Recomendación de noticias en el dataset MIND

Manuel Cifuentes, Diego Fernández, Juan Hernandez

Departamento de Ciencias de la Computación, Pontificia Universidad Católica de Chile

Motivación

Los sistemas de recomendación de noticias enfrentan desafíos particulares: los intereses de los usuarios cambian rápidamente, las noticias se vuelven obsoletas en poco tiempo, el volumen de datos puede ser muy alto y las noticias nuevas son complicadas de recomendar [1]. El dataset MIND (Microsoft News Dataset) [2] ofrece un escenario realista para abordar estos desafíos, lo que lo convierte en una referencia clave en la investigación sobre recomendación de noticias.

Problema

El tamaño y la complejidad de MIND suponen una gran dificultad para realizar recomendaciones efectivas con bajos recursos computacionales. Modelos como NRMS [3] y DKN [4] muestran buenos resultados, pero su costo computacional limita su aplicabilidad en entornos con recursos restringidos. Para abordar esto, se ha propuesto Fastformer [5], el estado del arte actual para este dataset, aunque todavía presenta un alto costo computacional. Por tanto, se buscará realizar modificaciones que permitan aplicar modelos similares (basados en *embeddings*) utilizando la menor cantidad de recursos computacionales posibles.

Datos

El dataset **MIND** es una colección de interacciones usuario-noticia publicada por Microsoft. Cada línea del archivo representa una *impresión*, e incluye información del usuario, el momento de la interacción, su historial de clics y un conjunto de noticias candidatas, marcadas con `1` (clic) o `0` (no clic).

Campo	Valor
`Impression ID`	10947
`User ID`	U1224
`Timestamp`	11/11/2019 10:02:31 AM
`History`	N98312, N23987, N53101
`Impressions`	N72635-1, N43882-0, N87312-0

Table 1:Ejemplo de una fila del archivo `behaviors.tsv`.

Metodología

Se implementaron dos arquitecturas basadas en la literatura: el modelo NRMS (Neural News Recommendation with Multi-Head Self-Attention) y una variante denominada **FastformerNRMS**, que reemplaza la atención multi-cabeza por bloques de atención jerárquica aditiva inspirados en *Fastformer*, con el objetivo de reducir la complejidad computacional.

Ambos modelos siguen una estructura encoder-encoder: el **codificador de noticias** procesa los títulos mediante atención sobre embeddings preentrenados GloVe [6], mientras que el **codificador de usuario** resume el historial aplicando atención sobre las representaciones de las noticias leídas.

Como la representación de los datos se aleja de la tradicional matriz user-item, permitiendo modelar mejor la dinámica de navegación real, se tuvo que ajustar los típicos modelos a este formato. Esto motivó el uso de la función de pérdida `CrossEntropyLoss`, que permite modelar la elección considerando el contexto completo de candidatos dentro de cada impresión.

Además, se implementó una función de inferencia personalizada que permite generar recomendaciones específicas por usuario, evaluando cuál noticia es más probable que sea clickeada entre las candidatas.

Todos los experimentos se realizaron en Google Colab (Código disponible en [7]) con GPU T4 y recursos estándar, demostrando la viabilidad del enfoque sin requerir infraestructura avanzada.

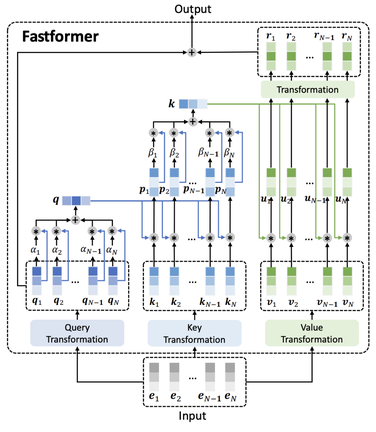
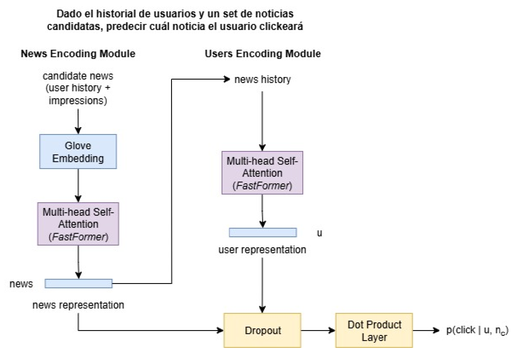


Figure 1:Arquitectura Fastformer

Arquitectura del modelo



Estructura encoder-encoder del modelo FastformerNRMS

Resultados

Table 2:Métricas de evaluación por método

Método	AUC	MRR	NDCG@5	NDCG@10
Random	-	-	-	0.0002
MostPopular	-	-	-	0.0000
MostPopular filtered	-	-	-	0.0031
NRMS (10%)$_a$	-	-	0.2256	-
FastFormerNRMS (10%)$_b$	-	-	0.2339	-
NRMS (BCE loss)$_c$	0.4859	0.2262	0.1983	0.2596
LSTUR	0.5374	0.2397	0.2160	0.2733
FastFormerNRMS (BCE loss)$_d$	0.5767	0.2756	0.2506	0.3105
FastFormerNRMS$_e$	0.6498	0.3363	0.3103	0.3712
FastFormerNRMS$_f$	0.6642	0.3506	0.3269	0.3882
FastFormerNRMS$_g$	**0.6681**	**0.3555**	**0.3304**	**0.3911**

[a] NRMS entrenado con 10% del dataset `small`. [b] FastFormerNRMS entrenado con 10% del dataset `small`. [c] NRMS entrenado con todo el dataset `small`. [d] FastFormerNRMS entrenado con todo el dataset `small`. [e] FastFormerNRMS entrenado con todo el dataset `small` y con el nuevo mecanismo de batches con embeddings pre-entrenados GloVe. [f] FastFormerNRMS entrenado con todo el dataset `large` y con el nuevo mecanismo de batches con embeddings pre-entrenados Word2Vec. [g] FastFormerNRMS entrenado con todo el dataset `large` y con el nuevo mecanismo de batches con embeddings pre-entrenados GloVe.

Resultado importante

Se logró correr el código por **2** épocas con una GPU T4, con uso en promedio de **18 GB** de RAM aumentada debido a la lectura de datos, produciendo resultados satisfactorios y probando que se podía ejecutar el modelo de noticias con menos recursos computacionales.

Conclusión

En este trabajo se abordó el problema de la recomendación de noticias en entornos con recursos computacionales limitados, proponiendo una variante eficiente del modelo NRMS, denominada FastformerNRMS. Esta arquitectura incorpora atención jerárquica aditiva inspirada en Fastformer y simplificaciones estructurales que permiten mantener un buen rendimiento con menor costo computacional. Los resultados muestran que FastformerNRMS supera a los otros modelos entrenados en métricas como AUC, MRR y NDCG.

Además, se verificó la viabilidad de entrenar este modelo en Google Colab con únicamente el aumento de RAM, lo que refuerza su aplicabilidad en contextos reales con recursos limitados. Como trabajo futuro, se podría explorar la incorporación de información semántica adicional como los *abstracts* y categorías, y la evaluación de nuevas variantes del mecanismo de atención.

Referencias

[1] Shaina Raza and Chen Ding. News recommender system: A review of recent progress, challenges, and opportunities. *arXiv preprint arXiv:2009.04964*, 2020.

[2] Microsoft Research. MIND: Microsoft News Dataset. https://msnews.github.io/. Accessed: 2025-05-08.

[3] Chuhan Wu, Fangzhao Wu, Tao Qi, Yongfeng Huang, and Xing Xie. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3026–3035. Association for Computational Linguistics, 2019.

[4] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. Dkn: Deep knowledge-aware network for news recommendation. In *Proceedings of the 27th International Conference on World Wide Web (WWW)*, pages 1835–1844. International World Wide Web Conferences Steering Committee, 2018.

[5] Shuo Wu, Zuohui Liu, Wenqing Xiao, Zhoujun Lin, Jianxin Wang, and Xu Sun. Fastformer: Additive attention can be all you need. *arXiv preprint arXiv:2108.09084*, 2021.

[6] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014.

[7] Manuel Cifuentes, Diego Fernández, and Juan Manuel Hernández. Repositorio del proyecto de sistemas de recomendación. https://github.com/JuanHernandez-uc/proyecto-recsys, 2025. Último acceso: 6 de junio de 2025.

[8] Chuhan Wu, Fangzhao Wu, Tao Qi, Yongfeng Huang, and Xing Xie. Neural news recommendation with attentive multi-view learning. In *Proceedings of the 27th ACM international conference on information and knowledge management (CIKM)*, pages 157–166. ACM, 2019.

[9] Andreea Iana, Goran Glavaš, and Heiko Paulheim. Newsreclib: A pytorch-lightning library for neural news recommendation, 2023.

[10] Microsoft Research. Mind dataset evaluation script. https://github.com/msnews/MIND/blob/master/evaluate.py, 2020. Último acceso: 6 de junio de 2025.

[11] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. Lstur: Neural news recommendation with long- and short-term user representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 336–345. Association for Computational Linguistics, 2019.