Recommendations for the Recommenders: Reflections on Prioritizing Diversity in the RecSys Challenge

Lucien Heitz
heitz@ifi.uzh.ch
Department of Informatics and Digital Society Initiative, University of Zurich
Zurich, Switzerland

Oana Inel
inel@ifi.uzh.ch
Department of Informatics, University of Zurich
Zurich, Switzerland

Sanne Vrijenhoek
s.vrijenhoek@uva.nl
AI, Media & Democracy Lab and Institute of Information Law, University of Amsterdam
Amsterdam, The Netherlands

ABSTRACT

The *RecSys Challenge 2024*, co-organized by the Danish news outlet *Ekstra Bladet*, called for participants to develop a news recommender system that accurately predicts which news article a reader is most likely to select from a given list. In the context of this challenge, the organizers explicitly stated their interest in beyond-accuracy objectives. However, the setup of the challenge did not facilitate pursuing these more normative goals: a missed opportunity, given the quality of the dataset. In this paper, we highlight the issues encountered in a submission that prioritized normative diversity. We reflect on the responsibility of conferences, RecSys in particular, when it comes to promoting beyond-accuracy objectives and provide recommendations for future challenge iterations.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Evaluation of retrieval results**; **Personalization**.

KEYWORDS

news diversity, beyond-accuracy objectives, recommender systems

ACM Reference Format:
Lucien Heitz, Oana Inel, and Sanne Vrijenhoek. 2024. Recommendations for the Recommenders: Reflections on Prioritizing Diversity in the RecSys Challenge. In *ACM RecSys Challenge 2024 (RecSys Challenge '24), October 14–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3687151.3687155

1 INTRODUCTION AND BACKGROUND

The *RecSys Challenge* is a yearly competition hosted as part of the ACM Conference on Recommender Systems.[1] Each year, an organization provides a recommender-relevant dataset and asks participants to develop recommender systems (RS) that tackle challenges specific to their use case. In 2024, the challenge was hosted by Ekstra Bladet (EB), a Danish tabloid newspaper. EB provided the **E**kstra **B**ladet **Ne**ws **R**ecommendation **D**ataset (*EB-NeRD*), containing six weeks of user interactions on their platform.

The goal of the 2024 edition of the RecSys Challenge is to predict which news articles users would click on from a set of candidate items.[2] In their description, EB explicitly states that they "embrace the normative complexities"[3] of RS and want to investigate how news recommendations resonate with editorial values and the effect of RS on the news flow by looking at beyond-accuracy objectives. Participants were provided an overview of beyond-accuracy objectives that are worth considering (for details, please see [18]). However, despite EB's interest in the normative aspects of news recommendation, the primary evaluation metric was area under the curve (AUC), a statistical measure expressing discriminative ability for accuracy optimization.

Academic research has long emphasized the importance of news recommendations on a societal level for creating a well-informed citizenry within democratic institutions [4, 15] and strengthening the right of informational self-determination [25, 33]. In the industry, similar to EB, both the BBC in the United Kingdom [26] and ZDF in Germany [10] have investigated beyond-accuracy objectives and note that more research needs to be done on assessing the societal impact of RS. The exact impact of a production recommender system on the news flow, however, has been hard to pinpoint, as there are only a handful of publicly available news datasets to work with [35]. Accordingly, researchers often needed to rely on interviewing practitioners on anticipated effects [24, 39], or collaborate with news outlets and aggregators [23]. An exception to this is the work of Einarsson et al. [8], who investigated the effect of the use of algorithmic RS for news on the EB platform during the 2022 Danish general election. They draw from "agenda-setting theory," which poses that *"it is the democratic role and responsibility of the media to prioritize matters of societal interest"*. While the effects were subtle, they found that *"personalized recommendations resulted in exposure to softer news content and a less diverse variation in the exposure of topics, actors, or issues."* We, therefore, see it as a crucial task to look at beyond-accuracy objectives when it comes to assessing the quality of news recommendations.

Against this backdrop, we present a diversity-focused challenge submission utilizing a random walk strategy [6, 28]. In the discussion of our approach, we highlight where the challenge either limits the promotion of beyond-accuracy objectives or neglects the potential for diversity-focused research. We end with a discussion of the responsibility of institutions, RecSys in particular, to prioritize and facilitate work on societal aspects of RS and make recommendations for future iterations of the RecSys Challenge.

[1] RecSys Challenge 2024 website: https://recsys.acm.org/recsys24/challenge/

[2] Ekstra Bladet website and dataset overview: https://recsys.eb.dk/

[3] RecSys Challenge 2024 website: https://www.recsyschallenge.com/2024/

RecSys Challenge '24, October 14–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-1127-5/24/10
https://doi.org/10.1145/3687151.3687155

2 CHALLENGE SUBMISSION

The method used in this submission is based on a long-tail diversity-driven three-hop graph random walk strategy RP^3_β, adapted from existing work [6, 28]. We chose this model for the challenge submission because the underlying random walk sampling technique of RP^3_β has the ability to optimize for both the **accuracy** as well as the **diversity** of recommendations [27, 28]. Using the random walks algorithm allows us to pursue not only the accuracy-focused goals of the RecSys'24 Challenge, but also the normative, beyond-accuracy objectives promoted by the organizers.

2.1 Data Pre-processing Steps

EB-NeRD (D) presents a collection of user-item interactions:

$$D = \{(u, i) | u \in U, i \in I\},$$

where u is one user of the collection of users U, and i is one item (i.e., one news article) of the collection of all items I. The interaction (u, i) denotes a given user having clicked on a news item. Within the scope of this submission, we followed the example of Paudel et al. [28] for creating the underlying graphs and defined a user clicking on a news item as a positive interaction:

$$I^+_{u_j} = \{i | (u_j, i) \in D\}.$$

This results in a user-item interaction matrix M of $m \times n$ for user $u_i (i \in 1...m)$ and item $i_j (j \in 1...n)$. As we use only positive interactions, every cell in M gets a value of 1 if the user in question has clicked on a given item, and 0 otherwise. As part of the data cleaning step, we removed all interactions from the graph with a `Read Time` of less than two seconds and/or a `Scroll Percentage` of **`null`**. The motivation behind doing so is that we want to focus on meaningful interactions, i.e., interactions where users not only clicked on articles, but also read them. The algorithm ran on the internal Informfully back end [12–14].

2.2 Item Re-ranking Procedure

By default, RP^3_β features a popularity-based discount function for diversifying the item recommendation:

$$\widetilde{p}^3_{ui} = \frac{p^3_{ui}}{d^\beta_{ii}}, \tag{1}$$

where the probability p of reaching an item node i from a starting user node u is discounted by the item node degree d^β, with $\beta = 2$ in our case. We do this with all entries in `history.parquet*` (i.e., a filtered version of `history.parquet` that resulted from the cleaning process outlined in Section 2.1) and get as a result a list of articles together with click probabilities for each user-item combination. In an optimal setting, RP^3_β is typically used as a model during the in-processing stage of the recommender pipeline to generate candidate items (cf. [11, 41]).[4] The current challenge setup, however, asks participants to predict what entry of `Inview Articles` (i.e., a curated list of articles) a user is most likely to select, severely limiting the item pool of potential candidates.

[4]A stand-alone version of the random walk algorithms is provided in our code repository, including all our hyperparameters: https://github.com/Informfully/Recommenders

The challenge task is more akin to a post-processing task of re-ranking an existing pool of candidate items [2, 30, 36], rather than creating such a pool by utilizing a model.[5] The difficulty introduced by this shift in focus of the recommendation problem is that the distance between the current user and any of the articles shown to the user might not be covered by the three hops of RP^3_β on the graph. Hence, if the algorithm cannot calculate any predictions (i.e., the user-item distance is larger than three hops), a default chronological ranking was applied (selecting the newest item).

2.3 Random Walk Results

Table 1 provides an overview of the results of the random walk approach (RP^3_β) with the EB recommender baseline (EBRec) as well as the average score for each metric of the top 3 challenge submissions (TOP3).[6] Results are listed separately for area under the curve (AUC), mean reciprocal rank (MRR), normalized discounted cumulative gain for the top 5 items (NDCG@5), intra-list diversity (ILD), novelty (NOV), and serendipity (SER), where the last three objectives are formalized according to Kaminskas and Bridge [18]. All results in Table 1 are from the official competition website.[7]

Table 1: Performance comparison of selected models.

Model	AUC	MRR	NDCG@5	ILD	NOV	SER
RP^3_β	0.5005	0.3167	0.3505	0.7549	**11.101**	**0.8069**
EBRec	0.5684	0.3517	0.3951	**0.8402**	3.071	0.7915
TOP3	**0.7643**	**0.5501**	**0.6117**	0.6255	5.771	0.7568

As expected, TOP3 shows the best performance when assessed with the accuracy-oriented metrics of AUC, MRR, and NDCG@5. Looking at the beyond-accuracy metrics of novelty and serendipity, we see a substantial improvement of RP^3_β over the top-ranking algorithms in both areas. EBRec has the best score when looking at the measurement of intra-list diversity.

With the setup of the challenge of restricting article selection to a small list of candidate items, the results of the random walk strategy are similar to random guessing. The only substantial difference of our modified random walks over competing models comes down to novelty increases, which is due to the heuristic rule of giving higher priority to newer articles. However, using a random walk strategy served as a three-fold purpose: 1) provide an example of focusing on beyond-accuracy objectives, 2) identify implementation challenges, and 3) understand performance trade-offs.

3 DISCUSSION

The results of Table 1 clearly indicate a trade-off between accuracy on the one side, and diversity, novelty, and serendipity on the other. Instead of ensuring that both accuracy and beyond-accuracy objectives are reflected in the recommendations, as the challenge organizers envisioned, we observe that the leaderboard is dominated by submissions that primarily focused on maximizing accuracy.

[5]One complicating factor here is that the curation process behind generating `Inview Articles` was not communicated nor how/where the item was displayed on screen.
[6]The top submissions are taken from the competition leaderboard: https://www.codabench.org/competitions/2469 in relation to their AUC results.
[7]Metrics and results taken from the official Codabench challenge website: https://www.codabench.org/competitions/2469/ (registration required).

This is not surprising, given that accuracy was the primary objective of the challenge. This trade-off, however, is not unavoidable. There are numerous studies in the news domain that show that multi-objective optimization for diversity and accuracy is possible [6, 21, 28, 31, 32]. To answer EB's call for utilizing beyond-accuracy objectives, our submission therefore tries to balance accuracy and beyond-accuracy objectives. In doing so, however, we encountered issues related to the definition of diversity, the dataset structure, and the overall recommendation task. Hence, we aim to shed light on potential improvements for the next normative challenge.

3.1 Defining Diversity

The first challenge encountered lies in the fact that it is unclear how, in the context of EB, normative diversity and other beyond-accuracy metrics should be conceptualized. The Codabench environment shows a metric corresponding to ILD (following the definition of Kaminskas and Bridge [18]). Yet, recent work has noted the discrepancy between "standard" RecSys notions of diversity, such as ILD, and diversity as it is conceptualized by news organizations themselves [39, 40]. Similar work has been carried out for serendipity metrics [34]. Furthermore, while ILD can be a useful proxy, it may not correspond to human perceptions of diversity [17].

Conceptualizing diversity in a way that does justice to its normative underpinning is a complicated process and is not something that is easily done by people that are not domain experts. It would have been helpful if the challenge organizers had explicitly stated *how* they wanted editorial values to manifest in the recommendations. For this, they could have leveraged the conclusions of Einarsson et al. [8], who found that the use of RS in news increased the amount of soft news people consumed and impacted the distribution of political actors in the articles read. Keeping these distributions representative of EB's brand, while optimizing for AUC, would have added an interesting spin to the challenge.

3.2 Dataset Structure

The dataset provided by EB is of high quality and contains more than 213 million interactions from over 1.1 million users with over 125,000 unique articles, over the span of six weeks. Challenge participants have access to the article's full body text, named entities have already been extracted, and there is a sophisticated taxonomy of categories, subcategories, and topics. However, it is unclear whether the dataset contains *all* articles published in that period, or only the articles shown to this subset of users. This is relevant to evaluate the diversity of recommendations to determine whether they are a good reflection of the organization's values [38].

Furthermore, it is unclear how the candidate list was generated. We assume there to be a relation with the article the user is currently reading. However, by reducing the scope of articles a RS can choose from to an average of around eleven articles greatly reduces its potential for a diversity-driven recommendation. It is also unclear *how* the items were displayed on screen. Were recommendations visualized in a way that attracts the user's attention, e.g., shown larger on the screen and/or had a more prominent position in the recommendation list? Were users actually interested in reading a selected news article, or do they always simply click the top item of the recommendation list?

In its structure and in the type of information that is present, EB-NeRD is similar to MIND [42]. This potentially lowers the threshold for participating in the challenge and makes it easier to provide baseline scores. However, Drunen and Vrijenhoek [7] have noted that MIND is in its setup particularly suitable for optimization of user engagement, as opposed to normative goals. Leveraging its structure makes researchers more likely to follow engagement optimization and in turn solve *Microsoft's* RS problems, rather than their own. Thus, even seemingly innocuous decisions like this have an impact on research agendas and the challenges that can be tackled by the scientific community.

3.3 Recommendation Task

Contrary to MIND, EB-NeRD contains data over a longer period of time. This makes the challenge dataset a suitable candidate for observing changes in the type of content users receive over time [37]. However, this potential is limited by the design of the recommendation task, which requires that only a single item from the candidate list is recommended. This, in return, also means that diversity can only be observed at an aggregate level (i.e., assessing whether the set of recommendations of *all users* is diverse), and not at an individual user level (i.e., whether *a particular user* receives diverse recommendations).

Many media organizations claim that it is their responsibility to offer a diverse selection, but that it is eventually up to the user to decide what they read [39]. Thus, designing the challenge in a way that only focuses on the top 1 recommendation does not adequately reflect the requirements many media organizations have for their recommendations. Instead, we should not only care about the content of the article that the user has read, but also about the properties of the list of items that was shown. Furthermore, such challenges should offer enough freedom for the challenge organizers to promote and focus on metrics that are relevant for their specific domain and organization.

Vrijenhoek [37] showed that the recommender systems trained on MIND were disproportionately likely to recommend articles from the "sports" and "lifestyle" category, and less likely to recommend "news" and "finance." This pattern was also observed by Einarsson et al. [8]. Given the comparatively low scores for ILD in the TOP3 challenge submissions, it is not unlikely that something similar would surface here. This leads to the general question of how the recommendations produced by the TOP3 challenge submissions are actually received by relevant stakeholders, regardless of their stellar performance in terms of AUC. How would news editors rate the quality of the recommendations? And what about the readers themselves?

The points raised here are not to be understood as criticism of the TOP3 participants. The organizers' interest in diversity was only mentioned in the challenge description, but not incorporated in the evaluation of the results. Practically, this choice makes sense. It is easier to determine a challenge winner with clear evaluation criteria. However, without explicit value being placed on beyond-accuracy objectives for the leaderboard and ranking, challenge participants are not incentivized to actually focus on these goals. If beyond-accuracy objectives are indeed considered to be a priority, then this needs to be reflected accordingly in the evaluation procedure.

4 CONCLUSION

In the general field of machine learning, terms such as "performance" and "accuracy" have almost become synonymous with more generic terms such as "success" and "progress" [5]. In opposition to this, the most-cited RecSys paper in the past five years notes that *"it is not even clear if slightly higher accuracy values are relevant in terms of adding value for recommendation consumers or providers"* [9]. This is especially true in the context of news recommendation, where researchers *and* practitioners are aware that the accuracy-driven view is very limited and have extensively published on other values that should be pursued [3]. Yet, the vast majority of the work at RecSys is performance-driven, and with the exclusive focus on predictive power we see the same dynamic reflected in the RecSys Challenge as well. To entice organizations to become challenge sponsors, they are promised *"the unique opportunity to showcase its support for advancing the knowledge [...] while also gaining exposure to a global audience of industry leaders and decision-makers."*[8] This is somewhat ironic: hosting organizations are promised they will "advance the field" while the field itself has repeatedly shown that this is perhaps not the type of advancement that is needed.

Furthermore, there is also a societal responsibility tied to conference challenges. Luitse et al. [22] found that challenges and competitions in the medical domain induce *"risks of bias, inequality, and power concentration"*. It is true that there are a number of important differences between the challenges analyzed by Luitse et al. [22] and the RecSys challenge. For one, there is relatively little prize money involved, and participants join the competition for a chance to work with a real-world dataset and the "street credibility." The barriers for hosting are sufficiently low to also make it feasible for a relatively small organization like EB to host, reducing the risk of power concentration. However, if media organizations have agenda-setting power, then so do academic conferences, such as RecSys [20], and by extension its challenges. Junior scholars and industry practitioners come to conferences to hear about new developments and promising research areas. What is being presented at the conference will then become the standard for what is considered "high quality" and "valuable." This finds its way into new research papers, projects, and implementations: a self-reinforcing loop. Hence, not including beyond-accuracy measures in the challenge makes it less likely to be picked up as a valuable topic to pursue. Simultaneously, researchers that have conducted high-quality work on this topic may choose to submit it to other venues, such as FAccT.[9]

Signaling what objectives are relevant—other than accuracy—could be a valuable by-product of the RecSys challenge without directly interfering with peer-reviewing of the main conference work. Given the importance of news recommendations on a societal level, this year's challenge—in our view—is a missed opportunity for the RecSys community to move beyond the "leaderboard-chasing culture" [16], and to promote research on societal values as well as on the implications of recommender systems by prioritizing beyond-accuracy objectives. As *"the premier international forum [...] in the broad field of recommender systems,"*[10] RecSys carries the responsibility to help the field forward in this regard.

[8]Official RecSys '24 challenge website: https://recsys.acm.org/recsys24/challenge/

[9]ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT): https://facctconference.org

[10]Official RecSys '24 website: https://recsys.acm.org/recsys24/

Finally, we would like to express our thanks to the challenge organizers of EB and the high-quality dataset they put together. We do not mean this piece as a criticism of how they organized the challenge. Rather, we want to encourage the RecSys Challenge Chairs to take the lead and help future hosts incorporate beyond-accuracy metrics.

RECOMMENDATIONS

We want to end this contribution with a list of recommendations on how to better facilitate beyond-accuracy objectives. These are not meant as definitive guidelines. We hope they can serve as inspiration for change. Concretely, we make the following recommendations to the RecSys Challenge Chairs:

- Incentivize submissions for beyond-accuracy objectives by 1) providing clear guidelines on and explicitly defining the types of beyond-accuracy factors to be taken into account, 2) providing the necessary contextual understanding on the subset of items that the dataset contains, and 3) introducing multiple leaderboards, such that the evaluation of the submitted systems is more easily performed.
- Promote and include online testing as a standard component of the RecSys challenge. In the past, this has been done during RecSys Challenge 2017 [1].[11] We recommend online evaluations to be performed on the top submissions for all available leaderboards.
- Streamline the organization and evaluation of submissions by choosing a suitable, single environment to host and organize the challenges over the years. Potential environments are for example, Codabench [43] which was already used in this year's challenge, CodaLab [29], or Dynabench [19].
- Promote forming inter- and cross-disciplinary participation in the challenge beyond designing, developing, and submitting recommender models.
- Encourage the challenge hosts to allow usage of the dataset for research purposes beyond the scope of the challenge submissions.

Facilitating change in an existing structure is challenging, especially for Challenge Chairs that only fulfill their role for a single year. We recommend appointing a person within the steering committee to oversee the Challenge for multiple years. This person could ensure that small improvements are made over time, and that responsibility does not solely fall on that year's chairs.

ACKNOWLEDGMENTS

Lucien Heitz was partly funded by the Digital Society Initiative (DSI) of the University of Zurich. Sanne Vrijenhoek is funded by the AI, Media & Democracy Lab.

REFERENCES

[1] Fabian Abel, Yashar Deldjoo, Mehdi Elahi, and Daniel Kohlsdorf. 2017. RecSys Challenge 2017: Offline and Online Evaluation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 372–373.

[2] Gediminas Adomavicius and YoungOk Kwon. 2011. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering* 24, 5 (2011), 896–911.

[3] Christine Bauer, Chandni Bagchi, Olusanmi A Hundogan, and Karin van Es. 2024. Where are the Values? A Systematic Literature Review on News Recommender Systems. *ACM Transactions on Recommender Systems* (2024).

[11]RecSys Challenge 2017 included both offline and online evaluation: https://www.recsyschallenge.com/2017/

[4] Abraham Bernstein, Claes de Vreese, Natali Helberger, Wolfgang Schulz, Katharina Zweig, Christian Baden, Michael A. Beam, Marc P. Hauer, Lucien Heitz, Pascal Jürgens, Christian Katzenbach, Benjamin Kille, Beate Klimkiewicz, Wiebke Loosen, Judith Moeller, Goran Radanovic, Guy Shani, Nava Tintarev, Suzanne Tolmeijer, Wouter van Atteveldt, Sanne Vrijenhoek, and Theresa Zueger. 2021. Diversity in News Recommendation (Dagstuhl Perspectives Workshop 19482). *Dagstuhl Manifestos* 9, 1 (2021), 43–61. https://doi.org/10.4230/DagMan.9.1.43

[5] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 173–184.

[6] Fabian Christoffel, Bibek Paudel, Chris Newell, and Abraham Bernstein. 2015. Blockbusters and wallflowers: Accurate, diverse, and scalable recommendations with random walks. In *Proceedings of the 9th ACM Conference on Recommender Systems*. 163–170.

[7] Max van Drunen and Sanne Vrijenhoek. [n. d.]. How datasets shape the way news is recommended, and how law could address dataset dependencies. ([n. d.]). Under review.

[8] Árni Már Einarsson, R Helles, and S Lomborg. 2024. Algorithmic agenda-setting: the subtle effects of news recommender systems on political agendas in the Danish 2022 general election. *Information, Communication & Society* (2024), 1–21.

[9] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM conference on recommender systems*. 101–109.

[10] Andreas Grün and Xenija Neufeld. 2021. Challenges experienced in public service media recommendation systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 541–544.

[11] Lucien Heitz. 2023. Classification of Normative Recommender Systems. In *Proceedings of the First Workshop on the Normative Design and Evaluation of Recommender Systems*.

[12] Lucien Heitz, Julian A Croci, Madhav Sachdeva, and Abraham Bernstein. 2024. Informfully - Research Platform for Reproducible User Studies. In *Proceedings of the 18th ACM Conference on Recommender Systems*.

[13] Lucien Heitz, Juliane A Lischka, Rana Abdullah, Laura Laugwitz, Hendrik Meyer, and Abraham Bernstein. 2023. Deliberative Diversity for News Recommendations: Operationalization and Experimental User Study. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 813–819.

[14] Lucien Heitz, Juliane A Lischka, Alena Birrer, Bibek Paudel, Suzanne Tolmeijer, Laura Laugwitz, and Abraham Bernstein. 2022. Benefits of diverse news recommendations for democracy: A user study. *Digital Journalism* 10, 10 (2022), 1710–1730.

[15] Natali Helberger. 2021. On the democratic role of news recommenders. In *Algorithms, Automation, and News*. Routledge, 14–33.

[16] Dietmar Jannach and Christine Bauer. 2020. Escaping the McNamara fallacy: Towards more impactful recommender systems research. *Ai Magazine* 41, 4 (2020), 79–95.

[17] Mathias Jesse, Christine Bauer, and Dietmar Jannach. 2023. Intra-list similarity and human diversity perceptions of recommendations: the details matter. *User Modeling and User-Adapted Interaction* 33, 4 (2023), 769–802.

[18] Marius Kaminskas and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 1 (2016), 1–42.

[19] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in NLP. *arXiv preprint arXiv:2104.14337* (2021).

[20] Benjamin Laufer, Sameer Jain, A. Feder Cooper, Jon Kleinberg, and Hoda Heidari. 2022. Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. 401–426.

[21] Feng Lu, Anca Dumitrache, and David Graus. 2020. Beyond Optimizing for Clicks: Incorporating Editorial Values in News Recommendation. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, 145–153.

[22] Dieuwertje Luitse, Tobias Blanke, and Thomas Poell. 2024. AI competitions as infrastructures of power in medical imaging. *Information, Communication & Society* (2024), 1–22.

[23] Mats Mulder, Oana Inel, Jasper Oosterman, and Nava Tintarev. 2021. Operationalizing framing to support multiperspective recommendations of opinion pieces. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 478–488.

[24] Lynge Asbjørn Møller. 2022. Recommended for You: How Newspapers Normalise Algorithmic News Recommendation to Fit Their Gatekeeping Role. *Journalism Studies* 23, 7 (2022), 800–817. https://doi.org/10.1080/1461670X.2022.2034522 arXiv:https://doi.org/10.1080/1461670X.2022.2034522

[25] Laurens Naudts, Natali Helberger, Marijn Sax, and Michael Veale. 2024. Toward Constructive Optimisation: a new perspective on the regulation of recommender systems and the rights of users and society. European Consumer Organisation (BEUC).

[26] Maria Panteli, Alessandro Piscopo, Adam Harland, Jonathan Tutcher, and Felix Mercer Moss. 2019. Recommendation Systems for News Articles at the BBC.. In *INRA@ RecSys*. 44–52.

[27] Bibek Paudel and Abraham Bernstein. 2021. Random walks with erasure: Diversifying personalized recommendations on social and information networks. In *Proceedings of the Web Conference 2021*. 2046–2057.

[28] Bibek Paudel, Fabian Christoffel, Chris Newell, and Abraham Bernstein. 2016. Updatable, accurate, diverse, and scalable recommendations for interactive applications. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 1 (2016), 1–34.

[29] Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2023. CodaLab Competitions: An Open Source Platform to Organize Scientific Challenges. *Journal of Machine Learning Research* 24, 198 (2023), 1–6.

[30] Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. 2021. Post-processing for individual fairness. *Advances in Neural Information Processing Systems* 34 (2021), 25944–25955.

[31] Shaina Raza, Syed Raza Bashir, and Usman Naseem. 2022. Accuracy meets diversity in a news recommender system. In *Proceedings of the 29th International Conference on Computational Linguistics*. 3778–3787.

[32] Shaina Raza and Chen Ding. 2020. A regularized model to trade-off between accuracy and diversity in a news recommender system. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 551–560.

[33] Holli Sargeant, Eliska Pirkova, Matthias C Kettemann, Marlena Wisniak, Martin Scheinin, Emmi Bevensee, Katie Pentney, Lorna Woods, Lucien Heitz, Bojana Kostic, et al. 2022. Spotlight on Artificial Intelligence and Freedom of Expression: A Policy Manual. *Organization for Security and Co-operation in Europe* (2022).

[34] Annelien Smets, Lien Michiels, Toine Bogers, and Lennart Björneborn. 2022. Serendipity in recommender systems beyond the algorithm: A feature repository and experimental design. In *CEUR Workshop Proceedings*, Vol. 3222. CEUR Workshop Proceedings, 46–66.

[35] Celina Treuillier, Sylvain Castagnos, Evan Dufraisse, and Armelle Brun. 2022. Being diverse is not enough: Rethinking diversity evaluation to meet challenges of news recommender systems. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 222–233.

[36] Saúl Vargas. 2011. New approaches to diversity and novelty in recommender systems. In *Fourth BCS-IRSG Symposium on Future Directions in Information Access (FDIA 2011) 4*. 8–13.

[37] Sanne Vrijenhoek. 2023. Do you MIND? Reflections on the MIND dataset for research on diversity in news recommendations. *arXiv preprint arXiv:2304.08253* (2023).

[38] 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*. 208–219.

[39] Sanne Vrijenhoek, Savvina Daniil, Jorden Sandel, and Laura Hollink. 2024. Diversity of What? On the Different Conceptualizations of Diversity in Recommender Systems. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 573–584.

[40] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a mission: assessing diversity in news recommendations. In *Proceedings of the 2021 conference on human information interaction and retrieval*. 173–183.

[41] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. 2023. In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data* 17, 3 (2023), 1–27.

[42] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind A large-scale dataset for news recommendation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 3597–3606.

[43] Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns* 3, 7 (2022), 100543.