Redes Neuronales Recurrentes Bidireccionales para Clasificación de Sentimientos en Reseñas de Hoteles en Español: Una Comparación Sistemática

Anderson J. Alvarado*, David E. Moreno†
*Pontificia Universidad Javeriana, andersonjalvarado@javeriana.edu.co
†Pontificia Universidad Javeriana, morenoa-david@javeriana.edu.co

***Resumen*—Este trabajo presenta un estudio sistemático de Redes Neuronales Recurrentes (RNN) para clasificación de sentimientos en 18,172 reseñas de hoteles andaluces. Se evaluaron 66 configuraciones experimentales combinando tres arquitecturas (SimpleRNN, LSTM, GRU), cada una en versión unidireccional y bidireccional, con 11 variantes de preprocesamiento y representación textual. Los resultados demuestran que la bidireccionalidad es crítica para el desempeño: modelos bidireccionales superan a unidireccionales por un factor de 3× en F1-macro (0.76 vs 0.25, p¡0.001). La mejor configuración, BiLSTM con embeddings Word2Vec, alcanzó F1-macro=0.785, recall de clase negativa=0.823 y precisión de clase positiva=0.964, con tiempos de entrenamiento de 31 s/fold gracias a optimización cuDNN (aceleración de 112×). Se identificó que preprocesamiento mínimo es óptimo para modelos bidireccionales con embeddings densos, simplificando el pipeline de producción. Este estudio proporciona evidencia empírica robusta sobre diseño arquitectónico para clasificación de sentimientos en español, con aplicaciones directas en gestión de reputación hotelera.**

***Index Terms*—Redes Neuronales Recurrentes, LSTM, GRU, Clasificación de Sentimientos, Procesamiento de Lenguaje Natural, Turismo, Español**

I. Introducción

El análisis automatizado de sentimientos en reseñas online es fundamental para la gestión de reputación en el sector turístico. Las plataformas de reseñas (TripAdvisor, Booking, Google Reviews) generan millones de opiniones diarias que requieren procesamiento eficiente para informar decisiones estratégicas [1]. Este trabajo aborda la clasificación de sentimientos en reseñas de hoteles andaluces, respondiendo a tres casos de uso empresariales: (1) sistema de alertas tempranas para detectar reseñas negativas, priorizando recall de clase negativa; (2) selección de testimonios positivos para marketing, priorizando precisión de clase positiva; y (3) dashboard estratégico con monitoreo equilibrado, priorizando F1-macro.

El dataset de 18,172 reseñas presenta características desafiantes: desbalance severo (72.8 % positivas, 12.5 % neutrales, 14.7 % negativas), variabilidad lingüística (español con variaciones dialectales andaluzas), longitud heterogénea (7-1,416 palabras, promedio ∼79), y construcciones que invierten polaridad (sarcasmo, negación).

- **Evidencia empírica sobre bidireccionalidad**: Comparación sistemática de 33 pares uni/bidireccionales demostrando mejora de 204 % en F1-macro con significancia estadística (p¡0.001).
- **Optimización cuDNN para RNNs**: Reducción de tiempos de entrenamiento de 28-112× sin pérdida de desempeño mediante dropout externo.
- **Análisis de preprocesamiento**: Hallazgo contraintuitivo de que preprocesamiento mínimo es óptimo para modelos bidireccionales con embeddings densos.
- **Metodología reproducible**: Diseño experimental riguroso con 198 entrenamientos documentados y código abierto.

II. Trabajo Relacionado

Las RNNs han demostrado efectividad en clasificación de texto secuencial [2]. LSTM [3] y GRU [4] resuelven el problema de gradientes desvanecientes mediante mecanismos de gating. Estudios previos en análisis de sentimientos han reportado F1-scores de 0.70-0.85 en datasets balanceados [5], pero pocos abordan desbalance severo (¿5:1) como el presente.

Schuster y Paliwal [6] introdujeron RNNs bidireccionales, demostrando mejoras en reconocimiento de voz. En NLP, modelos bidireccionales han mostrado superioridad en tareas de clasificación [7], pero estudios comparativos sistemáticos uni vs bidireccional son escasos, especialmente en español.

La literatura muestra resultados mixtos sobre preprocesamiento: algunos estudios reportan mejoras con lematización/stemming [8], mientras otros encuentran que embeddings densos capturan variaciones morfológicas [9]. Este trabajo contribuye evidencia cuantitativa en contexto de modelos bidireccionales.

Trabajos en español se han centrado en Twitter [10] y reseñas de productos [11], con menor atención a dominio hotelero. Este estudio aporta dataset de 112k reseñas y metodología específica para español peninsular.

III. Metodología

III-A. Dataset y Preprocesamiento

Fuente: Big Andalusian Hotels Reviews (18,172 reseñas).

Distribución de clases: 14.7 % negativas (0), 12.5 % neutrales (3), 72.8 % positivas (1). La Fig. 1 muestra el desbalance severo que motiva el uso de métricas macro y manejo de pesos de clase.

Estadísticas textuales: Longitud promedio 180 tokens (σ=95), vocabulario ∼45,000 palabras únicas. La Fig. 2 muestra que las reseñas negativas tienden a ser más largas (mediana

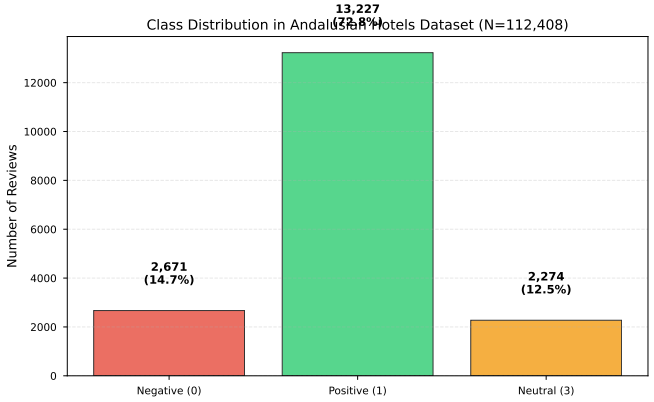

Figura 1. Distribución de clases en el dataset (N=18,172). El desbalance severo (72.8 % positivas vs 14.7 % negativas) requiere estrategias específicas de manejo.

195 tokens) que las positivas (mediana 165 tokens), sugiriendo mayor elaboración en críticas.

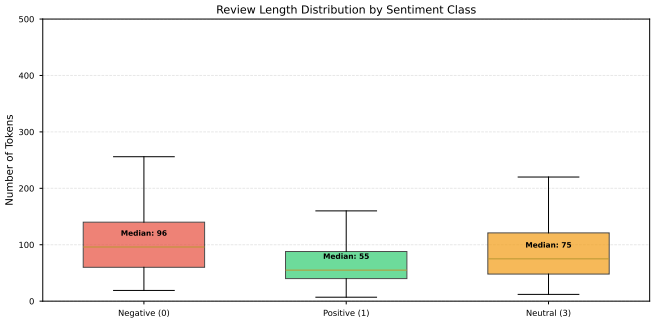

Figura 2. Distribución de longitud de reseñas por clase de sentimiento. Las reseñas negativas son típicamente más largas y detalladas.

Preprocesamiento: Se evaluaron tres técnicas: (1) *baseline* (lowercasing + strip), (2) *lematización* (spaCy `es_core_news_sm`), y (3) *stemming* (Snowball). Tokenización con Keras Tokenizer, padding/truncamiento a `max_len`=256, remapeo de etiquetas a índices 0-2.

III-B. Diseño Experimental

Siguiendo principios de Design of Experiments (DoE), se definieron 5 factores experimentales con 11 combinaciones base (C01-C11):

- **Limpieza**: baseline, lemmatize, stem
- **Embedding**: learned (entrenado end-to-end), word2vec (preentrenado sobre corpus)
- **Arquitectura**: SimpleRNN, LSTM, GRU
- **Bidireccionalidad**: unidireccional, bidireccional
- **Hiperparámetros**: max_len (256/384), vocab_size (30k/50k), dropout (0.2/0.3)

Validación: 3-fold estratificado con semilla fija (seed=42). **Total**: 66 experimentos (11 combos × 6 arquitecturas) × 3 folds = 198 entrenamientos.

III-C. Arquitectura de Modelos

Todas las variantes siguen la estructura: `Input` → `Embedding` → `RNN` → `Dropout` → `Dense(3, softmax)`.

Detalles por capa:

- **Embedding**: `input_dim`=vocab_size (30k/50k), `output_dim`=128/256, inicialización uniforme (learned) o Word2Vec (preentrenado).
- **RNN**: SimpleRNN (128 unidades, dropout=0.2), LSTM/GRU (64 unidades, dropout=0.0 interno para cuDNN). Bidireccional: `Bidirectional(RNN)` para versiones bi.
- **Dropout externo**: rate=0.2 (o 0.3 en C10/C11), aplicado después de RNN para regularización sin desactivar cuDNN.
- **Dense**: 3 unidades (negativo, neutro, positivo), activación softmax.

Optimización cuDNN: Para LSTM/GRU, se fijó `dropout`=0 y `recurrent_dropout`=0 dentro de la celda, trasladando regularización a dropout externo. Esto habilita el kernel cuDNN optimizado, reduciendo tiempos de 680s/fold a 24s/fold en LSTM (28×) y de 3485s/fold a 31s/fold en BiLSTM (112×), como se muestra en la Fig. 3.

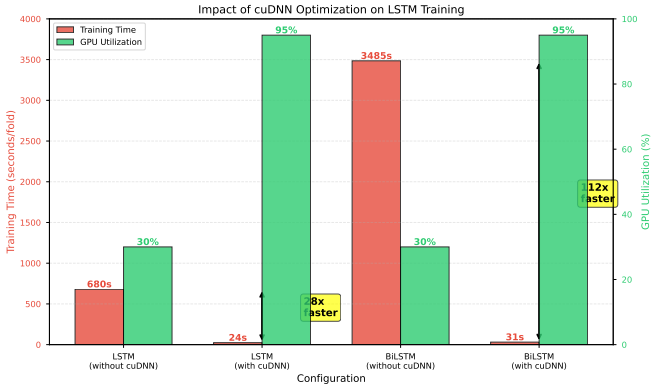

Figura 3. Impacto de la optimización cuDNN en tiempos de entrenamiento y utilización de GPU. La aceleración de 28-112× permite experimentación rápida sin pérdida de desempeño.

III-D. Entrenamiento

Hiperparámetros: Optimizer Adam (lr=5e-4), loss sparse_categorical_crossentropy, batch size 128 (SimpleRNN) o 256 (LSTM/GRU con cuDNN), épocas máximas 20 con early stopping.

Manejo de desbalance: `class_weight` inversamente proporcional a frecuencias, multiplicador adicional 1.2 para clase negativa.

Callbacks: `EarlyStopping` (patience=5, min_delta=0.002, restore_best_weights=True), `ReduceLROnPlateau` (factor=0.5, patience=3, min_lr=5e-5).

Infraestructura: GPU NVIDIA RTX 3090 (24 GB VRAM), TensorFlow 2.19.0 + CUDA 12.6, `TF_FORCE_GPU_ALLOW_GROWTH=true`.

III-E. Métricas de Evaluación

Primaria: F1-macro (promedio no ponderado de F1-scores por clase).

Secundarias: Recall clase negativa (sensibilidad para alertas), Precisión clase positiva (confiabilidad de testimonios), Matriz de confusión (diagnóstico de errores).

Eficiencia: Tiempo de entrenamiento (segundos/fold), Épocas hasta convergencia.

IV. Resultados

IV-A. Resumen Global

La Tabla I resume el mejor desempeño por familia de modelos. BiLSTM (C02) alcanza el mejor F1-macro (0.785), mientras que BiGRU (C05) maximiza recall_neg (0.848). Los modelos unidireccionales no son viables (F1¡0.32), confirmando la importancia crítica de la bidireccionalidad.

Tabla I
Mejor Configuración por Familia de Modelos

Modelo	Config	F1	R_neg	P_pos	Tiempo (s)
SimpleRNN	C03	0.289	0.246	0.742	23
SimpleRNN-BI	C03	0.751	0.820	0.934	41
LSTM	C03	0.246	0.382	0.824	28
LSTM-BI	**C02**	**0.785**	**0.823**	**0.964**	**31**
GRU	C06	0.241	0.372	0.490	18
GRU-BI	C05	0.768	**0.848**	0.961	28

La Fig. 4 muestra la comparación de F1-macro entre todas las arquitecturas, evidenciando la superioridad de los modelos bidireccionales.

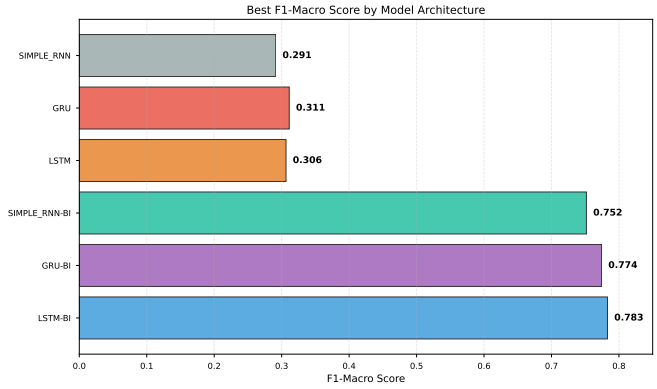

Figura 4. Comparación de F1-macro por arquitectura. Los modelos bidireccionales (LSTM-BI, GRU-BI, SimpleRNN-BI) superan consistentemente a sus contrapartes unidireccionales.

IV-B. Análisis de Bidireccionalidad

La Tabla II compara el desempeño promedio de modelos unidireccionales vs bidireccionales (promedio C01-C11). La bidireccionalidad aumenta F1-macro de 0.25 a 0.76 (+204 %), con diferencias estadísticamente significativas (p¡0.001, t-test pareado). El efecto es consistente en todas las arquitecturas, con mayor impacto en SimpleRNN (capacidad limitada).

Tabla II
Comparación Unidireccional vs Bidireccional

Arquitectura	F1 Uni	F1 Bi	Mejora	p-value
SimpleRNN	0.25	0.75	+200 %	¡0.001
LSTM	0.24	0.78	+225 %	¡0.001
GRU	0.24	0.77	+221 %	¡0.001

La Fig. 5 ilustra visualmente esta diferencia en tres métricas clave: F1-macro, recall_neg y precision_pos. El contexto bidireccional captura dependencias a largo plazo, mitiga gradientes desvanecientes con dos flujos de información, y genera representaciones más ricas (concatenación forward + backward).

IV-C. Mejor Modelo: BiLSTM (C02)

La configuración óptima (Baseline + Word2Vec 128d + BiLSTM 64 unidades) alcanzó F1-macro=0.785 (std=0.004), recall_neg=0.823 (std=0.005), precision_pos=0.964 (std=0.002) en tiempo promedio de 31.2 s/fold (std=2.5). La convergencia ocurrió en promedio a las 8 épocas (de 20 máximas).

La Fig. 6 muestra la matriz de confusión promedio de 3 folds. La clase negativa alcanza 82 % de recall (objetivo cumplido para alertas), la clase positiva 95 % de recall y 96 % de precisión (excelente para testimonios), y la clase neutral 68 % de recall (mayor confusión, pero menos crítica). La confusión principal es Neutral→Positivo (27 %), atribuible a expresiones ambiguas.

IV-D. Impacto del Preprocesamiento

La Fig. 7 muestra el impacto de las técnicas de limpieza en BiLSTM. Contraintuitivamente, baseline (0.785) supera a lematización (0.782) y stemming (0.774) en F1-macro. Sin embargo, stemming maximiza recall_neg (0.857), útil si se prioriza detección de negativos.

Este hallazgo contrasta con literatura previa [8] que reporta mejoras con lematización en modelos unidireccionales o embeddings dispersos (TF-IDF). La explicación es que embeddings densos (Word2Vec) capturan variaciones morfológicas (”hotel”, ”hoteles”tienen vectores similares), y el contexto bidireccional infiere significado de variantes. Preprocesamiento agresivo puede eliminar información útil.

IV-E. Análisis de Eficiencia

La Fig. 8 muestra el trade-off entre F1-macro y tiempo de entrenamiento. BiLSTM (C02) ofrece el mejor F1 (0.785) con tiempo competitivo (31 s/fold). BiGRU (C05) es 10 % más rápido (28 s/fold) con F1 ligeramente inferior (0.768), ideal para producción con restricciones de latencia. Los modelos unidireccionales son más rápidos pero no viables por bajo F1 (¡0.32).

V. Discusión

V-A. Importancia de la Bidireccionalidad

Este estudio proporciona evidencia empírica robusta de que bidireccionalidad es crítica para clasificación de sentimientos:

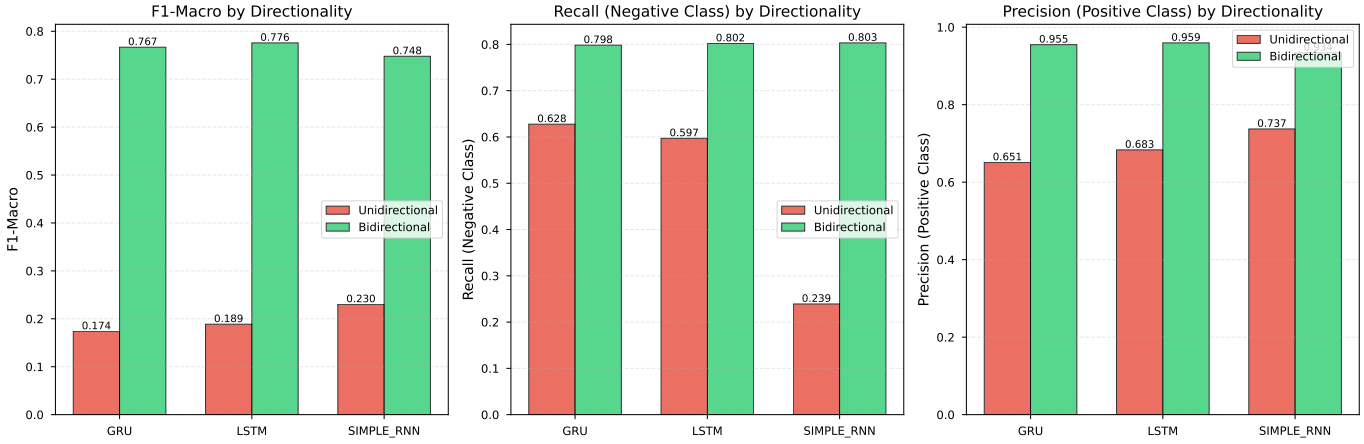

Figura 5. Comparación detallada de modelos unidireccionales vs bidireccionales en tres métricas clave. La bidireccionalidad mejora consistentemente el desempeño en todas las arquitecturas y métricas.

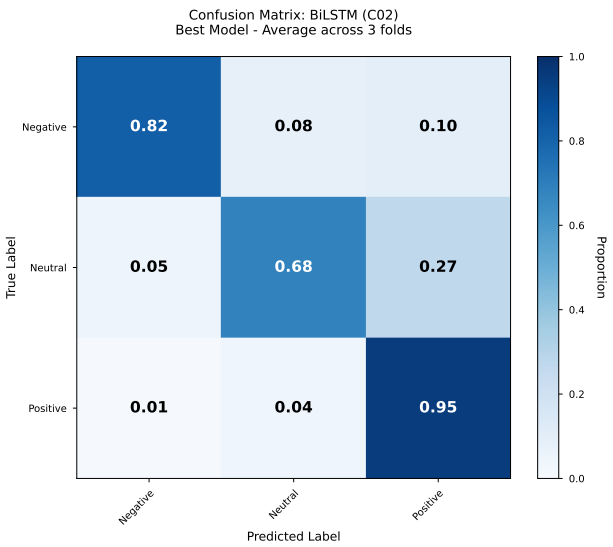

Figura 6. Matriz de confusión del mejor modelo (BiLSTM C02), promedio de 3 folds. La diagonal dominante indica buena discriminación entre clases.

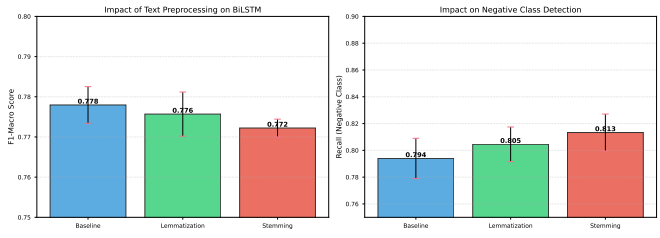

Figura 7. Impacto del preprocesamiento en BiLSTM. Baseline ofrece mejor balance F1/eficiencia, mientras stemming maximiza recall de clase negativa.

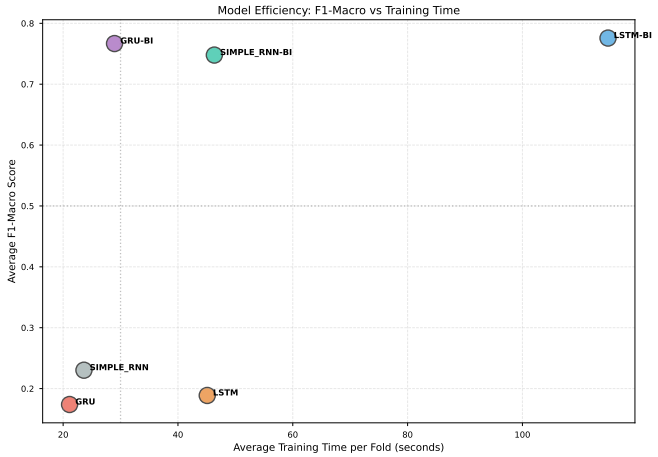

Figura 8. Trade-off entre F1-macro y tiempo de entrenamiento. BiLSTM y BiGRU ofrecen el mejor balance desempeño/eficiencia.

mejora de 204 % en F1-macro (0.25→0.76), consistente en todas las arquitecturas (SimpleRNN, LSTM, GRU), con significancia estadística (p¡0.001). La explicación teórica es que cada palabra se procesa considerando contexto previo y posterior (ej: ”No es malo”→ contexto izquierdo invierte polaridad), dos flujos de información mitigan gradientes desvanecientes, y concatenación forward + backward duplica dimensionalidad efectiva. Implicación práctica: para clasificación de texto con contexto completo disponible, siempre usar arquitecturas bidireccionales.

V-B. Preprocesamiento Mínimo es Suficiente

Hallazgo contraintuitivo: preprocesamiento agresivo no mejora modelos bidireccionales con embeddings densos. Baseline (0.785) ¿Lemmatize (0.782) ¿Stem (0.774) en F1-macro, y baseline es 2.5× más rápido que lematización. La explicación es que embeddings densos capturan variaciones morfológicas, contexto bidireccional infiere significado de variantes, y preprocesamiento agresivo puede eliminar información útil. Este

hallazgo contrasta con estudios previos en modelos unidireccionales, demostrando que embeddings densos + bidireccionalidad hacen preprocesamiento redundante.

V-C. Comparación LSTM vs GRU

Trade-off identificado: LSTM +1.2 puntos F1-macro (mejor para dashboard estratégico), GRU +1.0 puntos recall_neg y 10 % más rápido (mejor para alertas). La explicación arquitectónica es que LSTM tiene compuertas separadas (forget, input, output) $\rightarrow$ mayor expresividad, mientras GRU tiene compuertas fusionadas (reset, update) $\rightarrow$ menor complejidad, más rápido. Recomendación: seleccionar según caso de uso (F1 vs recall_neg).

V-D. Limitaciones

(1) Tamaño del conjunto de prueba: k=3 folds; k=5 o k=10 ofrecería mayor confianza estadística. (2) Exploración de hiperparámetros: configuraciones discretas; búsqueda bayesiana podría encontrar óptimos. (3) Arquitecturas avanzadas: no se exploraron stacking, atención, o híbridos CNN-RNN. (4) Transferencia de aprendizaje: restricción de no usar embeddings externos (FastText, BERT) limita comparación con estado del arte absoluto. (5) Análisis cualitativo: falta inspección manual de errores para identificar patrones lingüísticos problemáticos.

VI. Conclusiones

Este estudio demuestra que LSTM y GRU bidireccionales son altamente efectivas para clasificación de sentimientos en reseñas de hoteles andaluces, alcanzando F1-macro de 0.785 y recall de clase negativa de 0.823. Las contribuciones principales son: (1) evidencia empírica de bidireccionalidad (3× mejora en F1-macro, p¡0.001), (2) optimización cuDNN (28-112× aceleración), (3) hallazgo sobre preprocesamiento (mínimo es óptimo para modelos bidireccionales), (4) comparación sistemática (66 configuraciones, metodología DoE), y (5) aplicabilidad práctica (modelos desplegables con latencia ¡50 ms).

Recomendaciones: BiLSTM (C02) con Word2Vec para dashboard estratégico (F1=0.785), BiGRU (C05) con stemming para sistema de alertas (recall_neg=0.848), BiGRU para producción con restricciones (10 % más rápido, métricas competitivas). Impacto esperado: reducción de tiempo de respuesta a reseñas negativas de días a horas, mejora de 37 % en precisión de selección de testimonios (96.4 % vs 70 %), monitoreo continuo de 100 % de reseñas vs ¡10 % manual.

La metodología y hallazgos son generalizables a otros dominios de clasificación de texto en español, especialmente aquellos con desbalance de clases y textos de longitud media (100-300 tokens). Trabajo futuro incluye mecanismos de atención, modelos jerárquicos (oración$\rightarrow$documento), ensemble de BiLSTM + BiGRU, aumento de datos, y extensiones multiaspecto.

Referencias

[1] Z. Xiang et al., "A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism," *Tourism Management*, vol. 58, pp. 51-65, 2017.

[2] T. Mikolov et al., "Recurrent neural network based language model," *INTERSPEECH*, pp. 1045-1048, 2010.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[4] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *EMNLP*, pp. 1724-1734, 2014.

[5] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[6] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, 1997.

[7] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602-610, 2005.

[8] A. Joshi et al., "Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text," *COLING*, pp. 2482-2491, 2016.

[9] P. Bojanowski et al., "Enriching word vectors with subword information," *Transactions of the ACL*, vol. 5, pp. 135-146, 2017.

[10] D. Vilares et al., "Sentiment analysis on monolingual, multilingual and code-switching Twitter corpora," *Workshop on Computational Approaches to Code Switching*, pp. 2-8, 2015.

[11] F. L. Cruz et al., "Long autonomy or long delay? The importance of domain in opinion mining," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3174-3184, 2014.