**A) Análisis de sentimiento — columnas sí o sí**

* reviewId
* placeId
* title
* text
* publishedAtDate
* stars

**B) Análisis de sentimiento — columnas opcionales**

* textTranslated
* originalLanguage
* reviewOrigin
* (si existen) todas las reviewContext/...
* (si existen) todas las reviewDetailedRating/...

Nota: para el modelo de sentimiento la **entrada** es solo el texto (text o textTranslated si quieres unificar idioma). Las demás te ayudan a evaluar/segmentar resultados, no son insumo del modelo.

**C) EDA — columnas recomendadas**

* **Hotel / métricas globales:** totalScore, reviewsCount, hotelStars, price
* **Autor / señal de calidad:** isLocalGuide, reviewerNumberOfReviews, likesCount
* **Respuesta del hotel:** responseFromOwnerText, responseFromOwnerDate
* **Aspectos y sub-ratings:** todas las reviewContext/... y todas las reviewDetailedRating/...
* **Auditoría (si quieres lineage):** scrapedAt
* **(Opcional de control de calidad):** categoryName (si quieres verificar que todo es “Hotel/Lodging”)

**D) Lista completa recomendada (unión de A + B + C)**

* reviewId, placeId, title, text, publishedAtDate, stars
* textTranslated, originalLanguage, reviewOrigin
* totalScore, reviewsCount, hotelStars, price
* isLocalGuide, reviewerNumberOfReviews, likesCount
* responseFromOwnerText, responseFromOwnerDate
* todas las reviewContext/...
* todas las reviewDetailedRating/...
* scrapedAt
* *(opcional de control)* categoryName

**¿Por qué excluí las demás columnas?**

* **Ubicación fija o poco informativa:** address, street, neighborhood, city, state, postalCode, countryCode, location/lat, location/lng
  + En tu muestra todos los hoteles son de Buenaventura/Colombia; no aportan variación útil para sentimiento ni EDA temático.
* **Identificadores redundantes/volátiles:** fid, cid
  + placeId ya sirve como identificador estable; fid es menos estable y cid es prescindible si ya tienes placeId.
* **Campos de estado siempre falsos:** temporarilyClosed, permanentlyClosed
  + En tu dataset están en false para todos; no aportan información.
* **Campos multimedia o de navegación:** imageUrl, reviewImageUrls/..., reviewerPhotoUrl, reviewerUrl, url, reviewUrl
  + No aportan al análisis textual/estadístico y añaden peso/ruido y posibles temas de privacidad.
* **Duplicados o poco consistentes:** rating
  + Suele duplicar stars o venir vacío; mantén stars como fuente única.
* **Búsqueda/sistema:** searchString
  + Es la cadena/URL utilizada por el scraper; no agrega valor analítico.
* **Campos vacíos en tu muestra:** visitedIn
  + Si está totalmente vacío, no aporta. (Si más adelante aparece, re-evalúa.)
* **Categorías generales:** categories/... y, si no lo usas, categoryName
  + En tu caso casi todo es “Hotel/Lodging”; baja varianza. Puedes conservar solo categoryName como control de calidad (opcional).
* **Datos de identidad del revisor:** name, reviewerId
  + No son necesarios para el análisis; mejor evitar exponer identificadores personales. Si quieres una señal de “experiencia del revisor”, ya tienes reviewerNumberOfReviews.

Con esto tienes un set compacto para **sentimiento** y un set más rico para **EDA** sin cargar columnas que no aportan valor. Si luego decides cruzar con otras fuentes (TripAdvisor, etc.), reviewOrigin y categoryName te ayudarán a controlar la mezcla.