

# Informe Ejecutivo: Clasificación Automática de Opiniones Hoteleras

## Objetivo del proyecto:

Desarrollar un modelo automático de clasificación que determine si una crítica hotelera es favorable o desfavorable, utilizando reseñas de Booking.com como fuente de datos. Dado el desbalance en las clases (mayoría de críticas positivas), se priorizó el uso de métricas como ROC AUC en lugar de accuracy.

# Metodología Aplicada:



## 1. Exploración de Datos:

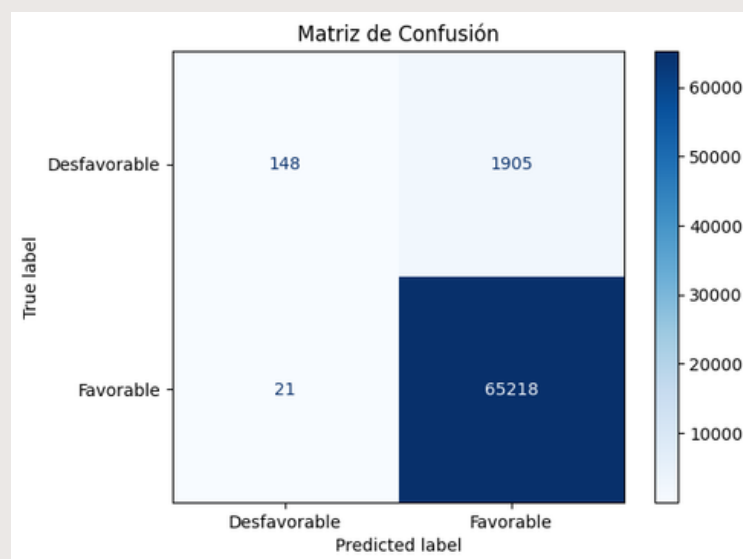
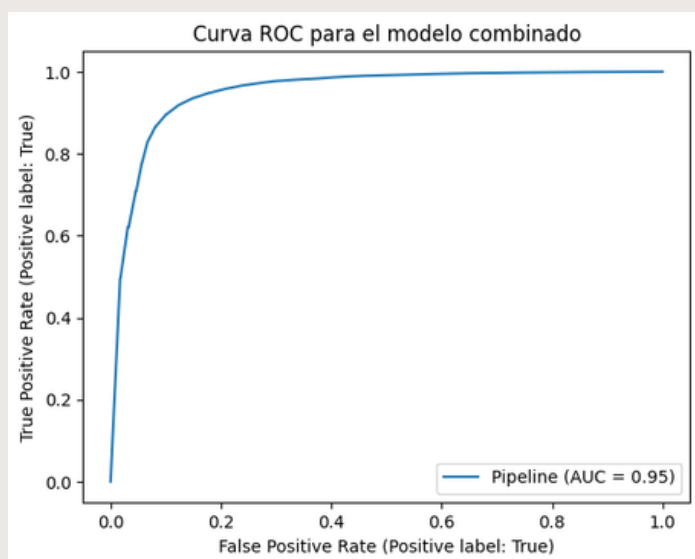
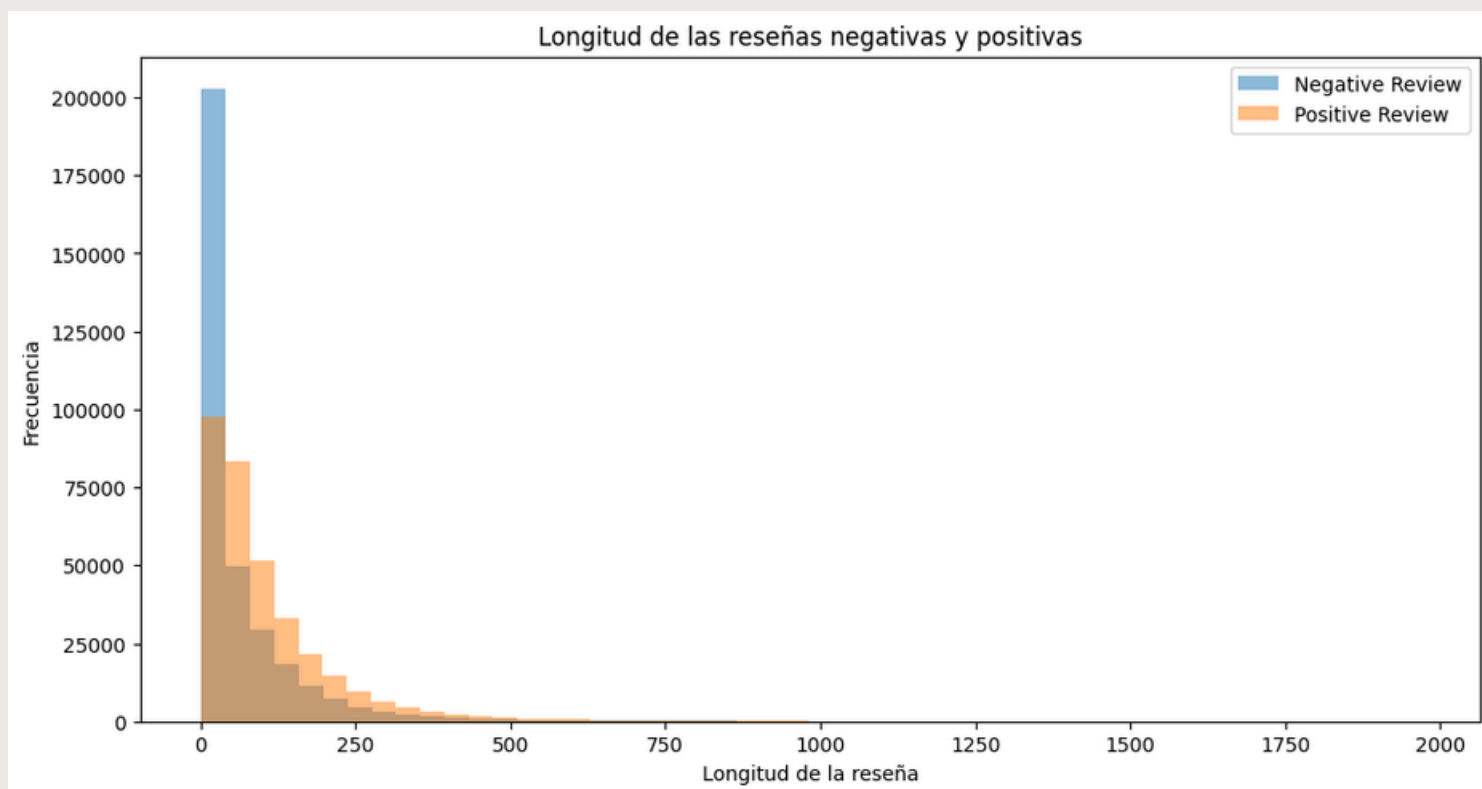
- Se analizaron más de 330.000 registros.
- Se verificó la ausencia de valores nulos y se observaron claras diferencias en la longitud de las críticas negativas y positivas.
- Se evidenció un gran desbalance: ~97% críticas positivas.

## 2. Modelos Implementados:

- Modelo basado en caracteres: Usando n-gramas de caracteres y regresión logística. ROC AUC: 0.9517.
- Modelo combinado por campo: Separación de textos positivos y negativos con pipelines especializados. ROC AUC: 0.9657.
- Modelo basado en tokens (palabras): Usando CountVectorizer y regresión logística. Accuracy: 98.30%.
- Modelo con análisis morfosintáctico (spaCy): Preprocesamiento con lematización + CountVectorizer. Accuracy: 98.39%.
- Modelo con Deep Learning (LSTM): Embeddings + LSTM + capas densas. Accuracy: 98.30%.
- Modelo combinado (TF-IDF + Random Forest): Mezcla de n-gramas de caracteres y palabras. ROC AUC: 0.9521.

# Resultados Clave:

Modelo	Métrica Principal	Resultado
Regresión Logística (caracteres)	ROC AUC	0.9517
Pipeline por campo (char-based)	ROC AUC	0.9657 
Tokenización (palabras)	Accuracy	0.9830
Análisis morfosintáctico	Accuracy	0.9839 
Deep Learning (LSTM)	Accuracy	0.9830
Modelo Combinado (TF-IDF + RF)	ROC AUC	0.9521



## Conclusiones

- El modelo basado en análisis morfosintáctico fue el más preciso.
- Se destacó la importancia de normalizar textos y utilizar técnicas como lematización o embeddings.
- El desbalance de clases fue mitigado con métricas apropiadas y buenos preprocesamientos.
- Se evidenció que técnicas más sofisticadas como redes LSTM aportan mejoras, aunque marginales respecto a métodos clásicos bien optimizados.