Fusión multimodal para tareas de recomendación

Pablo Pérez-Núñez[1] and Jorge Díez[1]

Centro de Inteligencia Artificial, Universidad de Oviedo
{pabloperez,jdiez}@uniovi.es

Resumen El contenido multimodal generado por los usuarios (texto e imágenes) se ha convertido en una fuente de información clave para los Sistemas de Recomendación. Para procesar estos datos, el enfoque más habitual es la Fusión Tardía. Aunque es computacionalmente eficiente, suele sufrir de "ceguera modal": no logra alinear semánticamente las modalidades ni detectar sarcasmos o contradicciones. Para superar esta limitación, este artículo evalúa un enfoque de Fusión Temprana que confía la integración multimodal a un Modelo de Visión-Lenguaje (VLM). En concreto, proponemos un esquema de instrucciones para que el VLM genere una representación semántica unificada de cada reseña.
Ambos paradigmas se evalúan sobre un conjunto de datos real de restaurantes de TripAdvisor en Gijón. Las representaciones extraídas se comparan en dos tareas: predicción de interacciones y análisis de sentimientos. Los resultados muestran que, si bien la Fusión Tardía es más robusta y eficiente para agregar perfiles colaborativos, la Fusión Temprana captura una semántica mucho más profunda y fiel a la intención del usuario. Sin embargo, la alta latencia de inferencia de los modelos generativos compromete su uso en tiempo real, limitándolos actualmente a procesos de indexación *offline*.

Keywords: Fusión Multimodal · Modelos de Lenguaje Grandes · Modelos de Visión-Lenguaje · Sistemas de Recomendación

1. Introducción

En los últimos años, los Sistemas de Recomendación han evolucionado integrando información contextual para modelar con mayor precisión las preferencias de usuarios e ítems. En plataformas de servicios locales, el Contenido Generado por el Usuario (UGC) en forma de reseñas multimodales (texto descriptivo e imágenes) se ha convertido en una fuente de información de gran valor [12,7], ya que captura matices que los metadatos estáticos ignoran, como la calidad del servicio o el ambiente.

A pesar de su utilidad, el debate sobre cómo integrar eficazmente estas modalidades sigue abierto. El enfoque dominante es la *Fusión Tardía Multimodal* [13,10], donde texto e imagen se procesan aisladamente mediante codificadores específicos (ej. BERT [1] y ResNet [4]) para luego agregar sus representaciones latentes. Aunque computacionalmente eficiente, este paradigma carece de

alineación semántica profunda: ignora interacciones complejas o contradictorias entre lo escrito y lo fotografiado, como el uso del sarcasmo.

Para capturar dichas interacciones se suele recurrir a la *Fusión Temprana Multimodal* [6], que integra las modalidades en etapas iniciales para extraer sus relaciones semánticas. Aunque tradicionalmente requería arquitecturas complejas, la reciente irrupción de los Modelos de Visión-Lenguaje (VLMs) permite delegar esta integración a la capacidad de razonamiento del modelo. Un VLM instruido puede analizar simultáneamente la imagen y el texto crudo para generar un resumen unificado y coherente de la experiencia del usuario, fácilmente codificable para su uso en recomendación.

El objetivo de este trabajo es comparar las diferencias de rendimiento entre ambos paradigmas (Fusión Temprana Multimodal con VLMs y Fusión Tardía Multimodal) al utilizarlos como métodos de codificación de reseñas en el ámbito de la recomendación de restaurantes.

En concreto, las contribuciones principales de este artículo son:

1. **Diseño de un esquema de instrucción multimodal**. Se propone y valida un *prompt* específico para VLMs diseñado para realizar una Fusión Temprana, extrayendo una representación semántica unificada que resuelve discrepancias entre texto e imagen.
2. **Evaluación en diferentes tareas**. Se compara empíricamente la calidad de la codificación propuesta frente a la Fusión Tardía en un dominio real (TripAdvisor), evaluando su rendimiento en la predicción de interacciones futuras (NDCG) y el análisis de sentimientos (F1 macro).
3. **Análisis comparativo de eficiencia y rendimiento**. Se discuten las implicaciones prácticas de usar VLMs como extractores de características, analizando el equilibrio entre la mejora predictiva y el incremento del coste y la latencia computacional.

2. Trabajo relacionado

Sistemas de Recomendación basados en Reseñas. La literatura ha demostrado ampliamente que incorporar UGC mejora la precisión predictiva frente al uso exclusivo de metadatos. Zheng et al. [12] evidenciaron el valor de los textos de las reseñas, y desde entonces, arquitecturas basadas en Transformers [1] se han estandarizado para este fin. Sin embargo, en dominios como la gastronomía, el texto por sí solo ofrece una visión parcial si se ignora la rica información visual que a menudo lo acompaña.

Fusión Tardía Multimodal. Para integrar dichas imágenes, modelos pioneros como VBPR [5] y arquitecturas posteriores como MMGCN [10] se han consolidado como el estándar de la industria. El denominador común de estos sistemas es el uso de la *Fusión Tardía Multimodal* [13]. Como se introdujo anteriormente, aunque este enfoque facilita la paralelización del cómputo, sufre de *ceguera modal* durante la extracción de características, siendo incapaz de identificar relaciones semánticas cruzadas críticas [6].

Fusión Temprana y Modelos de Visión-Lenguaje. Para superar esta independencia modal, la *Fusión Temprana* mediante espacios latentes entrelazados ha demostrado empíricamente capturar mejor la intención del usuario [11]. Recientemente, arquitecturas como MLLM-SRec [9] han evidenciado el potencial de los VLMs para extraer representaciones unificadas. No obstante, la mayoría se centra en fusionar metadatos estáticos de catálogo. Otros enfoques, como [8], analizan el texto con un LLM y las imágenes con un VLM de forma separada para luego concatenar los resúmenes, recayendo nuevamente en una fusión tardía sin una relación semántica cruzada real.

A diferencia de dichos trabajos, esta investigación se centra en el UGC, caracterizado por un alto nivel de ruido y contradicciones. Además, en lugar de proponer una nueva arquitectura *end-to-end*, este estudio aísla la extracción de características para ofrecer una comparativa empírica directa y un análisis de coste-beneficio entre la fusión temprana delegada a VLMs y la fusión tardía tradicional.

3. Codificación de las Reseñas Multimodales

Para dar respuesta a las limitaciones identificadas en la literatura y evaluar empíricamente la diferencia entre paradigmas, esta sección detalla la metodología diseñada para aislar y evaluar el proceso de integración multimodal.

El objetivo fundamental de esta etapa es transformar una reseña multimodal r, compuesta por un texto descriptivo $T \in \mathcal{T}$ y un conjunto de una o varias imágenes $I = \{i_1, i_2, \ldots, i_n\} \in \mathcal{I}^n$, en una representación vectorial densa v_r que encapsule su significado. Para garantizar una comparativa justa y aislar el impacto de la estrategia de fusión, en este trabajo empleamos el VLM *Gemma 4* [3] como arquitectura base para todos los enfoques. Al utilizar los mismos codificadores unimodales subyacentes, nos aseguramos de que cualquier diferencia en el rendimiento predictivo de los sistemas de recomendación se deba exclusivamente a la capacidad de cada método para capturar y explotar la relación semántica entre el texto y las imágenes, y no a diferencias arquitectónicas en la extracción de características.

Con este propósito, en primer lugar se detalla el comportamiento de los codificadores del modelo *Gemma 4 E2B* (versión del modelo seleccionada) a fin de fijar una notación matemática que sirva de base para, a continuación, detallar los dos métodos de codificación de reseñas propuestos.

3.1. Arquitectura de los Codificadores Unimodales

El procesamiento de información en *Gemma 4 E2B* se basa en la proyección de las modalidades en un espacio latente común de dimensión $d_{\text{emb}} = 1536$. Por un lado, el codificador de lenguaje transforma una secuencia de texto crudo T de longitud N en una matriz de *embeddings*, proceso formalizado mediante la función $E_{\text{txt}} : \mathcal{T} \to \mathbb{R}^{N \times d_{\text{emb}}}$. Por otro lado, el componente visual *Gemma-Vis*, basado en una arquitectura *Vision Transformer* (ViT) [2], procesa cada

imagen de entrada para proyectarla linealmente al mismo espacio dimensional. Este codificador de visión se define como la función $E_{\text{img}} : \mathcal{I} \to \mathbb{R}^{K \times d_{\text{emb}}}$, donde K representa el número de *soft tokens* visuales resultantes. De este modo, las características de una imagen dada i_k se expresan de forma compacta como $e_{i_k} = E_{\text{img}}(i_k) \in \mathbb{R}^{K \times d_{\text{emb}}}$.

Para una instrucción mixta P compuesta por un texto T y un conjunto de imágenes $I = \{i_1, i_2, \ldots, i_n\}$, el modelo extrae sus representaciones independientes asumiendo $e_T = E_{\text{txt}}(T)$ y $e_{i_k} = E_{\text{img}}(i_k)$. Aprovechando la equivalencia dimensional, la fusión inicial se realiza concatenando secuencialmente sus vectores en una matriz unificada $e_P = e_T \oplus e_{i_1} \oplus \cdots \oplus e_{i_n} \in \mathbb{R}^{(N+n \cdot K) \times d_{\text{emb}}}$. Esta secuencia continua se introduce directamente en el núcleo del modelo de lenguaje, donde los mecanismos de atención permiten el razonamiento cruzado intermodal sobre parches visuales y tokens textuales para generar, de forma autorregresiva, la respuesta. El proceso completo de inferencia multimodal se encapsula bajo la función generativa $Y = \mathcal{M}_{\text{Gemma}}(T, I)$, donde $\mathcal{M}_{\text{Gemma}} : \mathcal{T} \times \mathcal{I}^n \to \mathcal{T}$.

3.2. Método Base: Fusión Tardía Multimodal

Este método representa el paradigma tradicional de integración geométrica, donde las modalidades se procesan de manera estrictamente independiente, omitiendo la concatenación y el paso de la secuencia conjunta por los bloques del decodificador descritos anteriormente. Al no existir operaciones de atención cruzada entre los parches visuales y los tokens textuales, las representaciones se mantienen aisladas completamente. El proceso es el siguiente:

1. El texto T se procesa mediante E_{txt} para obtener su secuencia matricial de *embeddings* $\in \mathbb{R}^{N \times d_{\text{emb}}}$. Para condensar esta secuencia en un único vector representativo del texto se aplica la media aritmética sobre la dimensión N resultando en $\bar{e}_T \in \mathbb{R}^{d_{\text{emb}}}$.
2. Cada imagen del conjunto I se procesa de forma aislada mediante E_{img}. De manera análoga al texto, se promedia la dimensión espacial de los K *soft tokens* devueltos por el modelo para obtener un vector denso único por cada fotografía, $\bar{e}_{i_k} \in \mathbb{R}^{d_{\text{emb}}}$.
3. Dado que una reseña puede contener un número variable de fotografías (n), estas representaciones visuales individuales se agregan calculando la media para obtener el vector global representativo de las imágenes de la reseña, $\bar{e}_I = \frac{1}{n} \sum_{k=1}^{n} \bar{e}_{i_k}$ donde $\bar{e}_I \in \mathbb{R}^{d_{\text{emb}}}$.
4. Finalmente, el vector unificado que codifica la reseña completa, v_r, se obtiene calculando la media aritmética entre la representación textual aislada y la representación visual global ($v_r = \frac{\bar{e}_T + \bar{e}_I}{2}$).

En este enfoque, el vector resultante v_r no contiene información sobre las posibles interacciones o contradicciones semánticas entre el texto y las imágenes.

3.3. Método Propuesto: Fusión Temprana Semántica

A diferencia del método base, esta aproximación propone delegar la integración multimodal a las capacidades generativas y de atención cruzada inherentes

a la arquitectura *Gemma 4 E2B*. El objetivo principal es que el modelo interprete conjuntamente el texto descriptivo y el contenido visual para sintetizar una representación semántica unificada, resolviendo explícitamente posibles ambigüedades o contradicciones antes de la vectorización matemática final. El flujo de codificación propuesto consta de dos fases secuenciales:

1. La interacción con el VLM se estructura utilizando su paradigma nativo de roles conversacionales. Por un lado, se define una instrucción detallada bajo el rol de sistema (*system prompt*), denotada como P_{sys}, diseñada para condicionar el comportamiento lógico y analítico del modelo. Por otro lado, la información empírica de la reseña, compuesta por el texto original T y las imágenes I, se introduce conjuntamente bajo el rol de usuario (*user prompt*). Si definimos T_{in} como la secuencia textual internamente ensamblada por la plantilla del modelo para agrupar ambos roles (P_{sys} y T), el VLM evalúa este contexto junto con los tensores visuales para generar un resumen textual S que encapsula la esencia multimodal:

$$S = \mathcal{M}_{\text{Gemma}}(T_{\text{in}}, I) \tag{1}$$

2. El resumen sintético generado S se proyecta al espacio latente utilizando exclusivamente el codificador de lenguaje unificado. Se procesa S mediante E_{txt} para obtener su correspondiente matriz de *embeddings* $\in \mathbb{R}^{L \times d_{\text{emb}}}$, donde L es la longitud en tokens del texto generado. Finalmente, el vector denso que codificará la reseña en los posteriores sistemas de recomendación, $v_r \in \mathbb{R}^{d_{\text{emb}}}$, se extrae aplicando la media aritmética sobre la dimensión secuencial L:

$$v_r = \frac{1}{L} \sum_{j=1}^{L} [E_{\text{txt}}(S)]_j \tag{2}$$

La calidad de la representación latente v_r depende críticamente de las directrices proporcionadas en el rol de sistema P_{sys}. Por ello, el esquema de instrucción propuesto fuerza al VLM a actuar como un analista experto, exigiéndole resolver activamente las discrepancias entre las modalidades. Adicionalmente, se impone en P_{sys} la restricción de que la salida S adopte un formato estructurado en JSON. Esta restricción sintáctica no solo optimiza el rigor del modelo durante la decodificación, sino que facilita su extracción y evaluación computacional en los experimentos. El contenido exacto de esta instrucción de sistema se detalla en el fragmento de código 1.1.

4. Evaluación de la codificación

Para cuantificar la calidad y la riqueza semántica de las representaciones vectoriales obtenidas para cada reseña en la fase de codificación, se evalúa su rendimiento en dos tareas diferentes: recomendación de ítems y análisis de sentimientos. Para garantizar una evaluación realista y evitar la fuga de datos (*data leakage*), se establece un punto de corte temporal estricto, denominado t_{corte}.

Código 1.1: Prompt estructurado utilizado para la extracción semántica.

```
Eres un analista experto en sistemas de recomendación gastronómica y comportamiento del consumidor. Tu tarea
       es analizar de forma conjunta la reseña textual y las fotografías adjuntas proporcionadas por un
      usuario tras visitar un restaurante.

### Instrucciones de Análisis Multimodal:
1. Descripción Objetiva Visual: Lo primero que debes hacer es describir detalladamente y de forma rigurosa
      cada una de las imágenes proporcionadas. No inventes ni deduzcas nada que no sea estrictamente visible
       en ellas. Si no se adjunta ninguna imagen, debes indicarlo explícitamente.
2. Fusión Temprana: Tras la descripción, no analices el texto y la imagen de forma aislada, sino buscando su
       contexto cruzado. Presta especial atención a:
   - Confirmación: Cuando la imagen demuestra gráficamente lo que dice el texto (ej. 'ración enorme' y se ve
         un plato desbordado).
   - Contradicción o Sarcasmo: Cuando el texto dice una cosa pero la imagen muestra lo contrario (ej. 'buena
         pinta' pero la foto muestra comida quemada). En este caso, tu análisis final debe reflejar la
        realidad inferida de ambas fuentes.
   - Detalles Ocultos: Información visual útil (como la iluminación, la limpieza o el estilo del local) que
         el usuario no ha mencionado en el texto pero que es evidente en la fotografía.

### Formato de Salida:
Devuelve ÚNICAMENTE un objeto JSON válido con la siguiente estructura, sin texto adicional:
{ "comida_y_bebida": "Resumen semántico sobre la calidad, presentación y cantidad de la comida, fusionando
      lo visto y lo leído.",
  "ambiente_y_servicio": "Resumen semántico sobre el local, la atmósfera y el trato del personal.",
  "descripcion_imagenes": "Descripción detallada y puramente objetiva de cada imagen proporcionada, sin
        inventar nada. Si no hay imágenes, escribe exactamente: 'No se han proporcionado imágenes'.",
  "alineacion_multimodal": "Describe brevemente si el texto y la foto se apoyan mutuamente o si hay sarcasmo
       /contradicciones.",
  "resumen_unificado": "Un párrafo final y rico en semántica que capture la esencia real y completa de la
       experiencia del usuario. Este será el texto que defina la reseña.",
  "sentimiento_inferido": "Positivo, Neutral o Negativo"}
```

Toda la información generada antes de t_{corte} se utiliza para entrenar los modelos, mientras que las interacciones posteriores a t_{corte} conforman el conjunto de prueba.

4.1. Sistema de recomendación para la predicción de interacciones

Esta tarea inicial se plantea operativamente como un problema de clasificación binaria, cuya salida probabilística se utiliza para generar un ranking personalizado: dado un usuario y un conjunto de ítems, el sistema debe predecir la probabilidad de interacción para poder ordenarlos. Para lograrlo, primero es necesario disponer de una representación que resuma el historial de un usuario y las características de un ítem. En este estudio, los perfiles se construyen dinámicamente agregando las codificaciones de las reseñas publicadas hasta la fecha de corte de la siguiente forma:

- El *perfil del usuario* (P_u) se calcula como la media aritmética de los vectores que codifican todas las reseñas que el usuario u ha escrito antes del día t_{corte}.
- De la misma manera, el *perfil del ítem* (P_i) se calcula como la media aritmética de los vectores que codifican todas las reseñas que el ítem i ha recibido antes del día t_{corte}.

Cabe destacar que este proceso de agregación se ejecuta de forma paralela y aislada para los dos métodos de codificación descritos en la Sección 3.

Para construir el conjunto de entrenamiento, se extraen todas las interacciones registradas hasta el día t_{corte}, formando tuplas positivas estructuradas como $(P_u, P_i, 1)$. Dado que los conjuntos de datos de recomendación empleados solo

contienen retroalimentación implícita positiva, se aplica una técnica de muestreo negativo (*Negative Sampling*). Por cada interacción positiva, se selecciona aleatoriamente un par usuario-ítem que no haya interactuado en el conjunto de entrenamiento, generando una tupla negativa $(P_u, P_i, 0)$. Esto garantiza un conjunto de datos balanceado.

El modelo predictivo consiste en un Perceptrón Multicapa (MLP, por sus siglas en inglés). La red recibe como entrada la concatenación de los perfiles P_u y P_i, los procesa a través de varias capas ocultas no lineales, y finaliza en una capa de salida con función de activación sigmoide, la cual emite la probabilidad de que el usuario interactúe con el restaurante.

Para la fase de prueba, se utilizan todas las interacciones ocurridas a partir del día t_{corte}. Las probabilidades generadas por la red neuronal se emplean para ordenar los ítems candidatos para cada usuario. El rendimiento del modelo se evalúa utilizando la métrica estándar en tareas de ranking: la Ganancia Acumulada Descontada Normalizada (NDCG), evaluada habitualmente sobre los K ítems mejor posicionados.

4.2. Análisis de sentimientos

En la tarea anterior nos centramos en evaluar la capacidad de cada una de las codificaciones de reseñas en un escenario predictivo de recomendación. Sin embargo, la necesidad de agregar dichas codificaciones para crear perfiles históricos de usuarios e ítems introduce un factor de confusión: un bajo rendimiento predictivo podría deberse tanto a una representación semántica deficiente de la reseña original como a una pérdida de información durante el proceso de agregación.

Para aislar la calidad de la representación vectorial obtenida (v_r) y verificar empíricamente si la metodología propuesta logra capturar las relaciones intermodales intrínsecas, se introduce una segunda tarea de evaluación: el análisis de sentimientos a nivel de reseña individual.

El problema se modela como una tarea de clasificación binaria, donde el objetivo es predecir la polaridad del sentimiento expresado en la reseña a partir, exclusivamente, de su representación vectorial $v_r \in \mathbb{R}^{d_{\text{emb}}}$. Las etiquetas de clase se derivan de la valoración numérica original otorgada por el usuario en la plataforma: se asigna la clase positiva a las valoraciones mayores o iguales a 4 (sobre 5), y la clase negativa a las valoraciones menores o iguales a 3. Para garantizar la coherencia con la tarea de recomendación, se establece la misma división temporal estricta de los datos.

Para mitigar el desbalanceo extremo en las clases (en torno a un 80 % de reseñas positivas frente a un 20 % negativas) se aplica un sobremuestreo dinámico (*oversampling* aleatorio con reemplazo de la clase minoritaria) en cada época del entrenamiento. Por el contrario, en los conjuntos de validación y prueba no se realiza ninguna alteración, manteniendo estrictamente la distribución natural y desbalanceada de los datos con el fin de evaluar el rendimiento real del sistema.

Debido a este desbalanceo, el rendimiento se evalúa utilizando como métrica principal el *F1 macro*, el cual calcula la media aritmética no ponderada de la métrica F1 obtenida para la clase positiva y negativa de forma independiente.

5. Diseño Experimental

En esta sección se describen los detalles de la configuración de los experimentos, incluyendo una descripción del conjunto de datos, los criterios de filtrado aplicados y la configuración técnica de los modelos evaluados.

5.1. Conjunto de datos de TripAdvisor

Para los experimentos se ha seleccionado un subconjunto de datos reales tomado de la plataforma TripAdvisor con las reseñas disponibles hasta 2023 y centrado en el dominio de la restauración. Con el objetivo de acotar el dominio y garantizar una densidad de interacciones coherente para la recomendación, este subconjunto se restringe geográficamente a los restaurantes de la ciudad de Gijón. Este conjunto de datos es especialmente valioso debido a la alta presencia de reseñas multimodales donde el texto suele ir acompañado de una o varias fotografías capturadas por los comensales. Las fotografías aportadas suelen ser bastante diversas, centrándose no solo en la comida, sino que en ocasiones reflejan aspectos de la decoración o del ambiente del restaurante. Cabe mencionar que no todas las reseñas poseen imágenes; dado que la plataforma no obliga a aportarlas, se han descartado aquellas que poseen exclusivamente texto.

Para asegurar la significancia estadística y dotar a los modelos de suficiente historial para aprender representaciones semánticas, se ha aplicado un filtrado iterativo *2-core*. De este modo, se garantiza que todos los usuarios e ítems presentes en el conjunto final cuenten con un mínimo de 2 interacciones. Tras este preprocesamiento, el conjunto de datos final presenta las características resumidas en la Tabla 1 (columna *Preprocesado*).

Tabla 1: Estadísticas del conjunto de datos.

Nombre del atributo	**Original** (raw)	**Preprocesado** (img & 2-core)	**Train** ($\leq$2019)	**Val** (2020)	**Test** ($\geq$2021)
Usuarios	43,367	2,109	1,673	239	271
Ítems	792	511	450	218	254
Reseñas	82,142	8,281	5,616	523	711
Img. por reseña	0.480	2.616	2.539	2.838	2.968

5.2. Partición Temporal de los Datos

Siguiendo la metodología descrita, se ha huido de las particiones aleatorias tradicionales para evitar el sesgo de información futura (*data leakage*). Se ha establecido un punto de corte temporal (t_{corte}) correspondiente al día 1 de enero de 2021. De esta manera, el conjunto de entrenamiento comprende todas las

reseñas publicadas desde el inicio de la toma de datos hasta t_{corte}, mientras que el conjunto de prueba incluye todas las interacciones ocurridas entre t_{corte} y el fin del historial disponible.

A su vez, y con el fin de ajustar correctamente los valores de los hiperparámetros, se determinó un segundo punto de corte el día 1 de enero de 2020 (t_{val}) que se utiliza para dividir el conjunto de entrenamiento inicial y separar una porción del mismo como conjunto de validación.

Tras realizar estas particiones, se han eliminado de los subconjuntos de validación y prueba aquellas reseñas de ítems o usuarios que no aparecen previamente en el conjunto de entrenamiento, evitando así un escenario de arranque en frío (*Cold-Start*). Esto provoca que la suma total de reseñas de estos subconjuntos no coincida exactamente con las del conjunto preprocesado general. Las estadísticas de los subconjuntos resultantes se pueden visualizar en las últimas tres columnas de la Tabla 1.

5.3. Configuración de los Modelos y Entrenamiento

La implementación técnica de los experimentos se ha llevado a cabo bajo las siguientes especificaciones:

- **Extractor Multimodal**. Se ha empleado el ya mencionado modelo *Gemma 4 E2B* de 5.100 millones de parámetros, en concreto la versión instruida (*Instruction Tuned*). Para la generación de resúmenes, se ha desactivado el razonamiento avanzado del modelo a fin de acelerar el procedimiento de inferencia.
- **Arquitectura de Red Neuronal**. Tanto para la tarea de recomendación como para el análisis de sentimientos, se ha implementado una red neuronal prealimentada (*feed-forward*) compuesta por 4 capas ocultas de 1024, 256, 64 y 16 neuronas, respectivamente. Cada capa utiliza una función de activación ReLU y va seguida de una capa de normalización por lotes (*Batch Normalization*) y una de *Dropout*.
- **Ajuste de Hiperparámetros**. Para cada tarea y método de codificación distinto, se han evaluado 100 combinaciones diferentes utilizando el conjunto de validación. Seleccionada la mejor combinación en cada tarea, se ha reentrenado el modelo final 10 veces para mitigar posibles varianzas y se obtiene el rendimiento sobre el conjunto de test.
- **Entorno de Ejecución**. El entrenamiento de las redes neuronales se han implementado utilizando Tensorflow. Los experimentos se han ejecutado sobre dos GPU NVIDIA RTX 5090.

6. Resultados

En esta sección se presentan los resultados obtenidos para las dos tareas evaluadas. El análisis se centra en contrastar el rendimiento del método propuesto de Fusión Temprana frente al paradigma tradicional de Fusión Tardía.

Adicionalmente, se han incluido como líneas base las representaciones unimodales construidas empleando únicamente el texto y únicamente las imágenes, con el fin de evaluar la aportación individual de cada modalidad.

Los resultados globales, expresados en términos de media y desviación estándar tras 10 ejecuciones independientes, se resumen en la Tabla 2. Para dotar de rigor al análisis, las diferencias observadas entre las distintas codificaciones han sido sometidas a un test t de Welch de dos colas, confirmándose que las variaciones de rendimiento discutidas a continuación son estadísticamente significativas ($p < 0.05$).

Tabla 2: Rendimiento de los métodos de codificación en ambas tareas.

Codificación	**Tarea 1 (NDCG@10)**		**Tarea 2 (F1-Macro)**	
	Media	Desv. Est.	Media	Desv. Est.
Solo texto	0.1907	0.0255	0.7777	0.0270
Solo imágenes	0.2019	0.0121	0.5427	0.0402
Fusión Tardía	**0.2068**	0.0110	0.7630	0.0227
Fusión Temprana	0.1563	0.0175	**0.8123**	0.0595

En la tarea de recomendación de ítems, el paradigma de Fusión Tardía demostró ser significativamente superior, alcanzando un NDCG@10 de 0.2068. Este enfoque no solo superó a la Fusión Temprana por un amplio margen, sino que también mejoró ligeramente las representaciones basadas de forma aislada en texto e imagen. Este comportamiento sugiere que, en el proceso de construir perfiles históricos de usuarios e ítems mediante promedios, la separación estricta de las representaciones vectoriales originales preserva señales que resultan fundamentales para el filtrado colaborativo.

Por el contrario, el escenario se invierte de manera notable al evaluar la calidad semántica pura mediante la tarea de análisis de sentimientos a nivel de reseña individual. En este contexto, la Fusión Temprana obtiene el mejor resultado global (F1-Macro de 0.8123), superando de forma estadísticamente significativa tanto a la Fusión Tardía como a los modelos unimodales. Este sobresaliente rendimiento se explica, probablemente, porque la instrucción (Código 1.1) exige al VLM inferir el sentimiento explícitamente, convirtiendo al clasificador neuronal posterior en un decodificador de esta precisa decisión *zero-shot*. Lejos de ser una limitación, este hallazgo valida la hipótesis central de la investigación: utilizar el razonamiento de un LLM para la fusión temprana permite permite capturar y extraer información relevante adicional del texto y la imagen.

Los resultados revelan un equilibrio entre ambos paradigmas: mientras que la Fusión Temprana produce una representación más rica y alineada con la intención humana subyacente (reflejada en la clasificación de sentimientos), la Fusión

Tardía mantiene una geometría vectorial más robusta y eficaz para las dinámicas de agregación temporal exigidas por los sistemas de recomendación.

Más allá del rendimiento, el análisis de eficiencia revela una brecha crítica en la latencia de extracción por reseña: mientras que el procesamiento unimodal requiere apenas 0.0404 s para el texto y 0.1340 s para las imágenes, la Fusión Temprana con el VLM eleva el tiempo medio a ≈ 21 s debido a la generación de texto. Este factor evidencia el compromiso de usar VLMs como extractores de características: aportan una semántica superior y alineada (clave para el análisis de sentimientos), pero con un coste computacional y una latencia que comprometen su viabilidad en sistemas en tiempo real.

7. Conclusiones y trabajo futuro

Este trabajo ha evaluado la extracción de características multimodales en reseñas gastronómicas comparando la Fusión Tardía clásica frente a una Fusión Temprana delegada a un Modelo de Visión-Lenguaje. Los resultados revelan un claro compromiso: la Fusión Tardía resulta más robusta geométrica y computacionalmente para la agregación de perfiles en tareas de recomendación (NDCG). En contraste, la Fusión Temprana captura una semántica más profunda, superando al resto de enfoques en el análisis de sentimientos al resolver discrepancias entre texto e imagen, aunque su elevada latencia inferencial limita su viabilidad actual en entornos de producción en tiempo real.

Para afianzar estos hallazgos y ampliar la aplicabilidad de la investigación, el trabajo futuro se desarrollará en varias direcciones. En primer lugar, se extenderá la evaluación empírica a múltiples conjuntos de datos de diversa naturaleza, incorporando un abanico más amplio de métodos de codificación multimodal. Asimismo, el rendimiento de estas representaciones se analizará en una mayor variedad de tareas operativas, abarcando tanto enfoques adicionales de recomendación (como problemas de regresión o *ranking* avanzado) como tareas de aprendizaje no supervisado (por ejemplo, *clustering* de usuarios o ítems). Es plausible que en estos escenarios alternativos puedan destacar métodos de codificación que no resultaron óptimos bajo las métricas de filtrado de este estudio. Finalmente, se explorarán arquitecturas y estrategias algorítmicas más sofisticadas para la agregación y creación de los perfiles dinámicos de usuario e ítem, buscando superar las limitaciones de la pérdida de información inherente a los promedios aritméticos.

Agradecimientos La Agencia de Ciencia, Competitividad Empresarial e Innovación Asturiana (SEKUENS) ha colaborado en la financiación de este trabajo a través del proyecto IDE/2024/000696.

Conflictos de intereses Los autores declaran no tener conflictos de intereses en este artículo.

Referencias

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019), https://arxiv.org/abs/1810.04805
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Google DeepMind: Gemma 4: Our most capable open models to date. https://deepmind.google/models/gemma/gemma-4/ (2026), accessed: May 21, 2026
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), https://arxiv.org/abs/1512.03385
5. He, R., McAuley, J.: Vbpr: visual bayesian personalized ranking from implicit feedback. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. p. 144–150. AAAI'16, AAAI Press (2016), https://dl.acm.org/doi/10.5555/3015812.3015834
6. Hu, X., Yu, W., Wu, Y., Chen, Y.: Multi-modal recommendation algorithm fusing visual and textual features. PLOS ONE **18**(6), 1–18 (06 2023), https://doi.org/10.1371/journal.pone.0287927
7. Liu, J., Li, T., Yu, M., Yang, S., Tang, Z., Yang, Z.: A multi-factor collaborative prediction for review-based recommendation. In: Proceedings of the Nineteenth ACM Conference on Recommender Systems. p. 12–20. RecSys '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3705328.3748062
8. Wang, Z., Höpken, W., Jannach, D.: Beyond visit trajectories: Enhancing poi recommendation via llm-augmented text and image representations. In: Proceedings of the Nineteenth ACM Conference on Recommender Systems. p. 521–526. RecSys '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3705328.3748014
9. Wang, Z., Liu, B., Huang, W., Hao, T., Zhou, H., Guo, Y.: Leveraging multimodal large language model for multimodal sequential recommendation. Scientific Reports **15**(1), 28960 (08 2025). https://doi.org/10.1038/s41598-025-14251-1
10. Wei, Y., Wang, X., Nie, L., He, X., Hong, R., Chua, T.S.: Mmgcn: Multimodal graph convolution network for personalized recommendation of micro-video. In: Proceedings of the 27th ACM International Conference on Multimedia. p. 1437–1445. MM '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3343031.3351034
11. Yilma, B.A., Leiva, L.A.: Together yet apart: Multimodal representation learning for personalised visual art recommendation. In: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization. p. 204–214. UMAP '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3565472.3592964
12. Zheng, L., Noroozi, V., Yu, P.S.: Joint deep modeling of users and items using reviews for recommendation. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. p. 425–434. WSDM '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3018661.3018665
13. Zhou, H., Zhou, X., Zeng, Z., Zhang, L., Shen, Z.: A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions (2023), https://arxiv.org/abs/2302.04473