

Proyecto Final Deep Learning - Taller EDA – Fase 1

Anderson J. Alvarado*
David E. Moreno **

Pontificia Universidad Javeriana

Resumen Este documento presenta un análisis exploratorio del dataset de reseñas de hoteles andaluces, orientado al desarrollo de un sistema de clasificación de sentimientos mediante Redes Neuronales Recurrentes (RNN). El dataset contiene 18,172 reseñas en español con sus respectivas etiquetas de sentimiento, constituyendo un caso de clasificación multiclasa desbalanceado. Se analiza la dimensionalidad, distribución de clases, características lingüísticas y patrones textuales presentes en los datos, estableciendo las bases para el posterior procesamiento y modelado mediante arquitecturas RNN.

Keywords: Análisis de Sentimientos · Redes Neuronales Recurrentes · Procesamiento de Lenguaje Natural · Clasificación Multiclasa · Reseñas de Hoteles

1. Definición del Problema y Objetivos

1.1. Problema RNN a Analizar

El análisis de sentimientos en reseñas de hoteles representa un desafío en el procesamiento de lenguaje natural (PLN), particularmente cuando se trabaja con textos en español que presentan características lingüísticas específicas y variaciones regionales. Las Redes Neuronales Recurrentes (RNN) han demostrado ser especialmente efectivas para este tipo de tareas debido a su capacidad para capturar dependencias temporales y contextuales en secuencias de texto.

El problema específico abordado en este estudio consiste en desarrollar un modelo de clasificación multiclasa capaz de categorizar automáticamente las opiniones de los usuarios sobre hoteles andaluces. Este problema presenta varios desafíos técnicos:

- **Naturaleza secuencial del texto:** Las reseñas varían significativamente en longitud (7 a 1,416 palabras), requiriendo modelos capaces de procesar secuencias de longitud variable.

* Pontificia Universidad Javeriana, andersonjalvarado@javeriana.edu.co

** Pontificia Universidad Javeriana, morenoa-david@javeriana.edu.co

- **Complejidad lingüística:** El español andaluz puede contener modismos, expresiones regionales y estructuras gramaticales particulares que añaden complejidad al análisis.
- **Desbalance de clases:** Con un ratio de desbalance de 5.82:1, el dataset presenta un sesgo significativo hacia la clase mayoritaria.
- **Ambigüedad semántica:** Las opiniones pueden contener ironía, sarcasmo o sentimientos mixtos que dificultan la clasificación precisa.

1.2. Conjunto de Datos Elegido

El dataset seleccionado, *Andalusian Hotels Reviews Unbalanced* [2], comprende una colección de 18,172 reseñas de hoteles ubicados en Andalucía, España. Este conjunto de datos fue elegido por las siguientes razones:

Relevancia del dominio El sector turístico andaluz representa uno de los motores económicos más importantes de la región, generando miles de millones de euros anuales. La capacidad de analizar automáticamente las opiniones de los clientes permite:

- Identificar áreas de mejora en los servicios hoteleros
- Detectar tendencias y patrones en la satisfacción del cliente
- Facilitar la toma de decisiones basada en datos

Características del dataset El conjunto de datos presenta características que lo hacen ideal para el entrenamiento de modelos RNN:

Cuadro 1: Características principales del dataset

Característica	Valor
Total de reseñas	18,172
Número de características	7
Tamaño del vocabulario	62,639 palabras únicas
Longitud media de reseñas	78.8 palabras
Longitud mediana	61.0 palabras
Diversidad léxica	0.0437
Memoria utilizada	15.87 MB

Distribución geográfica Las reseñas provienen de 715 hoteles diferentes distribuidos en 33 ubicaciones dentro de Andalucía, incluyendo las principales provincias como Granada, Sevilla y Málaga (Costa del Sol). Esta diversidad geográfica asegura una representación amplia de las experiencias turísticas en la región.

2. Análisis de Dataset y Dimensionalidad

2.1. Análisis de Dimensionalidad

El análisis dimensional del dataset revela una estructura compleja que combina información textual, categórica y numérica. Con un total de 127,204 celdas de datos distribuidas en 18,172 muestras y 7 características originales, el dataset presenta una densidad de información del 97.15 %, con solo 3,632 valores nulos concentrados principalmente en las columnas de ubicación y hotel (9.99 % cada una).

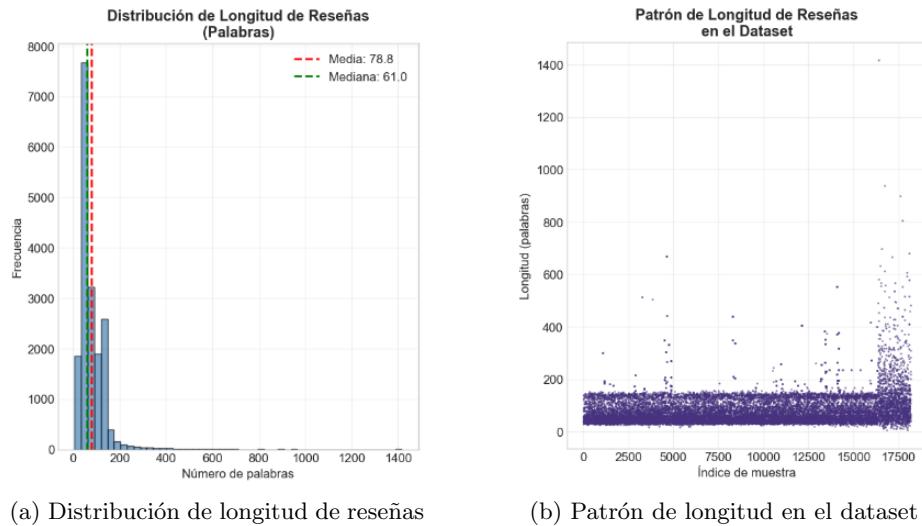
Distribución de tipos de datos La estructura del dataset se compone de:

- **4 columnas de tipo objeto** (57.1 %): Incluyen el texto de la reseña, título, ubicación y hotel
- **3 columnas de tipo entero** (42.9 %): Rating numérico, etiqueta de sentimiento e índice

Durante el proceso de análisis exploratorio, se generaron 3 características adicionales derivadas del texto:

- **longitud_chars**: Número de caracteres por reseña
- **longitud_palabras**: Número de palabras por reseña
- **longitud_oraciones**: Número de oraciones por reseña

Análisis de longitud textual La distribución de longitudes presenta características importantes para el diseño del modelo RNN:



(a) Distribución de longitud de reseñas

(b) Patrón de longitud en el dataset

Figura 1: Análisis de longitud de las reseñas

Como se observa en la Figura 1, la distribución de longitudes muestra:

- Una distribución asimétrica positiva ($\text{skewness} = 4.19$) con cola larga hacia la derecha
- Alta curtosis (42.87) indicando concentración alrededor de la mediana
- El 50 % de las reseñas contiene entre 42 y 103 palabras (rango intercuartílico)
- Presencia de outliers significativos con reseñas de hasta 1,416 palabras

Diversidad del vocabulario El análisis léxico revela:

- **Vocabulario total:** 62,639 palabras únicas
- **Corpus completo:** 1,432,453 palabras
- **Ratio de diversidad léxica:** 0.0437

Este ratio relativamente bajo sugiere un uso repetitivo de vocabulario común en el dominio hotelero, lo cual puede facilitar el aprendizaje de representaciones por parte del modelo RNN.

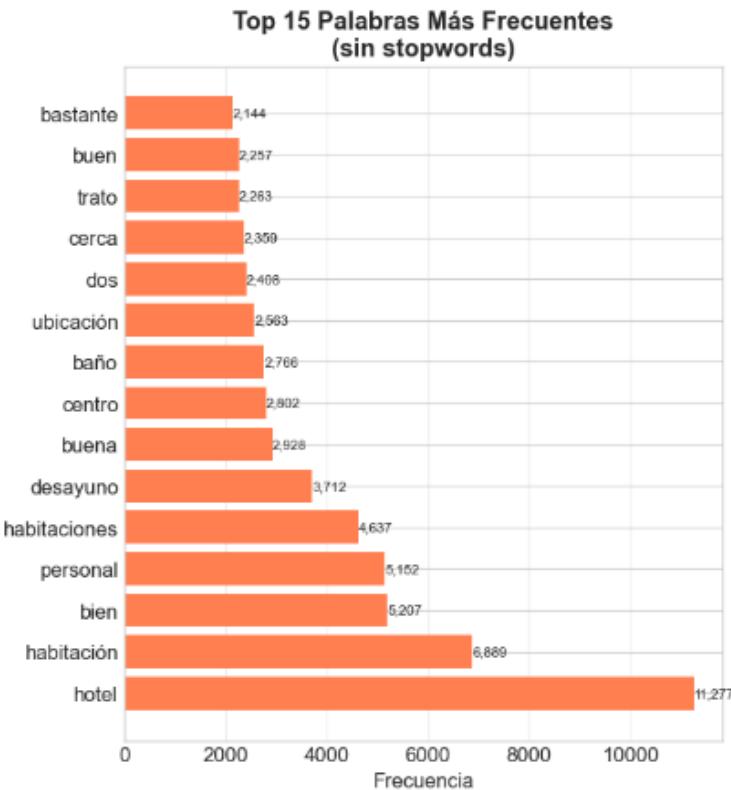


Figura 2: Top 15 palabras más frecuentes (sin stopwords)

2.2. Análisis de Variable Objetivo

La variable objetivo identificada es `label`, que representa el sentimiento asociado a cada reseña mediante una codificación numérica de tres clases.

Distribución de clases El dataset presenta un desbalance significativo en la distribución de clases:

Cuadro 2: Distribución de la variable objetivo

Clase	Frecuencia	Porcentaje	Descripción inferida
1	13,227	72.79 %	Sentimiento positivo
0	2,671	14.70 %	Sentimiento negativo
3	2,274	12.51 %	Sentimiento neutral/mixto

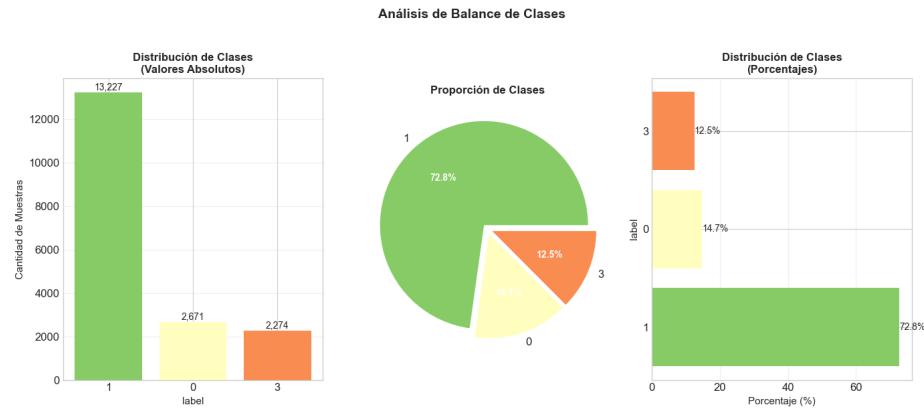


Figura 3: Análisis visual del balance de clases

Implicaciones del desbalance El ratio de desbalance de 5.82:1 entre la clase mayoritaria (1) y la minoritaria (3) presenta varios desafíos:

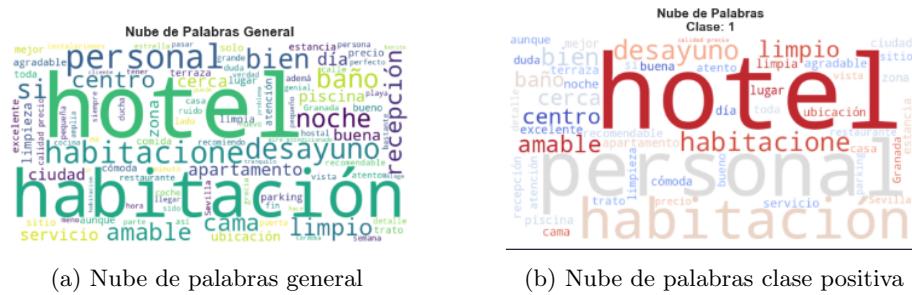
1. **Sesgo hacia la clase mayoritaria:** Los modelos RNN pueden tender a predecir predominantemente la clase 1
2. **Dificultad en el aprendizaje de clases minoritarias:** Las clases 0 y 3 pueden ser sub-representadas en el proceso de aprendizaje
3. **Métricas de evaluación sesgadas:** La precisión global puede ser engañosamente alta

Este desbalance requerirá estrategias específicas como:

- Técnicas de sobremuestreo (SMOTE, ADASYN)
 - Submuestreo de la clase mayoritaria
 - Ajuste de pesos en la función de pérdida
 - Uso de métricas balanceadas (F1-score macro, Cohen's Kappa)

2.3. Análisis de Características

Características textuales El campo `review_text` constituye la característica principal para el modelo RNN. El análisis revela patrones lingüísticos distintivos:



.) Nube de palabras general

(b) Nube de palabras clase positiva



(c) Nube de palabras clase negativa

Figura 4: Análisis visual del vocabulario por clase de sentimiento

Las nubes de palabras revelan términos clave asociados con experiencias hoteleras:

- **Términos positivos** **frecuentes**: “limpio”, “amable”, “bien”
 - **Aspectos evaluados**: servicio, habitaciones, cama, desayuno

Características categóricas Ubicación geográfica:

- 33 ubicaciones únicas
 - Concentración en provincias principales (Granada, Sevilla, Málaga)
 - 9.99 % de valores faltantes

Hoteles:

- 715 establecimientos únicos
- Distribución heterogénea con algunos hoteles sobre-representados
- Potencial para análisis de sesgo por establecimiento

Rating:

- Escala de 1 a 5
- Moda: 5
- Frecuencia de la moda: 9,005 (49.55 %)

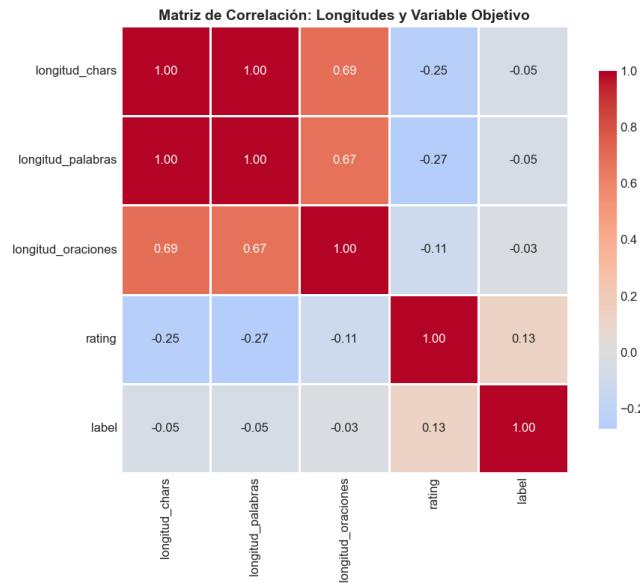


Figura 5: Matriz de correlación entre variables de longitud y objetivo

Análisis de correlaciones La matriz de correlación (Figura 5) muestra:

- Alta correlación entre medidas de longitud (>0.95)
- Correlación débil entre longitud y label
- Independencia relativa del rating respecto a la longitud textual

Análisis de frecuencias categóricas El análisis de frecuencias de las variables categóricas permite identificar patrones de concentración y diversidad en los datos. A continuación, se resumen los principales hallazgos:

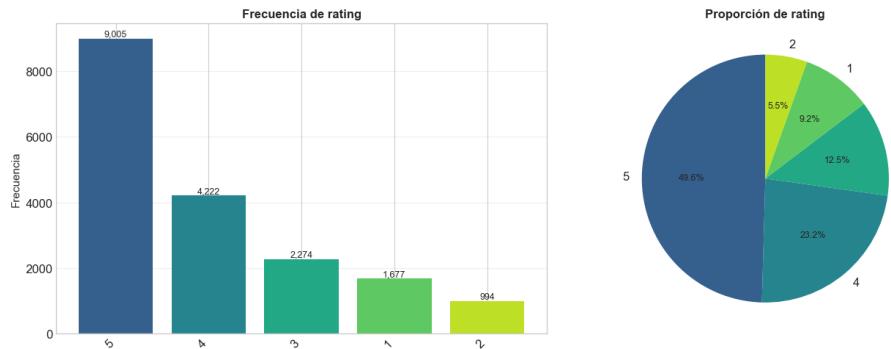


Figura 6: Análisis de frecuencia de variables categóricas

- **Variable title:** Se identificaron **10,262 categorías únicas**, lo que sugiere una alta variedad en los títulos de las reseñas. La moda corresponde a “Excelente” con una frecuencia de **245 (1.35 %)**. Las categorías más frecuentes incluyen: “Perfecto”, “Muy recomendable”, “Buen hotel”, “Recomendable” y “Muy bien”. No se registraron valores nulos.
- **Variable location:** Presenta **33 categorías únicas**, siendo la moda **Granada_Province_of_Granada_Aandalucia** con **3,426** ocurrencias (**20.95 %**). Otras ubicaciones destacadas incluyen: **Seville_Province_of_Seville_Aandalucia**, **Malaga_Costa_del_Sol**, y **Cordoba_Province_of_Cordoba_Aandalucia**. Se detectaron **1,816 valores nulos**. Esto indica una fuerte concentración geográfica de reseñas en pocas provincias andaluzas.
- **Variable hotel:** Se encontraron **715 categorías únicas**, con una moda en **Carmen_de_Ramilla** (**120** reseñas, **0.73 %**). Los siguientes hoteles más frecuentes incluyen: **Apartamentos_Mauror**, **Hotel_Cortijo_Del_Marques**, y **Casa_Olea**. También se identificaron **1,816 valores nulos**. Esto sugiere una amplia dispersión de hoteles, con algunos más representados por mayor actividad de reseñas.
- **Variable rating:** Consta de **5 categorías** (valores de 1 a 5). La categoría más frecuente es la **5 estrellas** con **9,005** reseñas (**49.55 %**), seguida de **4 estrellas (23.23 %)** y **3 estrellas (12.51 %)** (Figura 6). No hay valores nulos. Este patrón sugiere una tendencia general hacia evaluaciones positivas.

3. Plan de Procesamiento

3.1. Limpieza y Preparación de los Datos

El procesamiento a implementar seguirá un pipeline secuencial de cuatro etapas que transforma progresivamente el texto bruto en representaciones normalizadas optimizadas:

Cuadro 3: Pipeline de Procesamiento de Texto

Fase	Operaciones
Normalización	Eliminación de caracteres especiales, reducción de repeticiones, unificación de guiones, conversión a minúsculas y eliminación de espacios redundantes.
Tokenización	Segmentación del texto en unidades léxicas mínimas preservando la integridad semántica.
Eliminación de Stopwords	Filtrado de palabras comunes sin carga semántica discriminativa utilizando lexicones específicos para español.
Lematización	Reducción de palabras a su forma canónica para eliminar variaciones morfológicas y reducir la dimensionalidad.

4. Metodología

4.1. Modelos

4.1.1. Red Neuronal Recurrente LSTM

Para la implementación de la red recurrente se implementó el tipo LSTM de 128 neuronas de acuerdo al trabajo de Ríos [5]. A continuación se presenta la arquitectura de la red neuronal implementada.

Arquitectura e Implementación

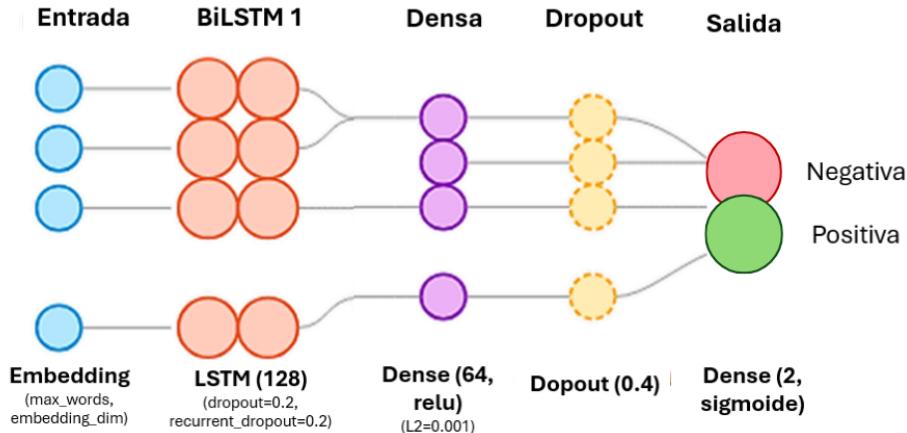


Figura 7: Modelo 1 RNN

La arquitectura implementada, ilustrada en la figura 7, integra los siguientes componentes obtenidos de la literatura y de los cuales se espera construir el primer modelo:

1. Capa de *embedding* que transforma palabras en vectores densos de 256 dimensiones utilizando un vocabulario de 15,000 términos [4].
2. Normalización de secuencias con longitud máxima de 150 tokens mediante operaciones de *padding* y truncamiento [1].
3. Capa *LSTM* bidireccionales con 128 unidades, para el procesamiento secuencial profundo.
4. Regularización mediante *dropout* (0.2 en LSTM, 0.4 en capa densas) y regularización L2 (0.001) para prevenir el sobreajuste [1,4].
5. Capa de salida con activación *sigmoide* para la clasificación binaria de sentimientos.

4.1.2. Red Neuronal Recurrente Bi-LSTM

Las redes LSTM representan una arquitectura especializada para procesamiento de datos secuenciales, diseñadas para capturar dependencias temporales en secuencias largas superando las limitaciones de RNNs tradicionales [1]. Para este modelo se quiere robustecer un poco el primer modelo y realizar un cambio en la función de la última capa. La implementación bidireccional permite procesar información en ambas direcciones (inicio-fin y viceversa), mejorando significativamente la comprensión contextual del sentimiento [4].

Arquitectura e Implementación

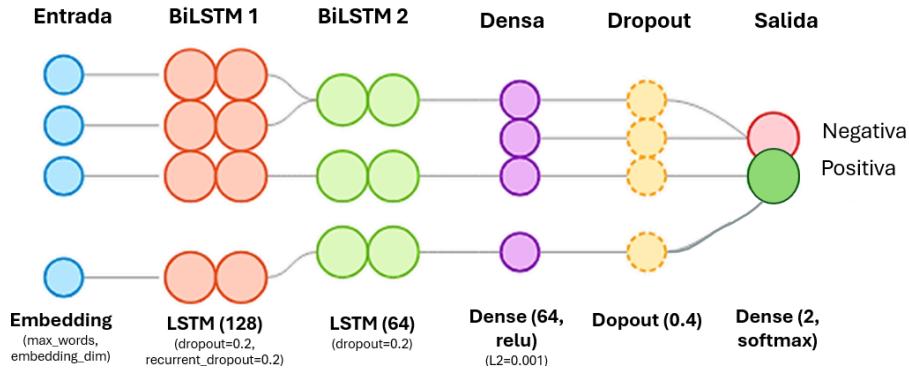


Figura 8: Modelo 2 RNN

La arquitectura implementada, ilustrada en la figura 8, integra los siguientes componentes obtenidos de la literatura y de los cuales se espera construir el primer modelo:

1. Capa de *embedding* que transforma palabras en vectores densos de 256 dimensiones utilizando un vocabulario de 15,000 términos [4].
2. Normalización de secuencias con longitud máxima de 150 tokens mediante operaciones de *padding* y truncamiento [1].
3. Capas *LSTM* bidireccionales con 128 y 64 unidades, respectivamente, para el procesamiento secuencial profundo.
4. Regularización mediante *dropout* (0.2 en LSTM, 0.4 en capas densas) y regularización L2 (0.001) para prevenir el sobreajuste [1,4].
5. Capa de salida con activación *softmax* para la clasificación binaria de sentimientos.

4.2. Métricas de Evaluación

La evaluación del rendimiento requiere métricas que capturen diferentes aspectos de la efectividad predictiva. Como establecen Moraes et al. [3], las métricas se fundamentan en la matriz de confusión que categoriza predicciones según su correspondencia con etiquetas verdaderas; por ende, se esperan utilizar las siguientes métricas de evaluación para tener una comparación más amplia de los modelos a implementar.

Matriz de Confusión y Métricas de Rendimiento La matriz de confusión organiza resultados en cuatro categorías fundamentales para clasificación binaria de sentimientos:

Cuadro 4: Matriz de Confusión para Clasificación de Sentimientos

	Predicción Positiva	Predicción Negativa
Clase Positiva Real	Verdaderos Positivos (TP)	Falsos Negativos (FN)
Clase Negativa Real	Falsos Positivos (FP)	Verdaderos Negativos (TN)

A partir de estos elementos se derivan las métricas principales utilizadas para evaluar rendimiento:

Exactitud (Accuracy): Proporción total de clasificaciones correctas, definida por Moraes et al. [3] como:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

Precisión (Precision): Proporción de predicciones positivas correctas:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

Sensibilidad (Recall): Capacidad para identificar casos positivos reales:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

F1-Score: Medida balanceada combinando precisión y recall mediante media armónica:

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

El F1-Score resulta especialmente valioso para comparación entre modelos cuando no existe preferencia clara entre minimizar falsos positivos versus falsos negativos, proporcionando equilibrio entre precisión y recall.

5. Conclusiones Preliminares

El análisis exploratorio del dataset de reseñas de hoteles andaluces ha revelado características que guiarán la implementación de los modelos RNN:

1. El dataset presenta un desbalance significativo que requerirá técnicas específicas de manejo
2. La variabilidad en la longitud de las reseñas (7-1,416 palabras) exigirá estrategias de padding o truncamiento
3. El vocabulario de 62,639 palabras únicas requerirá técnicas de reducción dimensional
4. La presencia de valores nulos en características geográficas necesitará imputación o exclusión
5. Los patrones lingüísticos identificados sugieren la viabilidad de un modelo RNN para capturar el sentimiento

Referencias

1. Behera, R.K., Jena, M., Rath, S.K., Misra, S.: Co-lstm: Convolutional lstm model for sentiment analysis in social big data. *Information Processing & Management* **58**(1), 102435 (2021). <https://doi.org/10.1016/j.ipm.2020.102435>
2. Chizhikchi: Andalusian hotels reviews unbalanced dataset. Kaggle (2023), <https://www.kaggle.com/datasets/chizhikchi/andalusian-hotels-reviews-unbalanced>, dataset for binary sentiment classification in Spanish hotel reviews
3. Moraes, R., Valiati, J.F., Gavião Neto, W.P.: Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Systems with Applications* **40**, 621–633 (2013). <https://doi.org/10.1016/j.eswa.2012.07.059>
4. Rehman, A.U., Malik, A.K., Raza, B., Ali, W.: A hybrid cnn-lstm model for improving accuracy of movie reviews sentiment analysis. *Multimedia Tools and Applications* **78**, 26597–26613 (2019). <https://doi.org/10.1007/s11042-019-07788-7>
5. Ríos Robby, D.F.: Implementación de análisis de sentimiento con evaluación numérica. Proyecto de grado para el título de ingeniero electrónico, Universidad de los Andes, Bogotá, Colombia (2023), <https://repositorio.uniandes.edu.co/server/api/core/bitstreams/54481b35-8459-48c1-9b77-25e54496defc/content>, asesor: Fredy Enrique Segura Quijano, Ph.D. Departamento de Ingeniería Eléctrica y Electrónica