Explainable Machine Learning for Content Personalization in Social Networks

Miguel Escarda-Fernández, Carlos Eiras-Franco, Brais Cancela, and Amparo Alonso-Betanzos

Departamento de Ciencias de la Computación, CITIC,
Universidade da Coruña, A Coruña, España
miguel.escarda@udc.es, carlos.eiras.franco@udc.es,
brais.cancela@udc.es, amparo.alonso.betanzos@udc.es

Resumen Aprendizaje automático explicable para la personalización de contenidos en redes sociales, entendidas en sentido amplio como grafos de usuarios, contenidos e interacciones. La investigación parte de un problema común en plataformas sociales, recomendadores y entornos digitales con contenido generado por usuarios: no basta con predecir qué relación usuario–contenido es probable o útil; también es necesario decidir cómo presentarla y explicarla de forma personalizada, eficiente y verificable. Partimos del trabajo PTER [7], que formula la explicación textual de nuevas díadas usuario–ítem como una tarea de ranking de reseñas personalizadas. Sobre esta base, la tesis ya cuenta con dos resultados propios: un artículo publicado sobre rendimiento y sostenibilidad de derivados de BERT en datos diádicos, y un trabajo sobre aprendizaje contrastivo para ranking de explicaciones. El objetivo doctoral es integrar estas líneas en un marco general para modelar interacciones usuario–contenido, seleccionar representaciones textuales eficientes, ordenar explicaciones post-hoc y extender la personalización desde díadas aisladas hacia estructuras de contenido más complejas, tomando como escenario final la recomendación explicable de itinerarios formativos. La contribución original no se restringe a una plataforma o dataset concreto: propone una metodología transferible para personalizar y explicar contenidos en redes heterogéneas, evaluando simultáneamente utilidad, calidad explicativa, robustez y coste computacional.

Keywords: Personalización de contenidos · redes sociales · datos diádicos · sistemas de recomendación · explicabilidad · aprendizaje contrastivo.

1. Datos

Doctorando	Miguel Escarda Fernández
Contacto	`miguel.escarda@udc.es`
Propuesta de título	*Explainable Machine Learning for Content Personalization in Social Networks*
Directora de tesis	Amparo Alonso Betanzos
Tutores de tesis	Carlos Eiras Franco; Brais Cancela Barizo
Departamento y universidad	Departamento de Ciencias de la Computación y Tecnologías de la Información, Universidade da Coruña
Fecha de inicio	16/10/2022
Régimen de dedicación	Tiempo parcial
Fecha prevista de defensa	Septiembre de 2027 (estimada)

La propuesta se sitúa en aprendizaje automático, procesamiento del lenguaje natural, sistemas de recomendación e inteligencia artificial explicable y responsable. En esta tesis, el término *social networks* se usa en sentido amplio: redes de usuarios, contenidos e interacciones, no solo plataformas de blogs o amistad explícita. Bajo esta definición, una reseña, una valoración o un comentario no son únicamente datos auxiliares; son evidencias de una relación que puede ser modelada, personalizada y explicada.

2. Planteamiento de la tesis

Los sistemas de recomendación han pasado a mediar una parte importante de la relación entre usuarios y contenidos, y la literatura sobre recomendación explicable ha mostrado que la transparencia puede afectar a la confianza y satisfacción del usuario [12]. Sin embargo, en muchos dominios la pregunta práctica no es solo *qué* ítem debe recomendarse, sino *cómo* debe presentarse y explicarse. Esta cuestión fue formulada de forma explícita por *Explain and Conquer*, donde PTER selecciona reseñas textuales que pueden actuar como explicación personalizada de una relación usuario–restaurante en TripAdvisor [7]. La tesis adopta esa idea como semilla conceptual y la extiende en dos direcciones: la evaluación eficiente de modelos de lenguaje para datos diádicos y el ranking contrastivo de explicaciones.

La hipótesis general es que las explicaciones en recomendación pueden estudiarse como objetos de ranking, una formulación alineada con trabajos recientes que relacionan explicación y recomendación mediante tareas de ordenación [6]. Una explicación no es útil por ser generada de forma aislada, sino porque resulta más adecuada que otras explicaciones posibles para una relación usuario–contenido concreta. De esta hipótesis se derivan tres hipótesis de trabajo:

1. **H1.** El texto asociado a las interacciones de una red usuario–contenido contiene señales suficientes para personalizar explicaciones de relaciones diádicas, incluso cuando los datos son dispersos.
2. **H2.** El ranking de explicaciones puede aprenderse de forma efectiva mediante representaciones de usuarios, ítems y explicaciones, especialmente cuando el entrenamiento compara explícitamente ejemplos positivos y negativos.
3. **H3.** La utilidad de un modelo explicable no debe evaluarse solo por métricas predictivas o de ranking, sino también por su coste de entrenamiento, inferencia y despliegue.

El objetivo principal de la tesis es desarrollar y evaluar métodos de aprendizaje automático explicable para personalización de contenido en redes usuario–ítem, considerando la explicación como una tarea central de ranking y no como un componente posterior meramente decorativo. Este objetivo se concreta en los siguientes subobjetivos:

- formalizar la explicación de relaciones usuario–contenido como un problema de ranking sobre datos diádicos textuales;
- analizar cómo cambia el compromiso entre rendimiento y coste al sustituir BERT por derivados más ligeros o alternativos en esta familia de tareas;
- diseñar un modelo contrastivo para ordenar explicaciones post-hoc en escenarios de recomendación;
- estudiar la transferencia de estas ideas a recomendación explicable de itinerarios formativos, donde los contenidos se relacionan mediante temas, habilidades, módulos o progresión pedagógica;
- integrar los resultados en una memoria de tesis que discuta precisión, explicabilidad, sostenibilidad y limitaciones de uso responsable.

3. Investigación previa y resultados obtenidos

3.1. Semilla: explicación textual de datos diádicos

PTER parte de una representación diádica de redes sociales donde usuarios e ítems se conectan mediante interacciones textuales. En su caso de estudio, los usuarios escriben reseñas sobre restaurantes, y el objetivo es seleccionar, entre las reseñas disponibles para un restaurante, aquellas que mejor encajan con el contexto de un usuario concreto [7]. La contribución importante para esta tesis no es el dominio turístico en sí, sino la formulación: explicar una relación recomendada equivale a recuperar y ordenar textos candidatos que hagan comprensible por qué esa relación es plausible.

Esta idea fija el hilo conductor de la tesis. En lugar de tratar la explicación como una frase generada al final del proceso, se la considera una entidad que puede aprenderse, compararse y evaluarse. A partir de ahí, la investigación se pregunta qué modelos de lenguaje son adecuados para representar esas interacciones y qué funciones de aprendizaje permiten ordenar explicaciones de forma robusta.

3.2. Resultado 1: rendimiento y sostenibilidad de BERT en datos diádicos

El primer resultado principal de la tesis es el artículo *Performance and sustainability of BERT derivatives in dyadic data*, publicado en *Expert Systems with Applications* [1]. El trabajo toma la tarea de explicación diádica inspirada en PTER y compara BERT, DistilBERT, ALBERT, RoBERTa y ELECTRA en un escenario de clasificación multi-etiqueta sobre reseñas de TripAdvisor.

La elección del encoder textual afecta tanto a la calidad como al coste, un aspecto coherente con la creciente preocupación por la sostenibilidad de la IA [9,10]. DistilBERT proporciona los menores tiempos de entrenamiento, consumo eléctrico y emisiones estimadas, mientras que RoBERTa obtiene los mejores resultados en balanced accuracy y recall en los experimentos reportados. La conclusión para la tesis es metodológica: en recomendación explicable, elegir el modelo más grande o el de mayor puntuación en un benchmark general no es necesariamente la mejor decisión. La evaluación debe registrar también latencia, memoria, energía y emisiones cuando el modelo se aplica a muchos pares usuario–ítem.

Este trabajo cubre la relación entre representaciones textuales, datos diádicos y sostenibilidad computacional. También establece una restricción para los trabajos posteriores: cualquier mejora de explicabilidad debe ser compatible con un despliegue razonable.

3.3. Resultado 2: aprendizaje contrastivo para ranking de explicaciones

El segundo resultado es *Contrastive Learning for Explanation Ranking* [2]. Este trabajo aborda de forma directa la pregunta de cómo ordenar explicaciones para una pareja usuario–ítem ya recomendada. CLER aprende embeddings de usuarios, ítems y explicaciones; combina la representación usuario–ítem y optimiza una función de pérdida contrastiva binaria balanceada, NT-BXent, diseñada para evitar que el gran número de ejemplos negativos domine el entrenamiento. La motivación general procede del aprendizaje contrastivo supervisado, donde la representación se aprende comparando instancias relacionadas y no relacionadas [4].

La evaluación sigue el benchmark EXTRA, con datasets de Amazon, TripAdvisor y Yelp. En estos experimentos, CLER mejora a los baselines comparados en métricas tradicionales de ranking, incluyendo NDCG@10, Precision@10, Recall@10 y F1@10. Además, el trabajo complementa estas métricas con indicadores como popularidad media, novedad y diversidad de las explicaciones seleccionadas. Su limitación principal también queda clara: al ser un ranker basado en identificadores y relaciones, puede acertar patrones del benchmark sin garantizar siempre alineamiento semántico o de polaridad en la explicación textual. Esta limitación orienta el cierre de la tesis.

3.4. Línea en curso: recomendación explicable de itinerarios formativos

El cierre de la tesis toma la recomendación explicable de itinerarios formativos como escenario de transferencia del marco desarrollado. La idea es pasar de la explicación de una relación usuario–ítem aislada a la explicación de una secuencia de contenidos que debe ser útil, coherente y justificable para un aprendiz concreto. En este contexto, el objeto recomendado ya no es solo un ítem, sino un posible camino formado por cursos, módulos, problemas, temas o habilidades.

La motivación procede de la literatura reciente sobre *learning path recommendation*, donde la personalización educativa se formula como la selección de recursos de aprendizaje a partir del historial, objetivos o estado del estudiante [8]. A diferencia de la recomendación clásica de ítems, un itinerario formativo debe considerar no solo preferencia o similitud, sino también progresión, cobertura de habilidades, dificultad, redundancia y posibles relaciones de prerrequisito.

Desde la perspectiva de esta tesis, la pregunta central es si la explicación como ranking de evidencias puede extenderse a este tipo de estructuras. Una explicación puede justificar no solo por qué un contenido es relevante, sino también por qué aparece en una posición determinada del camino: porque cubre una habilidad previa, porque conecta con un módulo posterior, porque evita una redundancia o porque existe evidencia textual que apoya su utilidad para un perfil concreto. Esta visión es compatible con enfoques basados en grafos de conocimiento, donde recursos, conceptos, objetivos y relaciones pedagógicas se modelan explícitamente para producir caminos más interpretables [5].

4. Metodología y plan de trabajo

El estudio final se apoyará en dos fuentes de datos complementarias, sin asumir que ambas definan exactamente la misma tarea. *FoundationalASSIST* se utilizará principalmente para estudiar señales de estado de aprendizaje y relaciones problema–habilidad, ya que contiene interacciones educativas con problemas, respuestas de estudiantes, corrección, uso de pistas y habilidades asociadas [11]. El dataset de Coursera permitirá estudiar recomendación de contenidos de mayor granularidad, a partir de cursos, reseñas, valoraciones de 1 a 5 y metadatos extraídos de los cursos, como temas, habilidades y módulos.

La formulación concreta se fijará tras analizar la densidad y calidad de los datos: por una parte, si existen trayectorias de usuario suficientes, se estudiará recomendación secuencial o predicción del siguiente curso; en caso contrario, el problema se formulará como ranking de itinerarios orientados a objetivos, donde dado un perfil, unos intereses o un conjunto de habilidades deseadas, el sistema ordenará cursos o módulos candidatos y generará una explicación basada en habilidades, temas, módulos o evidencia textual. Esta será la formulación mínima común del estudio.

Se compararán baselines simples, como popularidad, valoración media, similitud textual y cobertura de habilidades, con modelos más expresivos basados

en representaciones textuales, grafos de temas y habilidades, re-ranking eficiente o aprendizaje contrastivo. Las valoraciones y reseñas se tratarán como señales imperfectas de satisfacción o utilidad percibida, no como medidas directas de aprendizaje real [3].

La evaluación considerará tres dimensiones. Primero, calidad del ranking, mediante métricas como NDCG@k, Precision@k o Recall@k. La segunda dimensión consistirá en la calidad del itinerario, atendiendo a la coherencia temática, cobertura de habilidades, redundancia y progresión razonable. En tercer lugar, calidad explicativa y coste, incluyendo alineamiento semántico de las explicaciones, sesgo hacia contenidos populares, diversidad, latencia y coste computacional. El plan inmediato consiste en auditar los datos, definir la tarea final reproducible, comparar baselines con modelos más expresivos e integrar los resultados en la memoria doctoral.

Como hitos de la tesis, se distinguen resultados ya alcanzados y tareas pendientes. El primer hito, ya completado, es la evaluación de modelos de lenguaje eficientes para datos diádicos. El segundo, también completado, es el desarrollo de un modelo contrastivo para ranking de explicaciones. El tercer hito, actualmente en curso, es la transferencia del marco a recomendación explicable de itinerarios formativos. El cuarto hito será la integración de estos resultados en la memoria doctoral, incluyendo una discusión crítica sobre calidad explicativa, sesgos, sostenibilidad y límites de uso responsable.

5. Relevancia científica y transferencia

La relevancia científica de la tesis se encuentra en conectar la recomendación personalizada, explicabilidad y sostenibilidad computacional. No se trata la explicación como una frase generada al final del proceso, sino como una evidencia que puede representarse, compararse y ordenarse. Esta idea se ha estudiado ya en relaciones usuario–ítem y se extenderá ahora a contenidos educativos con estructura interna.

La recomendación de itinerarios formativos es un caso especialmente adecuado para esta extensión. En educación, no basta con saber qué contenido puede gustar a un usuario: también importa si el contenido cubre habilidades relevantes, si encaja con conocimientos previos, si evita redundancias y si forma parte de una progresión razonable. Esto permite estudiar explicaciones más ricas, por ejemplo basadas en habilidades cubiertas, módulos relacionados, dificultad esperada o evidencia textual procedente de reseñas.

La transferencia potencial incluye plataformas MOOC, sistemas de apoyo al aprendizaje, catálogos de formación continua y entornos donde los usuarios necesitan entender por qué se les recomienda una secuencia de contenidos. Al mismo tiempo, la tesis mantiene una perspectiva crítica: las valoraciones y reseñas pueden reflejar satisfacción o popularidad más que aprendizaje real, por lo que la evaluación deberá distinguir entre utilidad percibida, calidad del itinerario y adecuación explicativa.

6. Conclusión

La tesis *Explainable Machine Learning for Content Personalization in Social Networks* propone estudiar la personalización de contenido en redes usuario–ítem desde una perspectiva explicable y sostenible. Su hilo conductor procede de la formulación de PTER: explicar una relación personalizada mediante textos candidatos. Los dos resultados ya obtenidos cubren, respectivamente, la selección sostenible de encoders para datos diádicos y el aprendizaje contrastivo para ranking de explicaciones. El cierre de la tesis consistirá en integrar ambas líneas y completar un último estudio de transferencia a la recomendación explicable de itinerarios formativos. La decisión metodológica se apoyará en el análisis experimental de los datos, pero el objetivo se mantiene fijo: evaluar si el ranking de evidencias explicativas puede extenderse desde díadas usuario–ítem hacia secuencias educativas con estructura interna.

Referencias

1. Escarda, M., Eiras-Franco, C., Cancela, B., Guijarro-Berdínas, B., Alonso-Betanzos, A.: Performance and sustainability of BERT derivatives in dyadic data. Expert Systems with Applications **262**, 125647 (2025). https://doi.org/10.1016/j.eswa.2024.125647
2. Escarda-Fernández, M., Cancela, B., Eiras-Franco, C., Guijarro-Berdiñas, B., Alonso-Betanzos, A.: Contrastive learning for explanation ranking. Accepted for the ECML PKDD 2026 Journal Track (2026)
3. Gomez, M.J., Calderon, M., Sanchez, V., Garcia Clemente, F.J., Ruiperez-Valiente, J.A.: Large scale analysis of open mooc reviews to support learners' course selection. arXiv preprint arXiv:2201.06967 (2022). https://doi.org/10.48550/arXiv.2201.06967
4. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 18661–18673 (2020)
5. Li, A., Li, Y., Gao, X.: Personalized learning path recommendation based on knowledge graphs: A survey. Electronics **15**(1), 238 (2026). https://doi.org/10.3390/electronics15010238
6. Li, L., Zhang, Y., Chen, L.: On the relationship between explanation and recommendation: Learning to rank explanations for improved performance. ACM Transactions on Intelligent Systems and Technology **14**(2) (2023). https://doi.org/10.1145/3569423
7. López-Riobóo Botana, I., Bolón-Canedo, V., Guijarro-Berdínas, B., Alonso-Betanzos, A.: Explain and conquer: Personalised text-based reviews to achieve transparency (2022), `https://arxiv.org/abs/2205.01759`
8. Nasrin, A., Qian, L., Obiomon, P., Dong, X.: Enhancing learning path recommendation via multi-task learning. arXiv preprint arXiv:2507.05295 (2025)
9. Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for modern deep learning research. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13693–13696 (2020). https://doi.org/10.1609/aaai.v34i09.7123

10. Van Wynsberghe, A.: Sustainable AI: AI for sustainability and the sustainability of AI. AI and Ethics **1**(3), 213–218 (2021). https://doi.org/10.1007/s43681-021-00043-6
11. Worden, E., Heffernan, C., Heffernan, N., Sonkar, S.: Foundationalassist: An educational dataset for foundational knowledge tracing and pedagogical grounding of llms. arXiv preprint arXiv:2602.00070 (2026)
12. Zhang, Y., Chen, X.: Explainable recommendation: A survey and new perspectives. Found. Trends Inf. Retr. **14**(1), 1–101 (Mar 2020). https://doi.org/10.1561/1500000066, https://doi.org/10.1561/1500000066