



NEONEXUS

ANÁLISIS DE SENTIMIENTO EN REDES SOCIALES CON SPARK NLP

Minería de opinión en tiempo real usando Big Data y Deep Learning sobre Twitter



ORIGEN E IMPORTANCIA DEL PROYECTO



Redes Sociales como
Fuente de Opinion Pública



El desafio empresarial:
comprender al cliente



Solución humana y tecnológica:
análisis de sentimiento

Convertir millones de opiniones digitales
en decisiones humanas y empáticas

PREGUNTAS CLAVE PARA DEFINIR LA ESTRATEGIA DEL PROYECTO



¿Qué red social
nos permitiría
acceder a opiniones
en tiempo real?



¿Qué empresa
elegiríamos como
modelo para la
prueba de concepto?



¿Que herramienta
usaríamos para
extraer esos datos
de forma eficiente?

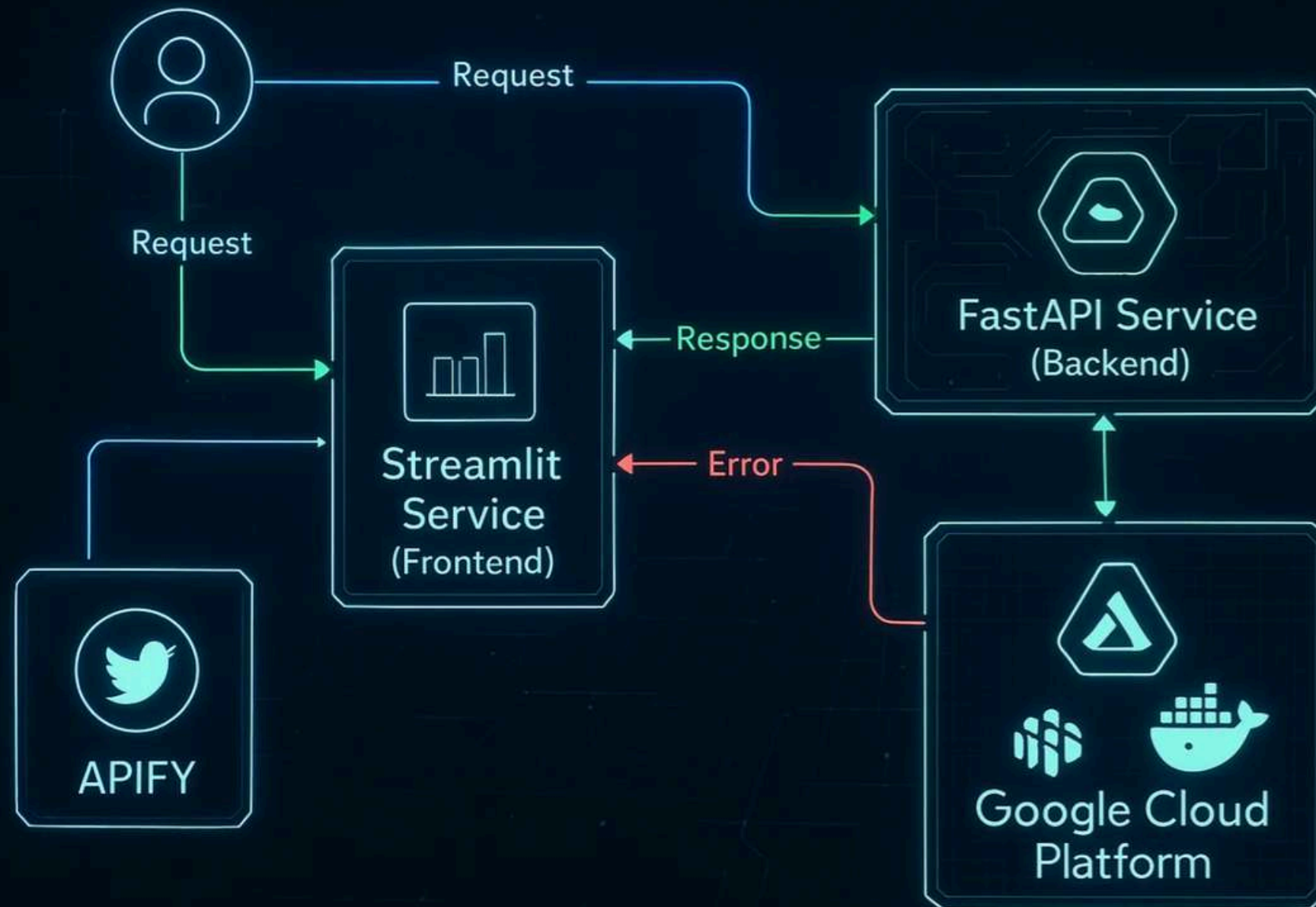
Empresa seleccionada para la prueba de concepto: NVIDIA



NLP



Twitter Sentiment Analysis



Analiza la percepción pública sobre tu marca en Twitter

Parámetros de búsqueda

Palabra clave o hashtag (ej: #Python, OpenAI)

Fecha de inicio

2021/02/24

Fecha de fin

2024/03/10

Opciones avanzadas de filtrado

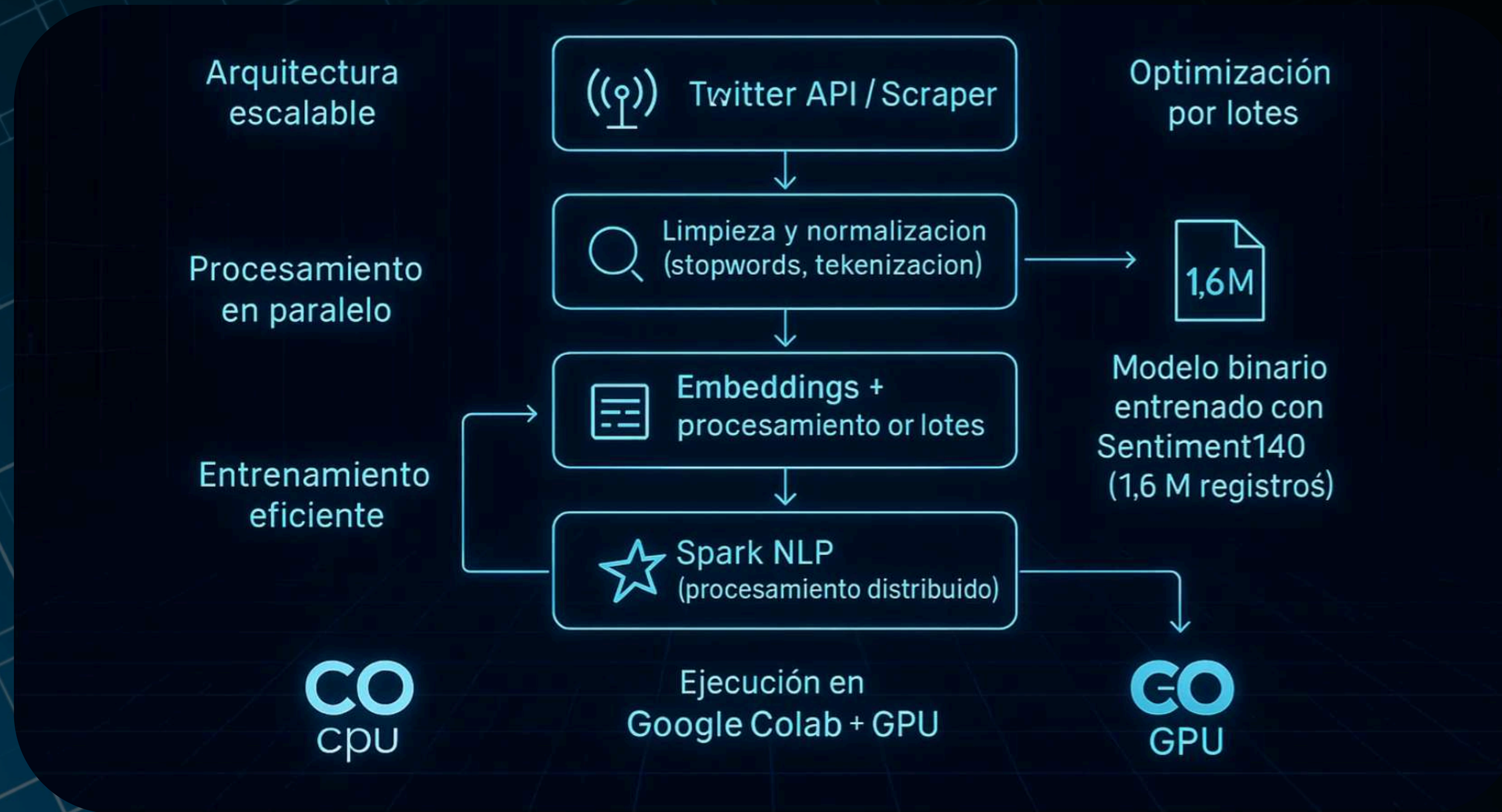
Analizar Tweets

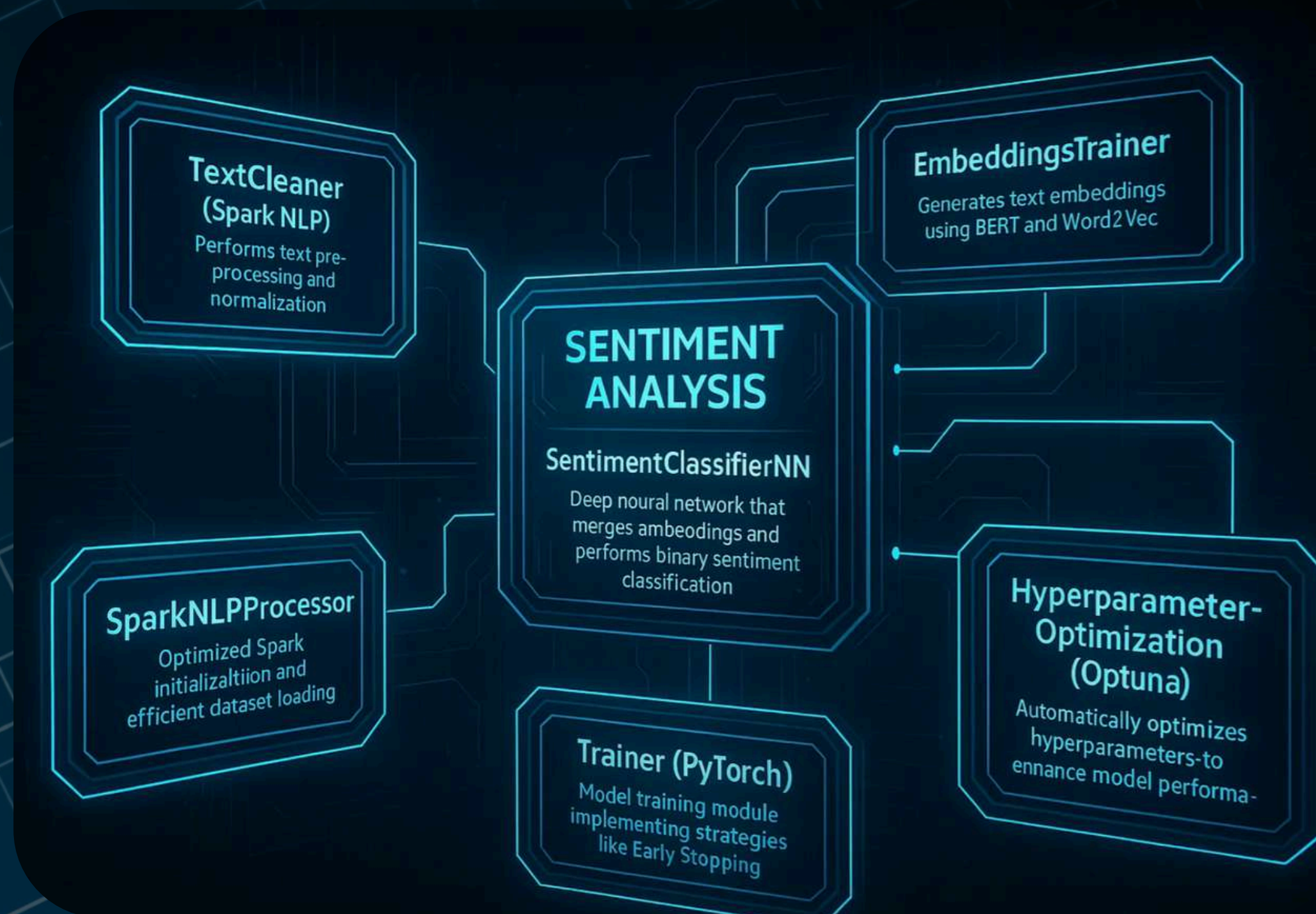
¿Cómo funciona este módulo?



DESCRIPCIÓN TÉCNICA DEL PROYECTO

NeoNexus





Fase 1: BERT + TF-IDF (con PCA)

- BERT: Representación semántica profunda.
- TF-IDF: Peso estadístico de términos relevantes.
- PCA: Reducción de TF-IDF de 5000 a 3000 dimensiones

Esta configuración fue efectiva en etapas tempranas de validación y sirvió como referencia para pruebas posteriores.

Fase 2: BERT + TF-IDF + Word2Vec

Se añadió WordVec (300 dimensiones, entrenado con Spark NLP) para enriquecer la representación. Sin embargo, el costo computacional superó los beneficios en rendimiento, lo que llevó a reconsiderar esta estrategia.

Fase 3: BERT + Word2Vec

Se eliminó TF-IDF, manteniendo BERT y Word2Vec. Esta configuración redujo la complejidad dimensional, manteniendo profundidad semántica y contexto distribuido. Junto a mejoras estructurales del modelo y un dataset balanceado de 400.000 registros, se lograron mejores resultados.

HYPERPARAMETER OPTIMIZATION

USING OPTUNA

LEARNING RATE 0.000124

WEIGHT DECAY 6.7e-06

DROPOUT 0.417

ACTIVATION FUNCTION Mish

OPTIMIZER AdamW

TRAINING LOSS

0.3818

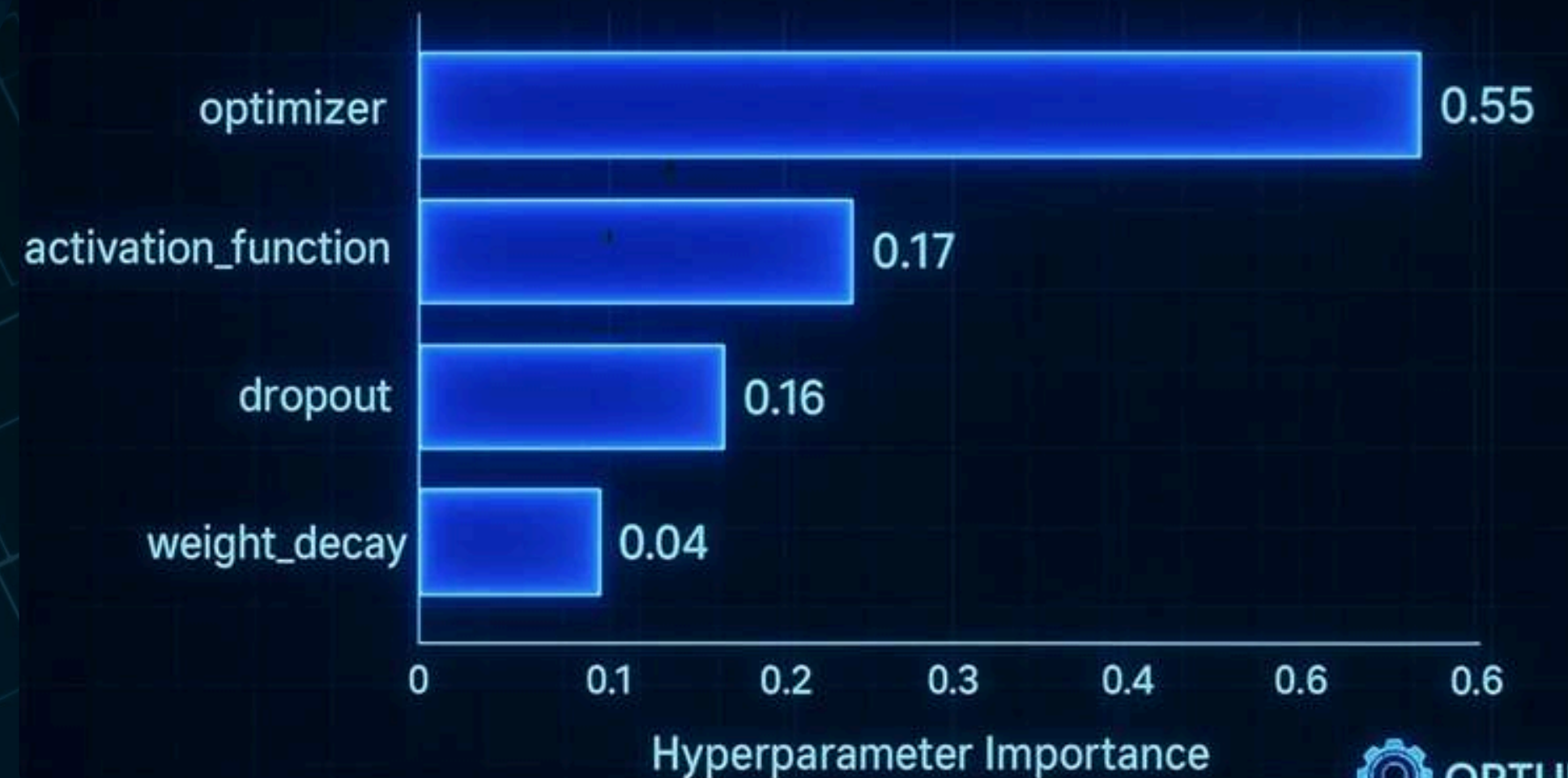
TRAINING AND VALIDATION LOSS



TRAINING AND VALIDATION ACCURACY



Importancia de Hiperparámetros



NeoNexus

MODEL NAME	F1-SCORE	ACCURACY	RECALL CLASS 0	RECALL CLASS 1	STRUCTURE
best_model_0.417099468279443.pth	0,8029	0,8029	0,7869	0,8195	Sentiment
best_model_0.417053810332437.pth	0,8029	0,8029	0,7873	0,8246	Sentiment
best_model_0.417973405104036.pth	0,8024	0,8028	0,7807	0,8245	Sentiment
best_model_0.419087691088377.pth	0,8024	0,8028	0,7802	0,8706	Sentiment
best_model_0.417270347679399.pth	0,8023	0,8028	0,7818	0,8244	Sentiment
best_model_0.416872384039581.pth	0,8023	0,8028	0,7804	0,8235	Sentiment
best_model_0.419353783763525.pth	0,8023	0,8028	0,7810	0,8710	Sentiment
best_model_0.417076307020645.pth	0,8022	0,8023	0,7902	0,8233	Sentiment

best_model_0.41709946827292443.pth

CONFUSION MATRIX

ACCURACY: 0.8029

F1-SCORE: 0.8029

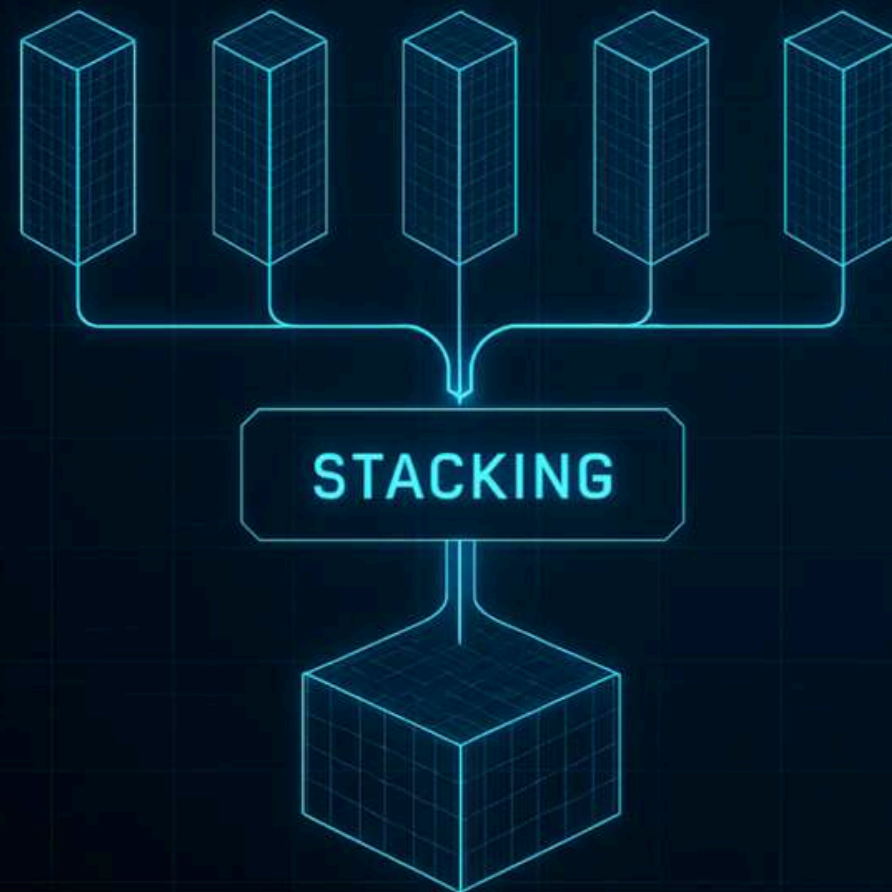
		PREDICTED	
		0	1
ACTUAL	0	157,371	42,629
	1	36,156	163,656



MODEL STACKING

FOR SENTIMENT CLASSIFICATION

INDIVIDUAL DEEP LEARNING MODELS



LOGISTIC REGRESSION (STACKING)

F1-score: 0,8025

Accuracy: 0,8025

Confusion matrix

Actual 0	121,173	28,827
Actual 1	30,403	119,456

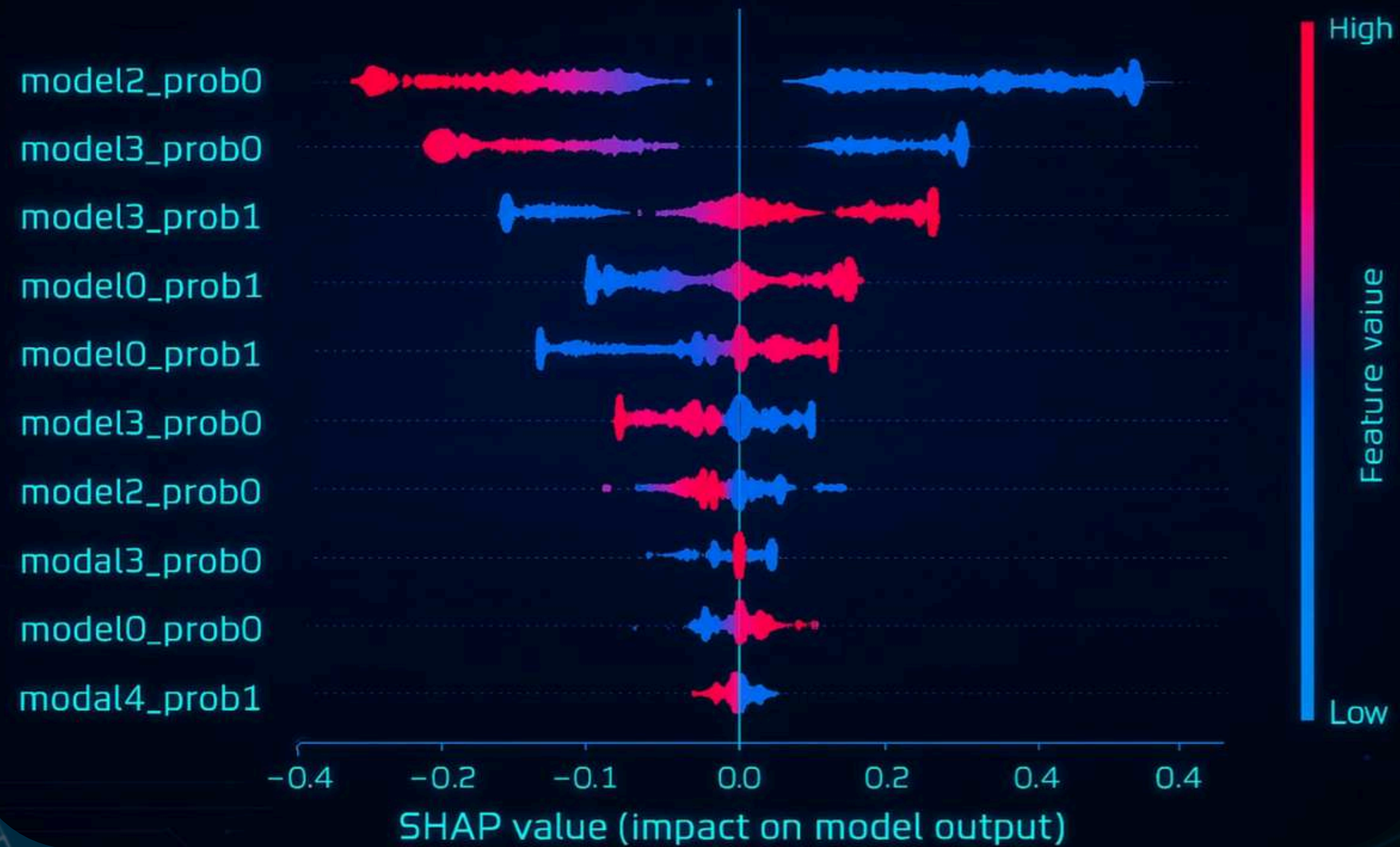
XGBOOST (STACKING)

F1-score: 0,8043

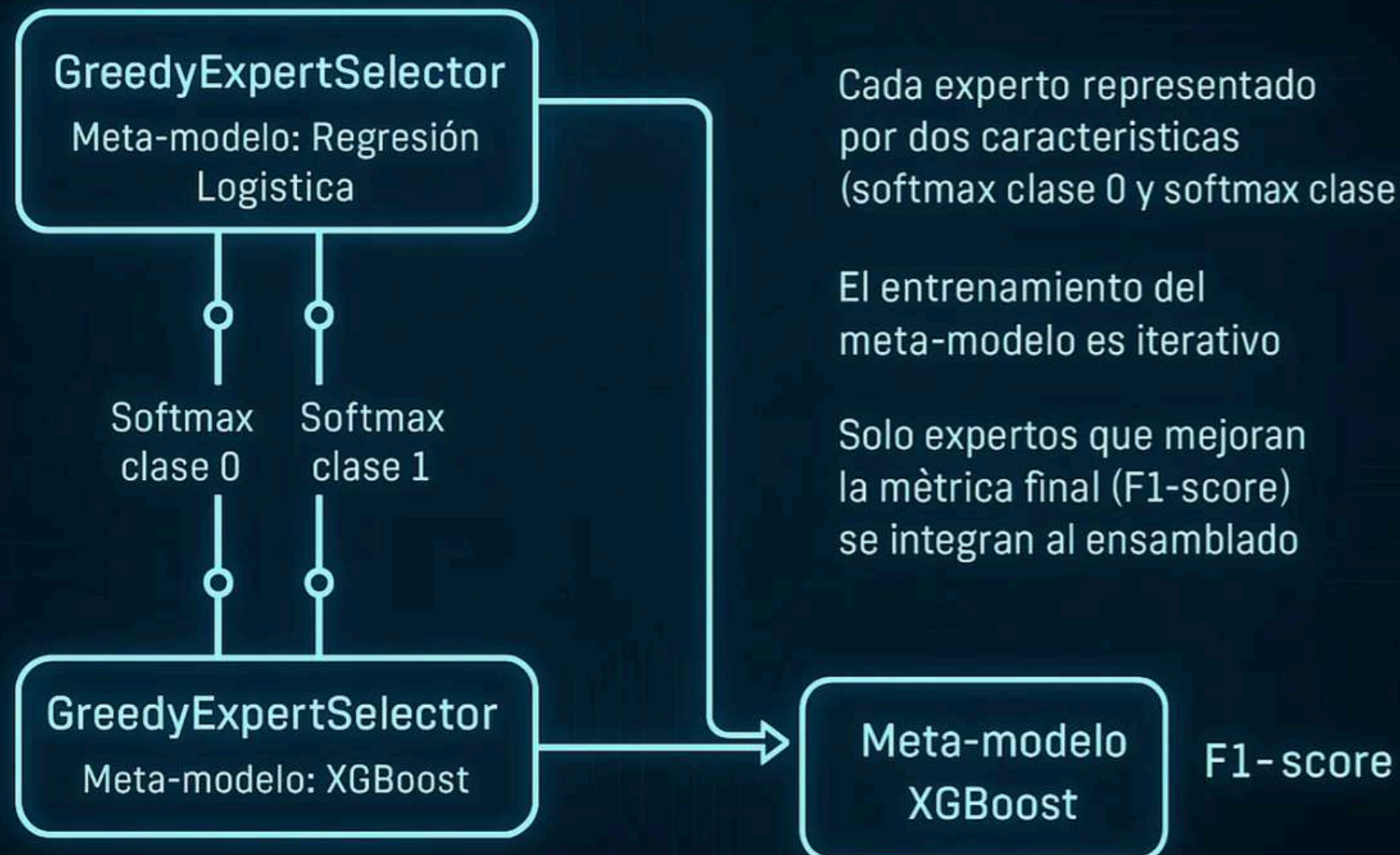
Accuracy: 0,8043

BEST PERFORMANCE DUE TO NON-LINEAR INTERACTION CAPTURE AND FLEXIBLE WEIGHTING

SHAP analysis for model interpretability (XGBoost stacking)



SELECCIÓN OPTIMIZADA DE EXPERTOS CON ESTRATEGIA GREEDY



RESULTADOS FINALES: COMPARATIVA DE META-MODELOS

REGRESIÓN LOGÍSTICA META-MODELO

F1-score: 0.8055
Accuracy 0.8055

MATRIZ DE CONFUSION

[162526	37474]
[40284	159528]

MODELO INTERPRETABLE
Y COMPUTACIONALMENTE EFICIENTE

XGBOOST - META-MODELO

F1-score: 0.8062
Accuracy 0.8063

MATRIZ DE CONFUSIÓN

[162333	37667]
[39796	160016]

MEJOR RENDIMIENTO GENERAL,
CAPTURA INTERACCIONES NÓ LINEALES

AMBOS MODELOS MUESTRAN RENDIMIENTO COMPETITIVO. XGBOOST
ES LIGERAMENTE SUPERIOR, PERO LA REGRESION LOGISTICA ES MAS
INTERPRETABLE Y ADECUADA EN ESCENARIOS CON RECURSOS LIMITADOS

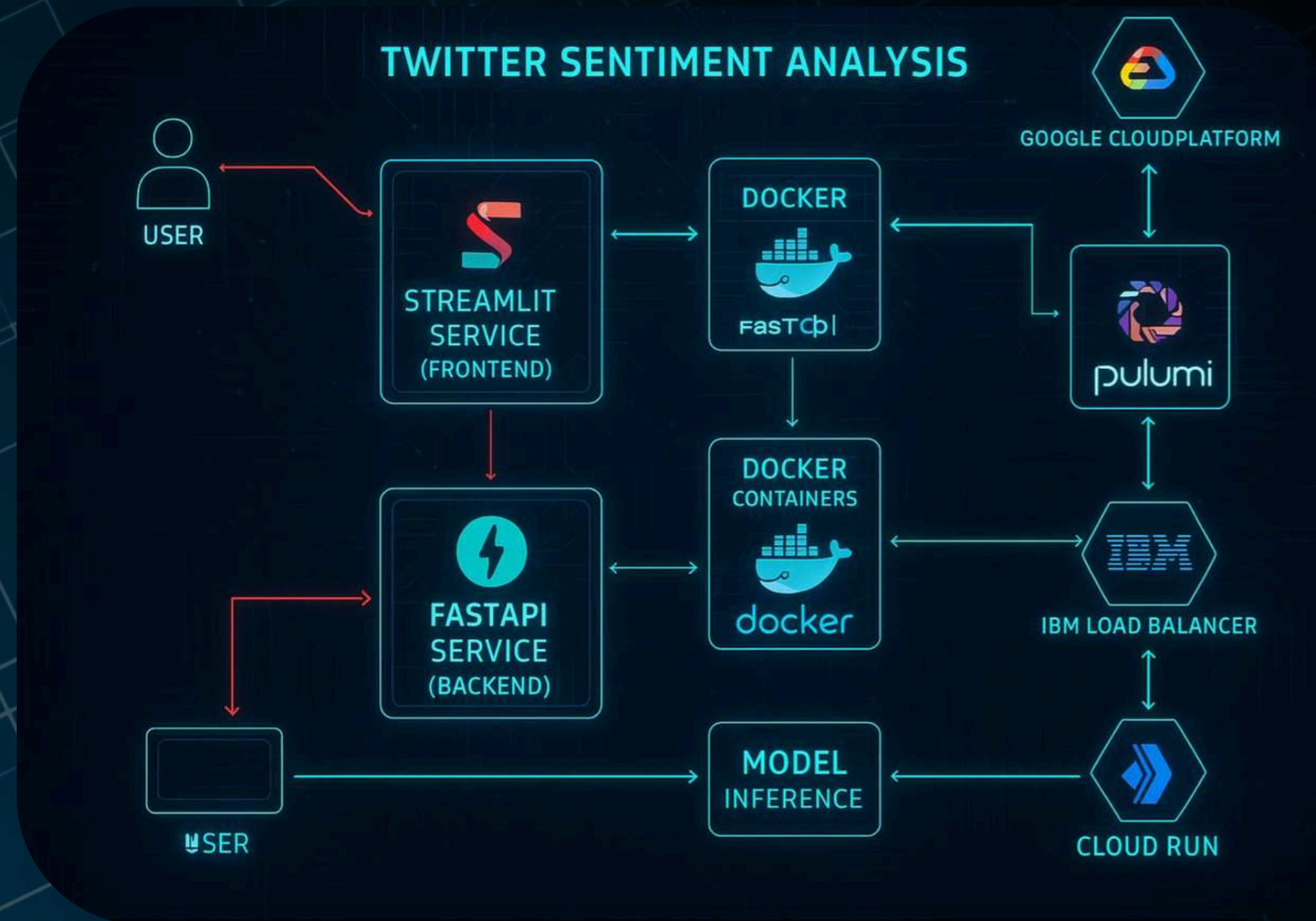
CONFUSION MATRIX

	POSITIVO	NEGATIVO
POSITIVO	386	25
NEGATIVO	69	69

F1-SCORE
0,89

ACCURACY
82.88%

DESPLIEGUE DE COMPONENTES Y MODELO



CONCLUSIONES Y FUTURAS MEJORAS

CONCLUSIONES



Big Data



Machine Learning



Cloud Computing

- Solución escalable, eficiente y en tiempo real
- XGBoost + Staking + Greedy garantizaron calidad y relevancia del análisis

FUTURAS MEJORAS



Incluir análisis de sentimiento neutral



Explorar nuevos modelos expertos



Optimizar hiperparámetros de XGBoost



Probar reducción de dimensionalidad (t-SNE + TF-IDF)



Añadir métricas avanzadas (ROC-AUC, Precision-Recall)



Análisis de palabras frecuentes en clases positivas/negativas



NeoNexus

Este proyecto no solo es un logro técnico, sino también una demostración clara del poder transformador de la inteligencia Artificial y el Machine Learning en ámbitos empresariales y sociales, acercando cada vez más la tecnología al servicio humano y empático con las personas.

**GRACIAS POR VUESTRA
ATENCIÓN**

