantes

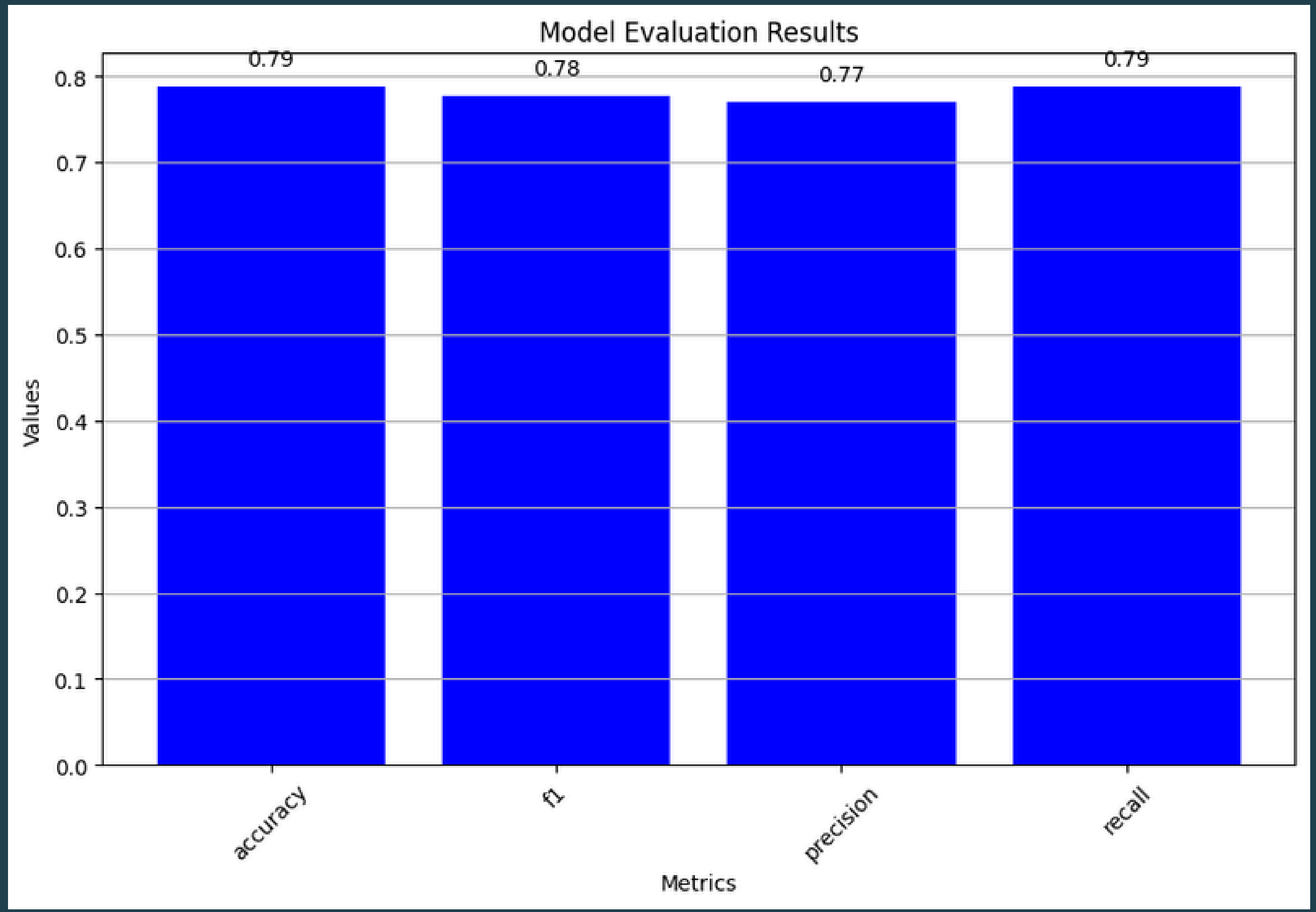
Resultados de

TRANSFORMERS

Precisión Entrenamiento: %78

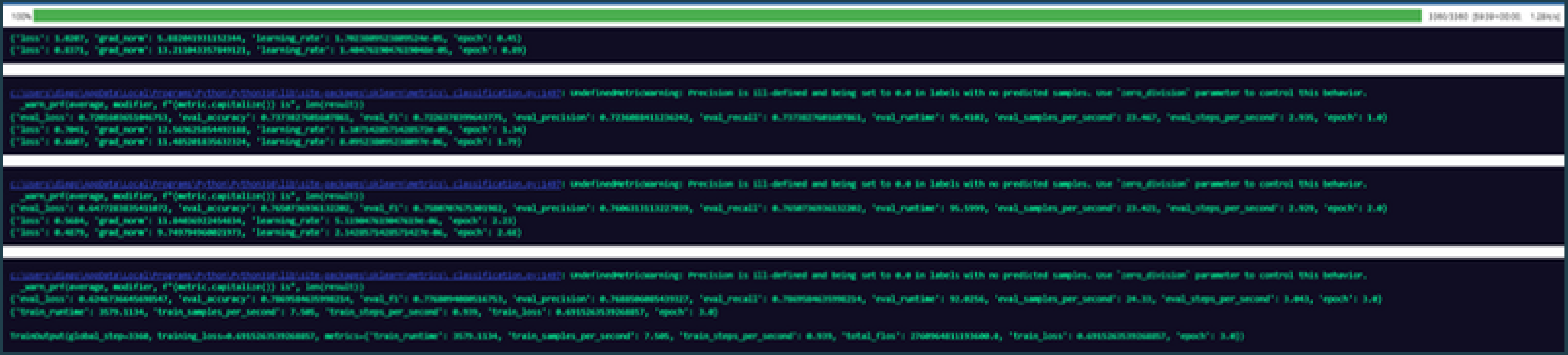
Precisión Validación: 79%





Valores de entrenamiento

TRANSFORMERS



REFERENCIAS

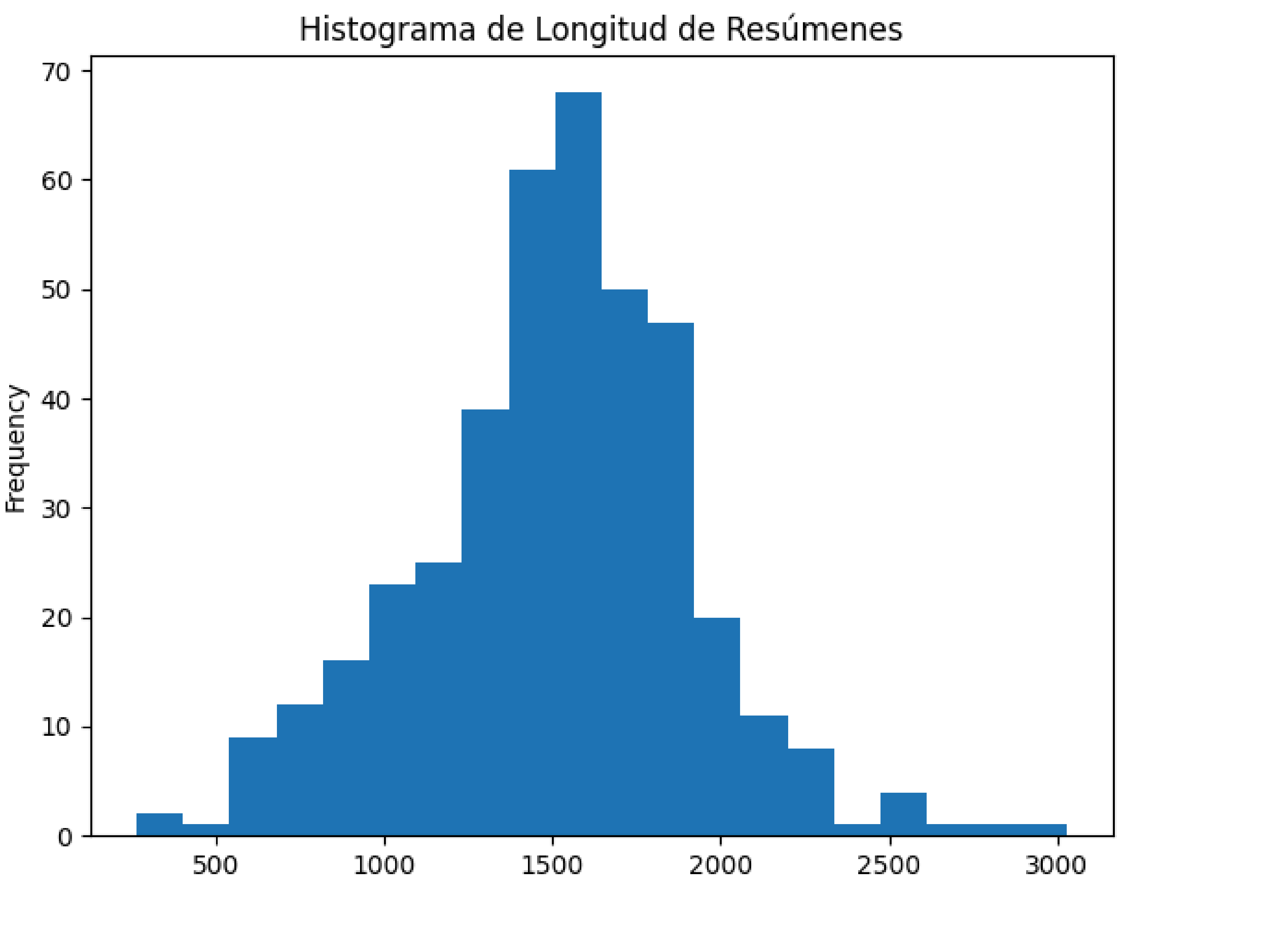
1. **Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J.** (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. Disponible en https://academic.oup.com/bioinformatics/article/36/4/1234/5566506?login=false.
2. **Muchene, L., & Safari, W.** (2021). Two-stage topic modelling of scientific publications: A case study of University of Nairobi, Kenya. *PLoS One, 16*(1), e0243208. Disponible en https://doi.org/10.1371/journal.pone.0243208.
3. **James, H.** (2022). RNNs and LSTMs. Disponible en https://web.stanford.edu/~jurafsky/slp3/9.pdf.

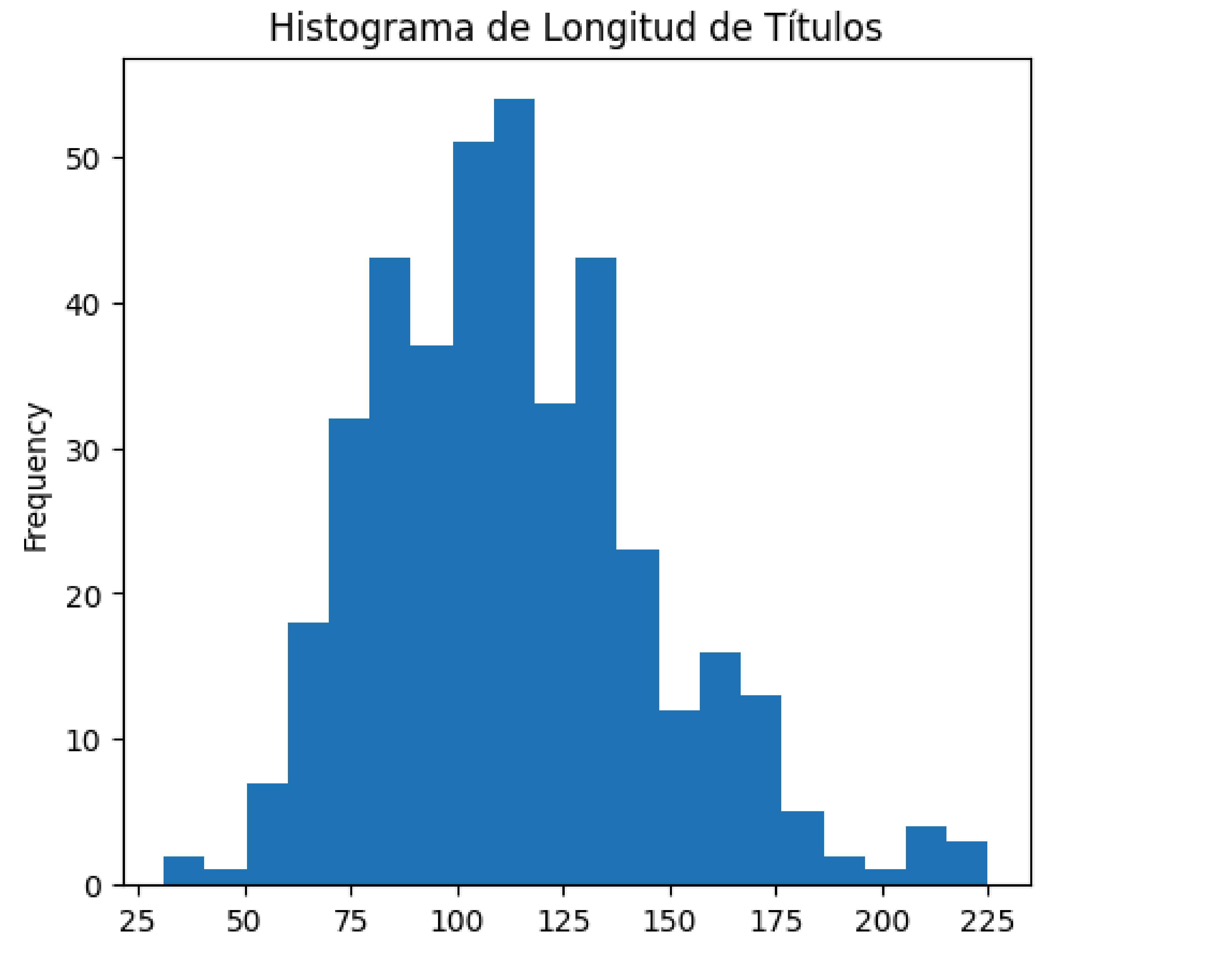
Analisis Exploratorio

ANALISIS DE LOS ABSTRACTS

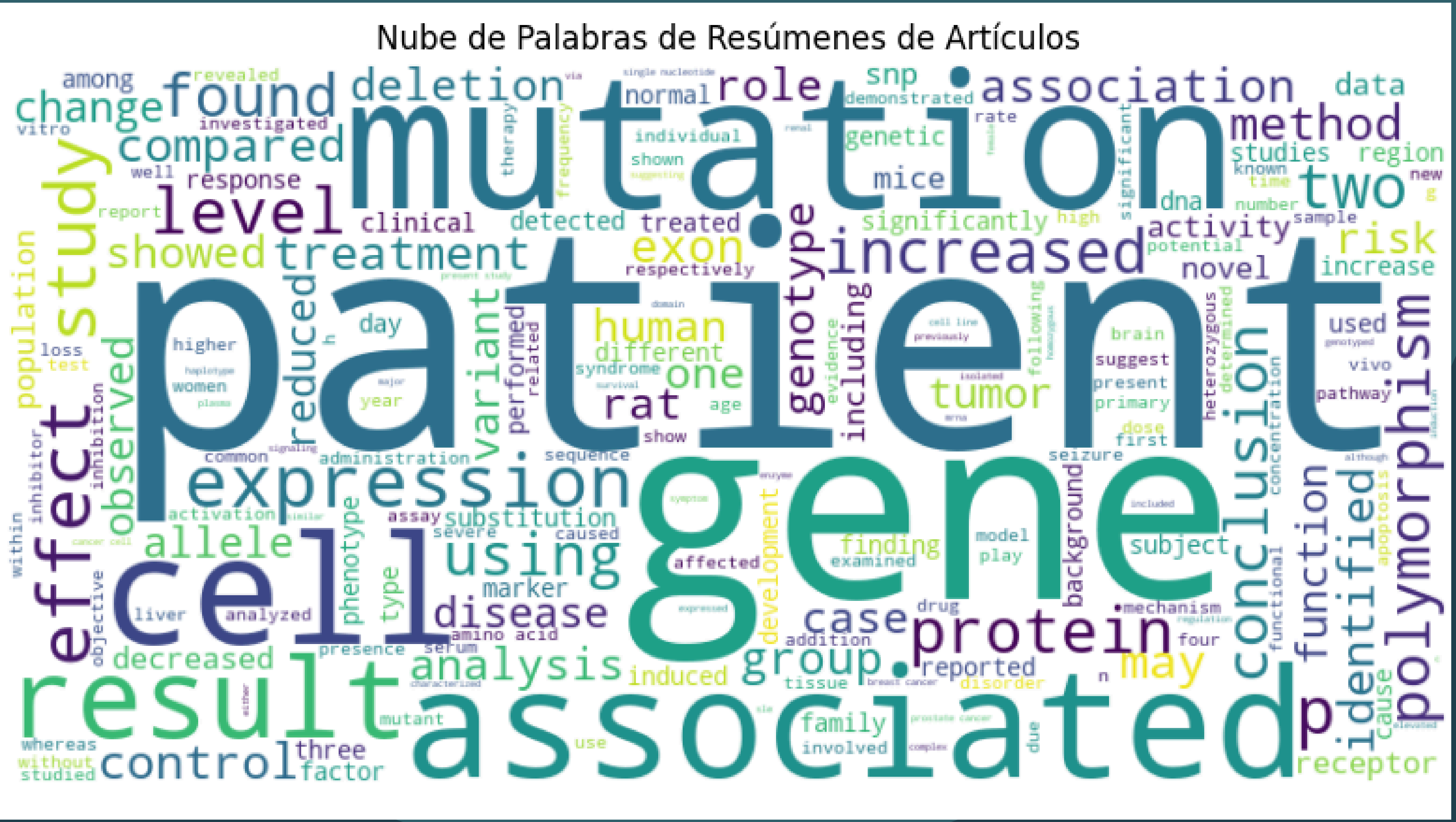
```
Resumen de variables:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 3 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   abstract_id  400 non-null    int64
 1   title        400 non-null    object
 2   abstract     400 non-null    object
dtypes: int64(1), object(2)
```

Analisis Exploratorio





Analisis Exploratorio

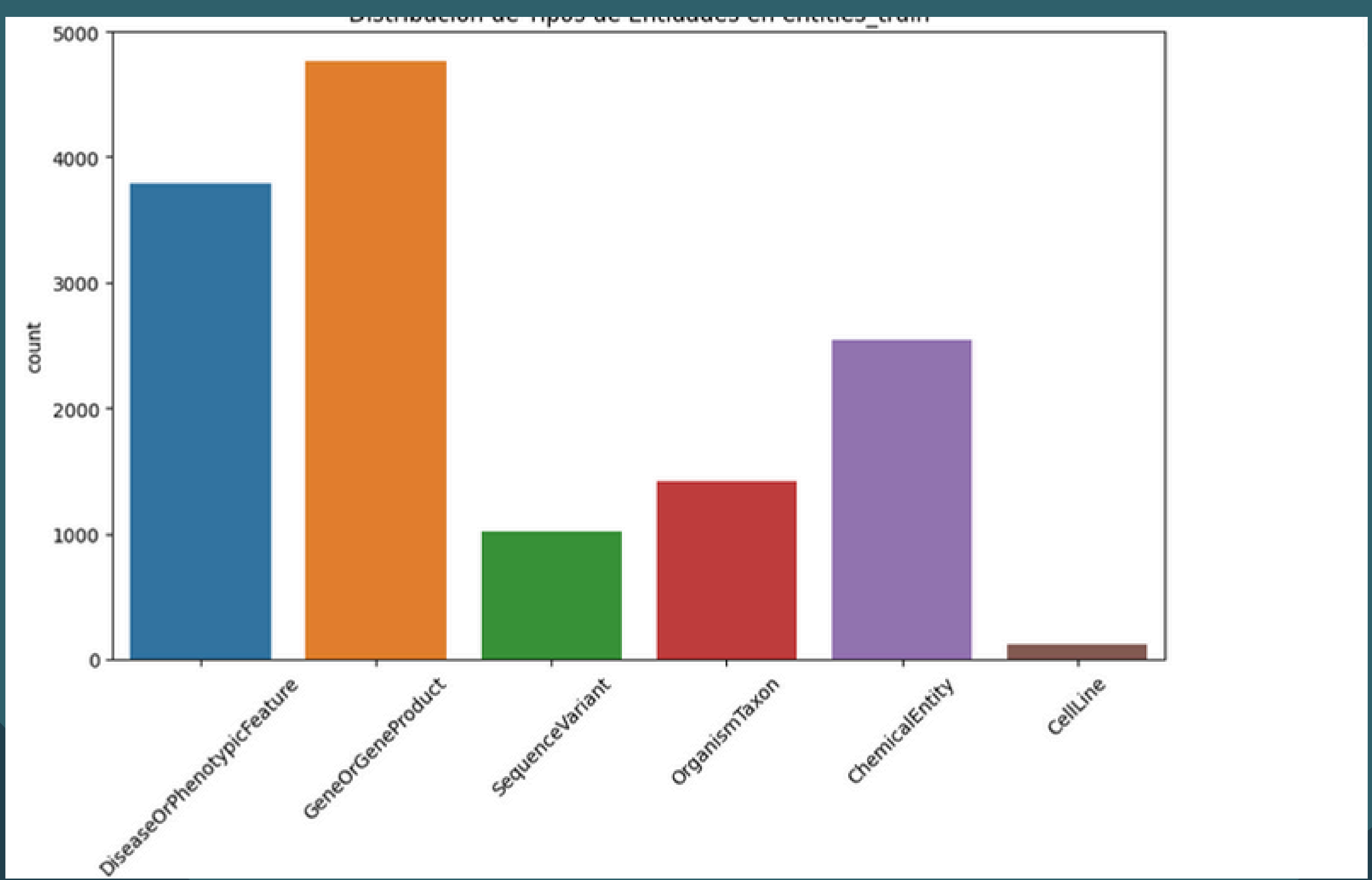


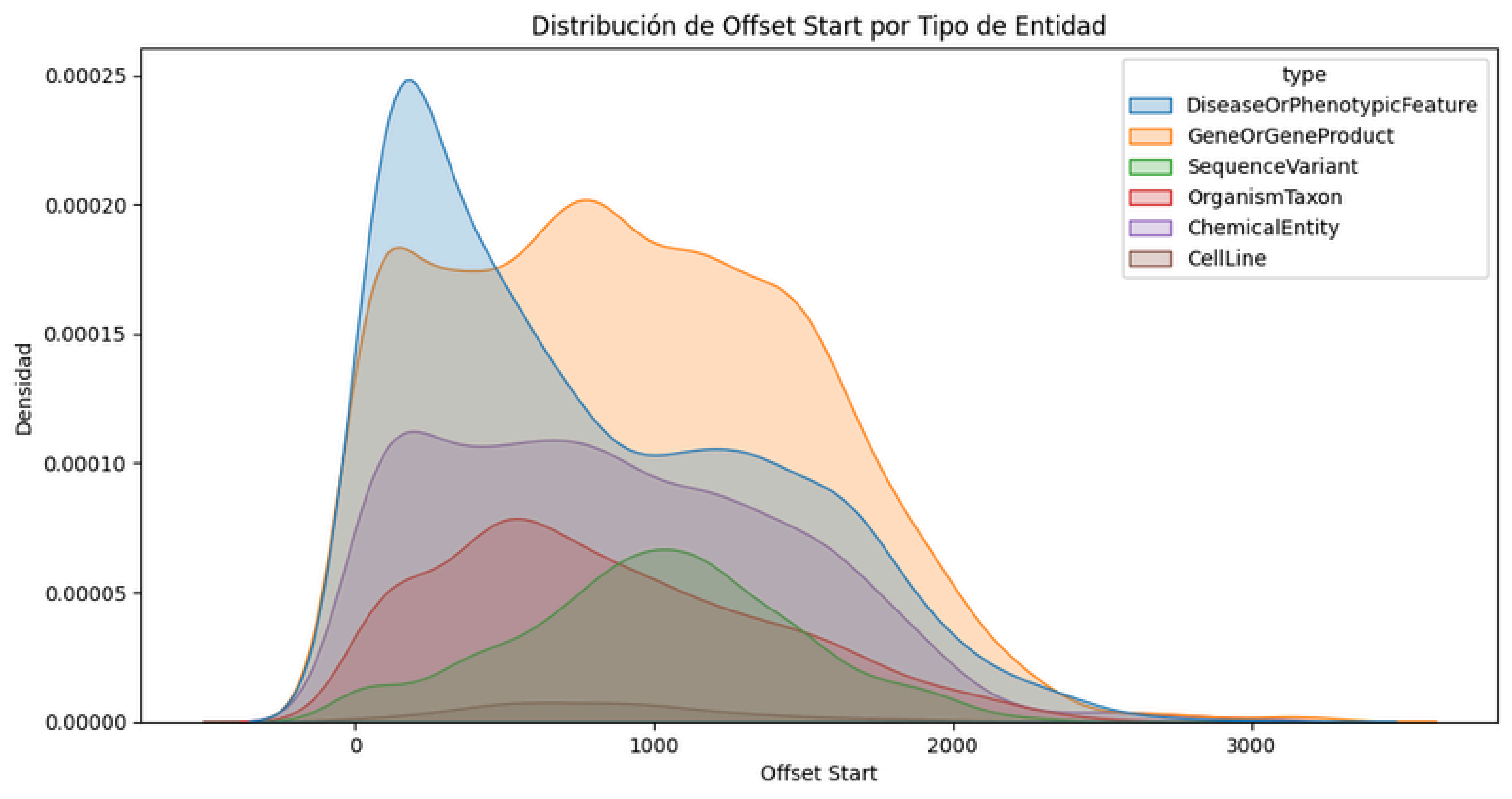
Analisis Exploratorio

ANALISIS DE LAS ENTIDADES

```
Resumen de variables:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13636 entries, 0 to 13635
Data columns (total 7 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   id             13636 non-null  int64 
 1   abstract_id    13636 non-null  int64 
 2   offset_start   13636 non-null  int64 
 3   offset_finish  13636 non-null  int64 
 4   type           13636 non-null  object
 5   mention        13636 non-null  object
 6   entity_ids     13636 non-null  object
dtypes: int64(4), object(3)
memory usage: 745.8+ KB
None
```

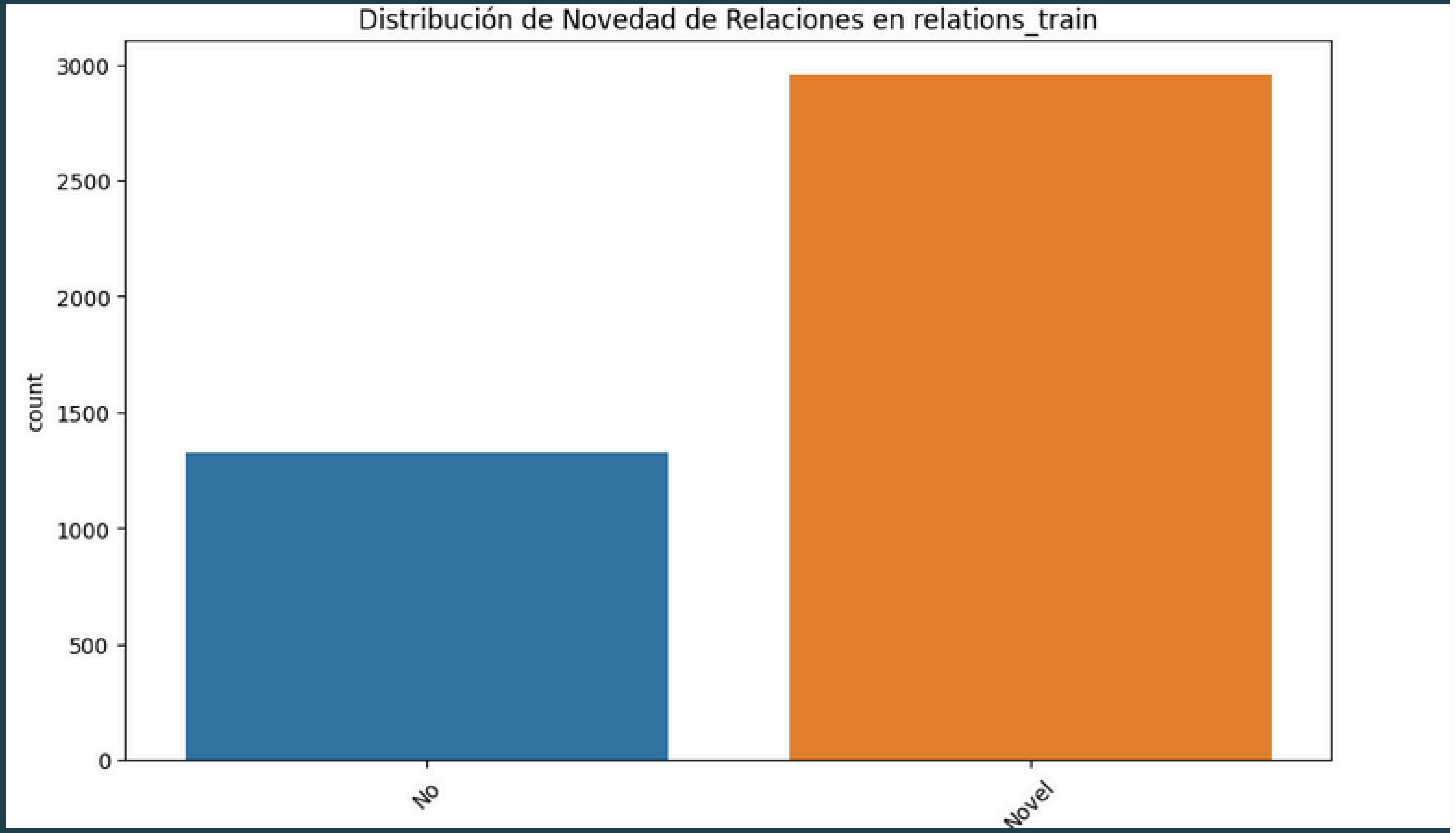
Ajustes en los datos

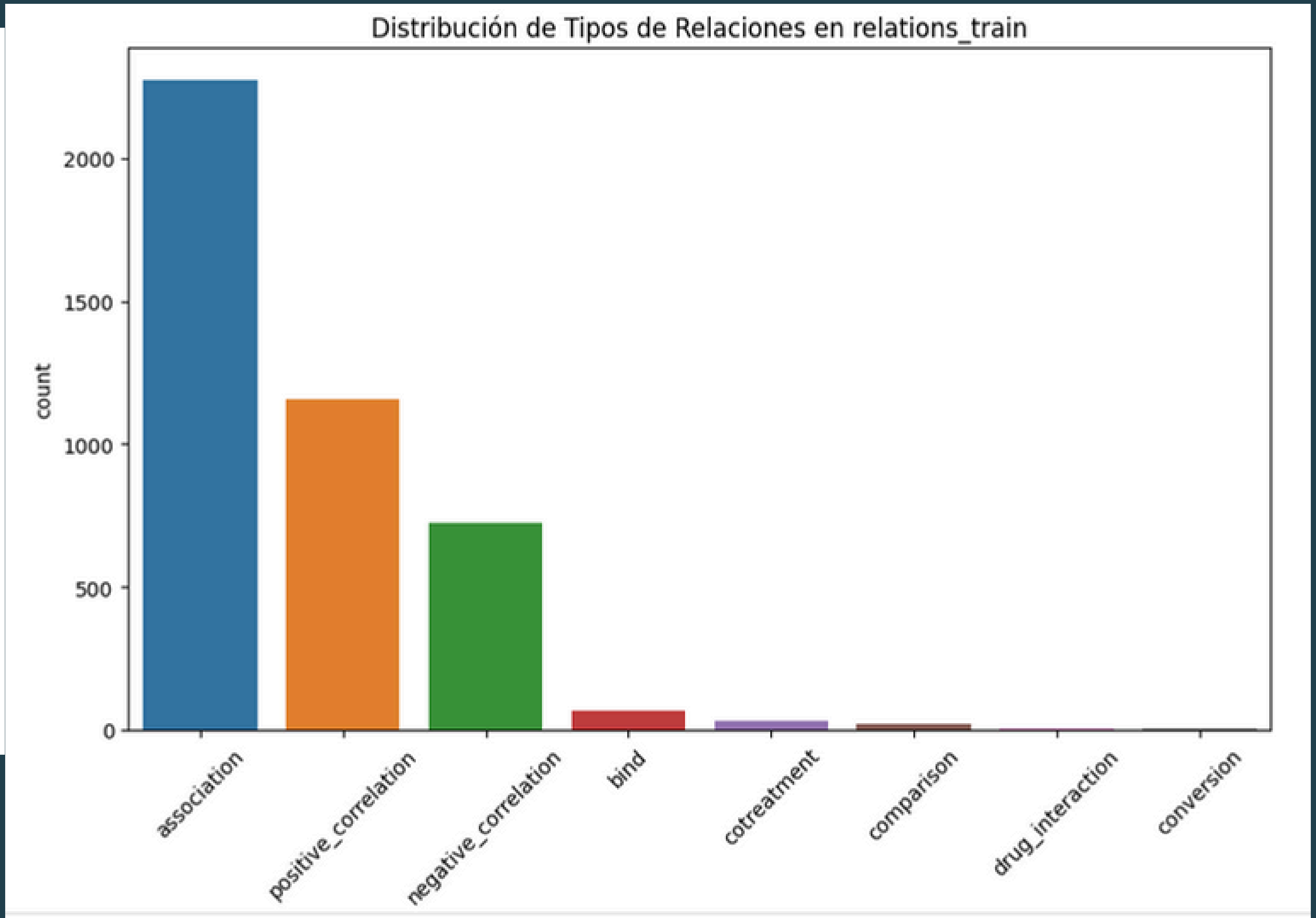




Distribución de Offset Start por Tipo de Entidad
type
DiseaseOrPhenotypicFeature
GeneOrGeneProduct
SequenceVariant
OrganismTaxon
ChemicalEntity
CellLine
Densidad
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
1000
2000
3000
Offset Start

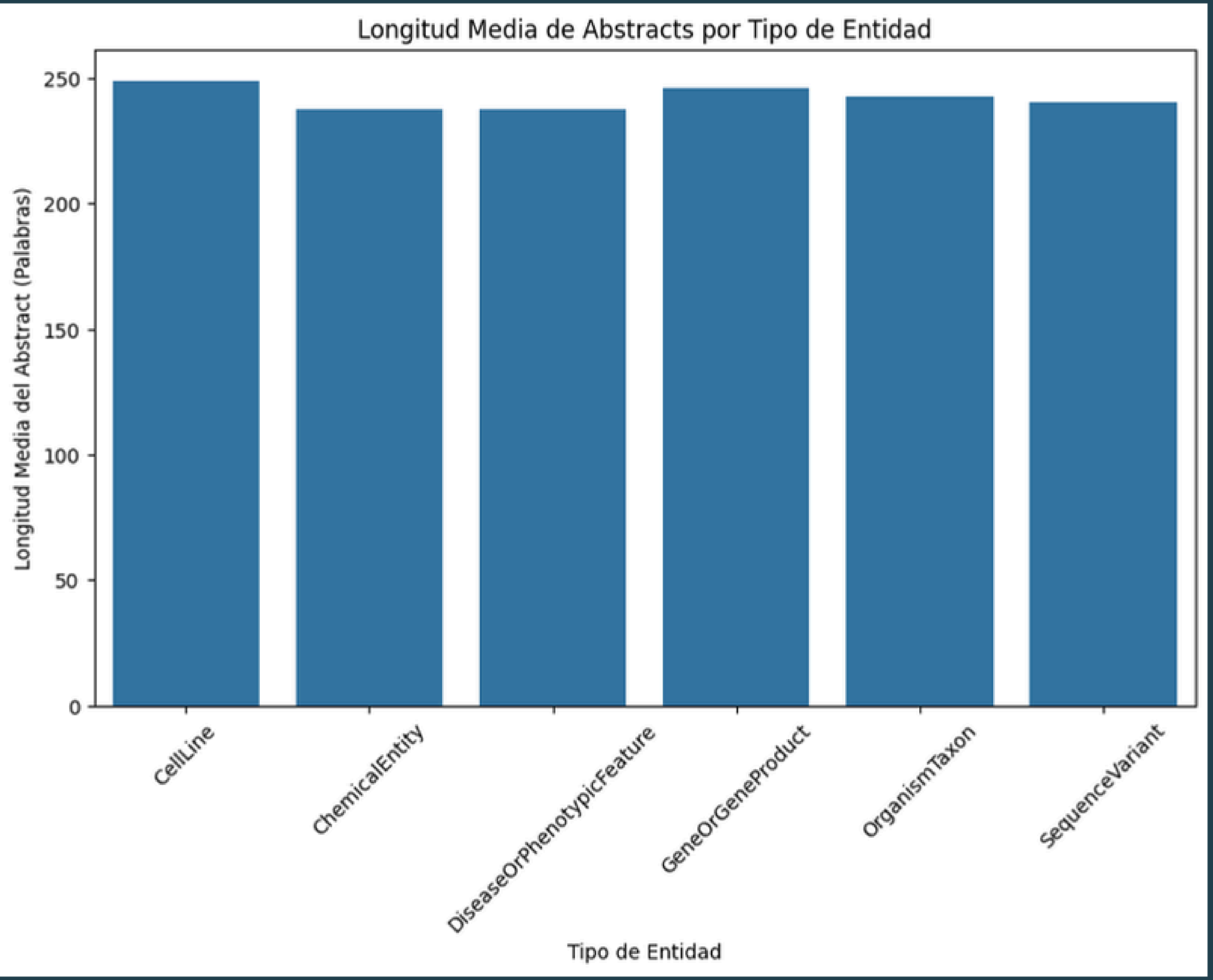
ANALISIS DE LAS RELACIONES

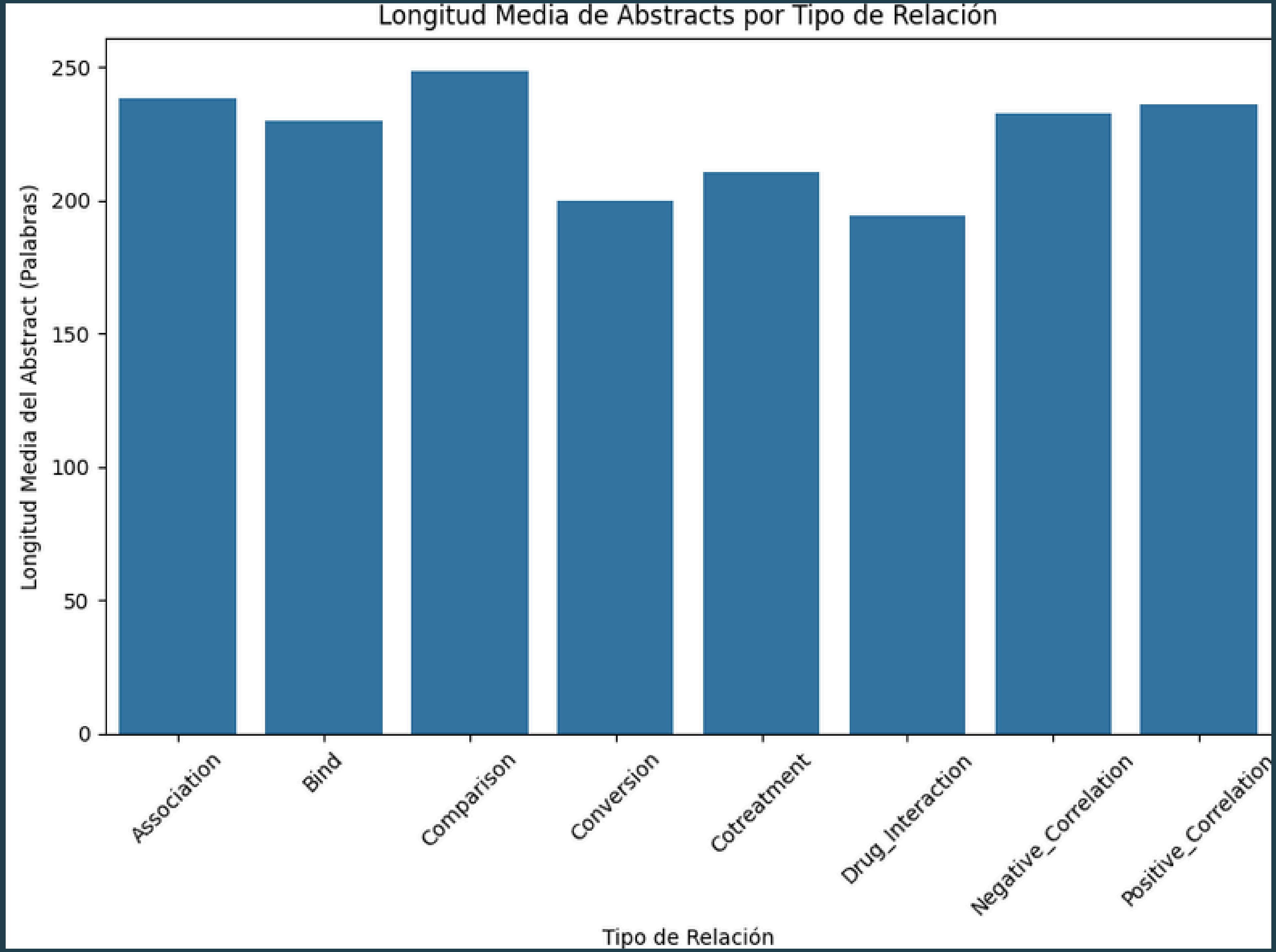

Distribución de Novedad de Relaciones en relations_train
count
0
500
1000
1500
2000
2500
3000
No
Novel

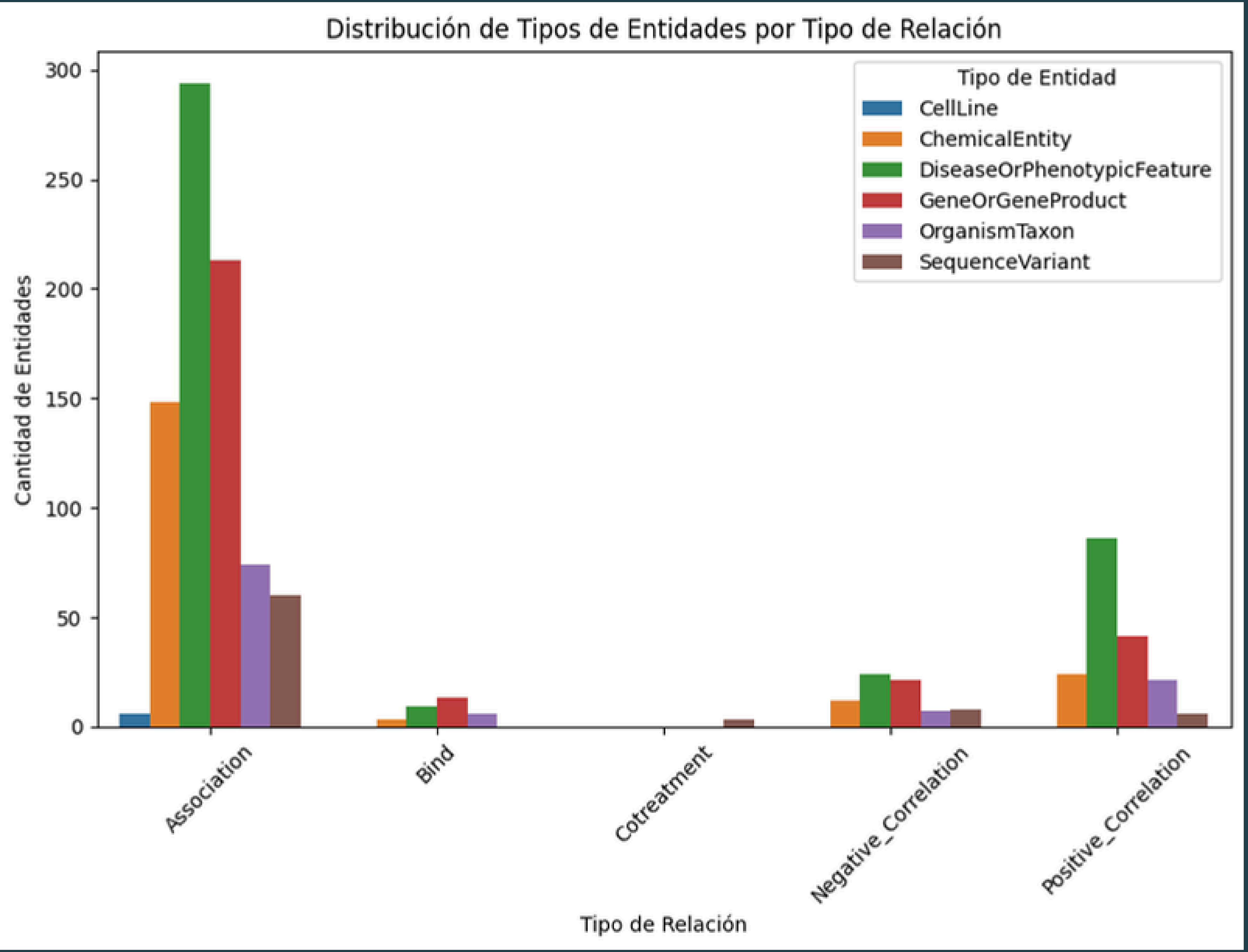

Distribución de Tipos de Relaciones en relations_train
count
0
500
1000
1500
2000
association
positive_correlation
negative_correlation
bind
cotreatment
comparison
drug_interaction
conversion

ANALISIS DE MEZCLADO

```
Resumen de variables:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4280 entries, 0 to 4279
Data columns (total 6 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   id            4280 non-null   int64 
 1   abstract_id   4280 non-null   int64 
 2   type          4280 non-null   object
 3   entity_1_id   4280 non-null   object
 4   entity_2_id   4280 non-null   object
 5   novel         4280 non-null   object
dtypes: int64(2), object(4)
memory usage: 200.8+ KB
None
```


Longitud Media de Abstracts por Tipo de Entidad
Longitud Media del Abstract (Palabras)
250
200
150
100
50
0
CellLine
ChemicalEntity
DiseaseOrPhenotypicFeature
GeneOrGeneProduct
OrganismTaxon
SequenceVariant
Tipo de Entidad


Longitud Media de Abstracts por Tipo de Relación
Longitud Media del Abstract (Palabras)
250
200
150
100
50
0
Association
Bind
Comparison
Conversion
Cotreatment
Drug_Interaction
Negative_Correlation
Positive_Correlation
Tipo de Relación



Distribución de Tipos de Entidades por Tipo de Relación
Cantidad de Entidades
300
250
200
150
100
50
0
Tipo de Entidad
CellLine
ChemicalEntity
DiseaseOrPhenotypicFeature
GeneOrGeneProduct
OrganismTaxon
SequenceVariant
Association
Bind
Cotreatment
Negative_Correlation
Positive_Correlation
Tipo de Relación

Resultados iniciales

GRACIAS