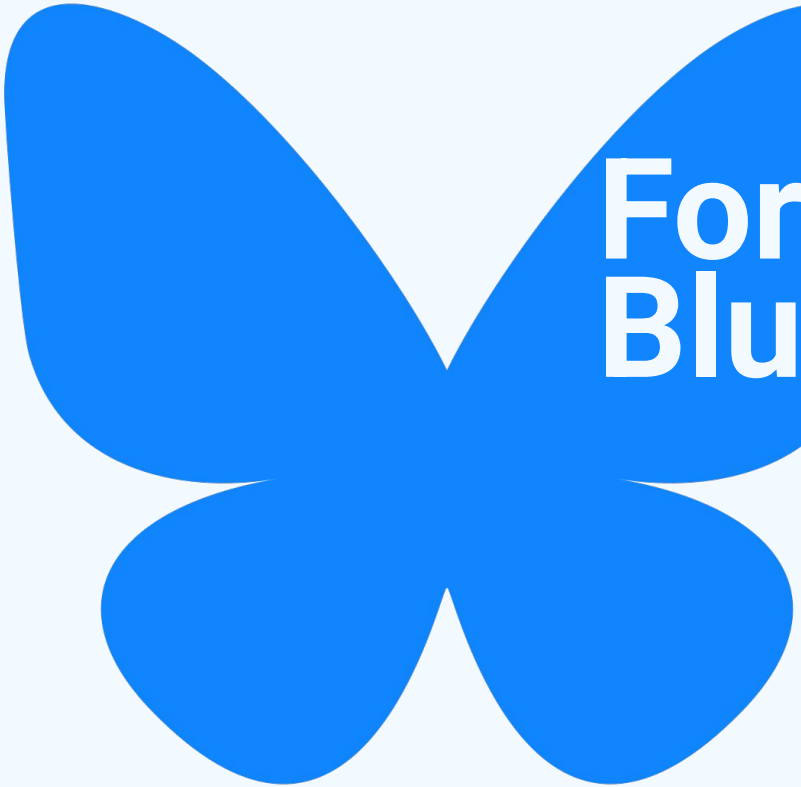


Forecasting Bluesky

Proyecto final del Bootcamp
Data Science & Machine Learning

Camino de Pablos

Predicción del crecimiento de usuarios a través del análisis de actividad y el impacto de noticias

Ver repositorio del proyecto >

Forecasting Bluesky
Data Science Lifecycle

Forecasting Bluesky
Data Science Lifecycle - 01. Identifying Problems

01

Identifying Problems

Bluesky
Definición de hipótesis
Búsqueda de referencias

02

Data Collection

Web Scraping
& Data Scraping

03

Data Processing

NLP
Feature Engineering
Series temporales
Escalado de datos
Categorización

04

Data Analysis

EDA Bluesky Stats
EDA noticias diarias
EDA correlación

05

Data Modeling

Impact Score
(NLP & XGBoost)
Linear Regression
Logistic Regression

06

Model Deployment

Streamlit
GitHub

contexto

¿Qué está pasando con los usuarios de X (Twitter)?

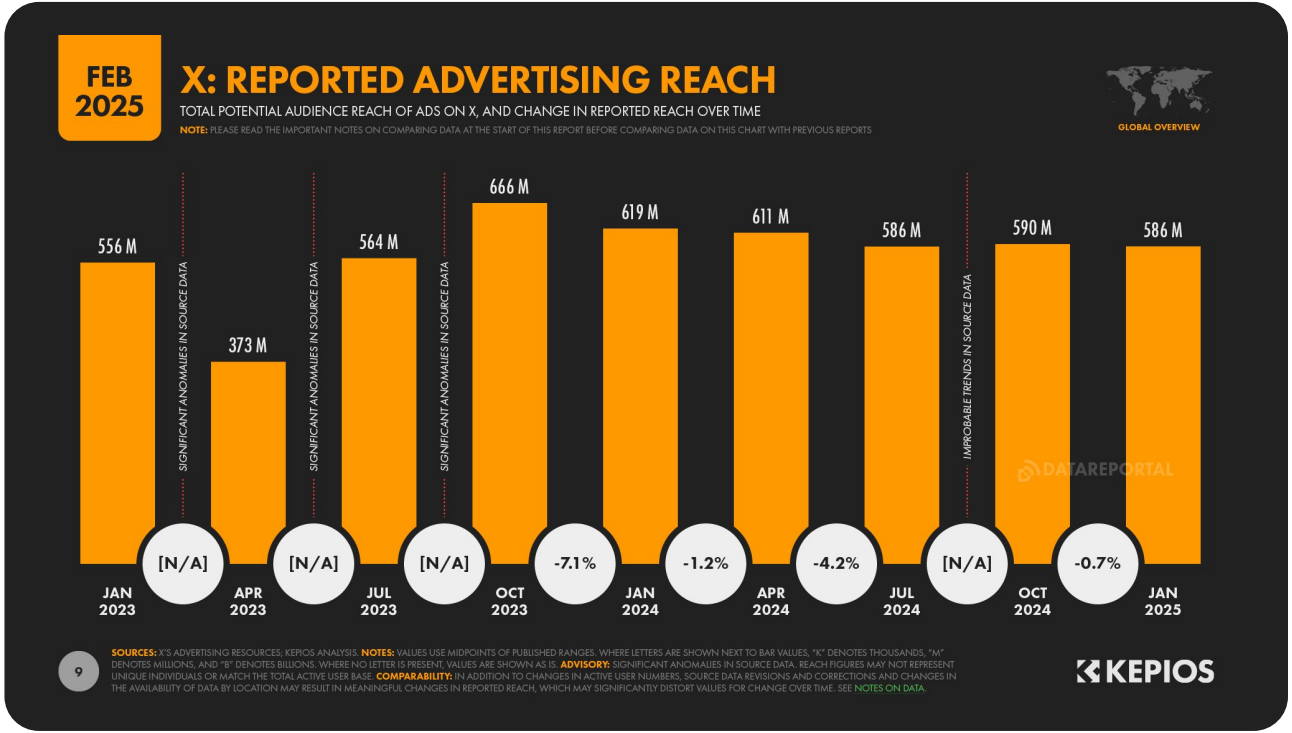
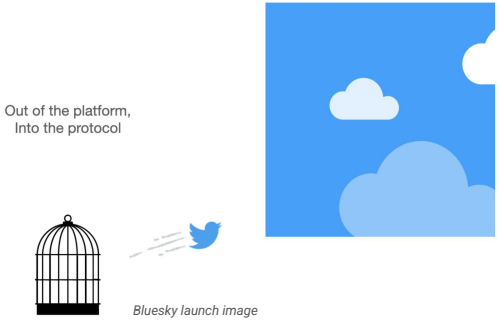


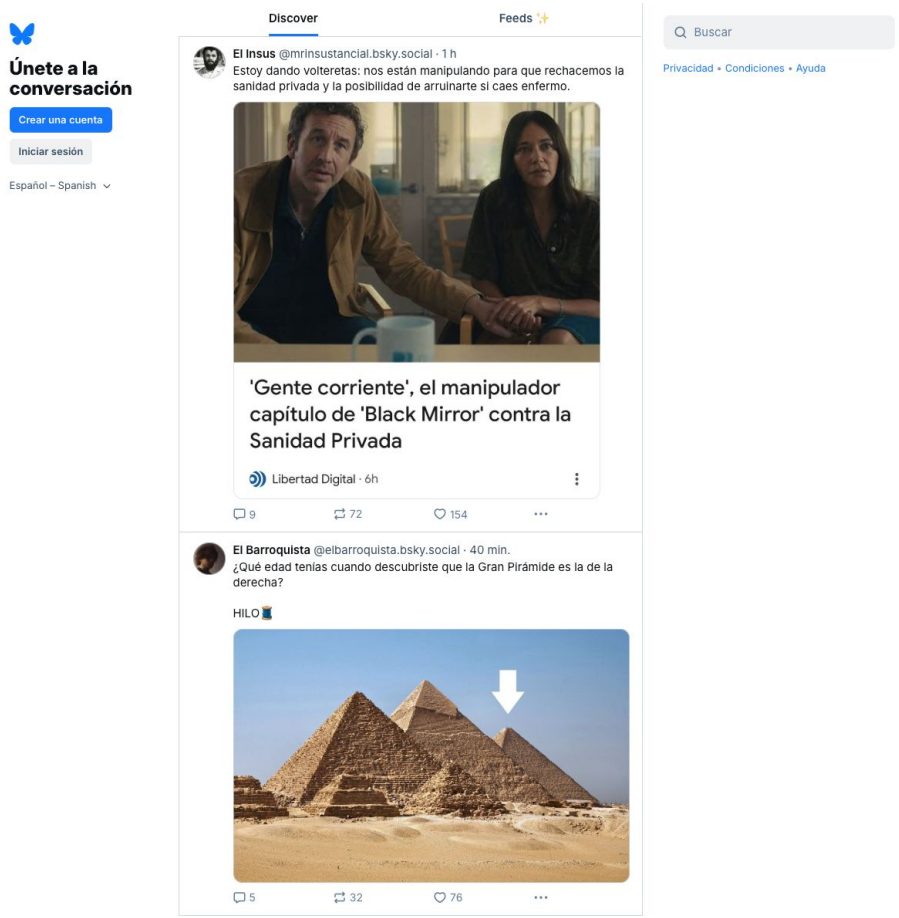
Gráfico: https://datareportal.com/essential-x-stats

qué es bluesky

The public deserves a thriving online commons. We're committed to building this space and ensuring that your social network can never be bought by a single individual or organization.



Bluesky launch image

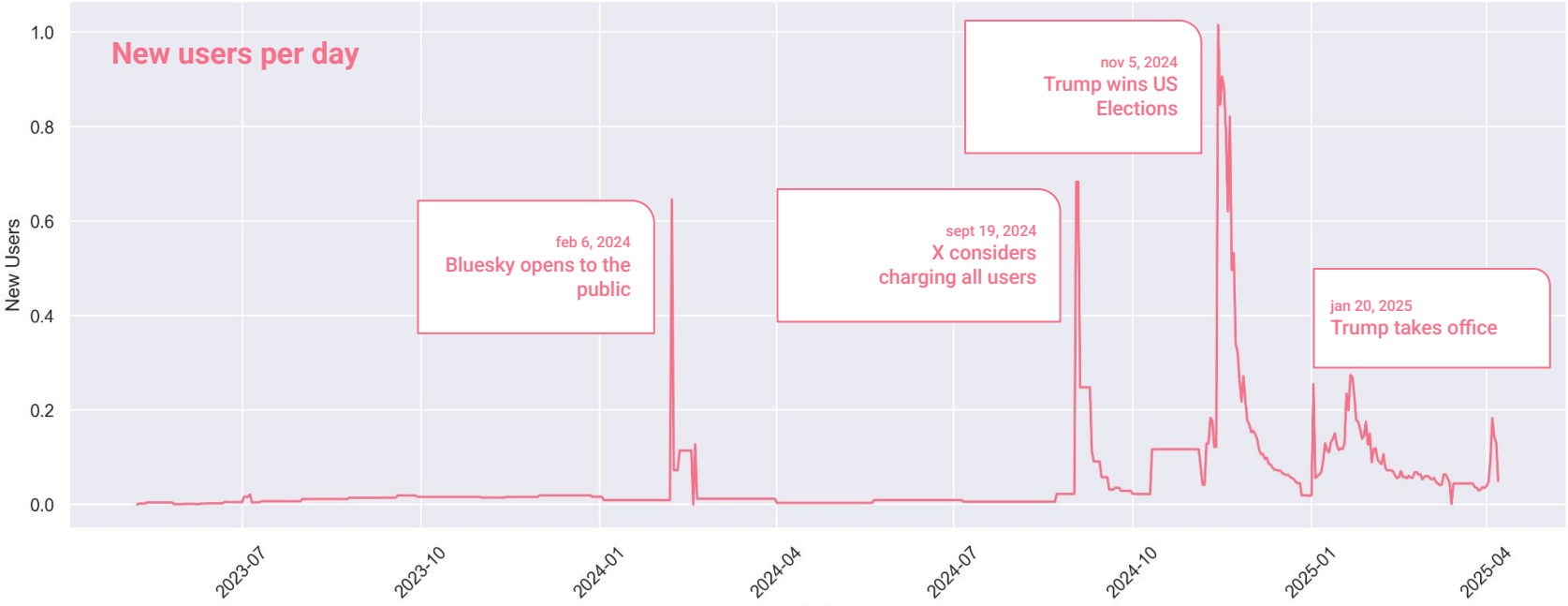


hipótesis y objetivos

Existe una correlación entre los cambios políticos y sociales y el aumento de usuarios en Bluesky.

Predicción del aumento de usuarios diarios en la red social.

Predicción del impacto de noticias en el crecimiento de la red social.



retos

- **Pocos datos**, Bluesky se hace público en mayo de 2023.
- **Serie temporal incompleta**, hasta nov de 2024 no tenemos registros diarios de nuevos usuarios pero sí de estadísticas y noticias.
- **Hipótesis**, buscamos hacer una correlación a partir de una hipótesis no probada.
- **La selección de noticias**, se complica por el acceso a las API y por la relevancia.

Forecasting Bluesky
Data Science Lifecycle - 02. Data Collection

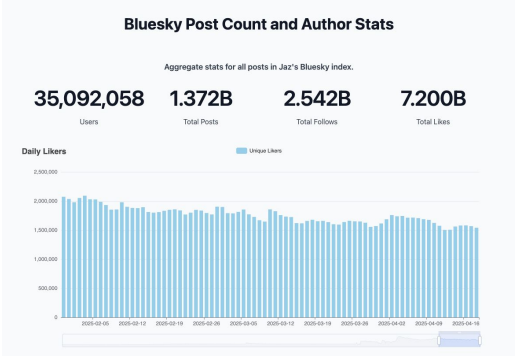
data collection stats

Actividad de Bluesky por día

Fuente:
Bluesky Stats by Jaz
bsky.jazco.dev/stats

Tamaño:
769x17

Herramientas:
javascript, json, os, sys

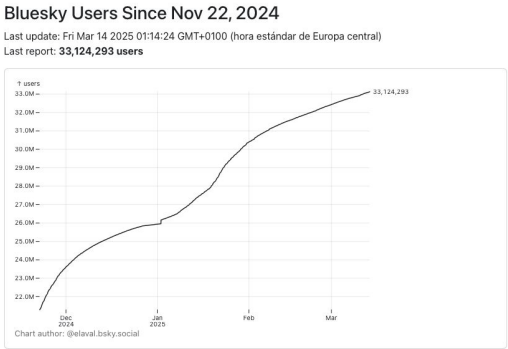


Número de usuarios agregado de Bluesky

Fuente:
https://elaval.github.io/bluesky-users/

Tamaño:
3172x7

Herramientas:
javascript, json, os, sys



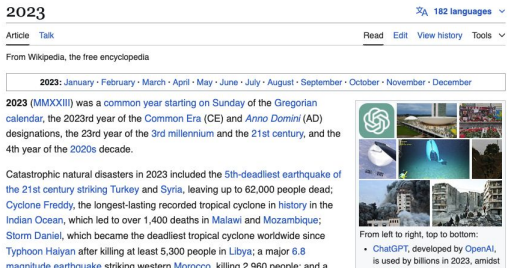
data collection news

Noticias de 2023 a 2025 más relevantes por temática

Fuente: Wikipedia
https://www.wikipedia.org/

Tamaño: 5068x10

Herramientas: os, sys

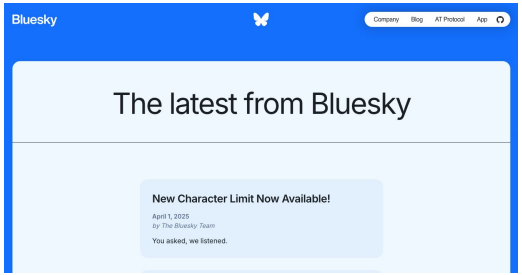


Noticias propias de Bluesky

Fuente: Bluesky Blog
https://bsky.social/about/blog

Tamaño: 39x6

Herramientas: Web Scraping con Beautiful Soup

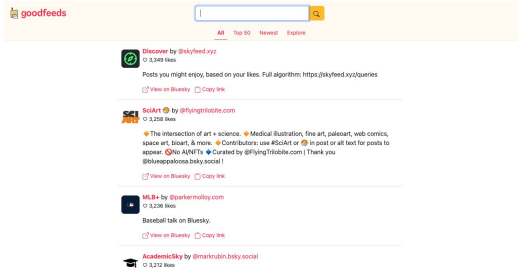


Feeds de Bluesky (de qué se habla)

Fuente: Goodfeeds
https://goodfeeds.co/all

Tamaño: 200x4

Herramientas: Web Scraping con Beautiful Soup



Forecasting Bluesky
Data Science Lifecycle - 03.Data Processing

stats data processing

Imputación temporal

Método: Interpolación lineal, asume que el cambio entre dos puntos sigue una línea recta

Objetivo: completar la serie temporal.

Series temporales

Lags: valor del día anterior.

Rolling Mean: promedio de los últimos 7 días.

Diferencias diarias: de nuevos usuarios y actividad.

Categorización

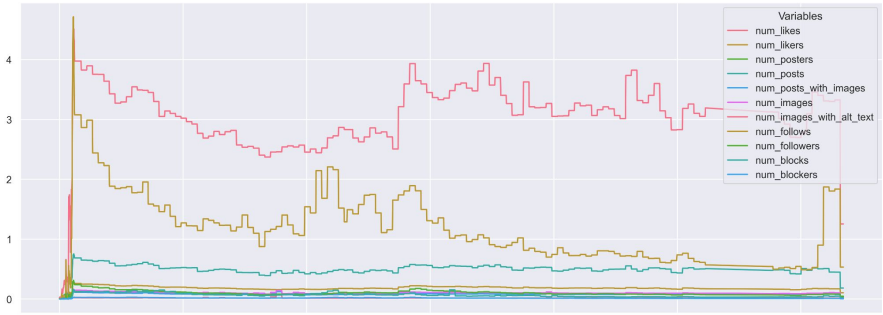
Categorización por quantiles de rangos de actividad.

De 0 (muy poca actividad) a 4 (actividad muy elevada).

PCA

‘Primary Component Analysis’ para las ‘activity stats’.

Se detecta una fuerte correlación entre todas las estadísticas de actividad en la red.



news data processing

Text Cleaning

Método: librería re

Objetivo: limpiar y unificar todos los headlines del dataset.

Stemming

Método: reducción de palabras a su raíz.

Objetivo: reducir el número de palabras y unificar el léxico.

TF-IDF

Term Frecuency - Inverse Document Frecuency

Vectorización de los headlines.

Zero-Shot-Classification

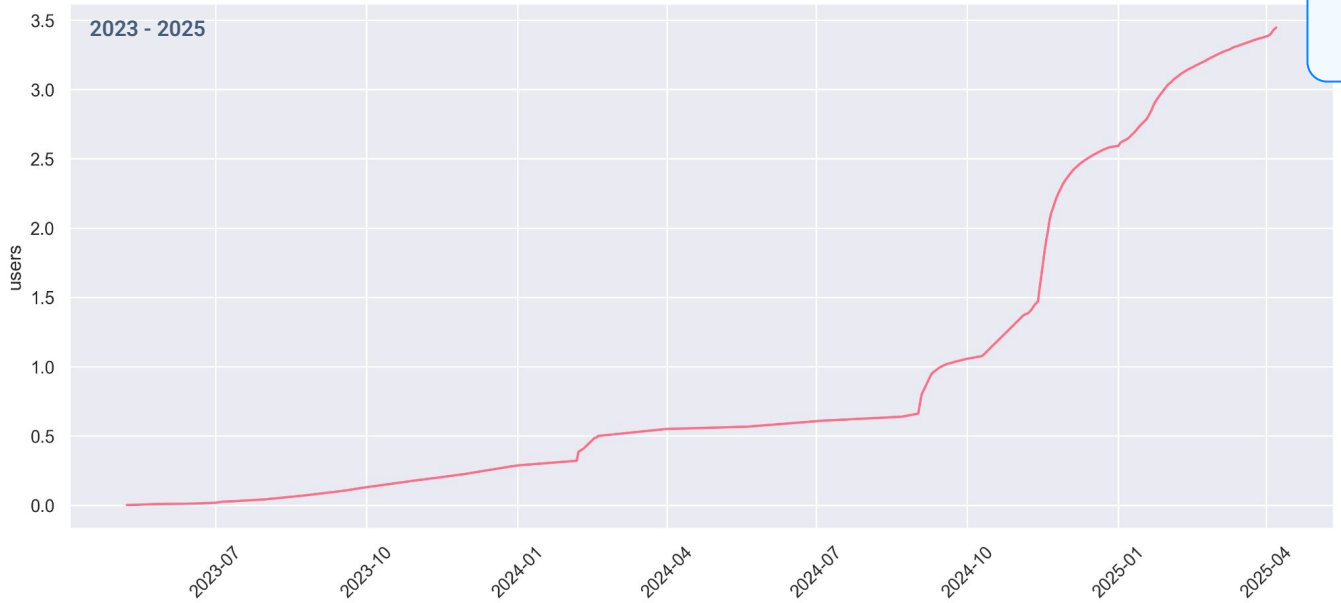
Método**:** modelo de NLP preentrenado

Modelo: facebook/bart-large-mnli (Hugging Face)

Objetivo: clasificar las noticias por *subject*.

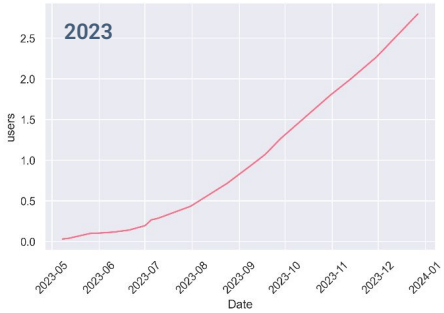
Forecasting Bluesky
Data Science Lifecycle - 04. Data Analysis

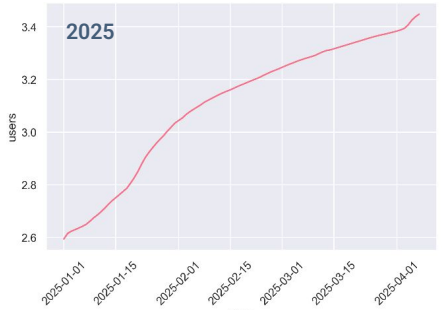
eda bluesky stats



35,211,016 users

Total users per day





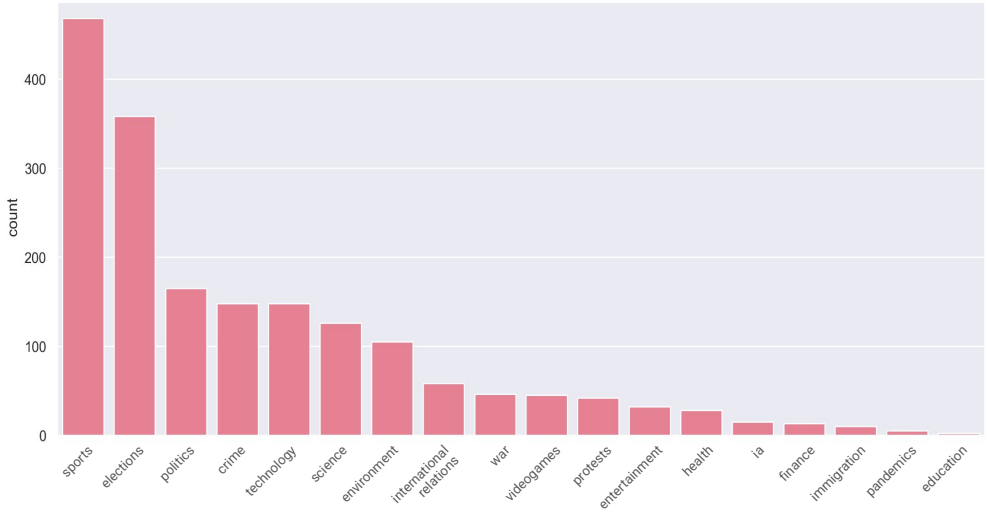
Total users per day, by year

eda bluesky stats

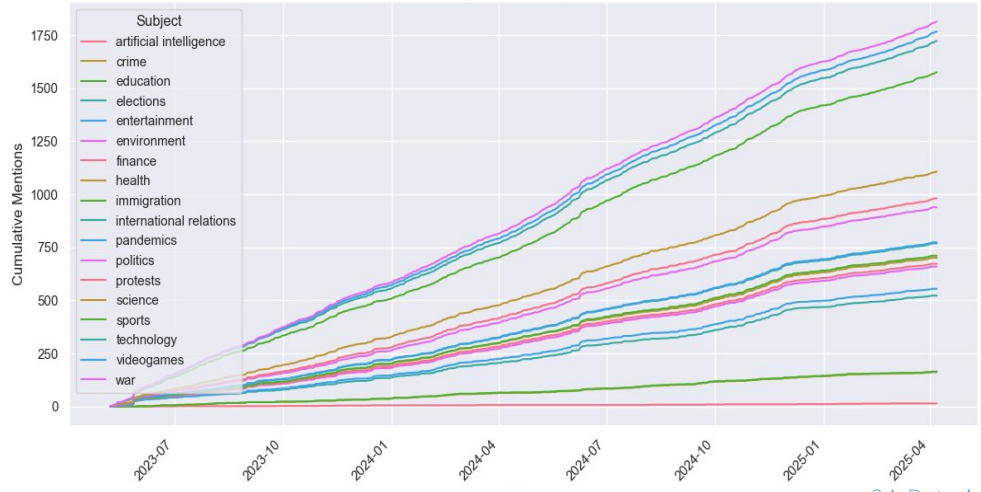
New users per day, by year

Activity per day, by year

eda news

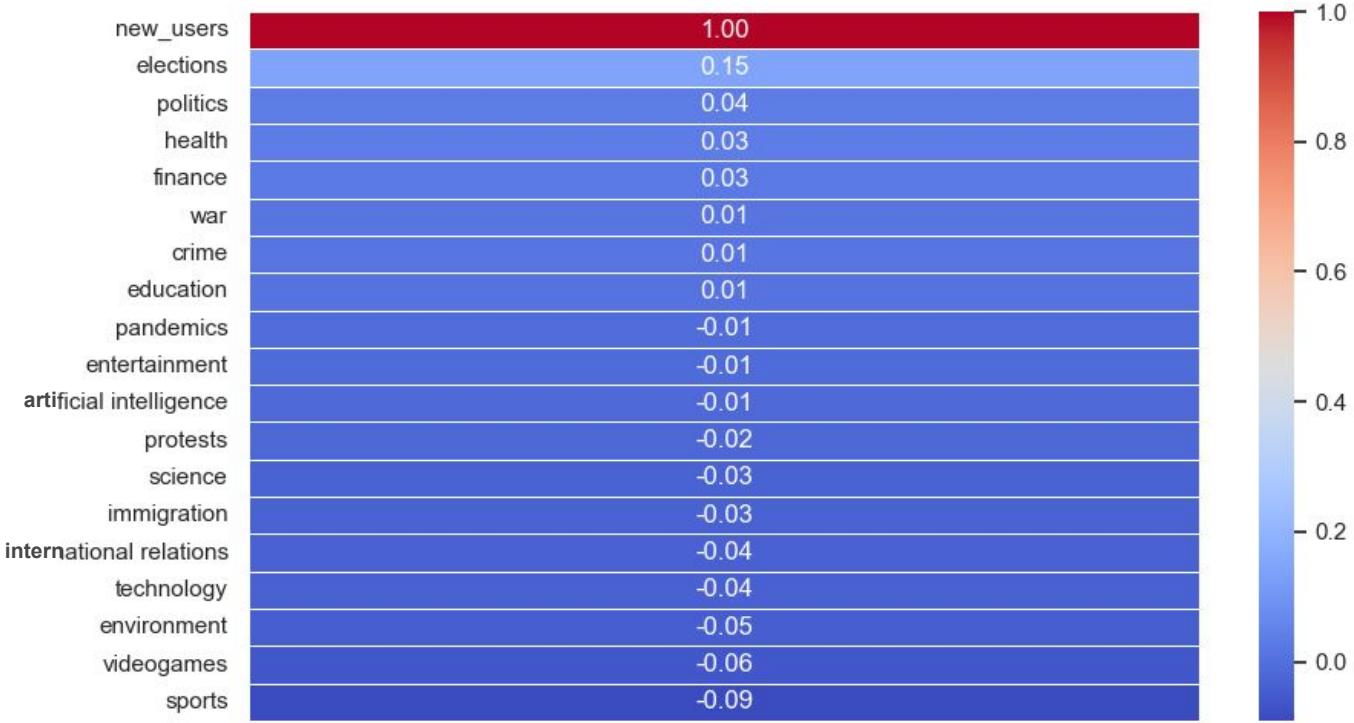


News Topics



Cumulative Topic Mentions over 2023-2025

Correlation:
New Users & News Topics



Forecasting Bluesky
Data Science Lifecycle - 05. Data Modelling

01 — 02 — 03 — 04 — **05** — 06

Identifying Problems

Bluesky
Definición de hipótesis
Búsqueda de referencias

Data Collection

Web Scraping
& Data Scraping

Data Processing

NLP
Feature Engineering
Series temporales
Escalado de datos
Categorización

Data Analysis

EDA Bluesky Stats
EDA noticias diarias
EDA correlación

Data Modeling

Impact Score (NLP & XGBoost)
Linear Regression
Logistic Regression

Model Deployment

Streamlit
GitHub

retos

- **Correlación noticias-Bluesky**, difícil de establecer de una forma directa.
- **Outliers de gran magnitud**, el conjunto de datos presenta una tendencia creciente suave y relativamente lineal, por lo que a los modelos les cuesta predecir los outliers.

.impact score

Predicción del **impacto del contenido de las noticias** sobre el crecimiento de usuarios en Bluesky, basada en datos reales históricos.

Componentes:

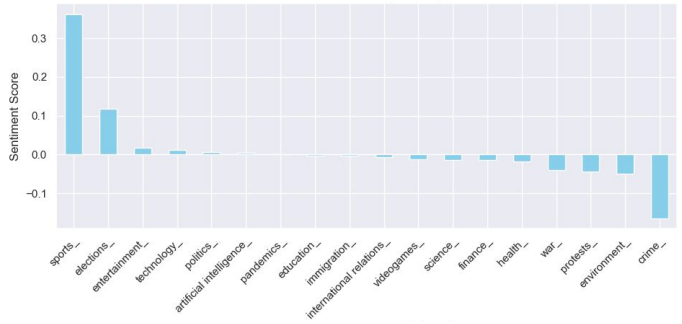
- Sentiment Score
- Category Impact
- NER Impact
- Novelty Score

XGBoost

componentes

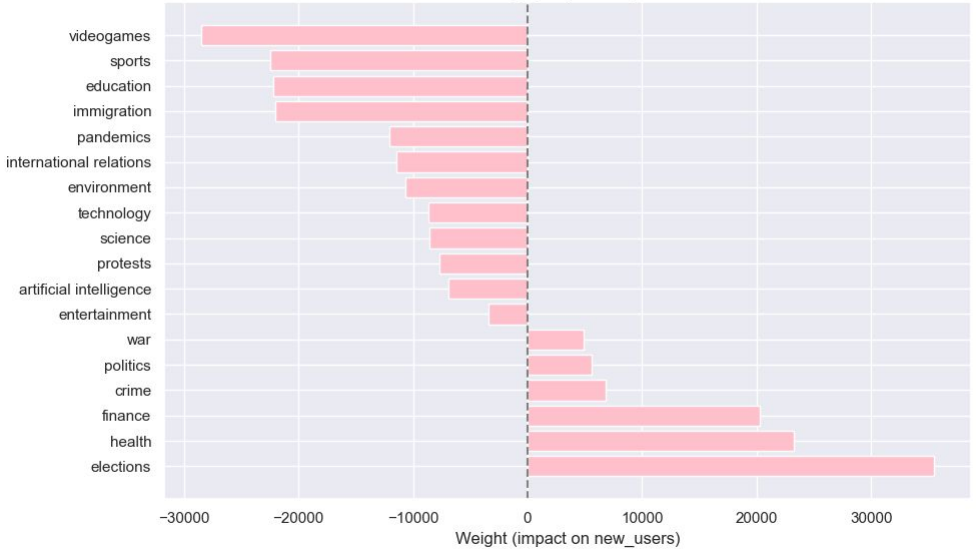
Sentiment Score
SentimentScoreExtractor()

- Análisis de sentimiento de las noticias con DistilBERT, HuggingFace.
- Devuelve un score de sentimiento por cada noticia.
- El score va de -1 (negativo) a 1 (positivo).



Category Impact
CategoryImpactExtractor()

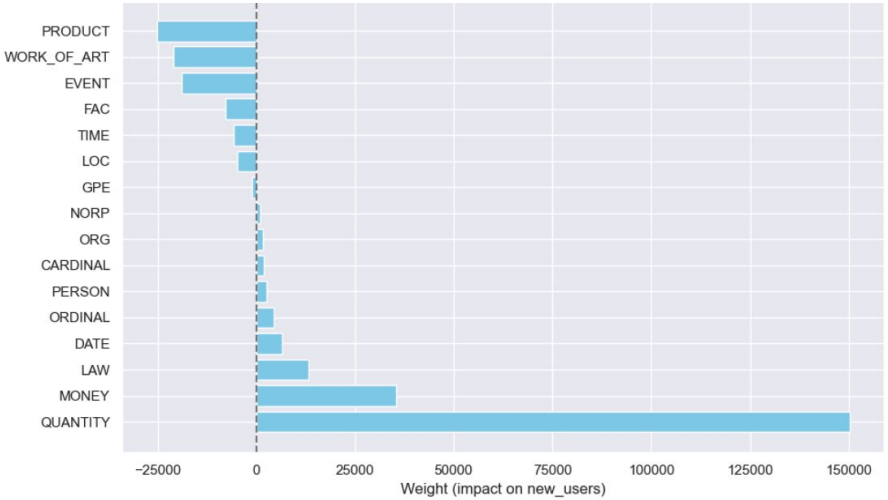
- Impacto de cada temática en el histórico de los datos (ej.: noticias políticas - impacto alto, noticias deportivas - impacto bajo).
- Otorga un score ya fijado a una categoría.
- El score se ha obtenido previamente a través de los coeficientes de una regresión lineal sobre el dataset original.
- El score va de -1 (muy poco impacto) a 1 (impacto elevado).



componentes

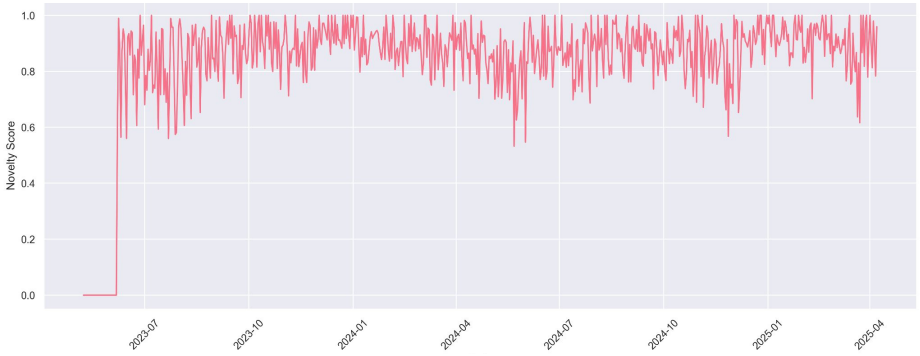
NER Score
NERScoreExtractor()

- NER: Named Entity Recognition con la librería SpaCy.
- Si encuentra un nombre de entidad o persona en el texto, le otorga un score en función a su frecuencia de aparición y a su label.
- El peso de las label se obtiene a través de una regresión lineal sobre el dataset original.
- El score va de -1 (muy poco impacto) a 1 (impacto elevado).



Novelty Score
NoveltyScoreExtractor()

- Cuán diferente es una noticia comparada con los últimos días.
- Se utiliza la matriz TF-IDF y el *cosine-similarity*.
- Devuelve un score de -1 (no novedoso) a 1 (muy novedoso).



Cómo se calcula el Impact Score

01

Score de componentes

Todas las noticias del dataset se evalúan sobre cada uno de los componentes.

```
impact_score_pipeline = FeatureUnion([
    ('sentiment',
     SentimentScoreExtractor()),
    ('category',
     CategoryImpactExtractor()),
    ('ner',
     NERScoreExtractor()),
    ('novelty',
     NoveltyScoreExtractor())
])
```

02

XGBoost Classifier

Se entrena un modelo de clasificación XGBoost con noticias de 2023 y 2024.

```
# X = matriz de scores de comportamiento
X = impact_score_pipeline.fit_transform(X_news)

# y = pico de actividad (0 no hay pico, 1 sí hay pico)
threshold = df['activity_growth_percentage'].quantile(0.93)
df['has_activity_peak'] = (df['activity_growth_percentage']
                           > threshold).astype(int)
```

03

Congelación del modelo

No se vuelve a entrenar para evitar colinealidad en el modelo.

```
xgb = XGBClassifier(scale_pos_weight=scale,
                    eval_metric='logloss',
                    n_estimators=200,
                    max_depth=3,
                    learning_rate=0.01,
                    n_jobs=1,
                    reg_alpha=0.1,
                    reg_lambda=1)

# Entrenamiento
xgb.fit(X_news_2023_2024)
```

04

Score de nuevas noticias

Se pasan las nuevas noticias por el modelo entrenado, el output será su impact score completo.

```
# Predicción = Impact Score
xgb.predict_proba(X_news_2025)
```

05

Imput de otros modelos

El Impact Score se utilizará como imput en los modelos de predicción de Bluesky representando la correlación entre noticias y actividad en la red.

xgboost

linear regression

Target: futuros nuevos usuarios

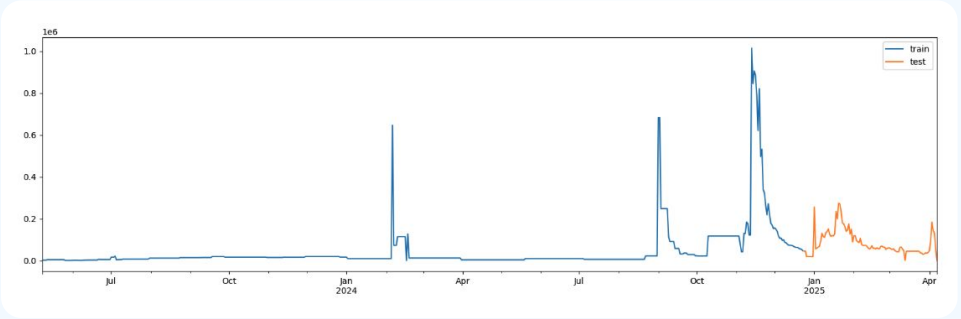
Features:

- Nuevos usuarios
- Total de usuarios
- PCA de actividad
- Series temporales
- Impact Score de noticias

Periodo: diario y semanal

Modelo: LinearRegression() / Ridge()

Train-Test Split: por periodos



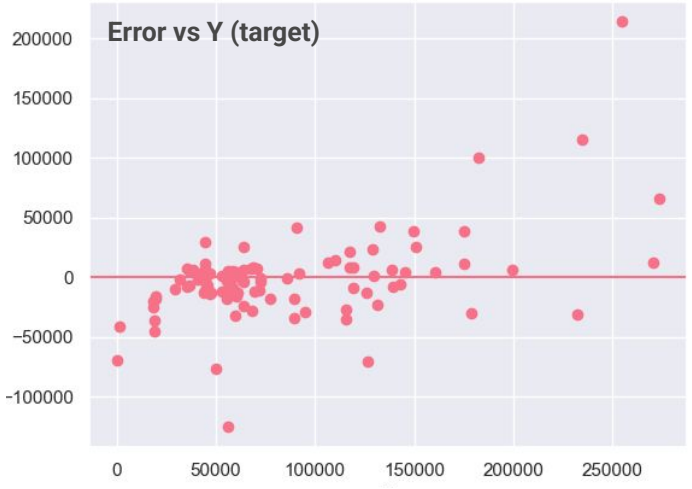
metrics for daily data

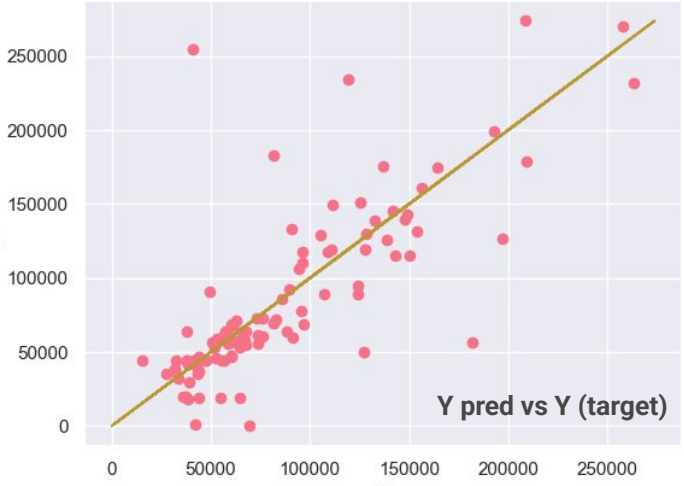
Métricas para el conjunto de train:

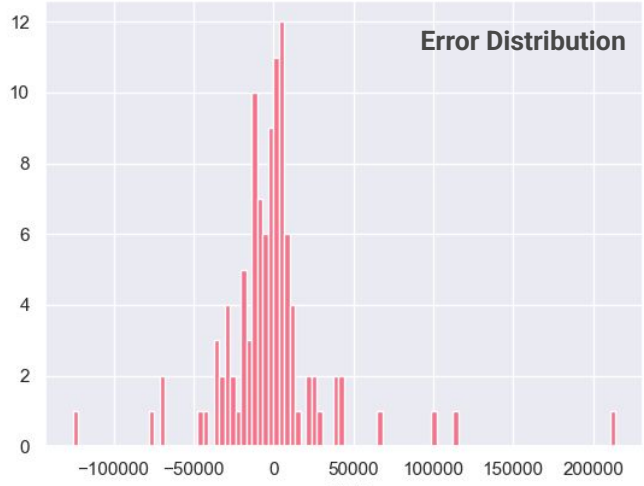
- MSE: 3526004337.32
- R^2: 0.72
- RMSE: 59380.17

Métricas para el conjunto de test:

- MSE: 1288143016.73
- R^2: 0.61
- RMSE: 35890.71







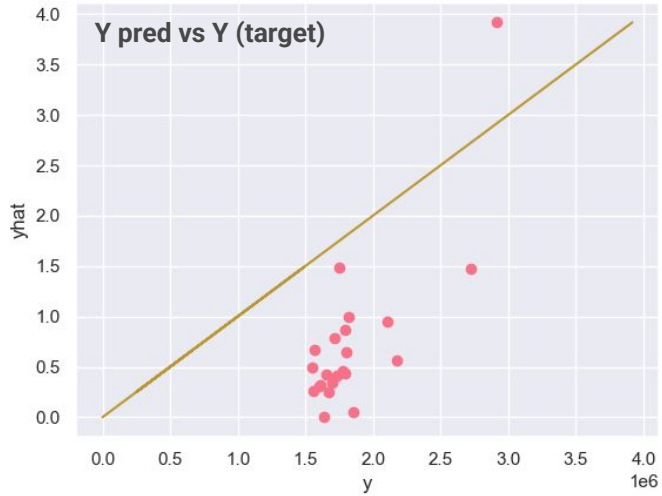
metrics for weekly data

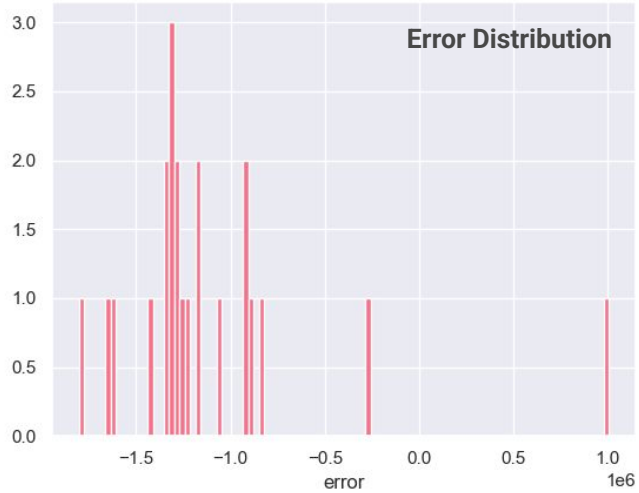
Métricas para el conjunto de train:

- MSE: 209312627848.60
- R^2: 0.33
- RMSE: 457506.97

Métricas para el conjunto de test:

- MSE: 1542096000508.01
- R^2: -1.46
- RMSE: 1241811.58





logistic regression

Target: futuros rangos de actividad

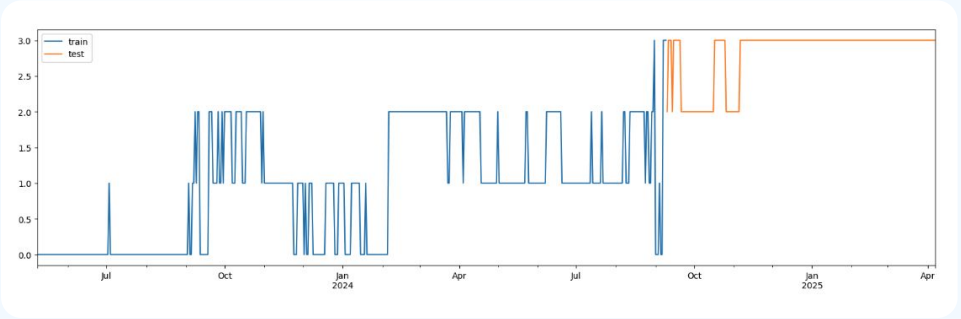
Features:

- Nuevos usuarios
- Total de usuarios
- PCA de actividad
- Series temporales
- Impact Score de noticias

Periodo: diario y semanal

Modelo: LogisticRegression()

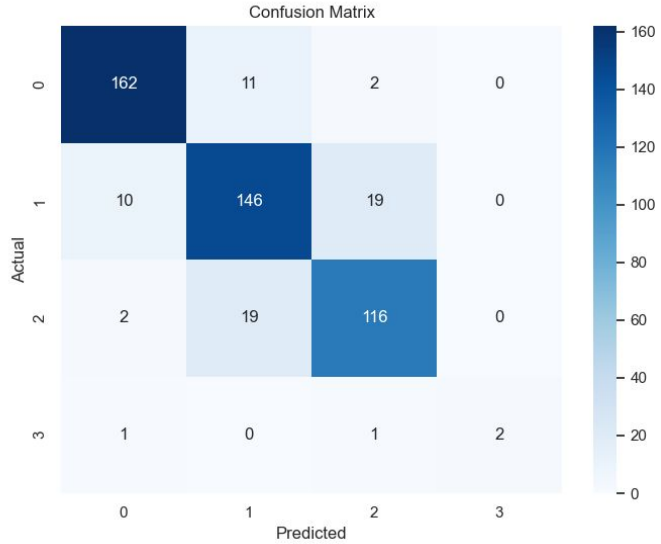
Train-Test Split: por periodos



metrics for daily data

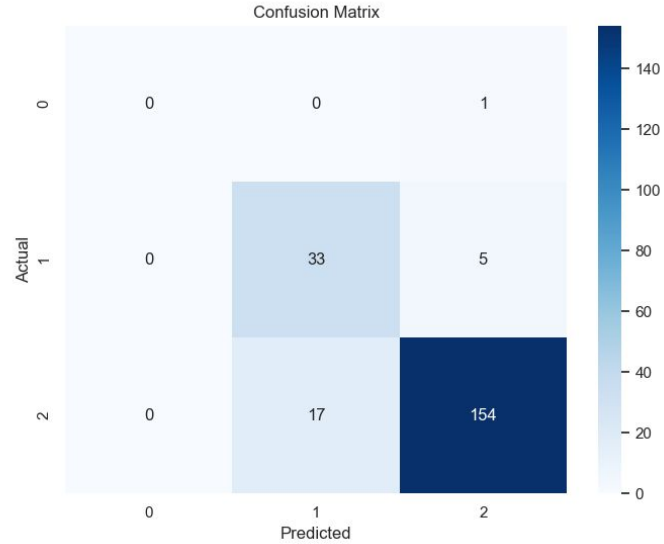
Métricas para el conjunto de train:

Report:	precision	recall	f1-score	
0.0	0.93	0.93	0.93	175
1.0	0.83	0.83	0.83	175
2.0	0.84	0.85	0.84	137
3.0	1.00	0.50	0.67	4
accuracy			0.87	491
macro avg	0.90	0.78	0.82	491
weighted avg	0.87	0.87	0.87	491



Métricas para el conjunto de test:

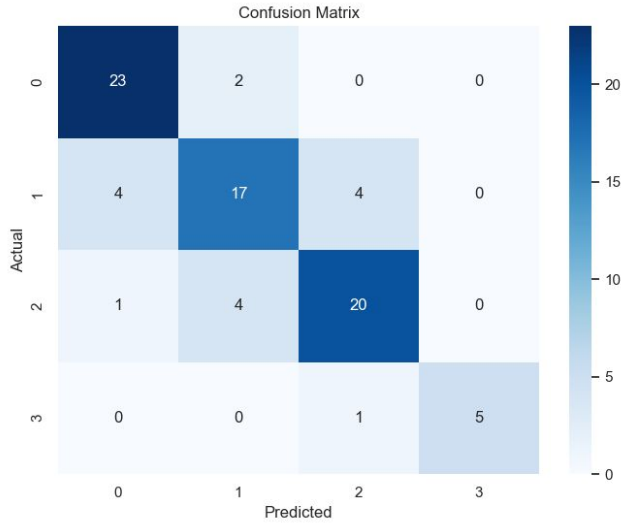
Report:	precision	recall	f1-score	
0.0	0.00	0.00	0.00	1
2.0	0.66	0.87	0.75	38
3.0	0.96	0.90	0.93	171
accuracy			0.89	210
macro avg	0.54	0.59	0.56	210
weighted avg	0.90	0.89	0.89	210



metrics for weekly data

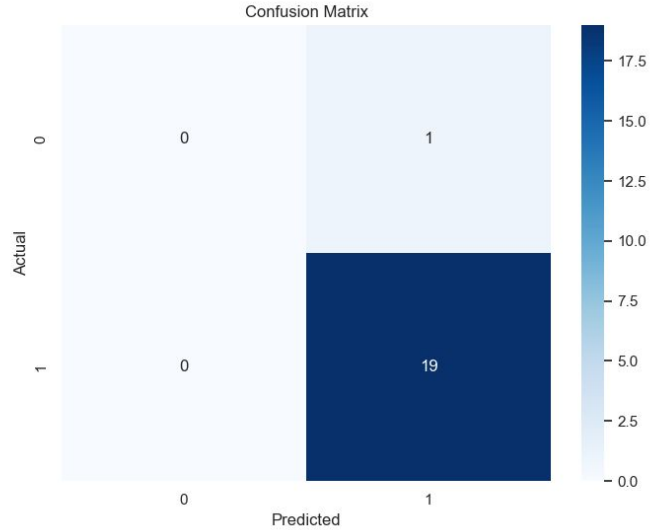
Métricas para el conjunto de train:

Report:	precision	recall	f1-score	
0.0	0.82	0.92	0.87	25
1.0	0.74	0.68	0.71	25
2.0	0.80	0.80	0.80	25
3.0	1.00	0.83	0.91	6
accuracy			0.80	81
macro avg	0.84	0.81	0.82	81
weighted avg	0.80	0.80	0.80	81



Métricas para el conjunto de test:

Report:	precision	recall	f1-score	
0.0	0.00	0.00	0.00	1
3.0	0.95	1.00	0.97	19
accuracy			0.95	20
macro avg	0.47	0.50	0.49	20
weighted avg	0.90	0.95	0.93	20



Forecasting Bluesky
Data Science Lifecycle - 06. Model Deployment

streamlit app

About this project

Forecasting Bluesky

Bluesky

About Me

Predict

Predict Tomorrow

Predict Next Week

Visitar app ⤴️

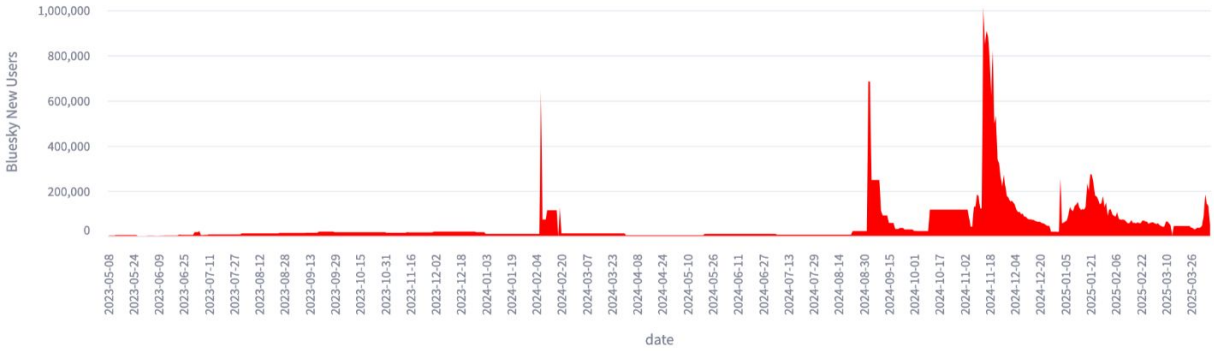
Fork

FORECASTING BLUESKY

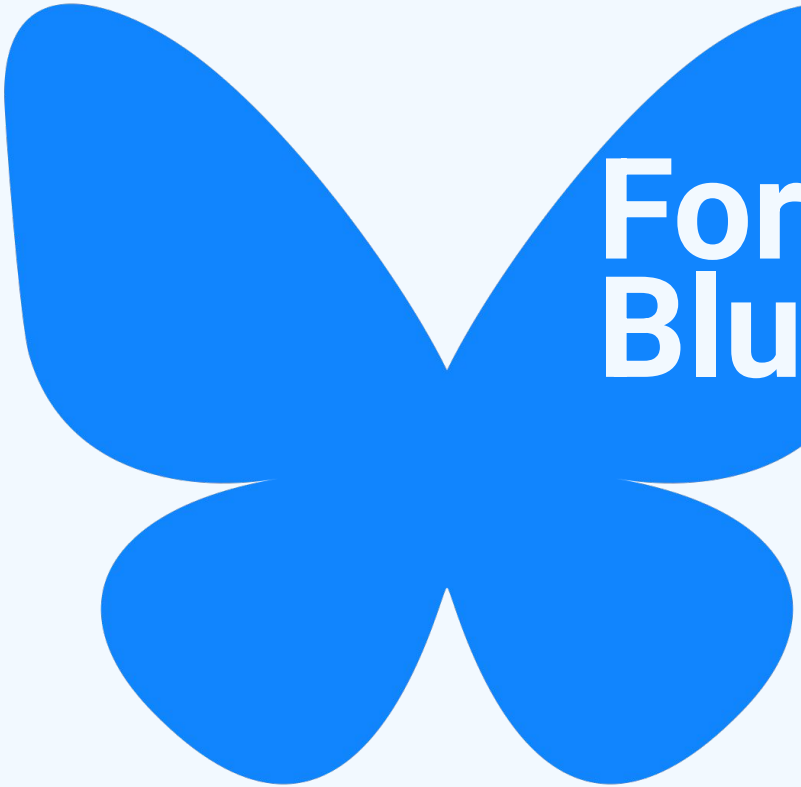
Predicción del crecimiento de usuarios a través del análisis de actividad y el impacto de noticias

Hipótesis: existe una correlación entre los cambios políticos y sociales y el aumento de usuarios en Bluesky.

New Users Activity



Detalle



Forecasting Bluesky...

...conclusiones y próximos pasos

conclusiones

- Los modelos tienen mucha **dificultad en la predicción de outliers** generados por factores exógenos.
- El **impact score** puede ser la clave para solucionar esto.
- Cambio en el enfoque: **predicción binaria**.

próximos pasos

- **Data Scraping**: automatización y mejora en el scraping de noticias.
- **Ampliar categorías de noticias**: incluir la posibilidad de que entren nuevas categorías inexistentes en el dataset original.
- **Nuevos componentes para el Impact Score**: Historic Virality (si un NER ha estado presente en un pico de usuarios), Word Importance…
- **Adaptación del Impact Score** hacia otras redes sociales.
- **Mejora de los modelos de regresión lineal**: feature engineering, búsqueda de hiperparámetros.

mejoras

- **Trabajo en equipo**: creo que es esencial para un feedback continuo, la detección de errores y para evitar bucles.
- **Gestión del tiempo**: dedicar menos tiempo al preprocesamiento de los datos me habría permitido mayor dedicación en los modelos.
- **Importancia de cada tarea**: una mejor decisión sobre la importancia real de cada tarea habría mejorado el desarrollo.

¡Gracias!

Si te interesa este proyecto, no dudes en contactarme

caminodepablos@gmail.com

github.com/caminodepablos

linkedin.com/in/caminodepablos