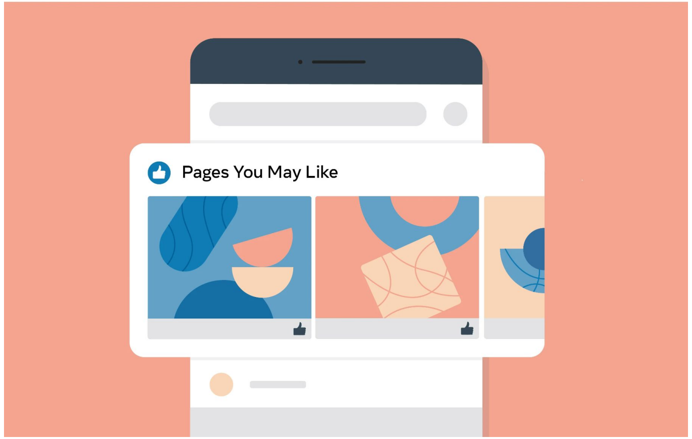
NYU | Center for Data Science

News Recommendation System with BERT

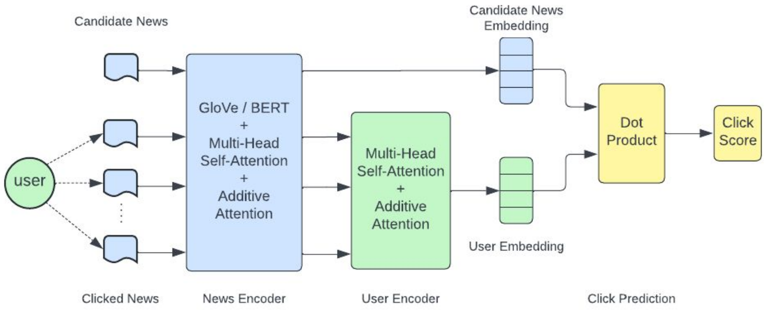
Ji Qi	Qihang Tang	Zhuojian Wei	Vivian Yan	Carla Zhao
jq2316@nyu.edu	qt2087@nyu.edu	zw2219@nyu.edu	qy620@nyu.edu	yz5996@nyu.edu

Background



- Previous news recommendation approaches include GRU, CNN, etc.
    - Challenge: learning accurate news and user representations.
- The **NRMS model** (Wu et al., 2019) uses multi-head self-attentions to encode news from news title and users from browsing history.
- The invention of **LLMs** offers the potential to deeply understand textual nuances and user contexts with a better initial point, with a possibility to enhance the recommendation quality.
    - Wu et al. (2021) replaced NRMS' multi-head self-attention with pre-trained **BERT** and fine-tune them with news recommendation task, and achieved better offline results.

NRMS Architecture



Additive Attention: learn more informative news and user representations

Apply **attention mechanism** to capture complex contextual and behavioral interactions.

- **News Encoder**: captures interactions between different words in news titles.
- **User Encoder**: captures the relatedness between news articles browsed by the same user.

Click Prediction: calculates the relevance of the candidates news to users.

Hypothesis

NRMS-BERT achieves higher accuracies in news recommendation than NRMS.

- **Hypothesis 1:** NRMS-BERT can better incorporate features (eg. category, popularity) in news representations than NRMS.

- **Hypothesis 2:** Features captured by model layers in NRMS-BERT can contribute to the effectiveness of news recommendation.

Probing Approaches

- Take the embeddings as feature inputs to classify news categories via a **logistic regression**.

- Use **t-SNE** to reduce news embeddings of the last layer to a two-dimensional space and **visualize** them with respect to specific features, such as news categories.

Experiments

- **Experiment 0**:

  We implement the methods (NRMS & NRMS-BERT) introduced by Wu et al. (2021) by adapting NRMS model to MIND data and empowering it with pre-trained language models.

- **Experiment 1**:

  We aim to employ linear probing techniques to explore whether specific features—news category, popularity, event time—are encoded in embeddings across multi-head attention layers.

- **Experiment 2**:

  We further explore the relationship between our target features and recommendation: “denoise” features from embeddings.

Experiment 0

Model	AUC	MRR	nDCG@5	nDCG@10
NRMS-baseline	0.6655	**0.3210**	0.3474	0.4044
NRMS-BERT	**0.6657**	0.3204	**0.3477**	**0.4055**

Table 1: Results of NRMS-baseline and NRMS-BERT on Test Set

NRMS-BERT generally performs better than NRMS-baseline on most metrics.

- Even though the improvement is not that significant, NRMS-BERT consumes less data to achieve the similar performance as NRMS-baseline. (converges more quickly)

Experiment 1 (News Topic)

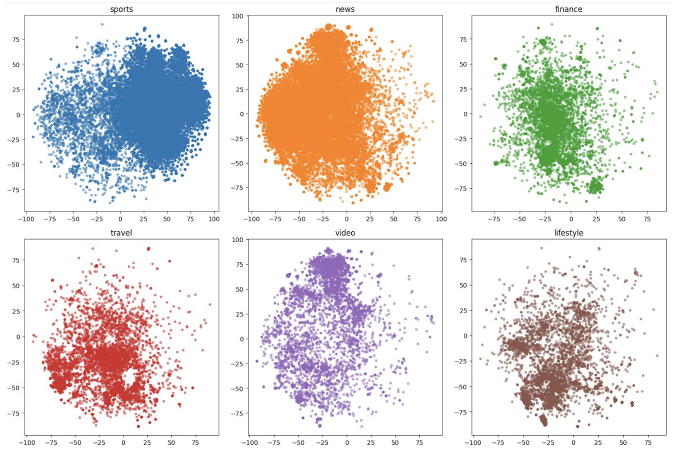
Category	**F1-Score (NRMS)**	**F1-Score (NRMS-BERT)**
Finance	0.44	0.58
Lifestype	0.46	0.56
News	0.73	0.78
Sports	0.87	0.93
Travel	0.37	0.49
Video	0.11	0.28
Overall Accuracy	0.72	0.78

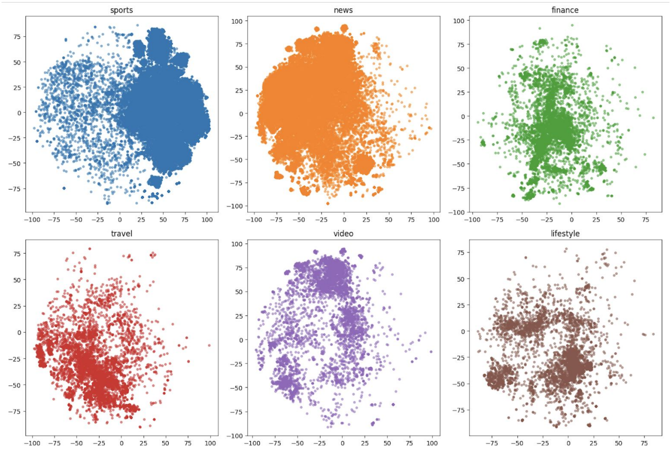
Table 2: Results of Two Embeddings v.s. Categories Using Logistic Regression

Take the embeddings to classify news topics via a logistic regression.

- The results of linear probing further shows the embeddings from NRMS-BERT outperforms the ones from NRMS-baseline in all categories, suggesting that NRMS-BERT is better at capturing category-related information.

Experiment 1 (News Topic)

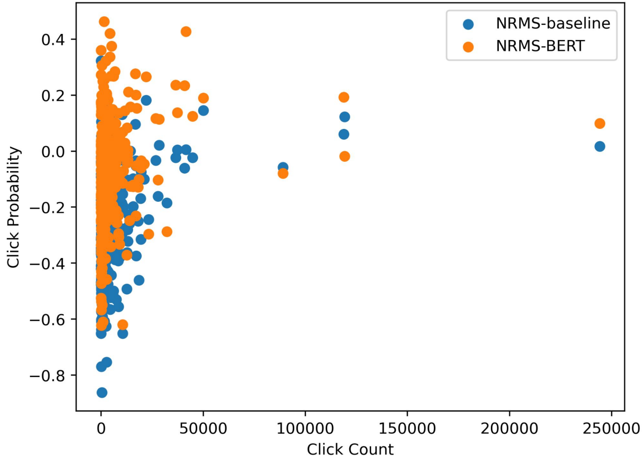




T-SNE visualizations of NRMS-baseline (left) and NRMS-BERT (right) embeddings by news topics:

- NRMS-BERT embeddings are better at differentiating the content inherent to each category as shown in more distinct and compact shapes of the six categories.

Experiment 1 (Popularity)



	Correlation	AUC (New User)
NRMS-baseline	0.23376	0.5843
NRMS-BERT	0.177776	0.5723

Scatter Plot (left) and Correlation (right) of Click Count v.s. Click Probability:

- Both models display no clear correlation between popularity and corresponding prediction of click probability, but both models tend to give relatively higher probability for popular news

NYU Center for Data Science

Experiment 2 (TBD)

- **Hypothesis:** News topics captured by model layers in NRMS-BERT can contribute to the effectiveness of news recommendation.

- **Approach:** “Debias” news embeddings and feed them into the model, hypothesis is true if accuracies declined significantly.
    - Subtract the “topic average vector” from each embedding?
    - Use SVM to find a “topic dividing subspace” and project the embeddings onto it?

Discussions & Limitations

- From the two news embedding we identify that the improvement of the news recommendation accuracy may come from the model's improved understanding of news categories.

- Our models only encode information from news titles, which is limited due to their short length and the insufficient clues they provide, even with larger models.

- News recommendation may need to incorporate article content to reach an accuracy breakthrough, but NRMS model alone would be insufficient at that time.

References

Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. 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 6389–6394, Hong Kong, China. Association for Computational Linguistics.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Empowering news recommendation with pre-trained language models. *Preprint*, arXiv:2104.07413.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, Online. Association for Computational Linguistics.

Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. Dkn: Deep knowledge-aware network for news recommendation. *Preprint*, arXiv:1801.08284.

Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. 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, Florence, Italy. Association for Computational Linguistics.

Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang, and Xiuqiang He. 2021. Unbert: User-news matching bert for news recommendation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*. Huawei Noah's Ark Lab.

Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1933–1942, New York, NY, USA. Association for Computing Machinery.

Thank You Q&A