Práctica 3

Grado en Ciencia e Ingeniería de Datos

Procesamiento de Lenguaje Natural 2025-26

Modelos neuronales para minería de opinión

Pareja 01 → Héctor Tablero Díaz y Álvaro Martínez Gamo.

1. Embeddings de Documento y Configuración

Para la primera fase de la práctica, centrada en el uso de **embeddings estáticos**, se optó por un enfoque basado en representaciones preentrenadas para transformar los tokens de las reseñas en vectores densos.

Configuración de Embeddings:

- **Modelo Preentrenado:** `glove-wiki-gigaword-100`.
- **Dimensionalidad:** 100 dimensiones por vector.
- **Longitud de secuencia:** Se fijó un máximo de 200 tokens.
- **Estrategia de Agregación:** Para la red FNN, los embeddings de palabra se agregaron (promedio) para formar el embedding del documento. En el caso de la RNN, se mantuvo la secuencia temporal.

Configuración de Entrenamiento:

Se establecieron 10 épocas con un *batch size* de 32 para ambos modelos neuronales desarrollados.

2. Análisis de Resultados de Clasificación (Ejercicios 1 y 2)

En esta sección se comparan los resultados obtenidos mediante arquitecturas neuronales básicas (FNN y RNN) utilizando embeddings estáticos, frente al ajuste fino (*fine-tuning*) de un modelo de lenguaje preentrenado (BERT).

2.1. Redes Neuronales con Embeddings Estáticos (FNN vs RNN)

Se evaluaron dos arquitecturas distintas sobre el conjunto de test:

1. **Feedforward Neural Network (FNN):** Obtuvo un **F1-Score de 0.4810** y un Accuracy de 0.4881. El modelo mostró debilidades significativas en la clase *neutral*, con una precisión de apenas 0.38 y un recall de 0.31.
2. **Red Neuronal Recurrente (Bi-LSTM):** Esta arquitectura superó a la FNN, alcanzando un **F1-Score de 0.5291** y un Accuracy de 0.5412. La capacidad de la LSTM bidireccional para capturar dependencias secuenciales mejoró la identificación de clases, elevando el F1-score de la clase *negativa* a 0.61 y la *positiva* a 0.62.

2.2. Fine-tuning con BERT

Para el Ejercicio 2, se realizó un *fine-tuning* del modelo `bert-base-uncased` durante 3 épocas con un *learning rate* de 0.0002.

Resultados en Test:

- **Accuracy:** 0.6022
- **F1-Score:** 0.6063

El modelo BERT superó ampliamente a los modelos basados en embeddings estáticos (GloVe) y también a los modelos clásicos (SVM con Bigramas) de la práctica anterior.

Análisis por clase (BERT):

La matriz de confusión revela una mejora notable en la separación de clases. Aunque la clase *neutral* sigue siendo la más difícil de clasificar (F1: 0.50), las clases polarizadas (*negative* y *positive*) alcanzaron un F1-Score de 0.66 y 0.65 respectivamente.

2.3. Comparativa Global

Modelo	Representación	F1-Score	Observaciones
FNN	GloVe (Static)	0.4810	Rendimiento base bajo, especialmente en neutros.
Bi-LSTM	GloVe (Static)	0.5291	Mejora al capturar contexto secuencial.
BERT	Contextual (Fine-tuned)	**0.6063**	Mejor rendimiento global y mayor capacidad de generalización.

El salto de calidad entre la Bi-LSTM y BERT (aprox. +7.7 puntos porcentuales) evidencia la superioridad de los embeddings contextualizados frente a los estáticos para tareas de análisis de sentimiento complejas donde el contexto modifica la polaridad de las palabras.

3. Generación de Resumen de Opinión con LLM

Para el tercer ejercicio, se seleccionó la **Opción 3a: Generación de resumen de opinión**. El objetivo fue sintetizar cualitativamente las críticas de un juego de mesa a partir de un conjunto de reseñas.

3.1. Metodología y Prompting

- **Juego Analizado:** Game ID 13 (Catan).
- **Datos:** 15 reseñas con distribución equilibrada (5 positivas, 5 neutras, 5 negativas).
- **Método:** Prompting utilizando Ollama (local) y simulación de estructura para ChatGPT/Gemini.

Se diseñaron dos niveles de prompts:

1. **Prompt Simple:** Instrucción directa para resumir.
2. **Prompt Detallado:** Se proporcionó al LLM el contexto del experto, estadísticas del juego (rating promedio 5.73) y las reseñas estructuradas con metadatos (rating y sentimiento explícito).

3.2. Estructura del Prompt

El prompt final enviado al modelo incluyó las reseñas completas y solicitó una salida estructurada en cuatro bloques: Opinión General, Puntos Fuertes, Puntos Débiles y Conclusión.

Ejemplo de evidencia extraída por el LLM:

El modelo fue capaz de identificar matices específicos más allá de las palabras clave. Por ejemplo, detectó el problema de mecánica conocido como *"runaway leader problem"* mencionado en las reseñas neutrales y la discrepancia entre la importancia histórica del juego (1995) y su jugabilidad considerada "mediocre" bajo estándares modernos.

3.3. Resultados Obtenidos

El LLM generó un resumen coherente que reflejaba la dualidad de las opiniones:

- **Aspectos Positivos:** Destacó la importancia histórica y cómo las expansiones (fan-developed) revitalizan el juego.
- **Aspectos Negativos:** Identificó correctamente la frustración por la eliminación implícita de jugadores ("losing way too early") y la duración excesiva para los estándares actuales.

Esta capacidad de síntesis semántica demuestra que, mientras BERT es útil para asignar una etiqueta numérica (clasificación), los LLMs aportan una capa de explicabilidad y resumen cualitativo indispensable para entender el *porqué* de dichas etiquetas.

> Se puede ver el prompt y su correspondiente respuesta por parte del LLM en:
> https://github.com/HectorTablero/pln-neural-network-sentiment-analysis