Comparativa EDA: Full vs Sample 30%

Este informe resume los resultados de la comparativa entre el dataset completo de reseñas y una muestra estratificada del 30%. El objetivo es validar si la muestra mantiene la representatividad de las principales distribuciones (sentimiento, categorías, longitudes y tokens).

# 1. Distribución de Sentimiento

- Resultado chi²: No hay diferencias significativas (p=0.9169).  
- La muestra mantiene las proporciones de sentimiento (negativo, neutro, positivo).  
- Conclusión: El muestreo estratificado funcionó.

# 2. Distribución de Categorías

- Se compararon las proporciones por secondary\_category.  
- Diferencias observadas: en general pequeñas (cercanas a 0).  
- Conclusión: La muestra es representativa también en categorías.

# 3. Longitud de Textos

- Diferencias en caracteres:  
 • Media: +0.45  
 • Mediana: 0  
 • P95: +2  
 • P99: -1  
- Conclusión: No hay sesgo en longitudes, distribuciones equivalentes.

# 4. Distribución de Tokens

- χ² detecta diferencias significativas (p≈0), debido a outliers/extremos.  
- Diferencias en la práctica:  
 • Media: +0.09  
 • Mediana: 0  
 • P95: 0  
 • P99: +1  
 • Máx.: -500 (outlier ausente en el sample)  
- Conclusión: La muestra es representativa hasta p99. Las diferencias se concentran en los outliers.

# Conclusión Global

La muestra estratificada del 30% mantiene bien la representatividad del dataset completo en sentimiento, categorías y longitudes. En tokens, las diferencias significativas provienen de casos extremos (outliers), pero en el rango principal (hasta el percentil 99) las distribuciones son prácticamente idénticas. Por tanto, la muestra es adecuada para entrenar y evaluar modelos de NLP en el marco del TFM.