RADio – Rank-**A**ware **D**ivergence Metr**i**cs to Measure N**o**rmative Diversity in News Recommendations

Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, Maarten de Rijke

16th ACM Conference on Recommender Systems, 2022

Group 3:

- Mathias Lambert V.
- Jaime Pérez.
- Matías Ossul.

The Role of News Recommender Systems

- Increase engagement (Nic, 2018).
- Help raise informed citizens (Eskens, 2017).
- Could even:
  - Foster tolerance and understanding (Ferrer-Conill, 2018).
  - Counter so-called filter bubbles or echo chambers (Möller, 2018).

The Problem: Beyond Clicks

- News recommenders taking over the role of human editors in news selection.
- Merely optimizing for click-through rates and engagement may (Tenenboim,2015):
  - Promote sensationalist content.
  - Spread of misinformation.
- Difficulties in translating their editorial norms into concrete metrics that can inform recommender system design (Boididou, 2021).

Descriptive (General-Purpose) Diversity

- Typically defined as the “opposite of similarity” (Bradley, 2001).
- Its goal is to prevent users from being shown the same type of items in their recommendations list and is often expressed as intra-list-diversity (ILD) (Castells, 2015).

$$ILD(R) = \sum_{i \in R} \sum_{j \in R} d(i, j)$$

- Diversity is most often implemented as a descriptive distance metric such as cosine similarity between two bag-of-words models or word embeddings (Kunaver, 2017).

Normative Diversity

- We define a normatively diverse news recommendation as one that succeeds in informing the user and supports them in fulfilling their role in democratic society (Helberger, 2019).

- Four different models:
  - **Liberal model**, which aims to enable personal development and autonomy.

  - **Participatory model**, which aims to enable users to fulfill their role as active citizens in a democratic society.

  - **Deliberative model**, which aims to foster discussion and debate by equally presenting different viewpoints and opinions in a rational and neutral way.

  - **Critical model**, which aims to challenge the status quo and to inspire the readers to take action against existing injustices in society.

Mathematical Foundation

- Distance metrics should satisfy the axions:
  - Identity. $D(x,y) \Leftrightarrow x = y$
  - Symmetry. $D(x,y) = D(y,x)$
  - Triangle inequality. $D(x,z) \leq D(x,y) + D(y,z)$

- f-Divergence measures diversity as a comparison between two probability distributions: **Recommendations (Q)** and **Context (P).**

$$D_f(P,Q) = \sum_x Q(x) \cdot f\left(\frac{P(x)}{Q(x)}\right)$$

$$f_{KL}(t) = t \cdot \log(t)$$

$$f_{JS}(t) = \frac{1}{2}\left[(t+1)\log\left(\frac{2}{t+1}\right) + t \cdot \log(t)\right]$$

Incorporating Rank-Awareness

- Reformulation of f-Divergence metric.
- Inspired by LTR metrics (Chakrabarti, 2008) like Mean reciprocal rank (MRR) and Normalized discounted cumulative gain (NDCG).

- The discrete probability distribution of a ranked recommendation set Q^*, given each item i in the recommendation list R:

$$Q^*(x) = \frac{\sum_i w_{R_i} \cdot 1_{i \in x}}{\sum_i w_{R_i}}$$

- Where w_{R_i} is the weight of a rank for item i.
  - For MMR, $w_{R_i} = \frac{1}{R_i}$.
  - For NDCG, $w_{R_i} = \frac{1}{\log_2(R_i+1)}$

Normative Diversity metrics as Rank-Aware f-Divergences

- S: The list of news articles the recommender system could make its selection from, also referred to as the "supply."
- R: The ranked list of articles in the recommendation set.
- H : The list of articles in a user's reading history, ranked by recency.
- $R_u^i \in \{1, 2, 3, \ldots\}$ refers to the rank of an item i in a ranked list of recommendations for user u.

- **Calibration**: measures to what extent the recommendations are tailored to a user's preferences.

$$\sum_c Q^*(c|R) \cdot f\left(\frac{P^*(c|H)}{Q^*(c|R)}\right)$$

- **Fragmentation:** reflects to what extent we can speak of a common public sphere, or whether the users exist in their own bubble.

$$\sum_c Q^*(e|R^v) \cdot f\left(\frac{P^*(e|R^u)}{Q^*(e|R^v)}\right)$$

Normative Diversity metrics as Rank-Aware f-Divergences

- **Activation**: Absolute sentiment score as a proxy for emotional intensity and Activation level in a single article.

$$\sum_{c} Q^*(k|R) \cdot f\left(\frac{P(k|S)}{Q^*(k|R)}\right)$$

- **Representation**: aims to approximate a notion of viewpoint diversity where the viewpoints are expressed categorically.

$$\sum_{c} Q^*(p|R) \cdot f\left(\frac{P(p|S)}{Q^*(p|R)}\right)$$

- **Alternative Voices**: It captures viewpoint diversity based on the speaker, not the content. Focuses on whether the viewpoint holder belongs to a protected group.

$$\sum_{c} Q^*(m|R) \cdot f\left(\frac{P(m|S)}{Q^*(m|R)}\right)$$

Experimental Setup

- MIND dataset (Wu, 2020): MSN News between October 12 and November 22, 2019.

# News	161,013	# Users	1,000,000
# News category	20	# Impression	15,777,377
# Entity	3,299,687	# Click behavior	24,155,470
Avg. title len.	11.52	Avg. abstract len.	43.00
Avg. body len.	585.05		

- Models used:
  - NPA (Neural News Recommendation with Personalized Attention) (Wu, 2019).
  - NAML (Neural News Recommendation with Attentive Multi-View Learning) (Wu, 2019).
  - LSTUR (Neural News Recommendation with Long- and Short-term User Representations) (An, 2019).
  - NRMS (Neural News Recommendation with Multi-Head Self-Attention) (Wu, 2019).
  - Random
  - Most popular

Experimental Setup

- **Metadata Enrichment Pipeline:** Developed NLP pipeline to extract features.
    - **Complexity analysis**: Each item is assigned a complexity score based on the Flesch-Kincaid reading ease test.
    - **Story clustering**: The individual news items are clustered into so-called news story chains, which means that stories about the same event will be grouped together.
    - **Sentiment analysis:** Assign a sentiment polarity score to each article using the open-source NLP library TextBlob.
    - **Named entity recognition:** Using spaCy, identify the people, organizations, and locations mentioned in the text, and count their frequency.
    - **Named entity augmentation**: Using fuzzy name matching, attempt to link the entities identified in the previous step to their corresponding Wikidata entries, in order to determine whether individuals are politicians and whether organizations are political parties.

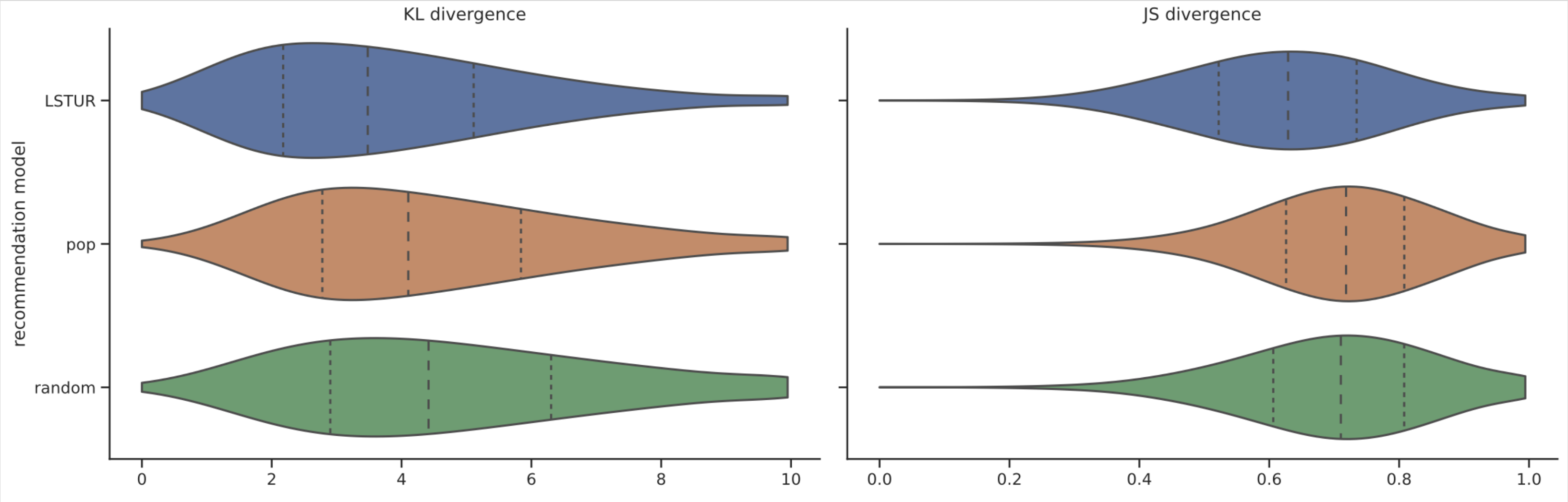
- **Evaluation:** RADio with JS divergence and rank-awareness @10 as default.

Results - Overview

Algorithm	Calibration (topic)	Calibration (complexity)	Fragmentation	Activation	Representation	Alternative voices	NDCG
LSTUR	0.5847	0.3632	0.9046	0.1819	0.1261	0.0409	0.4134
NAML	0.5709	0.3593	0.8836	0.1842	0.1230	0.0384	0.4091
NPA	0.5838	0.3619	0.8979	0.1841	0.1359	0.0390	0.4068
NRMS	0.5662	0.3548	0.8872	0.1794	0.1278	0.0362	0.4163
Most popular	0.6526	0.3477	0.8923	0.1949	0.1268	0.0342	0.2750
Random	0.6636	0.3981	0.9439	0.2715	0.2578	0.0698	0.2949

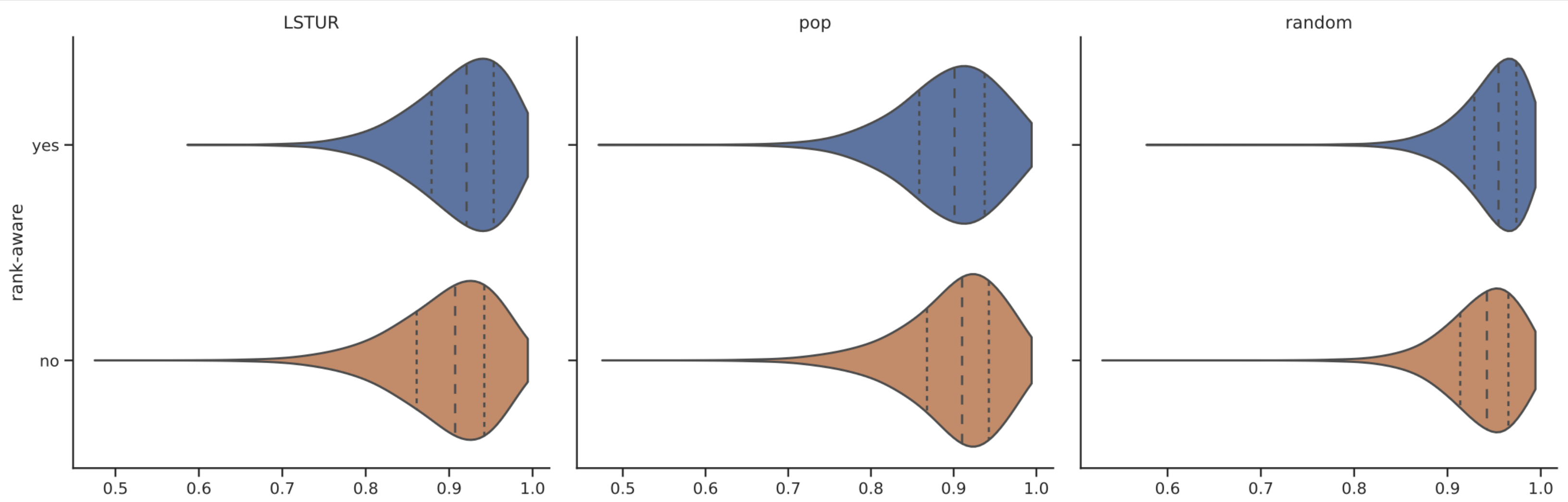
- Random recommender scores highest divergence across all metrics.
- Neural recommenders (LSTUR, NAML, NPA, NRMS) generally have lower divergence scores than baselines and similar NDCG.
- Neural recommenders are more Calibrated to user history than baselines.

Results - Sensitivity Analysis



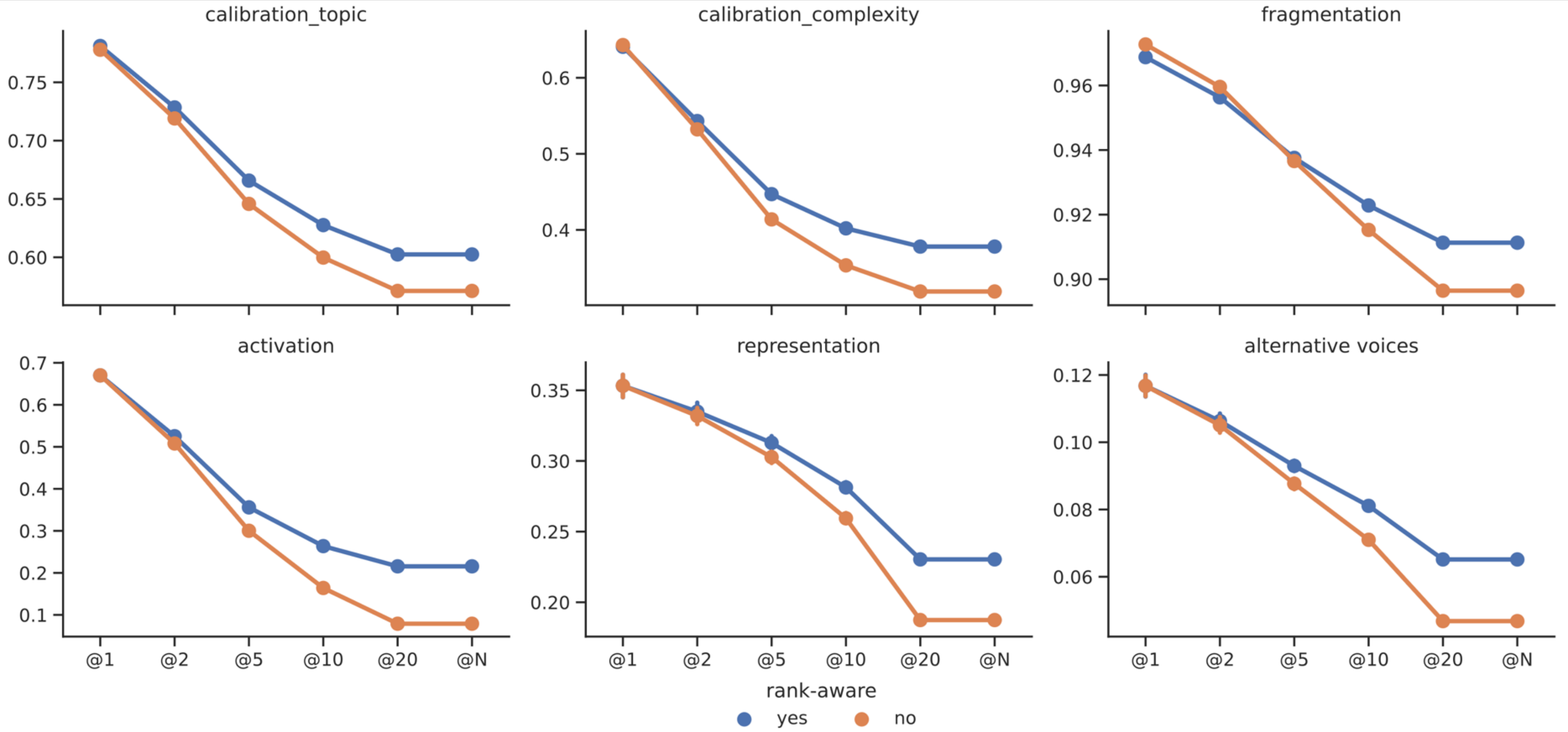
- JS divergence provides a more centered distribution and better contrast between neural and naive methods compared to KL divergence.

Results - Sensitivity Analysis



- The figure shows the effect of removing the rank-awareness (in blue) on Fragmentation and compare to the original rank-aware Fragmentation (in orange). Rank-awareness allows for better differentiation between methods:
  - LSTUR and “most popular” seem to be similarly distributed without a rank discount.
  - Introducing rank-awareness shifts LSTUR towards a larger divergence, whereas “most popular” remains largely the same.

Results - Sensitivity Analysis



- Figure demonstrates that the effect of rank-awareness increases with cutoff, and divergence scores stabilize after around 10 recommendations due to the heavy MRR discount.

Normative Evaluation & Discussion

- RADio allows comparing algorithms based on their alignment with normative goals.
- For Participatory goals (low Fragmentation, low Activation), neural recommenders appear more suitable than baselines.
- For Critical goals (high Representation, Alternative Voices), the random recommender scores highest divergence, but this doesn't mean it's *suitable*.

- **Limitations:**
  - f-Divergence doesn't account for semantic relationships between categories (e.g., similar political parties).
  - Requires discretizing continuous values into potentially arbitrary bins.
  - The quality of metadata from the NLP pipeline is an approximation and hard to evaluate without ground truth.
  - Influence of dataset content (soft vs. hard news) needs investigation.

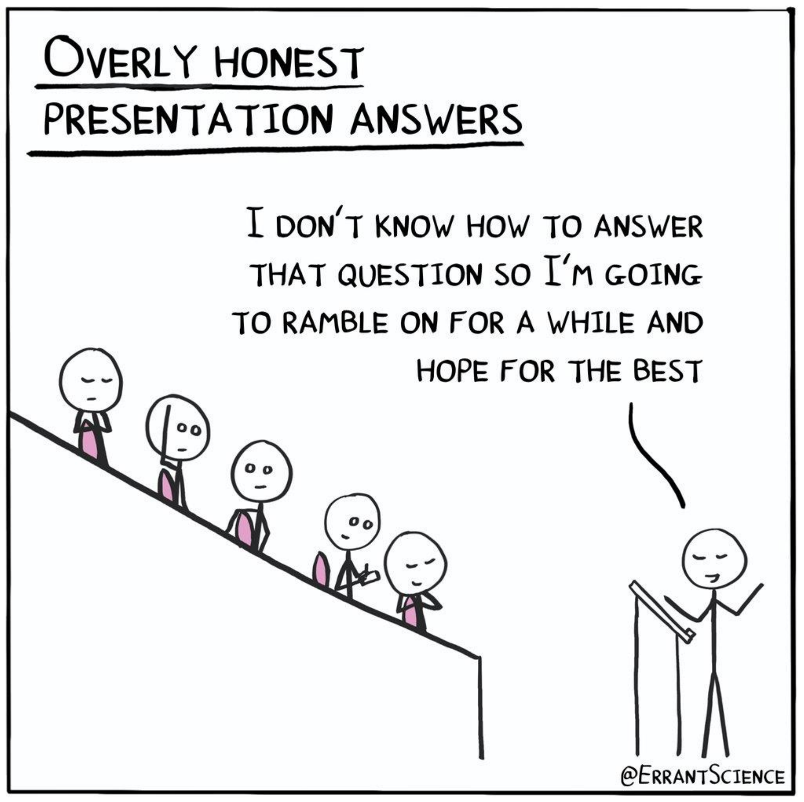
Conclusion

- RADio bridges the gap between technical evaluation and journalistic norms.
- It is mathematically grounded, with JS divergence being the preferred choice.
- RADio provides insights into how different algorithms influence news exposure from a normative perspective.
- RADio metrics should **supplement** standard evaluation metrics, not replace them.
- The goal is to foster **interdisciplinary discussion** and informed decision-making.

References

- **Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten De Rijke. 2022. RADio – Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations. In Proceedings of the 16th ACM Conference on Recommender Systems (RecSys '22). Association for Computing Machinery, New York, NY, USA, 208–219.**
- Sarah Eskens, Natali Helberger, and Judith Moeller. 2017. Challenged by News Personalisation: Five Perspectives on the Right to Receive Information. Journal of Media Law 9, 2 (2017), 259–284.
- Raul Ferrer-Conill and Edson C. Tandoc Jr. 2018. The Audience-Oriented Editor. Digital Journalism 6, 4 (2018), 436–453.
- Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 2018. Do Not Blame It on the Algorithm: An Empirical Assessment of Multiple Recommender Systems and their Impact on Content Diversity. Information, Communication & Society 21, 7 (2018), 959–977.
- Newman Nic, Richard Fletcher, Antonis Kalogeropoulos, David AL Levy, and Rasmus Kleis Nielsen. 2018. Reuters Institute Digital News Report 2018. Reuters Institute for the Study of Journalism (2018), 39.
- Ori Tenenboim and Akiba A Cohen. 2015. What Prompts Users to Click and Comment: A Longitudinal Study of Online News. Journalism 16, 2 (2015), 198–217.
- Christina Boididou, Di Sheng, Felix J Mercer Moss, and Alessandro Piscopo. 2021. Building Public Service Recommenders: Logbook of a Journey. In Fifteenth ACM Conference on Recommender Systems (Amsterdam, Netherlands) (RecSys '21). Association for Computing Machinery, New York, NY, USA, 538–540.
- Keith Bradley and Barry Smyth. 2001. Improving Recommendation Diversity. In Proceedings of the 12th irish conference on artificial intelligence and cognitive science (Maynooth, Ireland) (AICS'01). 85–94.
- Pablo Castells, Neil J Hurley, and Saul Vargas. 2015. Novelty and Diversity in Recommender Systems. In Recommender Systems Handbook. Springer, 881–918.
- Matevz Kunaver and Tomaz Pozrl. 2017. Diversity in Recommender Systems – A Survey. Knowledge-Based Systems 123 (2017), 154 – 162.
- Natali Helberger. 2019. On the Democratic Role of News Recommenders. Digital Journalism 0, 0 (2019), 1-20.
- Soumen Chakrabarti, Rajiv Khanna, Uma Sawant, and Chiru Bhattacharyya. 2008. Structured Learning for Non-Smooth Ranking Losses. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Las Vegas, Nevada, USA) (KDD '08).
- Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu and Ming Zhou. MIND: A Large-scale Dataset for News Recommendation. ACL 2020.
- Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: Neural News Recommendation with Personalized Attention. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2576–2584.
- Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Attentive Multi-view Learning. arXiv preprint arXiv:1907.05576 (2019).
- 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. 336–345.
- 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. 6389–6394.

End



Thank you for listening!

Questions?

Normative Diversity Overview

	Context	Type	Distribution of
Calibration (topics)	Reading history	Categorical	article subcategories as provided in the MIND dataset
Calibration (complexity)	Reading history	Continuous	article complexity **(1)** as calculated with the Flesch-Kincaid reading ease test
Fragmentation	Other users	Categorical	recommended news story chains **(2)**, which are identified following the procedure in [16]
Activation	Available articles	Continuous	affect scores, which is approximated by the absolute value of a sentiment analysis score **(3)**
Representation	Available articles	Categorical	the presence of political actors **(4)**
Alternative Voices	Available articles	Continuous	the presence of minority voices versus majority voices. We identify someone as a ’minority voice’ when they are identified as a person through the NLP pipeline **(5)**, but cannot be linked to a Wikipedia page.[7]

	Calibration (topic)	**Calibration (complexity)**	**Fragmentation**	**Activation**	**Representation**	**Alternative voices**
Liberal	Low	Low	High	–	–	–
Participatory	High	Low	Low	Medium	Reflective	Medium
Deliberative	–	–	Low	Low	Equal	–
Critical	–	–	–	High	Inverse	High

Neural News Recommendation with Personalized Attention

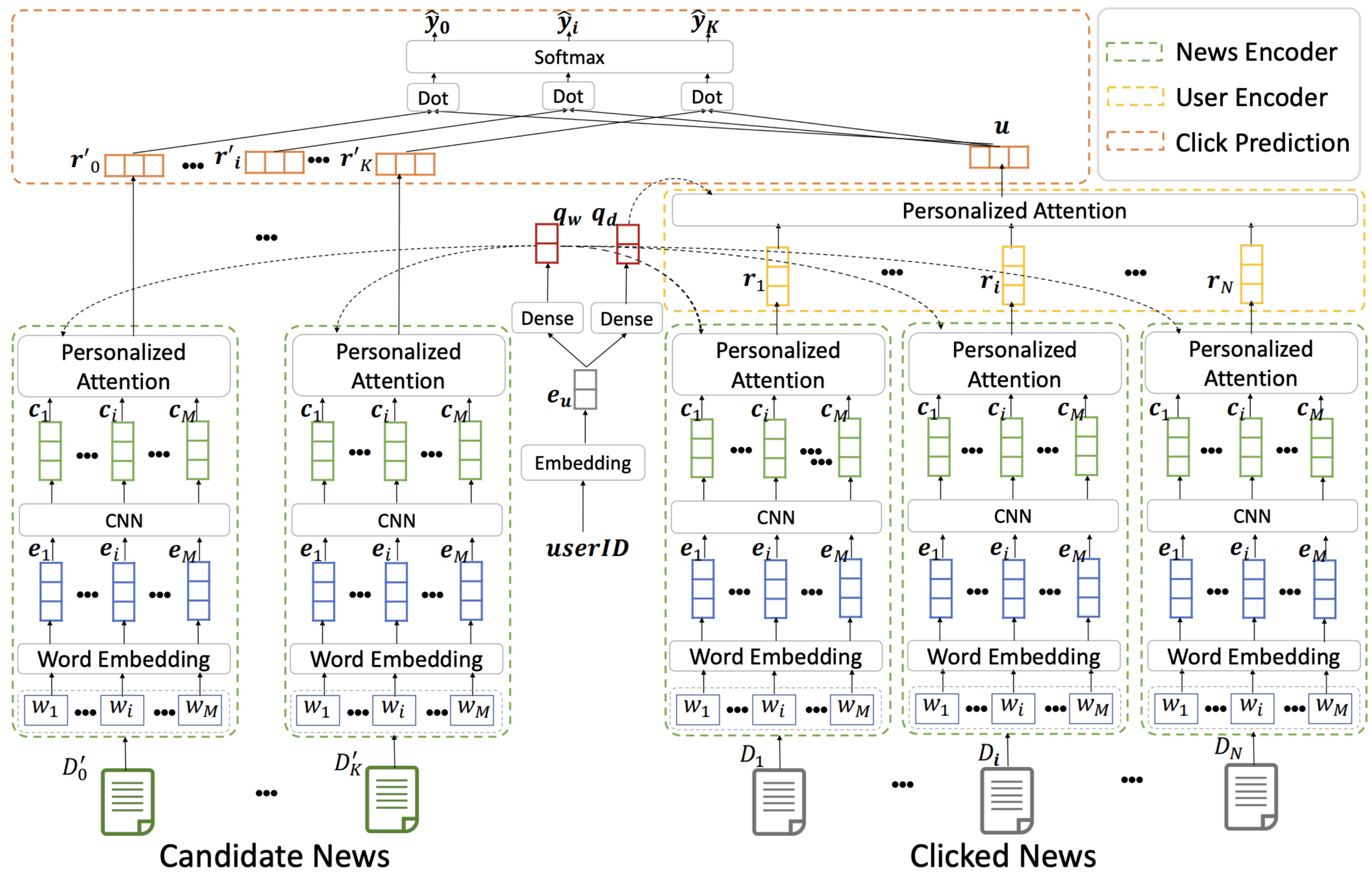


Figure 2: The framework of our NPA approach for news recommendation.

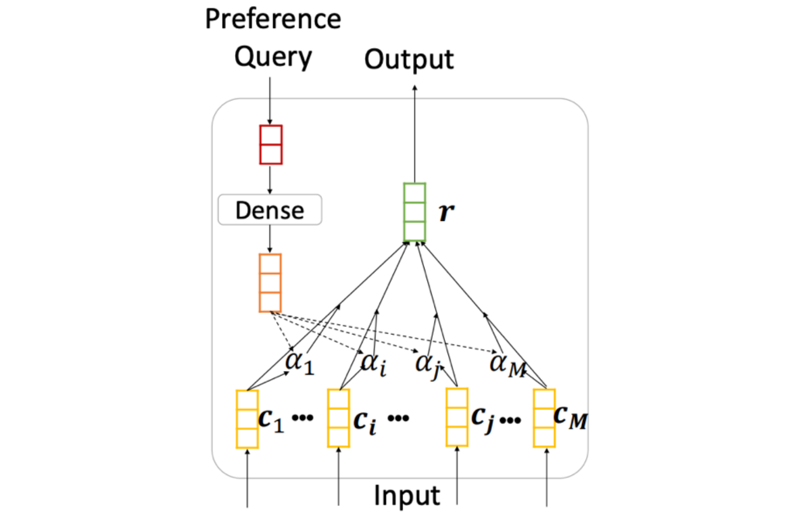


Figure 3: The architecture of the personalized attention module in our *NPA* approach.

Neural News Recommendation with Attentive Multi-View Learning

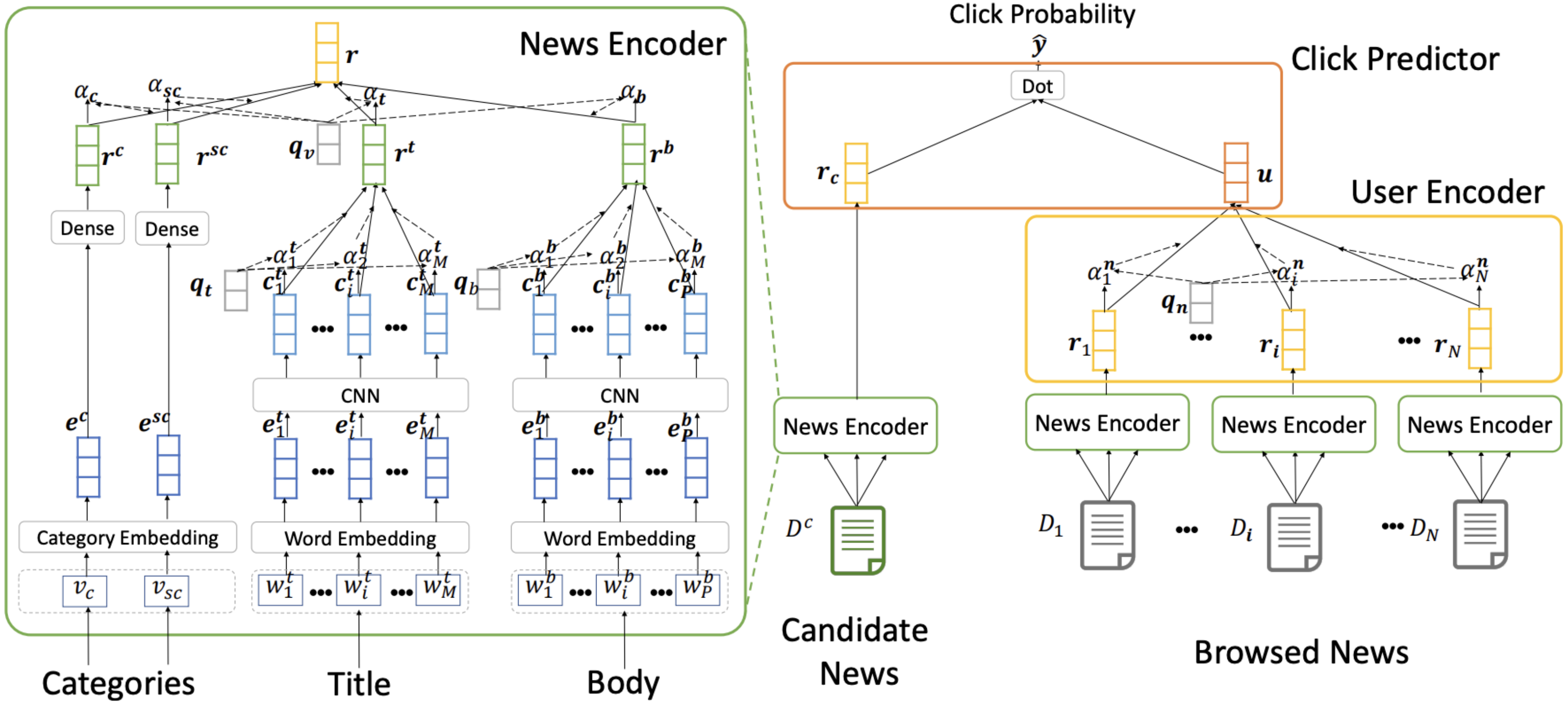


Figure 2: The framework of our *NAML* approach for news recommendation.

Neural News Recommendation with Long- and Short-term User Representations

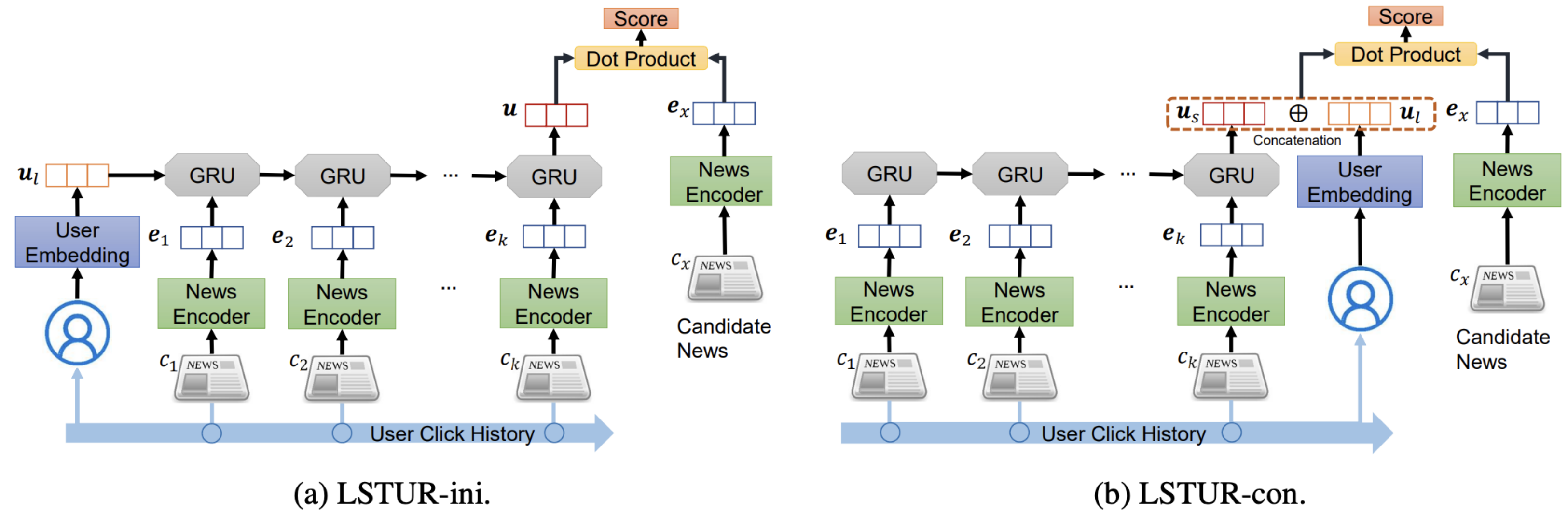


Figure 3: The two frameworks of our LSTUR approach.

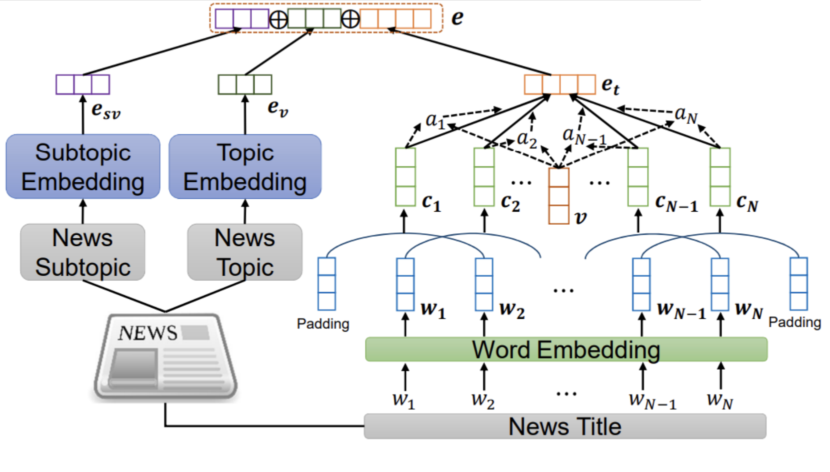


Figure 2: The framework of the news encoder.

Neural News Recommendation with Multi-Head Self-Attention

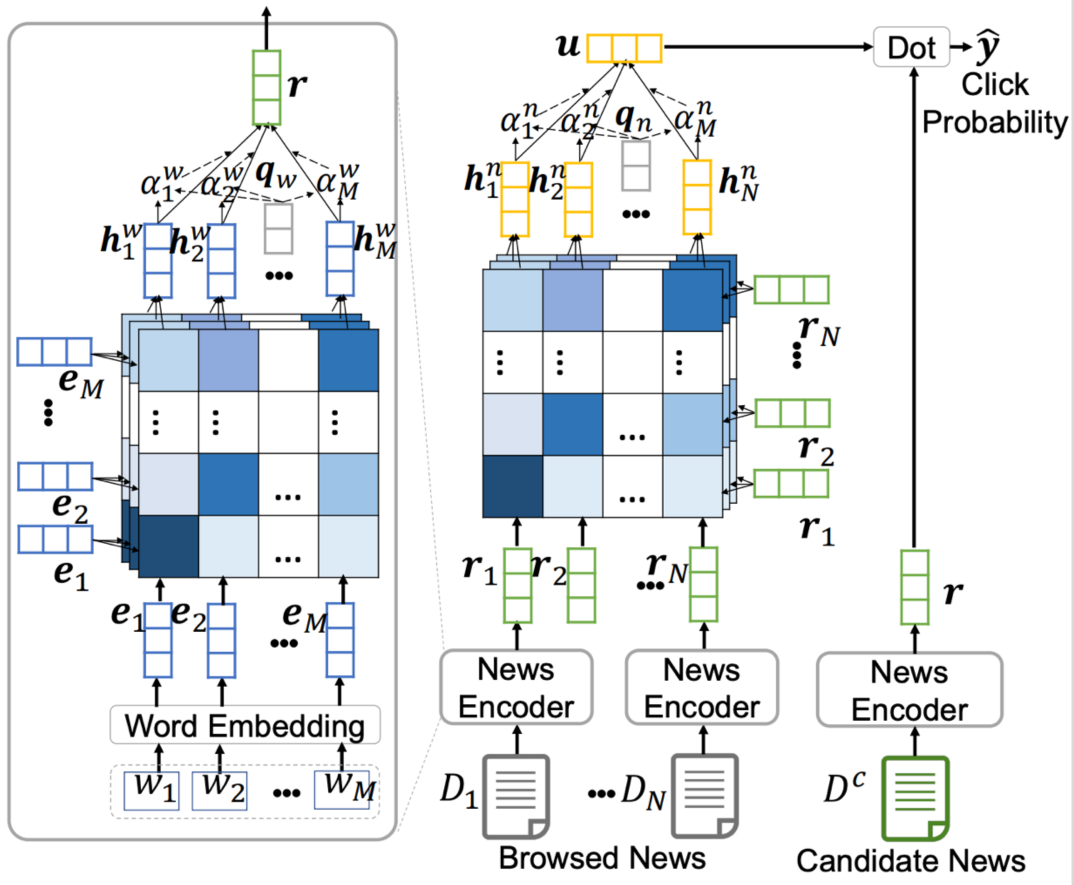


Figure 2: The framework of our *NRMS* approach.

Flesch–Kincaid readability tests

- designed to indicate how difficult a passage in English is to understand.
- developed under contract to the U.S. Navy in 1975 by J. Peter Kincaid and his team.

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

Score	School level (US)	Notes
100.00–90.00	5th grade	Very easy to read. Easily understood by an average 11-year-old student.
90.0–80.0	6th grade	Easy to read. Conversational English for consumers.
80.0–70.0	7th grade	Fairly easy to read.
70.0–60.0	8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
60.0–50.0	10th to 12th grade	Fairly difficult to read.
50.0–30.0	College	Difficult to read.
30.0–10.0	College graduate	Very difficult to read. Best understood by university graduates.
10.0–0.0	Professional	Extremely difficult to read. Best understood by university graduates.