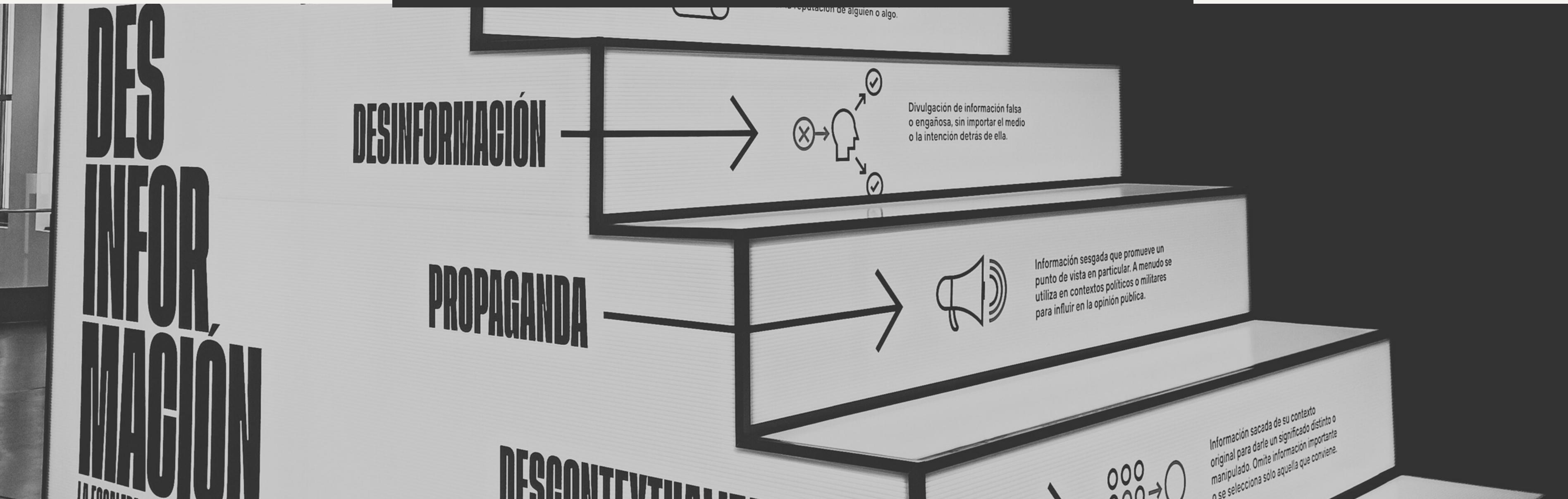


FAKE NEWS

PRESENTADO POR NATALIA SÁNCHEZ Y PABLO SÁNCHEZ



”

BUSCAN EL IMPACTO, NO LA VERDAD

Se transmite una experiencia apasionada al receptor generando una conexión emocional y provocando así una respuesta intensa.
La emoción gana a la razón.

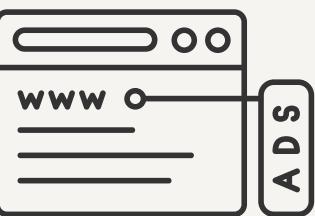
“
OBJETIVOS
PRINCIPALES

03



**DETECTAR SI UNA NOTICIA ES
VERDADERA O Falsa**

**PREDECIR QUÉ PERIÓDICO HA
PUBLICADO LA NOTICIA**



**CREAR UNA INTERFAZ QUE
PERMITA EJECUTAR EL
MODELO DE PREDICCIÓN
MÁS ÓPTIMO**



FASES *DEL PROYECTO*

05 Recogida de datos
08 Análisis de los datos
09 Limpieza y preprocesamiento
13 Evaluar diferentes técnicas de análisis de texto
15 Entrenar y evaluar modelos
27 Comparar y elegir los modelos con mejores resultados
27 Crear una interfaz que permita detectar noticias falsas



Economía

MERCADOS · VIVIENDA · FORMACIÓN · MIS DERECHOS · NEGOCIOS · CINCO DÍAS · RETINA · ÚLTIMAS NOTICIAS

Ibex **-0,30 +** Euro Stoxx **0,31 +** Dow Jones **0,52 +** Cac **0,17 +** FTSE **0,13 +** S&P **0,40 +** Nikkei **-0,06 +** Nasdaq **0,14 +**

E Naturgy tendrá que lograr la autorización del próximo Gobierno para su plan de escisión

JESÚS SERVULO GONZÁLEZ / IGNACIO FARIZA

La reforma del 'escudo antiopas' por parte del Ejecutivo de coalición endurece las inversiones extranjeras



RECOGIDA DE DATOS

Opinión

E Insolidaridad de las grandes corporaciones

La mitad de las mayores empresas posee el 70% de todos los activos, frente al 72% de sus socios

ANDRÉU MISSÉ

E Disciplina fiscal y

FAKE NEWS

E El próximo Gobierno

El gasto en pensiones supera por primera vez los 12.000 millones de euros, un 10,8%

EDITION 2023

La fragmentación de los objetivos ya no es viable y la solución pasa por Europa

EL PAÍS



WEB SCRAPING PERIÓDICOS

Búsqueda de datos a partir de páginas web de periódicos de noticias **verdaderas y falsas**

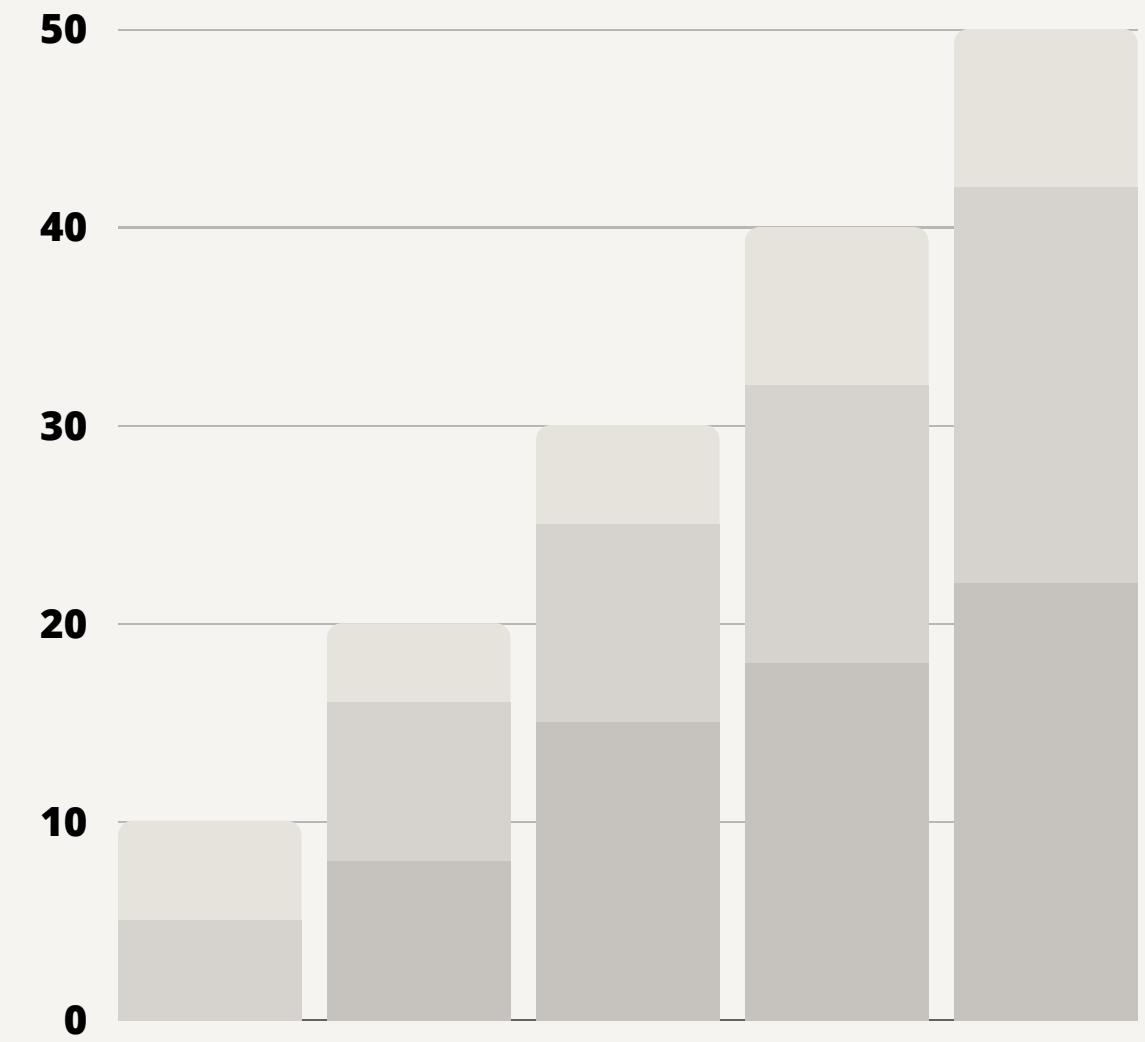
[4]:

	Periódico	Hipervínculo	Fecha publicación	Titular	Subtítulo	Cuerpo	Categoría	Target	Texto
0	ElMundo	https://www.elmundo.es/economia/macroeconomia/...	2022-04-04	Los sindicatos denuncian que el Gobierno ha ge...	CCOO, UGT y CSIF critican duramente la falta d..	"La situación de la Seguridad Social es de sin...	economia	0	Los sindicatos denuncian que el Gobierno ha ge...
1	HayNoticia.es	https://haynoticia.es/el-77-de-los-cuñados-afi...	2020-05-22	El 77% de los "cuñados" afirma que la vacuna y...	NaN	Ante la crisis sanitaria que acontece en la ac...	actualidad	1	El 77% de los "cuñados" afirma que la vacuna y...
2	El Mundo Today	https://www.elmundotoday.com/2012/05/muere-un-...	2012-05-22	Muere un machista de misoginia	SUS ÚLTIMAS PALABRAS FUERON "TODAS PUTAS"	Andrés Gutiérrez, vecino de Huelva de 33 años,...	NaN	1	Muere un machista de misoginia Andres Gutiérre...
3	El Mundo Today	https://www.elmundotoday.com/2020/08/millones-...	2020-08-31	Millones de españoles vuelven al trabajo a la ...	GRANDES RETENCIONES EN EL PASILLO. SEGÚN LA DGT	Con mucha pereza y pocas ganas, así han empre...	NaN	1	Millones de españoles vuelven al trabajo a la ...

De cada noticia se ha extraído el **periódico**, el **hipervínculo**, la **fecha de publicación**, el **titular**, el **subtítulo**, el **cuerpo** y la **categoría**

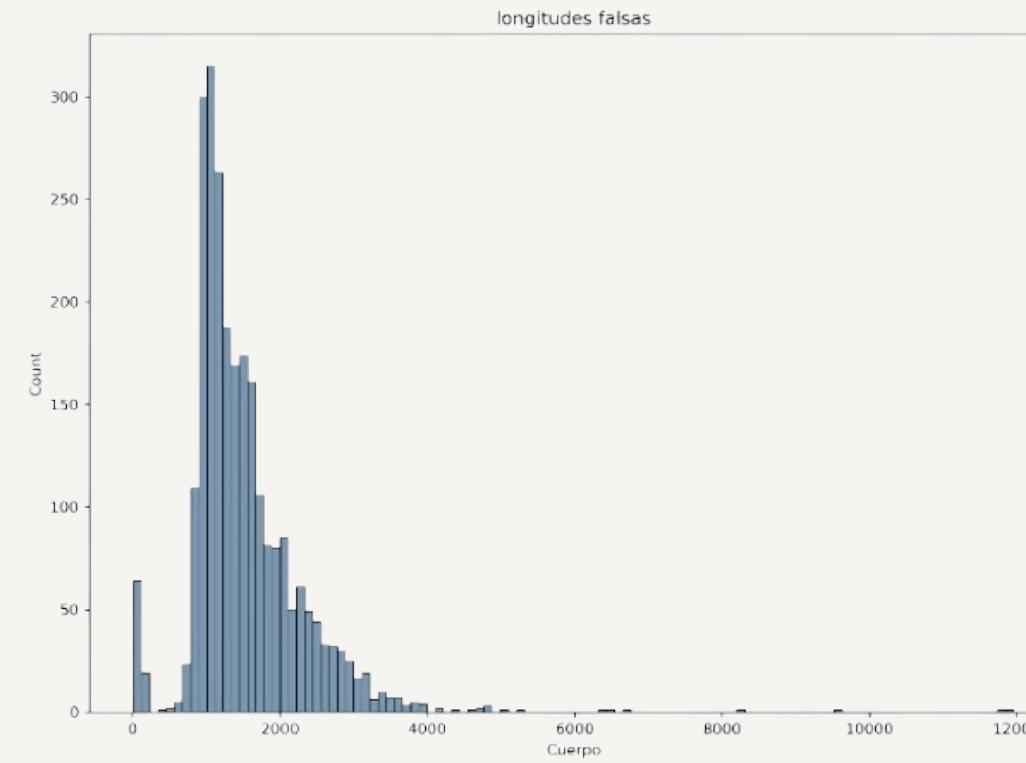
DATOS RECOGIDOS

Verdaderas	El Mundo, El País, ElDiario.es	2506
Falsas	ElMundoToday, Haynoticia.es	2473

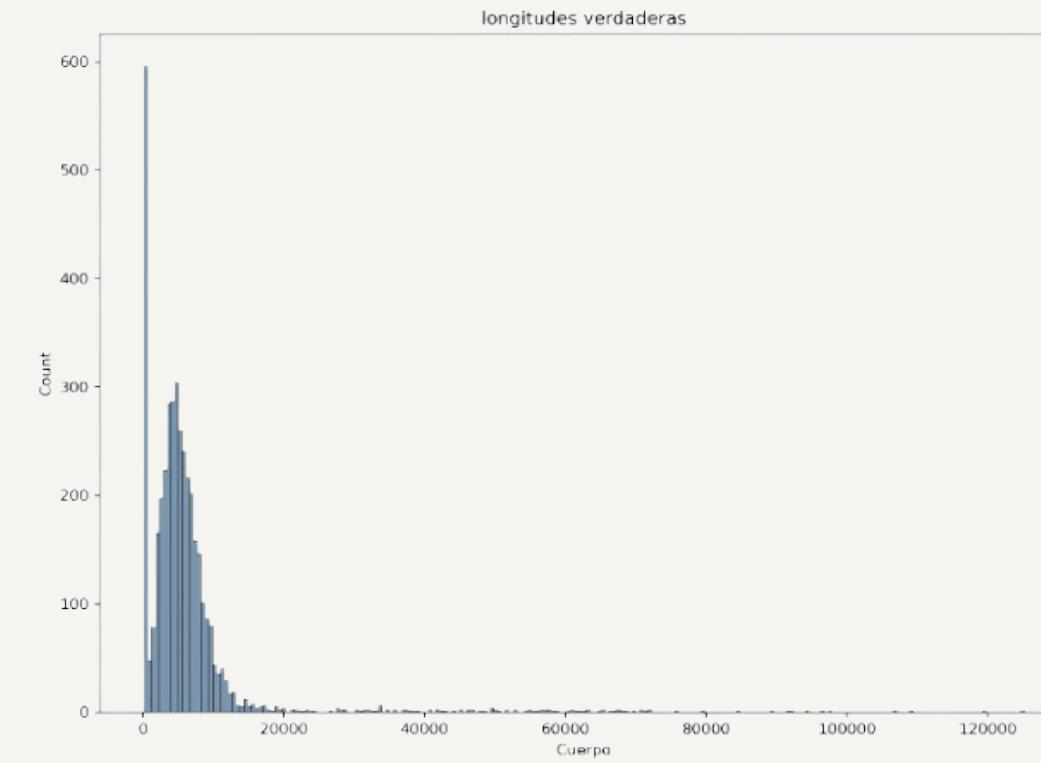


ANÁLISIS Y PROCESAMIENTO DE LOS DATOS

LONGITUD CUERPO DE LA NOTICIA



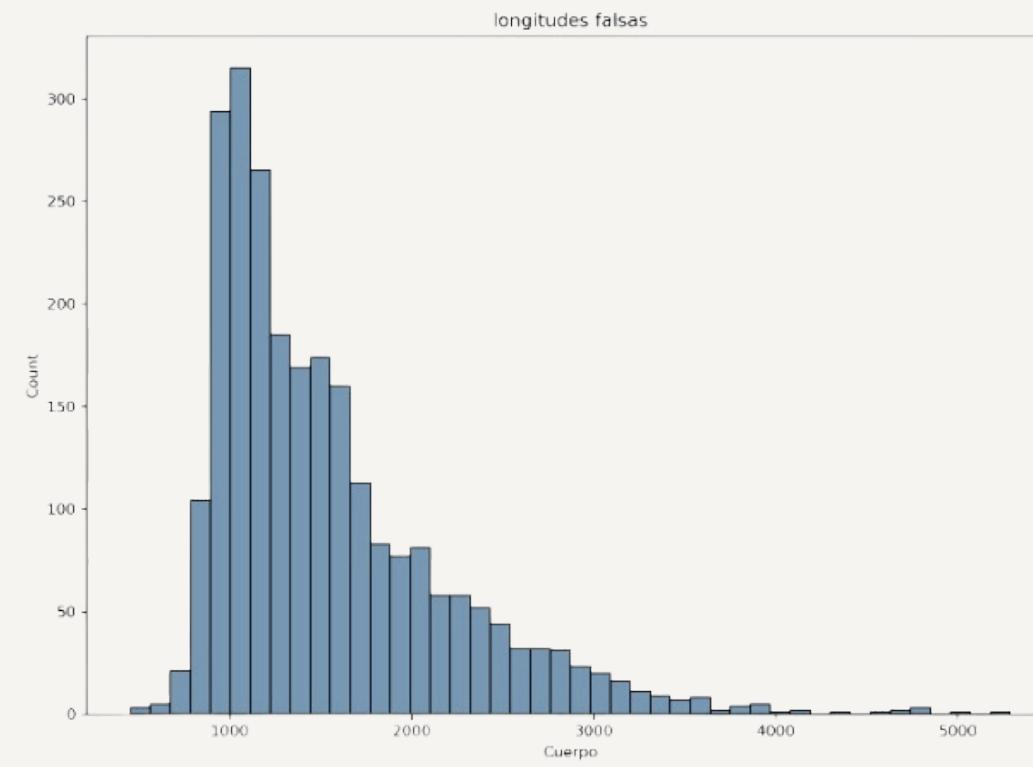
Falsas
< 1% de noticias: longitud > 4000 caracteres



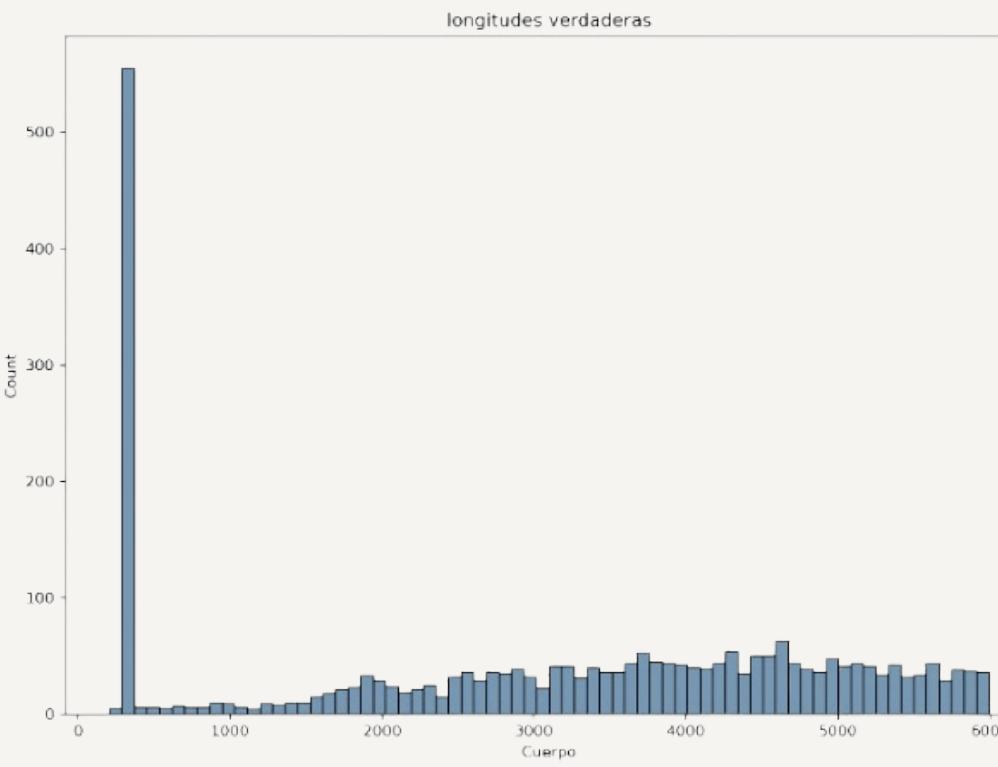
Verdaderas
60% de noticias: longitud > 4000 caracteres

LONGITUD CUERPO DE LA NOTICIA

Tras la primera limpieza



Falsas

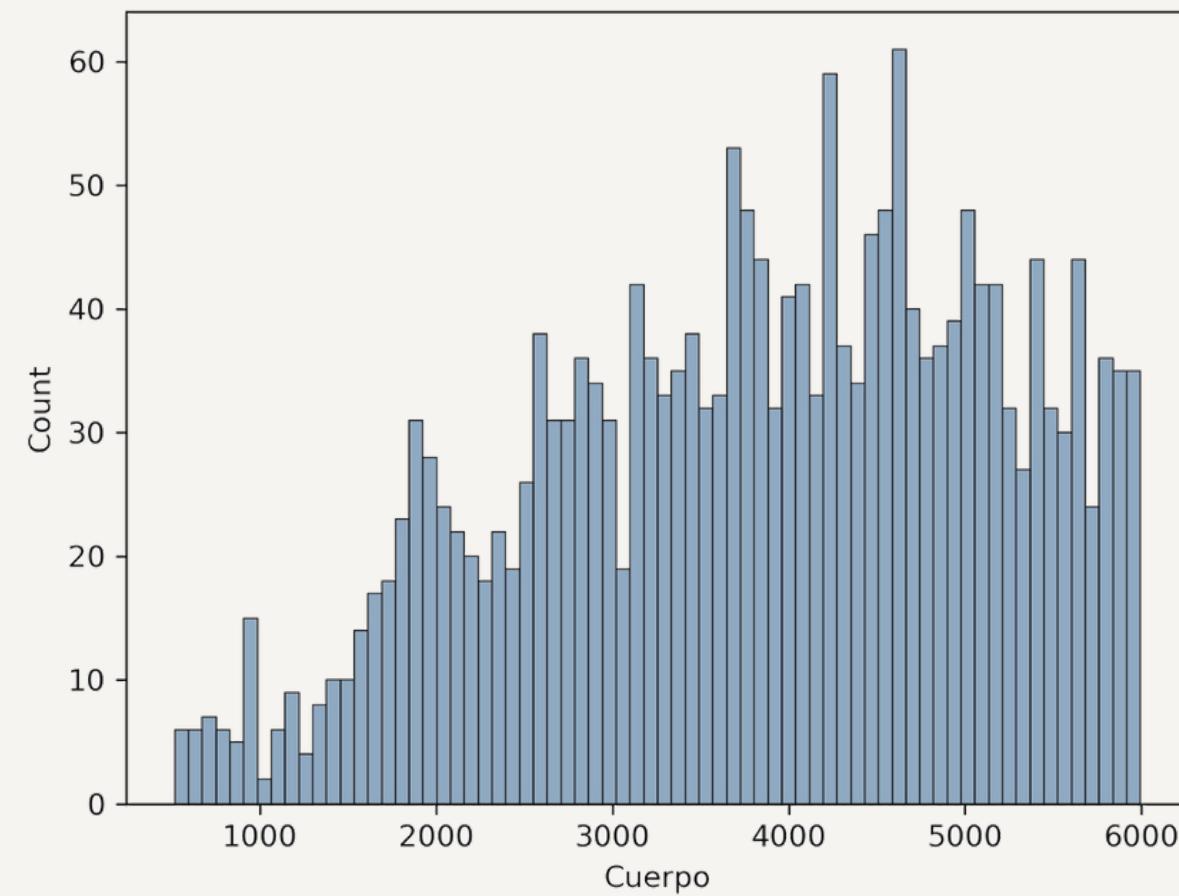


Verdaderas

- Cantidades más similares de un tipo y otro
- Longitudes medias mas parecidas
- Menor asimetría en la distribución de las longitudes de las verdaderas
- Menor diferencia en la dispersion de ambas longitudes

- Eliminamos grupo de noticias cortadas a partir de cierto punto por la fórmula: 'Hazte Premium...'
- Distribución más homogénea
- Volver a hacer web scraping

LONGITUD RESULTANTE



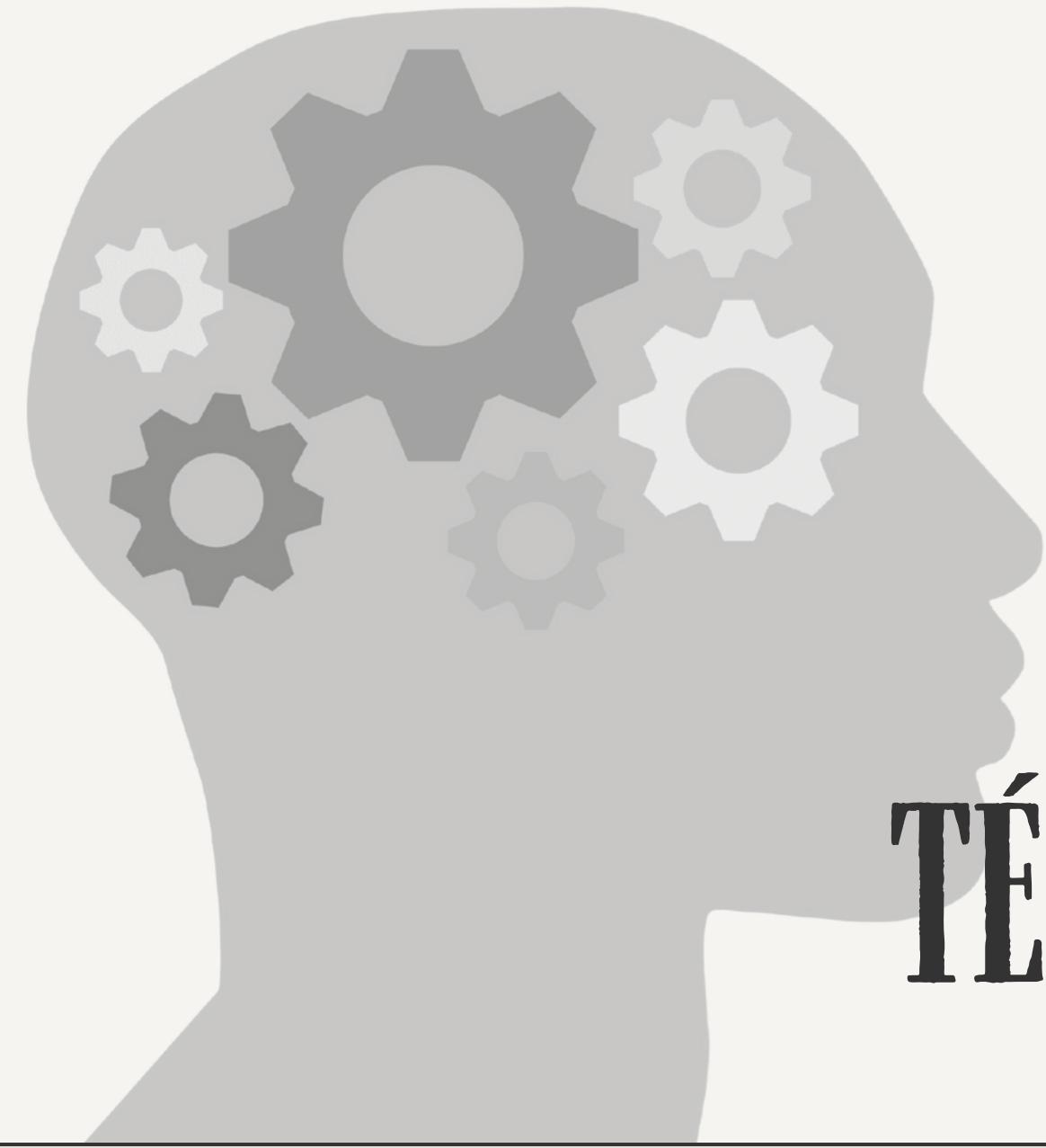
Noticias Falsas

ahora
casa
años
mañana
ser
sido
dos
vez
así

Noticias Verdaderas

madrid
años
año
se
parte
ahora
país
después
sido

PALABRAS MÁS
FRECUENTES



TÉCNICAS DE ANÁLISIS DE TEXTO



MODELOS DE CLASIFICACIÓN DE TEXTO (NLP)

MACHINE LEARNING

Tf-idfVectorizer

- Regresión logística
- Multinomial Naive Bayes
- Árbol de decisión
- Gradient Boosting
- Ada boost

REDES NEURONALES DEEP LEARNING

Word embeddings

- Doc2Vec

Transformers

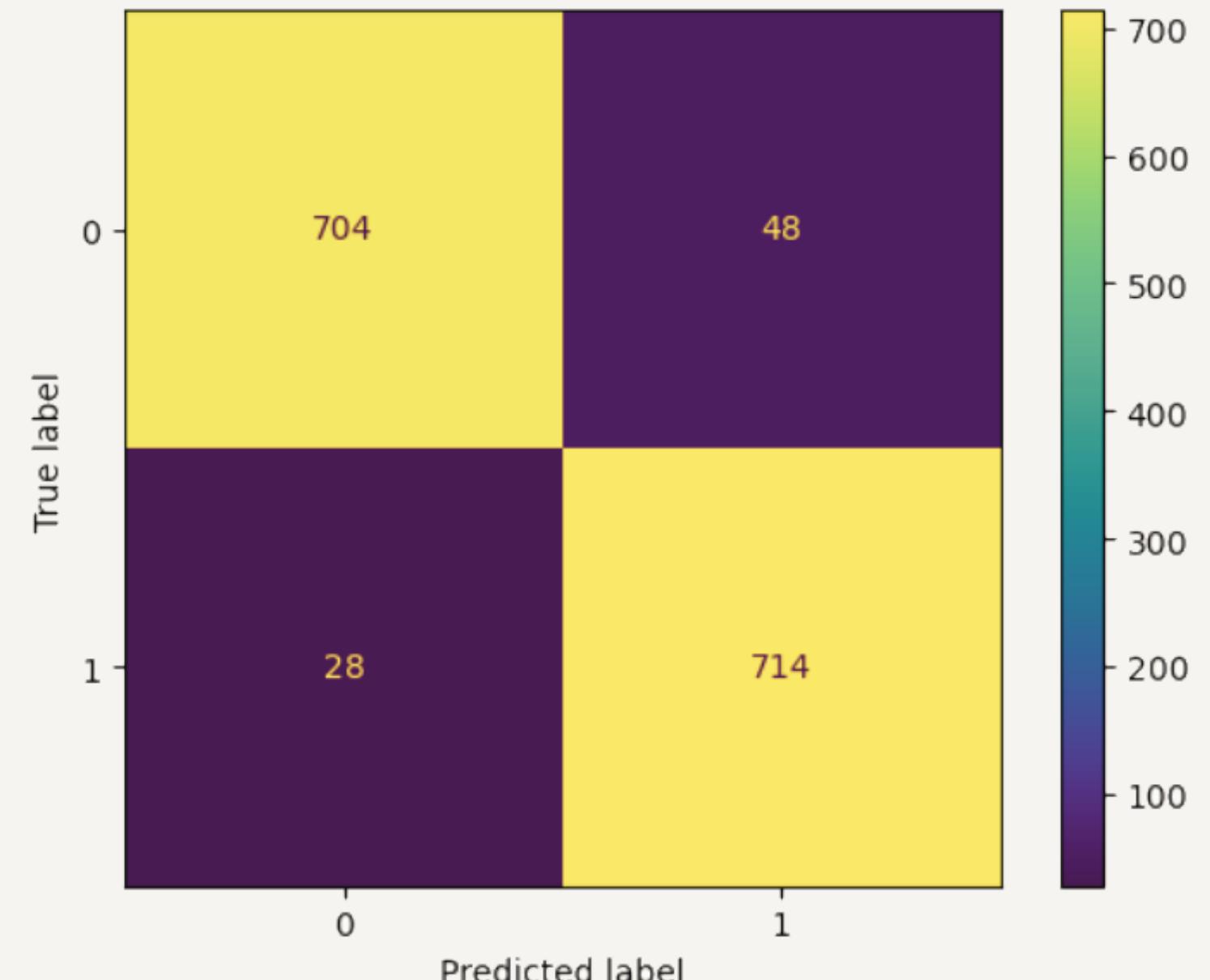
- BERT
- ALBERT
- DistilBERT

REGRESIÓN LOGÍSTICA

TFIDFVECTORIZER

Se basa en el concepto de una función logística para estimar la probabilidad de que una instancia pertenezca a una clase determinada.

	Predictión verdadera o falsa	Predictión periódico
Accuracy	0.949	0.809
	Predicción verdadera o falsa	Predicción periódico



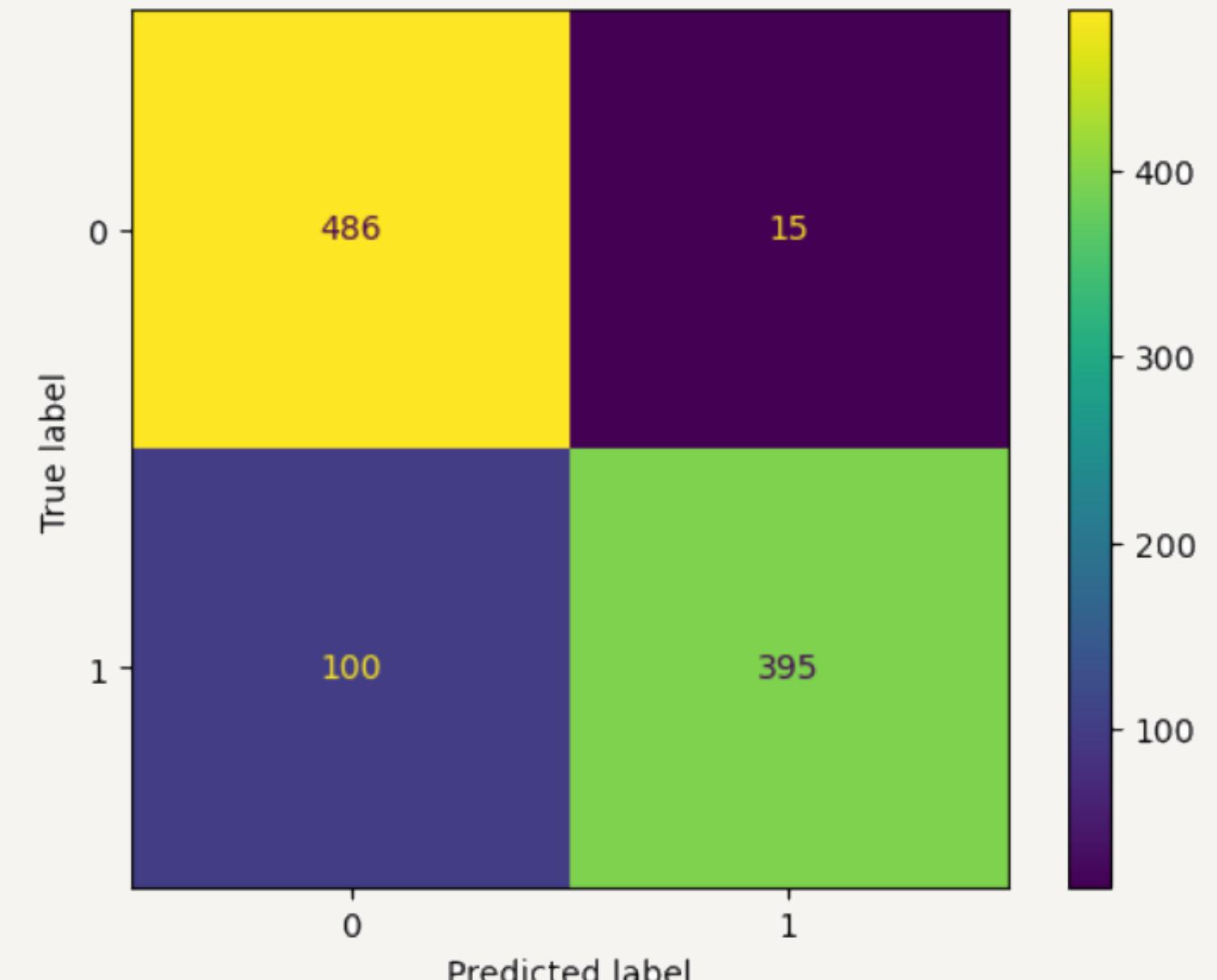
Matriz de confusión
Predicción verdadero o falso

MULTINOMIALNB

TFIDFVECTORIZER

Variante del clasificador Naive Bayes que se utiliza comúnmente para problemas de clasificación con atributos discretos o conteos, como la frecuencia de palabras en un documento.

Predictión verdadera o falsa	Titular + Cuerpo	Cuerpo
Accuracy	0.884	0.888



Matriz de confusión

Predicción verdadero o falso

ÁRBOL DE DECISIÓN

TFIDFVECTORIZER

Divide el conjunto de datos en subconjuntos más pequeños mediante reglas de decisión basadas en características, formando una estructura en forma de árbol.

Predicción periódico
Accuracy train

Titular	Cuerpo	Titular + Cuerpo
0.343	0.609	0.614

Accuracy test

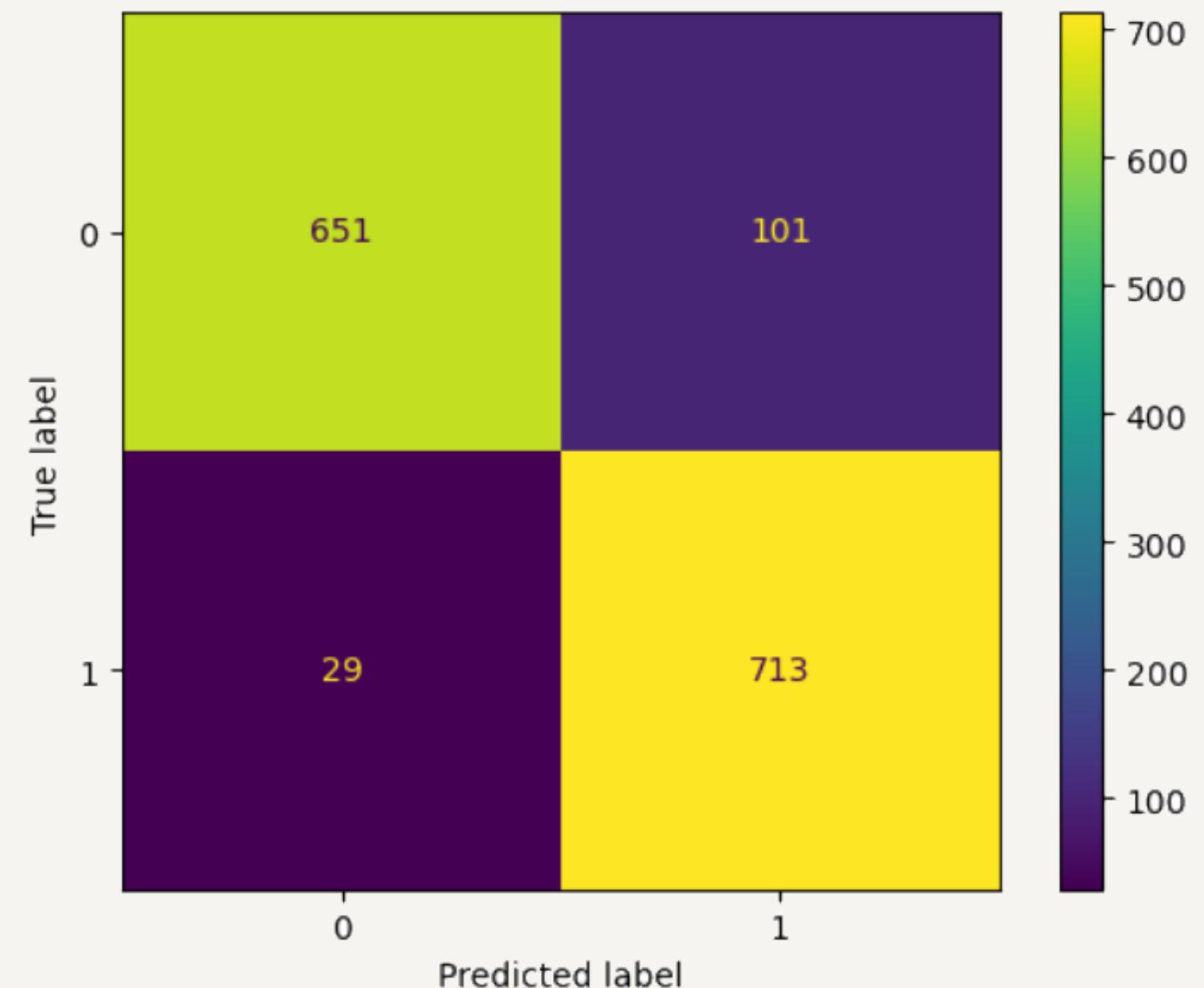
Titular	Cuerpo	Titular + Cuerpo
0.334	0.574	0.573

GRADIENT BOOSTING

TFIDFVECTORIZER

Es un conjunto de algoritmos de aprendizaje que se utilizan para problemas de clasificación y regresión. Funciona construyendo varios árboles de decisión en secuencia, donde cada árbol intenta corregir los errores del anterior.

Predictión verdadera o falsa	Titular + Cuerpo	Cuerpo
Accuracy	0.912	0.917



Matriz de confusión

Predicción verdadero o falso

Accuracy train

Titular	Cuerpo	Titular + Cuerpo
0.440	0.761	0.752

Accuracy test

Titular	Cuerpo	Titular + Cuerpo
0.385	0.708	0.702

Funciona mediante la combinación de varios clasificadores débiles para formar un clasificador fuerte, donde cada clasificador débil se enfoca en corregir los errores de los clasificadores anteriores.

TFIDFVECTORIZER

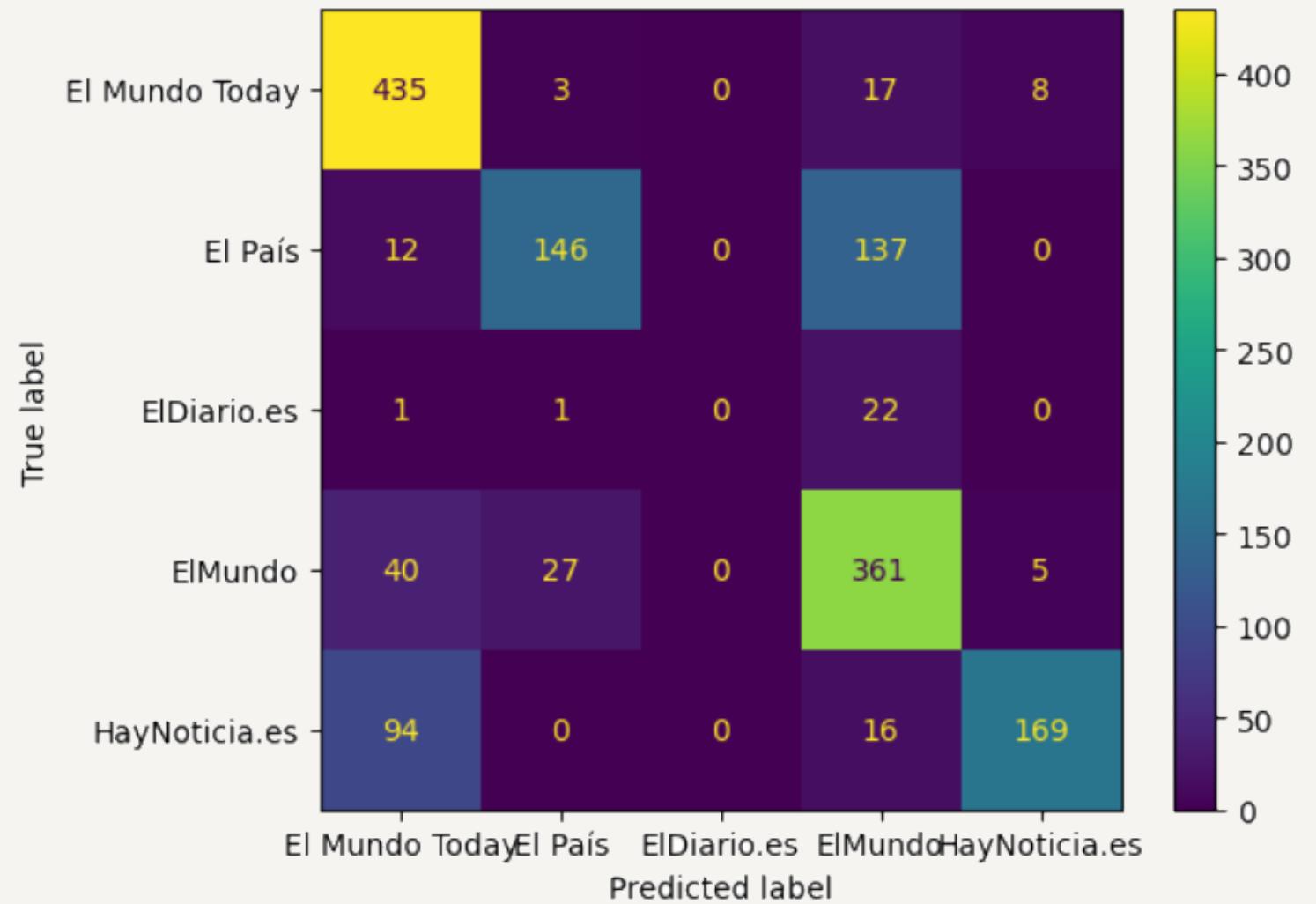
ADA BOOST

”

CONCLUSIONES

MODELOS TF-IDF

REGRESIÓN LOGÍSTICA

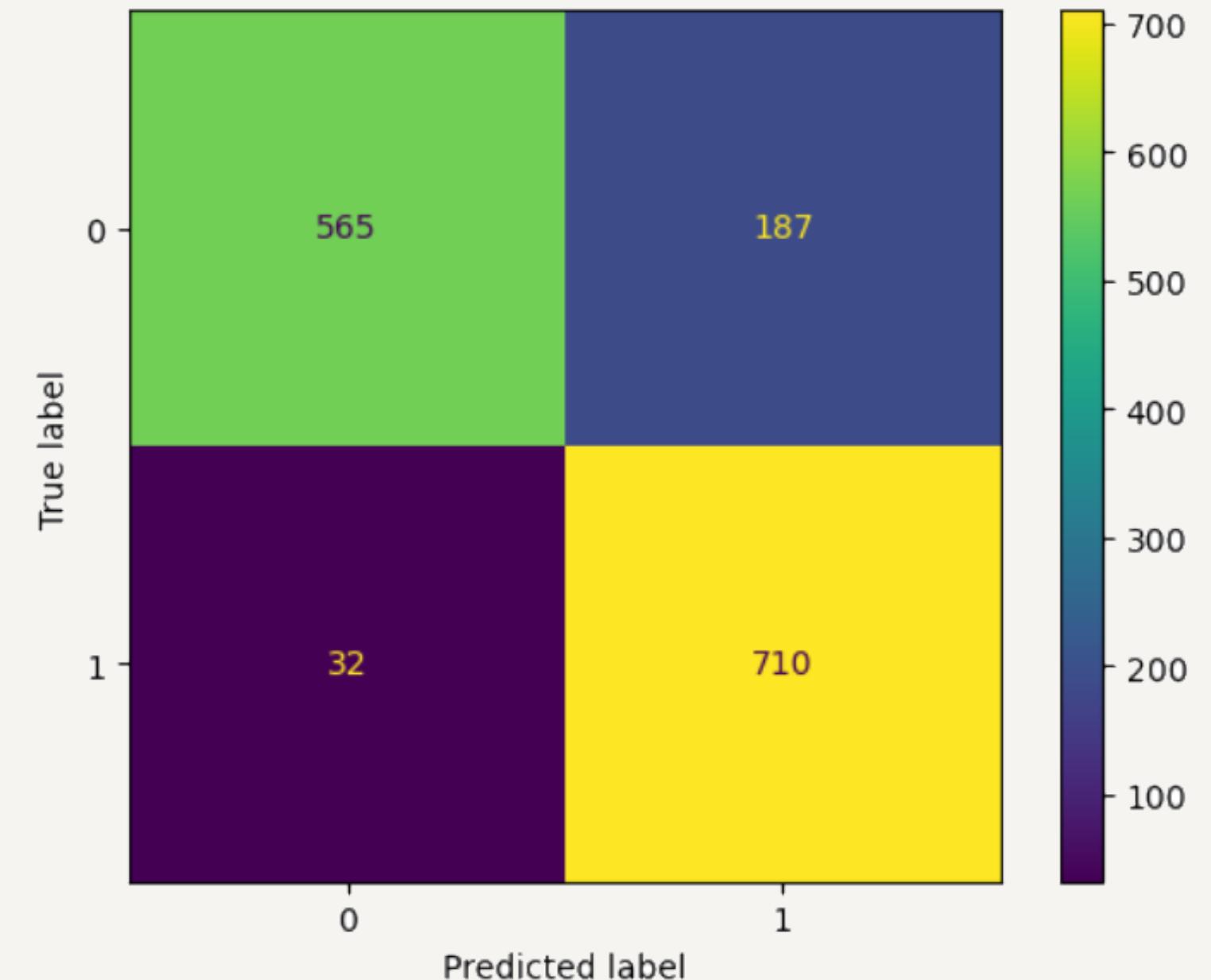


Matriz de confusión
Predicción periódico - Regresión logística

DOC2VEC

WORD EMBEDDINGS

Usamos el modelo Doc2Vec preentrenado para representar cada noticia como un vector denso ya que este modelo permite utilizar técnicas más avanzadas con documentos de texto completos.



BERT

TRANSFORMERS

(Bidirectional Encoder Representations from Transformers)

El pre-entrenamiento de BERT se lleva a cabo en grandes corpus de texto y, después de este proceso, el modelo adquiere una comprensión profunda del lenguaje. Luego tomamos BERT como punto de partida y se adapta a la clasificación de texto mediante el proceso llamado fine-tuning.

Predicción periódico

```
tokenizer = AutoTokenizer.from_pretrained('bert-base-uncased')
```

```
train_encodings = tokenizer(X_train.tolist(), truncation=True, padding=True)
test_encodings = tokenizer(X_test.tolist(), truncation=True, padding=True)
```

Configurar los argumentos de entrenamiento

```
training_args = TrainingArguments(
    per_device_train_batch_size=8,
    per_device_eval_batch_size=64,
    num_train_epochs=3,
    evaluation_strategy="epoch",
    output_dir=".//results" )
```

	bert-base-cased	bert-base-uncased
Accuracy	0.913	0.938

Predicción verdadero o falso

```
tokenizer = AutoTokenizer.from_pretrained('dccuchile/albert-tiny-spanish')
```

```
train_encodings = tokenizer(X_train.tolist(), truncation=True, padding=True)
test_encodings = tokenizer(X_test.tolist(), truncation=True, padding=True)
```

Configurar los argumentos de entrenamiento

```
training_args = TrainingArguments(
    per_device_train_batch_size=8,
    per_device_eval_batch_size=64,
    num_train_epochs=3,
    evaluation_strategy="epoch",
    output_dir=".//results" )
```

Es una versión más liviana y eficiente del modelo BERT original, que utiliza la compartición de parámetros y la segmentación de oraciones para reducir la cantidad de parámetros y mejorar la eficiencia, mientras mantiene un alto rendimiento en tareas de NLP.

ALBERT

TRANSFORMERS

Predicción verdadero o falso

Epochs = 3 Batch_size_train = 16

Acc = 0.996

[702/702 02:37, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Accuracy
1	No log	0.097129	0.971888
2	No log	0.074418	0.982329
3	0.103900	0.069552	0.982329

Epochs = 3 Batch_size_train = 6

Acc = 0.997

[1869/1869 02:48, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Accuracy
1	0.171300	0.084252	0.983133
2	0.056900	0.074454	0.984739
3	0.032600	0.057849	0.987149

Epochs = 4 Batch_size_train = 8

Acc = 0.998

[1868/1868 04:17, Epoch 4/4]

Epoch	Training Loss	Validation Loss	Accuracy
1	No log	0.093696	0.980723
2	0.139100	0.072183	0.985542
3	0.045800	0.047083	0.989558
4	0.019000	0.042895	0.989558

DISTILBERT

TRANSFORMERS

Al igual que ALBERT, el objetivo principal de DistilBERT es reducir el tamaño y la complejidad del modelo original de BERT para que sea más ligero y eficiente, manteniendo al mismo tiempo un rendimiento competitivo en tareas de procesamiento del lenguaje natural.

El enfoque de DistilBERT para lograr la eficiencia es la destilación del conocimiento (knowledge distillation) desde el modelo BERT completo hacia un modelo más pequeño y más rápido.

```
# Datos:  
# Epochs: 3  
# per_device_train_batch_size=6  
# per_device_eval_batch_size=8  
  
print('accuracy del train:')  
trainer.evaluate(train_dataset)['eval_accuracy']
```

accuracy del train:

[343/343 00:47]

0.9974433893352812

```
# Datos:  
# Epochs: 1  
# per_device_train_batch_size=6  
# per_device_eval_batch_size=8
```

```
print('accuracy del train:')  
trainer.evaluate(train_dataset)['eval_accuracy']
```

accuracy del train:

[343/343 08:03]

0.9737034331628927

”
CONCLUSIONES

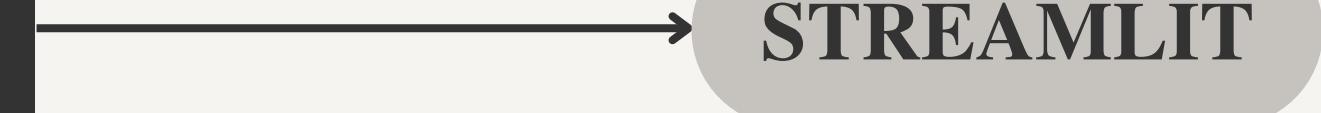
MODELOS DEEP LEARNING

REGRESIÓN LOGÍSTICA

BERT

ALBERT

DISTILBERT



STOP FAKE NEWS

Natalia Sánchez
Pablo Sánchez