

¿Es realmente el chicharrón bueno para la salud?



Modelos de ML y TL para detectar noticias falsas en el contexto colombiano.

IT ACADEMY



Buscar



Enviar



Comentar



Compartir



Vida Moderna

Sorprendente e consumir chich saludable que a

Visita nuestro
canal aquíACTUALIDAD
PORCINA[Home](#) > [Noticias](#) > Estudios internacionales aseguran que el chicharrón de cerdo es saludable

El chicharrón ha sido un el
culturas alrededor del mun

NOTICIAS

🕒 23 NOVIEMBRE, 2023

🕒 17 0 0

El chicharrón es más saludable que algunas verduras, según un nuevo estudio

Así lo concluyeron investigadores de la Escuela de Medicina de la Universidad de Boston en un estudio recientemente publicado.

de la gastronomía.



BULOS INTERNET >

No, el chicharrón no es más saludable que las verduras y no existe ningún estudio que lo demuestre

Dietistas, nutricionistas y divulgadores científicos desmienten una noticia falsa que ha inundado las redes, los periódicos, los canales de televisión y las radios de Colombia, Latinoamérica y España en la que se asegura que la piel de cerdo es mejor para la salud que las espinacas, las zanahorias y la coliflor

“Todo es una gran mentira, una noticia falsa replicada sin pensar por los medios para ganar clicks y audiencia”

Juan Camilo Mesa - Nutricionista
El País 14/12/2023

Objetivo

Comprobar la utilidad de las herramientas de machine learning (ML) y de transfer learning (TL) para detectar noticias falsas en español



Data Science & Fake News

“

**a news article which is delusive
both intentionally
and verifiably and could
misinform the readers**

Ansar & Goswami, 2021, p. 3

Perspectivas de detección

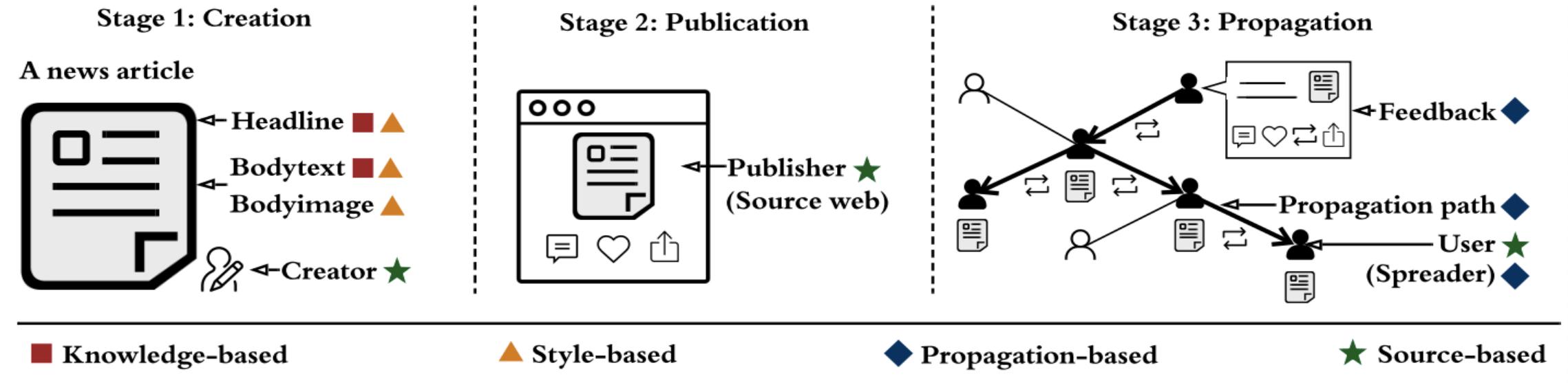


Fig. 1. Fake News Life Cycle and Connections to the Four Fake News Detection Perspectives Presented in this Survey

A Comparative Study of the Fake News Detection using Statistical and Machine Learning Approaches.

Model	Description	Methods Used	Data Set(s)	Performance
Mukherjee et al. (2013)	Style oriented approach to Fake News Detection	SVM	Mechanical Turk and Yelp	Accuracy=0.861, F1-Score=0.857
Kwon et al. (2013)	Detecting Fake News based upon temporal, structural and linguistic characteristics	SVM, Random Forest and Decision Tree	Microblogs	Accuracy=0.897, F1-Score=0.878 (Random Forest)
Castillo et al. (2013)	Determination of newsworthiness and credibility of microblog propagation	LR and Random Forest	Microblogs	Mean F-Score=0.824, Mean ROC area=0.816
Ma et al. (2015)	Detecting rumors based upon social context	SVM	Microblogs	Mean Accuracy =0.871, Mean F1=0.87
Rubin & Lukoianova (2015)	Fake News Detection through discourse structure analysis	RST-VSM	Mechanical Turk	NA
Jin et al. (2016a)	Detecting relations among tweets and constructing a credibility-network	Topic Modeling	Microblogs	Accuracy=0.84
Rubin et al. (2016)	Using satirical cues for Fake news Detection	SVM and TF-IDF	News Articles	F-Score=0.87
Granik & Mesyura (2017)	Fake News Detection using NB	NB	Posts on Facebook	Accuracy=0.74
Tacchini et al. (2017)	Automated detection of hoaxes	LR and HBLC	Posts on Facebook	Mean Accuracy =0.992 (HBLC)
Gravanis et al. (2019)	Fake News Detection using content-based features	SVM, AdaBoost and Bagging	BuzzFeedNews, BS Detector and PolitiFact	Mean Accuracy=0.787
Shrivastava et al. (2020)	A mathematical model to analyze the spread of fake news in social media	Basic Reproduction Number	Online Social Networks	NA
Verma et al. (2021)	A Fake News Detection Approach through Word Embedding over Linguistic Features	TF-IDF and CV	WELFake	Accuracy=0.967

Note: NB: Naive Bayes Theorem, LR: Logistic Regression, HBLC: Harmonic Boolean-Label-Crowdsourcing, RST-VSM: Rhetorical-Structure-Theory - Vector-Space-Model, SVM: Support Vector Machine, TF-IDF: Term Frequency - Inverse Document Frequency, CV: Count Vectorizer

Detección de Fake News en español

AUTORES	DESCRIPCIÓN	DATASET	BEST ML MODEL	SCORE
Posadas-Durán, J. P., Gómez-Adorno, H., Sidorov, G., & Escobar, J. J. M. (2019)	Detection of fake news in a new corpus for the Spanish language. México - España	491[0] 480[1]	BOW + POS RF	Acc = 76.94%
Mafla, N., Flores, M., Castillo-Páez, S., & Andrade, R. (2022)	Automatic Detection of Fake News in Spanish: Ecuadorian Political Satire	1075[0] 553[1]	TF-IDF SVM lineal	F-Score = 95.81%
Tretiakov, Arsenii & Martín García, Alejandro & Camacho, David. (2022)	Detection of False Information in Spanish Using Machine Learning Techniques. Solo castellano	2750[0] 2742[1]	SVM	Acc = 0.872
Flores Quinayás, J. E., & Montaño Morcillo, J. G. (2022)	Modelo para la detección de noticias falsas en formato texto en la red social Twitter, aplicado al contexto político colombiano de las elecciones presidenciales de 2022	317[0] 317[1]	TF-IDF SVM	Acc = 0.9133

Resultados Flores & Montaño

	Modelos	Métrica - Accuracy
0	Random Forest	0.897638
1	Naive Bayes	0.905512
2	SVC	0.913386
3	Logistic Regression	0.905512
4	XGBoost	0.874016
5	Neural Network	0.913386

	Modelos	Métrica - Accuracy
0	Random Forest	0.708459
1	Naive Bayes	0.699396
2	SVC	0.737160
3	Logistic Regression	0.728097
4	XGBoost	0.714502
5	Neural Network	0.658610



Extracción y preprocesamiento

Proceso de extracción





CHEQUEOS

INVESTIGACIONES

ESPECIALES

PODCAST

ZOOM



16 de Febrero de 2024



By Viral 24
@ByViral24
ATENCIÓN. Se conoce la identidad de quienes intentan ingresar a la fuerza a la @CorteSuprema se trata de José Cuesta Novoa, exguerrillero del grupo M19 y Concejal de Bogotá y miembros de @fecode

Reacción Nacional
@RNacional_News
URGENTE | OFICIALMENTE hay ruptura institucional en Colombia. Magistrados de la Corte Suprema de Justicia tienen que huir en helicóptero por el ataque ordenado por el gobierno. Los cocaleros milicianos petristas fracasan en su GOLPE DE ESTADO. Petro debe ir a cárcel.

César Gaviria
@CesarGaviria4
Confirmado. Varias de las diez empresas más grandes de Colombia compran sentencias a la @CorteSuprema, especialmente a los millonarios magistrados de la Sala Civil Familia. La reforma a la justicia propuesta por el gobierno de @petrojgustavo debe tener en cuenta esta carne.

Las mentiras del 8F: del rescate en un helicóptero a la supuesta afiliación política de magistrados

Antes y durante las manifestaciones en el Palacio de Justicia de Bogotá, cuentas desinformadoras movieron narrativas contra la Corte Suprema, objetivo de la protesta, y el gobierno Petro, que la...

[Ver más +](#)



EL ESPECTADOR

Feb. 17 - 2024

Últimas Noticias Opinión Política Judicial Economía Mundo Bogotá Deportes Colombia + 20 Empleos

Suscriptores



Suscriptores Cine y TV

El Dios de una sociedad que sobrevivió a la nieve

Hace 4 minutos



Política

Gobierno radicó proyecto para

ESTE DÍA

El comercio puede ser la herramienta para mover los acuerdos climáticos

La angustia porque los acuerdos climáticos mundiales se quedan en retórica se puede contrarrestar si se apela a una diversidad de...



Foto: Cristian Garavito

Columnistas

Caricaturas



Suscriptores

La tiranía de la mediocridad

Mauricio García Villegas

Hace 20 horas



“Hawthorne” por Henry James

Julio César Londoño

Hace 20 horas



¿Techo, lecho y mesa?

Catalina Uribe Rincón

Hace 20 horas



Cortázar, 40 años

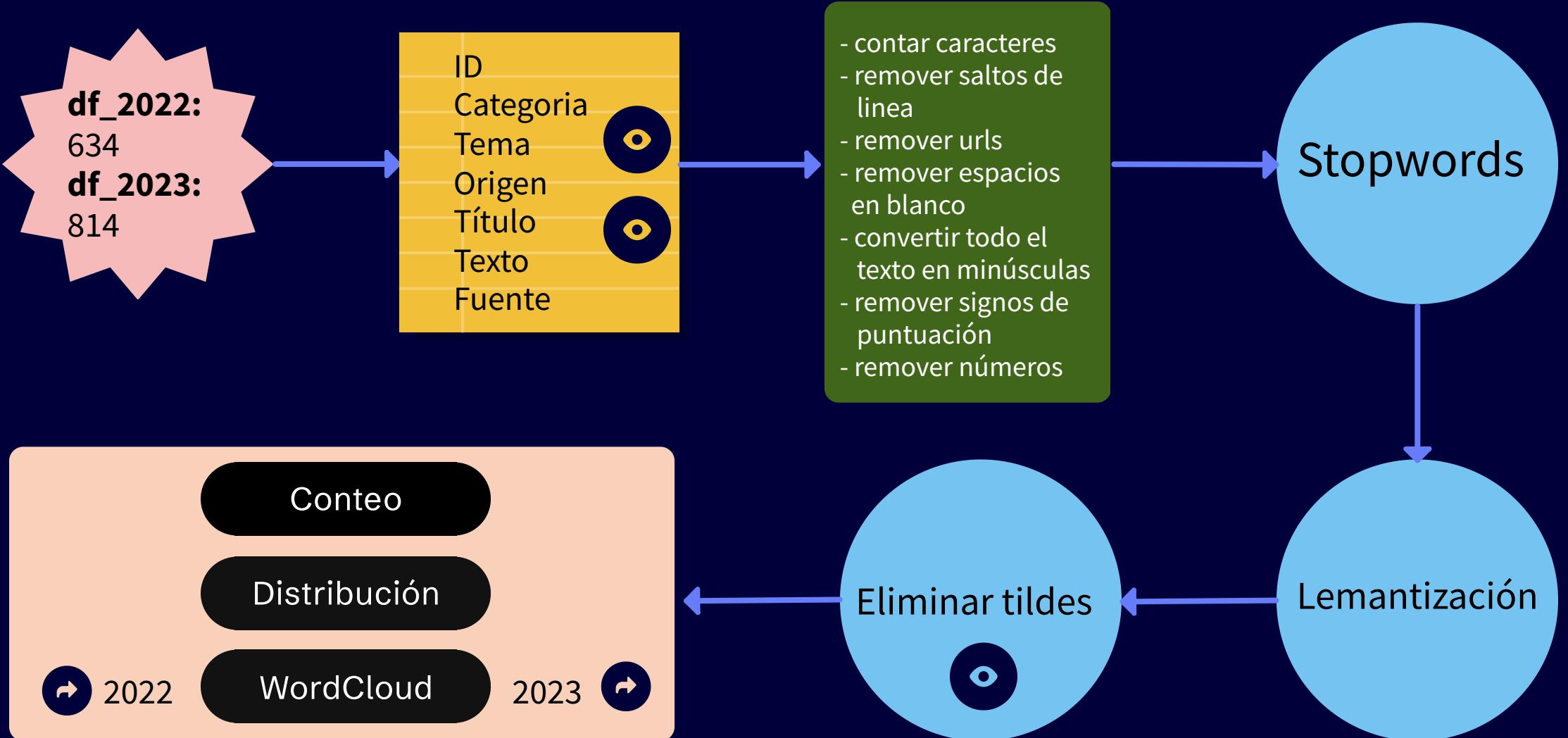
Santiago Gamboa

Hace 20 horas

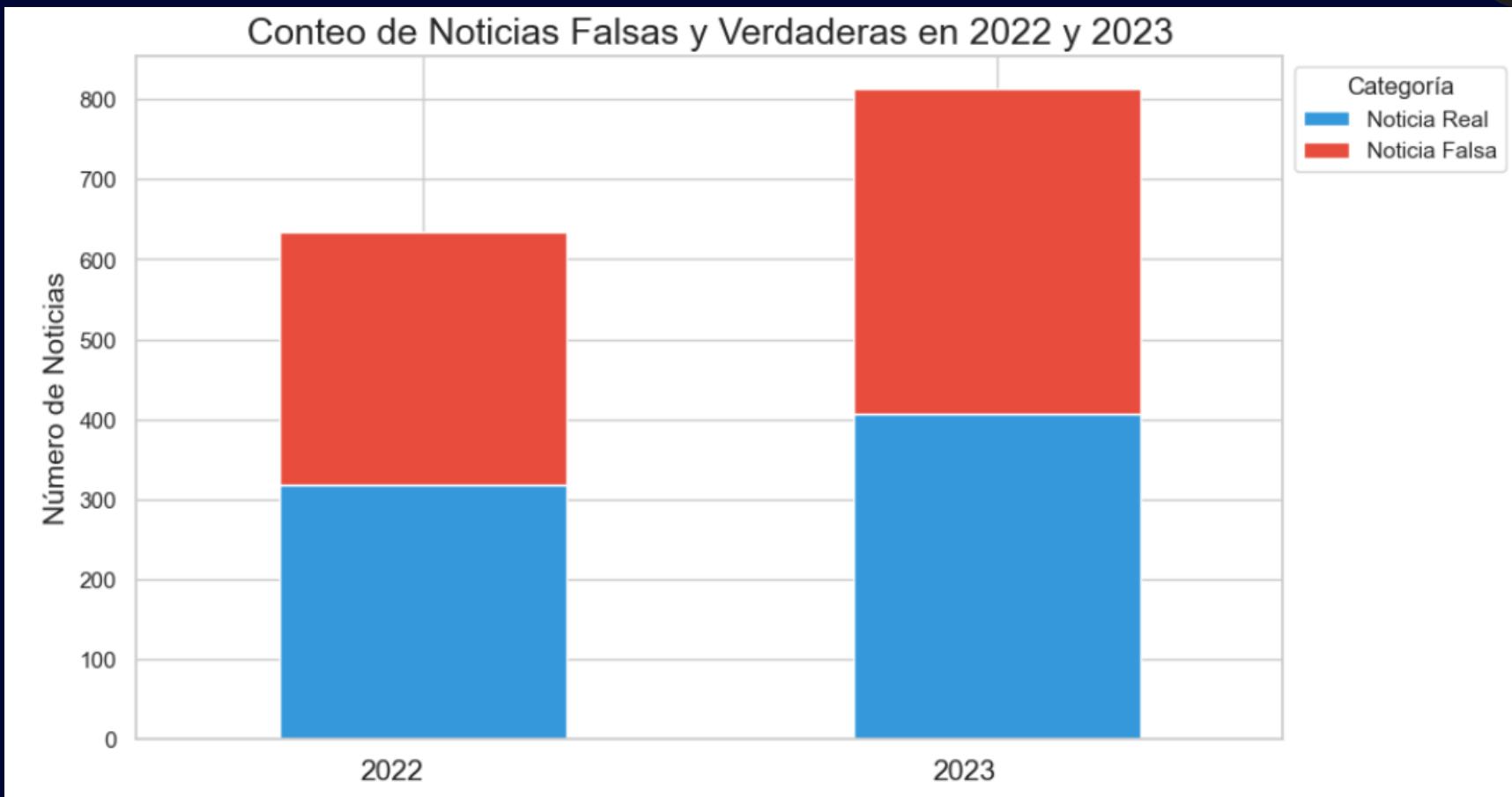


[Más voces >](#)

Flujo de preprocesamiento

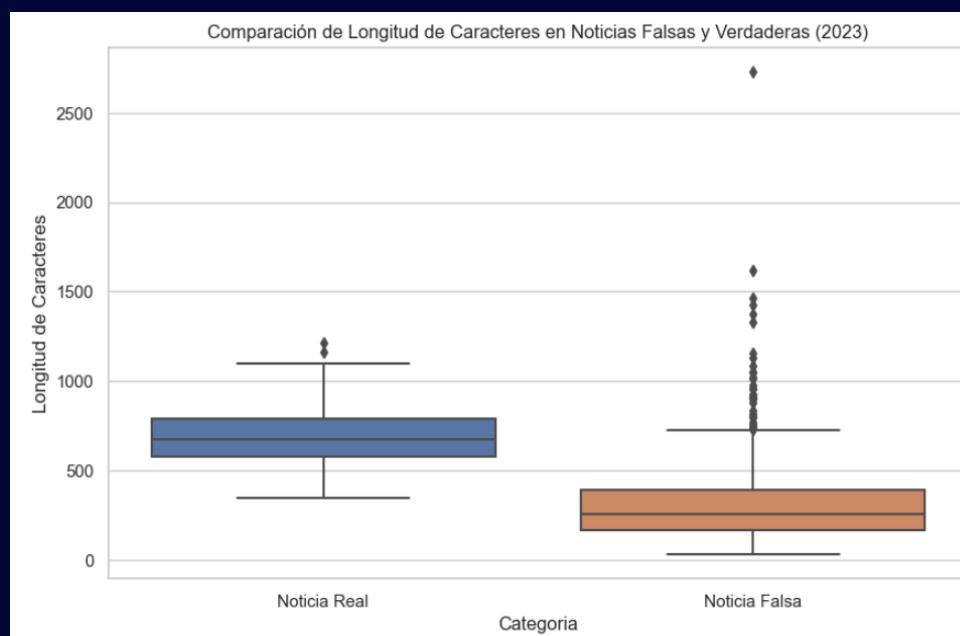
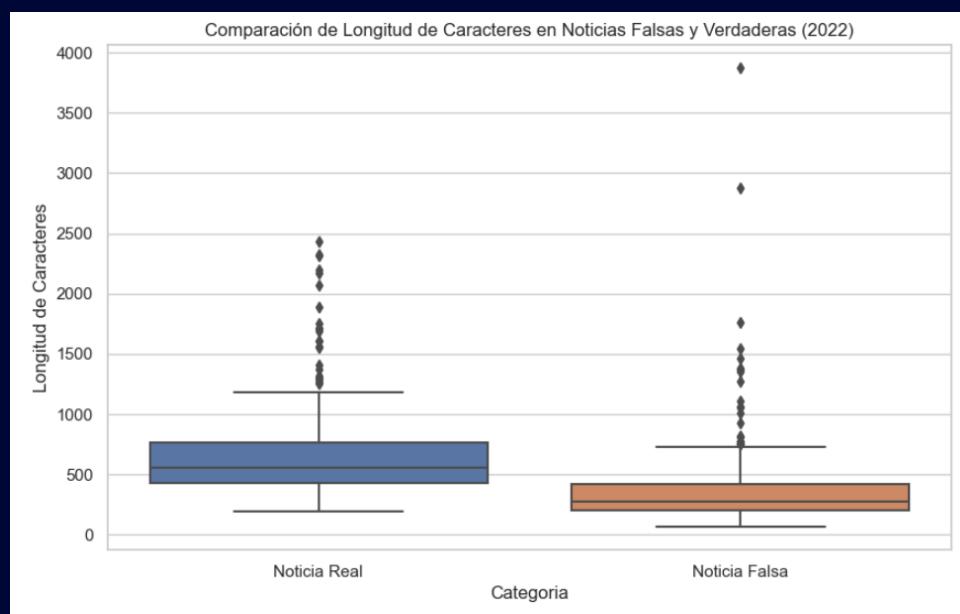


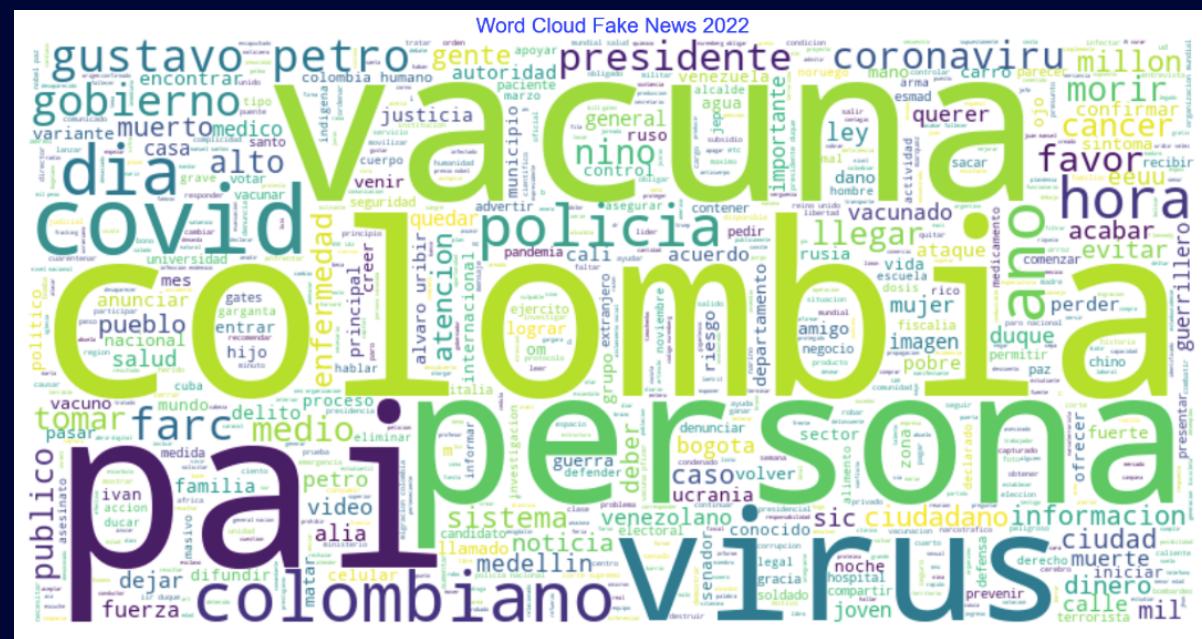
X

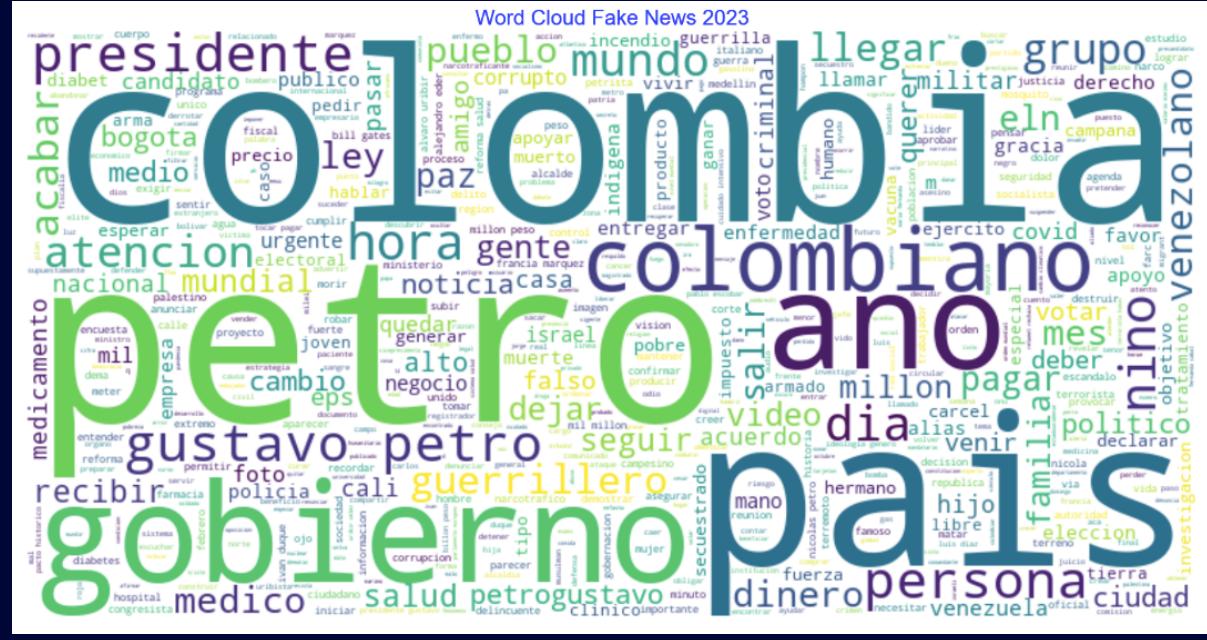




	Categoría	Texto	longitudCaracteres	Processed_Text	label
1443	Noticia Falsa	Esta imagen me dejó marcado, yo la recuerdo y siento ira, M19 LA VIOLARON, LA EMBARAZARON DIZQUE PARA HACERLE UN HIJO DEL PUEBLO,Y LUEGO LA ASESINARON. Gloria Lara secuestrada, violada, torturada y, finalmente, asesinada vilmente por el M-19 de Gustavo Petro y Antonio Navarro. Fue hallada muerta en estado de embarazo causado por sus captores del M-19	353	imagen dejar marcado recuerdo siento ira m violar embarazar dizque hacer el hijo puebloy asesinar gloria laro secuestrado violado torturado finalmente asesinado vilmente m gustavo petro antonio navarro hallado muerto embarazo causado captor m	1
1444	Noticia Falsa	En Alemania el gobierno subió el precio del combustible y en sólo una hora de tiempo la gente abandonó sus coches en las calles y avenidas y se fueron caminando a casa. Más de un millón de coches abandonados. Tuvieron qué bajar el precio del combustible. Cuando el pueblo es inteligente, los corruptos no logran concretar sus metas	331	alemania gobierno subir precio combustible solo horo gente abandonar coche call avenida caminar casa mas millon coche abandonado que bajar precio combustible pueblo inteligente corrupto lograr concretar meta	1
1445	Noticia Falsa	Gustavo Petro es un demonio que le ha dado el alma al diablo. hoy el Petro y sus secuaces comandan un narcoestado que atenta en contra de la seguridad del pueblo estadounidense. 5-feb-2023 ya marchamos ya plantoneamos y el pacto diabólicos continúa destruyendo nuestro país	273	gustavo petro demonio alma diablo petro secuases comandir narcoestado atento seguridad pueblo estadounidense feb marchamo plantoneamo pacto diabolico continuar destruir pais	1







Word Cloud Fake News 2023



Word Cloud Real News 2023



ID	Categoría	Tema	Origen	Título	Texto	Fuente
0 1	Noticia Real	salud	el espectador	Pese a las advertencias, en diciembre ya van 440 lesionados por pólvora.	<p>Pese a las advertencias, en diciembre ya van 440 lesionados por pólvora.\n\nLa situación empieza a ser muy preocupante, pues hay 50 casos más que en el mismo período del 2022. \n\nCada vez que llega diciembre, las advertencias se repiten: no hay que usar pólvora para celebrar las fiestas. Mucho menos hay que hacerlo mientras se ingiere alcohol y hay que evitar que los niños y niñas manipulen estos elementos. Pero las peticiones, al menos este año, no han tenido el efecto deseado. El último boletín del Instituto Nacional de Salud (INS) muestra números muy inquietantes: desde el 1 de diciembre ha habido 440 lesionados por pólvora pirotécnica, es decir, 50 más que en el mismo período de 2022</p>	https://www.elspectador.com/salud/pese-a-las-advertencias-en-diciembre-de-2023-ya-van-440-lesionados-por-polvora/

X

Categoría	Texto
0 Noticia Real	El presidente Iván Duque condenó fuertemente el atentado terrorista que se presentó en la noche de este viernes, 7 de enero, contra miembros del Escuadrón Móvil Antidisturbios (Esmad) en la ciudad de Cali.
1 Noticia Real	Gustavo Petro se reunirá con Pedro Sánchez, presidente de España Hasta ahora es el único precandidato que se reuniría con un primer mandatario. Lo hace en el marco de su campaña internacional.



Vectorización y despliegue de los modelos

Vectorización - Tfifdf

01

¿Qué es?

+info

02

¿Cómo
funciona?

+info

03

Hiperpara
metrización

+info

04

Implemen
tación

+info



```
def tifdf_features (X_train, X_test, y_train, y_test,
                    ngram_range, max_df, min_df, max_features, norm, sublinear_tf):
    tfidf = TfidfVectorizer(encoding='utf-8',
                           ngram_range=ngram_range,
                           stop_words=None,
                           lowercase=False,
                           max_df=max_df,
                           min_df=min_df,
                           max_features=max_features,
                           norm=norm,
                           sublinear_tf=sublinear_tf)
    features_train = tfidf.fit_transform(X_train)
    features_test = tfidf.transform(X_test)
    labels_train = y_train
    labels_test = y_test
    return features_train,features_test,labels_train,labels_test,tfidf
```

- Función para vectorizar
- Argumentos: conjuntos de entrenamiento y prueba e hiperparámetros
- Retorna: conjuntos de test y prueba vectorizados y método instanciado
- Implementamos la función tanto para 2023 como para concat
- test_size = 0.20
- retorna vectorizadores diferentes para 2023 y para concat



Term Frequency (TF):

- Mide la frecuencia con la que una palabra específica aparece en un documento.

- Fórmula:

$$TF(t, d) = \frac{\# \text{ de veces que } t \text{ aparece en el documento } d}{\# \text{ total de palabras en el documento } d}$$

- El valor de TF varía de 0 a 1.

TF-IDF

- Producto de TF e IDF
- Fórmula:

$$TF-IDF(t, d, D) = TF(t, d) * IDF(t, D)$$

- Cuantifica la importancia de una palabra en un documento específico en relación con su importancia en el conjunto de documentos.

Inverse Document Frequency (IDF):

- Mide la importancia de una palabra en el conjunto de documentos.
- Fórmula:

$$IDF(t, D) = \log \left(\frac{\# \text{ total de documentos en } D}{\# \text{ total de documentos que contienen } t + 1} \right) + 1$$

- El valor de IDF tiende a ser más alto para palabras que son menos comunes en el conjunto de documentos.



```
param_grid = {  
    'tfidf_max_df': [0.7, 0.8, 0.9],  
    'tfidf_min_df': [0.001, 0.01, 0.1],  
    'tfidf_max_features': [1000, 2000, 3000, 4000],  
    'tfidf_norm': ['l1', 'l2'],  
    'tfidf_sublinear_tf': [True, False],  
    'tfidf_ngram_range': [(1, 1), (1, 2), (1, 3)]  
}  
  
#instanciamos el vectorizador y el modelo sobre los que vamos a probar los resultados  
tfidf_vectorizer = TfidfVectorizer()  
logreg = LogisticRegression(random_state=SEED)  
  
#Creamos un pipeline con el vectorizador y el modelo  
pipeline = Pipeline([  
    ('tfidf', tfidf_vectorizer),  
    ('classifier', logreg)  
])  
  
#Creamos el objeto grid_search  
grid_search = GridSearchCV(pipeline, param_grid, cv=5, scoring='accuracy', n_jobs=1)  
  
#Entrenamos el modelo sobre el conjunto completo de datos  
grid_search.fit(X, y)  
  
#Imprimimos los mejores parámetros  
best_params = grid_search.best_params_  
print("Mejores hiperparámetros:", best_params)
```

- X=df_news_concat['Process ed_Text']
- y=df_news_concat['label']
- Realicé la hiperparametrización del vectorizador sobre el conjunto de datos más grandes (df_news_concat)
- Los hiperparámetros se buscan a partir de una RL (modelo menos complejo)
- max_df = 0.7
- min_df = 0.001
- max_features = 4000
- norm: l2
- sublinear_tf = True
- ngram_range = (1, 3)

Resultados ML

Modelos:

RL

SVM

NB

RF

XGB

test_size = 0.20

Cada uno hiperparametrizado tanto para 2023 como para concat. RF y XGB con Random Search el resto con GridSearch

Modelo	ACC_2023	ACC_Concat
RL	0.8834	0.9103
SVM	0.8711	0.9137
NB	0.8220	0.8758
RF	0.8650	0.8758
XGB	0.8343	0.8517

Concurso

¿Cuál es el mejor modelo para detectar noticias falsas?



SVM Concat



K-Means 2023



VHS Concat

SEND



Modelos de transfer learning (BERT)

Transfer Learning



Intuición básica

Para alguien que sepa conducir un coche, será más fácil aprender a conducir un camión



Implementación

Es necesario reentrenar las capas finales del modelo pre-entrenado para obtener resultados (fine-tuning)

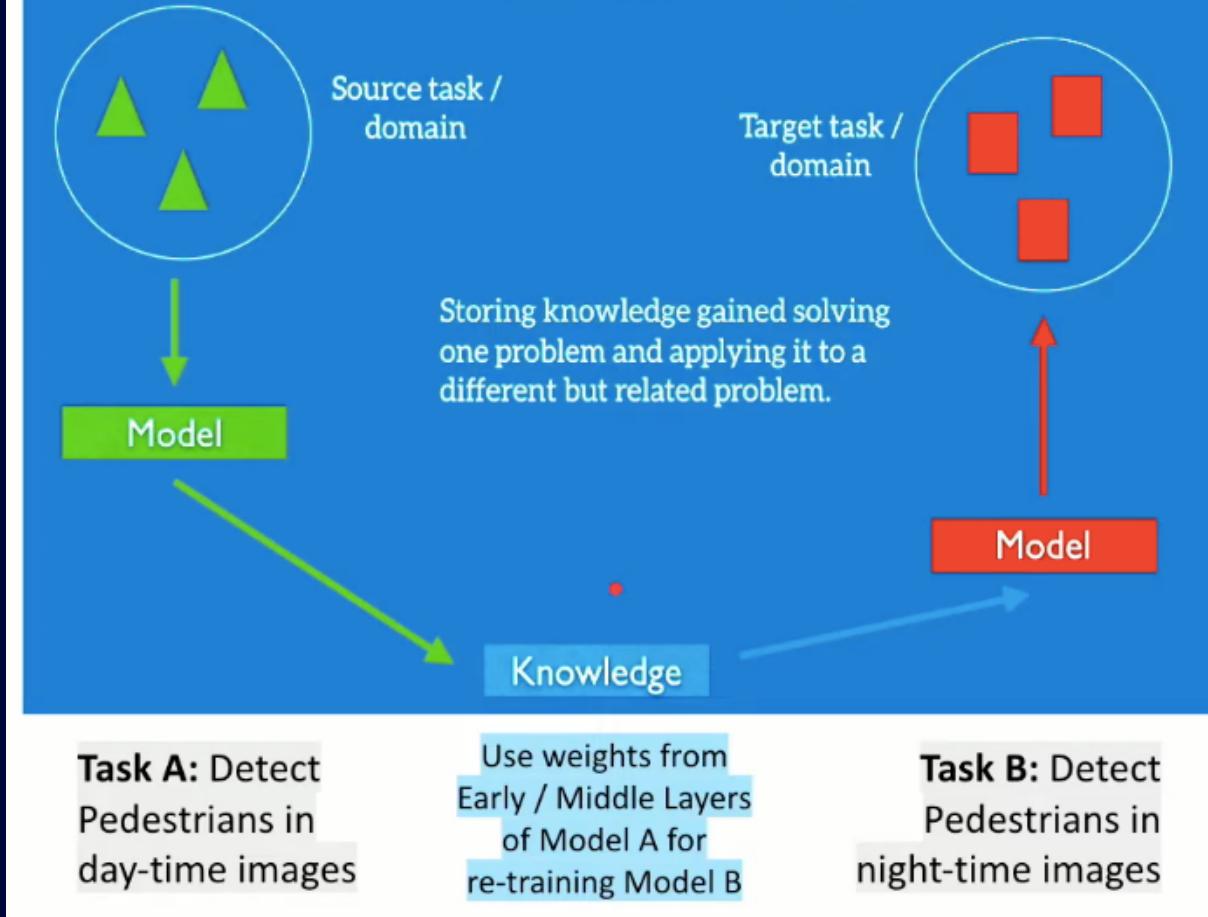


Aplicaciones

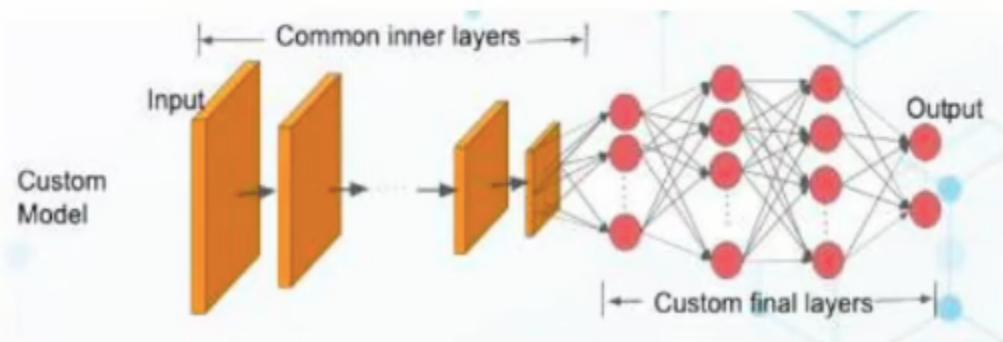
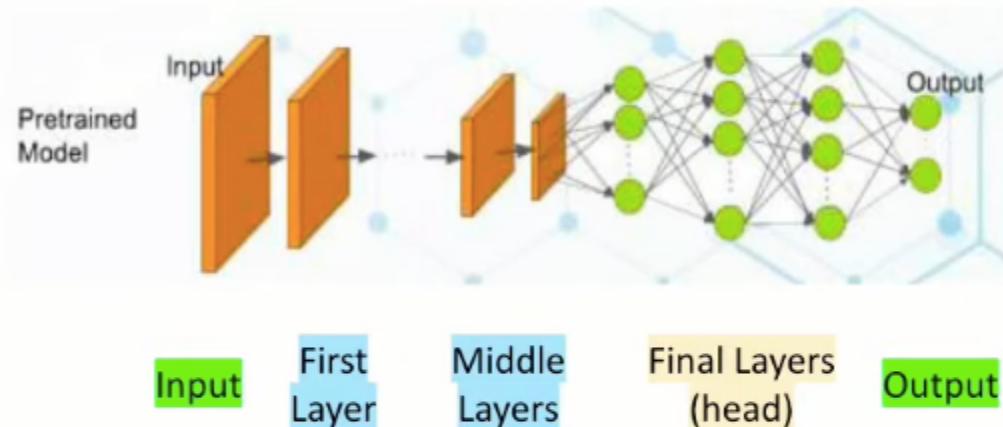
Ahorro de recursos energéticos y capacidad de potenciar datasets pequeños

A small white 'X' inside a dark circular button.

Transfer learning



X





Natural Language Processing

- **BERT**, GPT-2, FastText
- Linguistic Characteristics
- Applications:
 - next word prediction
 - question answering
 - machine translation



Computer Vision

- ResNet, VGG, Xception
- Complex Image Features
- Applications:
 - image recognition
 - object detection
 - image noise removal



Speech / Voice Recognition

- ContextNet
- Audio/Speech Recognition
- Applications:
 - speech recognition
 - speech-to-text
 - translation



BERT

**Bidirectional Encoder Representations
from Transformers**

- Google (2018)
- Wikipedia BookCorpus
- Comprensión contextual

+ INFO



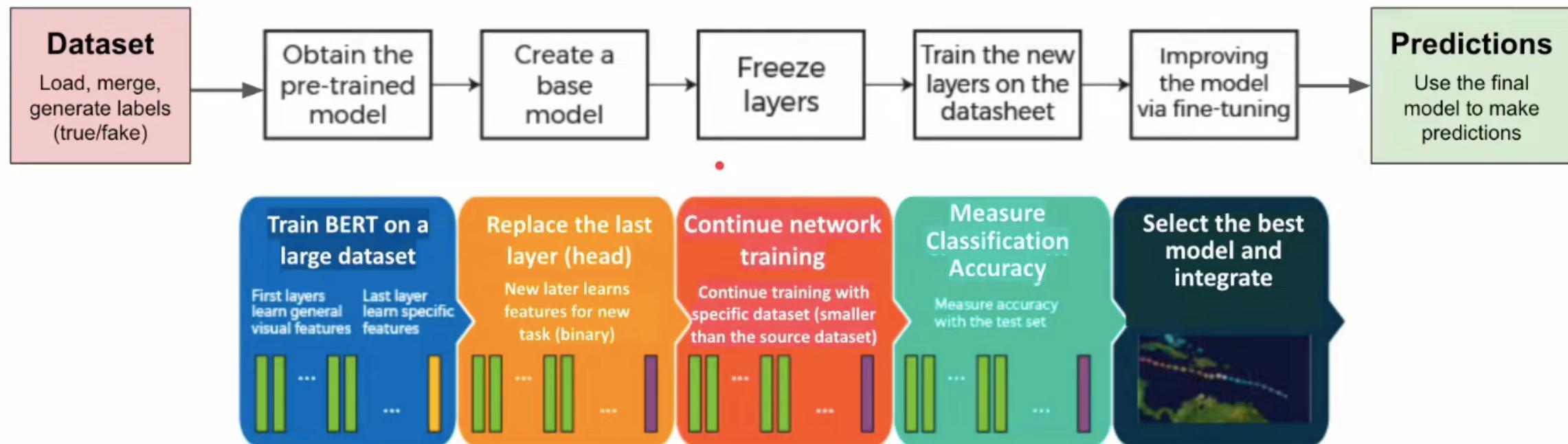
BETO

**Bidirectional Encoder Representations
from Transformers**

- 2020
- Spanish Wikis, ParaCrawl
- Número de líneas: 300904000 (300M)
- Número de tokens: 2996016962 (3B)

+ INFO

Implementación del modelo





Comparación de resultados

Métricas

0.91

SVM Concat

+ INFO

Precisión similar para [0] y
para [1]

F-Score similar para [0] y
para [1]

0.92

BETO

+ INFO

Mejor precisión para [1] que
para [0]

F-Score similar para [0] y para
[1]

Confusion Matrix

Label_Test

Noticias verdadera

135

12

Noticias Falsa

13

130

Noticias verdadera

Noticias Falsa

Predicted

X

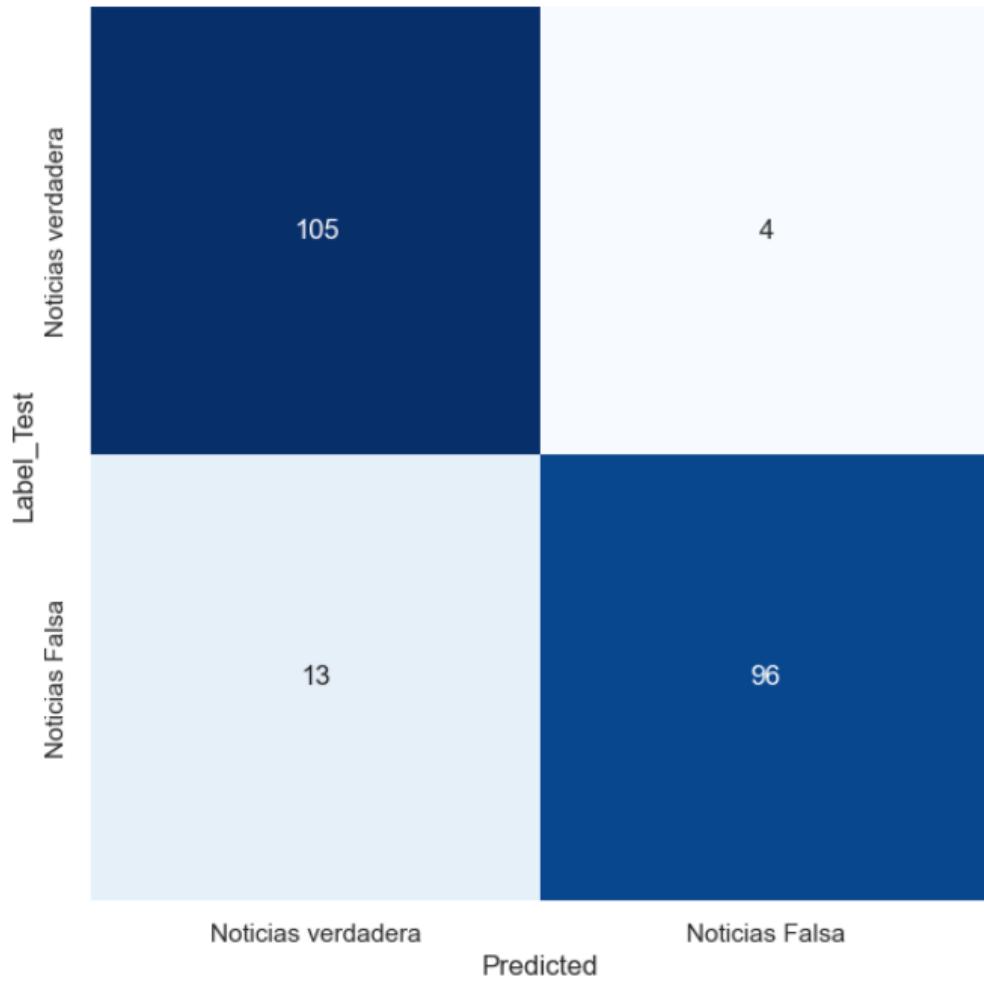
Exactitud (Accuracy): 0.9137931034482759

Informe de clasificación:

	precision	recall	f1-score	support
0	0.91	0.92	0.92	147
1	0.92	0.91	0.91	143
accuracy			0.91	290
macro avg	0.91	0.91	0.91	290
weighted avg	0.91	0.91	0.91	290

X

Confusion Matrix



	precision	recall	f1-score	support
0	0.89	0.96	0.93	109
1	0.96	0.88	0.92	109
accuracy			0.92	218
macro avg	0.92	0.92	0.92	218
weighted avg	0.92	0.92	0.92	218



**¿Y entonces?
¿Cambiamos el brócoli
por chicharrón?**

Dataset con noticias nuevas

		titulo	label	tipo
0		Nuevo uniforme de la Policía será reversible	1	sátira
1		Botiquines de vehículos deberán incluir frasco de ayahuasca	1	sátira
2	Ayudenos a pasar esta información! En los próximos días habrá un Domo Atmosférico de Calor		1	evidentemente falsa
3	Fallo de la Corte Suprema de Estados Unidos: Las vacunas contra el Covid NO son vacunas		1	evidentemente falsa
4	Petro reconoce responsabilidad de su Gobierno y el de Duque en pérdida de Panamericanos		0	evidentemente cierta
5	Capturan a alias Tato y a alias Gordo, dos de los hombres más buscado de Antioquia		0	evidentemente cierta
6	El chicharrón es más saludable que algunas verduras, según un nuevo estudio		1	salud
7	El director de la OMS pide acabar con el estigma y la discriminación de la lepra		0	salud

Resultados SVM

- Las noticias satíricas (0 y 1) en un alto porcentaje las asume como ciertas
- El acierto porcentual es bastante alto con las noticias falsas de redes sociales.
- A la noticia del chicharrón le asigna una probabilidad porcentual muy alta de ser verdadera (solo el 34.27% de probabilidad de ser falsa).
- A la noticia cierta 4 le da una posibilidad un poco más alta que el azar de ser falsa. A la noticia 5 sí que la califica categóricamente con verdadera.

ID	TIPO	Label_TRUE	Label_PRED	PROB[1]
0	Sátira	1	0	22.60%
1	Sátira	1	0	27.40%
2	Ev Falsa	1	1	91.97%
3	Ev Falsa	1	1	85.48%
4	Ev Cierta	0	1	54.54%
5	Ev Cierta	0	0	2.70%
6	Chicharrón	1	0	34.27%
7	Salud - Cierta	0	0	17.95%

Resultados BETO

- Las noticias sátricas las clasifica como verdaderas pero en los dos casos está muy cerca de clasificarlas como falsas. Podría decir que estas noticias cumplen su papel de confundir al modelo de predicción.
- Las noticias que son evidentemente falsas tienen un porcentaje muy alto de ser falsas (especialmente la noticia 2)
- Las noticias verdaderas (4 y 5) tienen una probabilidad muy alta de ser verdaderas.
- La noticia del chicharrón, aunque la califica como verdadera en este modelo tiene más probabilidad de ser falsa (39.48%) que con el modelo de ML (34.27%).
- La noticia 7 que es verdadera en este modelo tiene mayor probabilidad de ser falsa (45.35%) que en el modelo ML (17.95%)

ID	TIPO	Label_TRUE	Label_PRED	PROB[1]
0	Sátira	1	0	48.09%
1	Sátira	1	0	46.64%
2	Ev Falsa	1	1	81.29%
3	Ev Falsa	1	1	57.17%
4	Ev Cierta	0	0	7.26%
5	Ev Cierta	0	0	6.84%
6	Chicharrón	1	0	39.48%
7	Salud - Cierta	0	0	45.35%

Conclusiones

01

Unificar los criterios de extracción con miras a realizar investigaciones cooperativas

02

Esto es un párrafo de texto listo para escribir un contenido genial

03

No somos nadie sin los verificadores, pero...

Bibliografía

Ansar, W., & Goswami, S. (2021). Combating the menace: A survey on characterization and detection of fake news from a data science perspective. International Journal of Information Management Data Insights, 1(2), 100052. ISSN 2667-0968.

Flores Quinayás, J. E., & Montaño Morcillo, J. G. (2022). Modelo para la detección de noticias falsas en formato texto en la red social Twitter, aplicado al contexto político colombiano de las elecciones presidenciales de 2022 (Proyecto de grado). Universidad ICESI, Facultad de Ingeniería, Maestría en Ciencia de Datos, Santiago de Cali.

Hernández Bonilla, J. M. (14 de diciembre de 2023). No, el chicharrón no es más saludable que las verduras y no existe ningún estudio que lo demuestre. El País. <https://elpais.com/america-colombia/2023-12-14/no-el-chicharron-no-es-mas-saludable-que-las-verduras-y-no-existe-ningun-estudio-que-lo-demuestre.html> (Consultada el 24 de enero de 2024)

Mafla, N., Flores, M., Castillo-Páez, S., & Andrade, R. (2022). Automatic Detection of Fake News in Spanish: Ecuadorian Political Satire. Revista Politécnica, 50(3), 7-16.

Núñez Arroyo, A. (2022). Sistema de Detección de Noticias Falsas en Formato de Texto para Idioma Español Ajustado al Contexto Boliviano (Proyecto de grado). Universidad Católica Boliviana "San Pablo", Sede La Paz, Facultad de Ingeniería, Carrera de Ingeniería Mecatrónica. Tutor: Guillermo Sahonero Álvarez. La Paz, Bolivia.

Posadas-Durán, J. P., Gómez-Adorno, H., Sidorov, G., & Escobar, J. J. M. (2019). Detection of fake news in a new corpus for the Spanish language. Journal of Intelligent & Fuzzy Systems, 36(5), 4869-4876.

Ortega Riveros, J. A., & Quintero Perozo, D. Y. (2020). Detección automática de noticias falsas en español con técnicas de machine learning. Trabajo de grado dirigido por Haydemar María Núñez Castro. Universidad de los Andes, Facultad de Ingeniería, Ingeniería de Sistemas y Computación. Bogotá, Colombia.

Tretiakov, Arsenii & Martín García, Alejandro & Camacho, David. (2022). Detection of False Information in Spanish Using Machine Learning Techniques. 10.1007/978-3-031-21753-1_5.

Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Comput. Surv. 1, 1, Article 1 (January 2020)



¡Muchas gracias!

<https://github.com/juanviatela>