RADio* – An Introduction to Measuring Normative Diversity in News Recommendations

SANNE VRIJENHOEK, Institute for Information Law, University of Amsterdam, The Netherlands
GABRIEL BÉNÉDICT, University of Amsterdam and RTL Nederland B.V., The Netherlands
MATEO GUTIERREZ GRANADA and DAAN ODIJK, RTL Nederland B.V., The Netherlands

In traditional recommender system literature, diversity is often seen as the opposite of similarity and typically defined as the distance between identified topics, categories, or word models. However, this is not expressive of the social science's interpretation of diversity, which accounts for a news organization's norms and values and which we here refer to as *normative* diversity. We introduce RADio, a versatile metrics framework to evaluate recommendations according to these normative goals.

RADio introduces a rank-aware Jensen Shannon (JS) divergence. This combination accounts for (i) a user's decreasing propensity to observe items further down a list and (ii) full distributional shifts as opposed to point estimates. We evaluate RADio's ability to reflect five normative concepts in news recommendations on the Microsoft News Dataset and six (neural) recommendation algorithms, with the help of our metadata enrichment pipeline. We find that RADio provides insightful estimates that can potentially be used to inform news recommender system design.

CCS Concepts: • **Information systems** → **Recommender systems**; • **Social and professional topics** → **Systems analysis and design;**

Additional Key Words and Phrases: News recommendation, diversity, divergence, normative framework

ACM Reference format:
Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, and Daan Odijk. 2024. RADio* – An Introduction to Measuring Normative Diversity in News Recommendations. *ACM Trans. Recomm. Syst.* 3, 1, Article 5 (August 2024), 29 pages.
https://doi.org/10.1145/3636465

5

S. Vrijenhoek and G. Bénédict contributed equally.
This article extends Reference [79]. The additional contributions of this article are and additional section on the effects of metric design choices, extensive discussions of the normative theory the metrics are built on, mathematical proofs, a benchmark analysis with Intra-List Distance, elaborations on the use of KL divergence, an additional table of results with robust estimates (median and IQR), and additional plots for a more granular sensitivity analysis.
This research was partially funded by Bertelsmann SE & Co. KGaA; by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research (https://hybrid-intelligence-centre.nl); the SIDN Fonds (https://www.sidnfonds.nl/projecten/algorithms-for-freedom-of-expression-and-a-well-informed-public).
Authors' addresses: S. Vrijenhoek, Institute for Information Law, University of Amsterdam, Amsterdam, The Netherlands; e-mail: s.vrijenhoek@uva.nl; G. Bénédict, University of Amsterdam and and RTL Nederland B.V., Amsterdam, The Netherlands; e-mail: gabriel.benedict@rtl.nl; M. Gutierrez Granada and D. Odijk, RTL Nederland B.V., Hilversum, The Netherlands; e-mails: mateo.gutierrez.granada@rtl.nl, daan.odijk@rtl.nl.

2770-6699/2024/08-ART5 $15.00
https://doi.org/10.1145/3636465

1 INTRODUCTION

For centuries, the interplay between journalists and news editors has shaped how news items are created and how they are shown to their readers [92]. With the digitization of society, much has changed: While before, people would typically limit themselves to reading one type of newspaper, they now have a wealth of information available at the click of a button [68]—more than anyone could possibly be expected to read or make sense of. News recommender systems can filter the enormous amount of information available to just those news items that are in some way interesting or relevant to their users [8, 56]. The use of news recommender systems is widespread, not just for *personalized* news recommendations, but also to automatically populate the front page of a news website [57] or present the reader of a particular news article with other articles about the same topic, but from a different perspective [58]. The use of news recommender systems has a wide range of benefits. They can increase engagement [59] and help raise informed citizens [30]. A news recommender system may broaden the horizons of their users through diverse recommendations, including items different from what they are used to or expect seeing. They could even foster tolerance and understanding [31, 72] and counter so-called filter bubbles or echo chambers [56, 62].

To realize the potential benefits of news recommender systems, much attention has been given to generating recommendations that reflect the user's interests and preferences [42]. However, with news recommenders taking over the role of human editors in news selection, they are becoming gatekeepers in what news is shown to audiences and have thus a democratic role to play in society [48]. As such, their evaluation has different requirements than those of other types of recommender systems [4, 5, 82, 85]. Recent controversies have shown that merely optimizing for click-through rates and engagement may promote sensationalist content [74] and is particularly conducive to the spread of misinformation.[1] This observation is not limited to the academic literature—an increasing number of media organizations, both public service and commercial, have acknowledged the difficulties in translating their editorial norms into concrete metrics that can inform recommender system design [9, 34]. News recommender systems exist in a complex space consisting of many different areas and disciplines, each with their own goals and challenges; think of balancing diversity and accuracy [61], financial incentives [12], nudging [54] or even identifying user preferences [6, 53] and biases [84]. In this article, we focus on the process of translating normative theory (i.e., what it means for a recommendation to be diverse) into metrics that are usable and understandable for both technical and editorial purposes. We build on the work of Helberger [36], who provides a theoretical foundation for conceptualizing diversity, and of Vrijenhoek et al. [80], who propose a new set of metrics (DART) that reflect this theory. The DART metrics represent a first step towards a normative interpretation of diversity in news recommendations. We identify a number of improvements on these metrics: more consideration for the theory of metrics and distance functions, generalization to other normative concepts, unification under one framework, and rank-awareness. In this article, we focus on the mathematical aspects of a rank-aware metric, versatile to different normative concepts. We refer to our framework as the ***Rank-Aware Divergence metrIcs to measure nOrmative diversity*** **(RADio)**.

Our contribution consists of a diversity metric that is (i) versatile to any normative concept and expressed as the divergence between two (discrete) distributions; (ii) rank-aware, taking into account the position of an item in a recommendation set; and (iii) mathematically grounded in distributional divergence statistics. We demonstrate the effectiveness of this formulation of the metrics by defining a **natural language processing (NLP)** metadata enrichment pipeline (e.g., sentiment analysis, named entity recognition) and running it against the MIND dataset [90].

[1]See, for example, the alleged role Facebook played in the storming of the Capitol: https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/

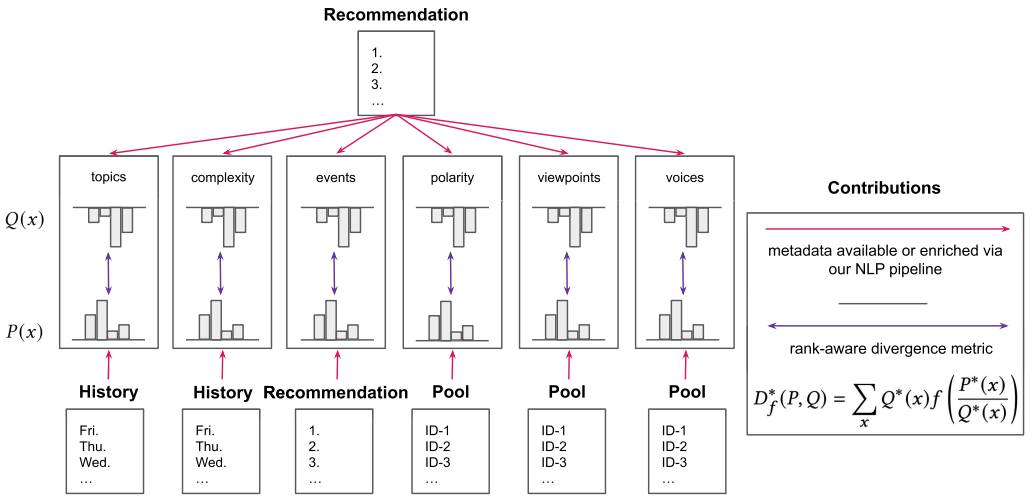


Fig. 1. Comparing discrete diversity distributions in the context of news recommendations. First, metadata is collected in the news dataset or retrieved via our NLP pipeline (red). Discrete distributions of that metadata are then compared via a rank-aware divergence metric (purple). Recommendation set Q and the context articles P are compared with rank-aware f-Divergence.

Figure 1 illustrates the operationalization. The pipeline and the code produced for metadata enrichment and metric computation are available online.[2] The goal of RADio is not to serve as thresholds or strict guidelines for "diverse recommendations," but to provide developers of recommender systems with the tools to evaluate their systems on normative principles.

The extended version of the paper for *Transactions on Recommender Systems* contains additional experiments, highlighting the effect of different design choices when operationalizing a metric. It also contains more extensive discussion of the normative concepts at the origin of the metrics, and the paper is as such less reliant on prior knowledge of Vrijenhoek et al. [80]. We added the **Intra List Distance (ILD)** baseline and included it in our analysis to underline the strengths of the approach taken in RADio; included proofs that KL and JS Divergences are f-Divergences; and added two more analytical plots that provide more in-depth comparisons (with appropriate discussions). Last, we added an additional section to the Discussion on how practitioners looking to put the metrics in production could approach this.

2 RELATED WORK

We first highlight recent work on the formal mathematical work on diversity in news recommendation before citing related work on the normative aspect of diversity. Finally we describe the gap that exists between descriptive and normative diversity.[3]

2.1 Descriptive (General-purpose) Diversity

Diversity is a central concept in Information Retrieval literature [19, 66], albeit with a different interpretation than the normative diversity described in the previous section. During the development of news recommender systems, there is currently a large focus on the predictive power

[2] https://github.com/svrijenhoek/RADio

[3] This dichotomy is oftentimes referred to as normative (*what ought to be*) and positive (*what is*) statements [38] but can easily be confused with concepts such as positivenegative examples in Machine Learning. We thus opt for the more explicit normative/descriptive duo.

of an algorithm. However, this may unduly promote content similar to what a user has interacted with before and lock them in loops of "more of the same." To tackle this, "diversity" is introduced, which is typically defined as the "opposite of similarity" [11]. 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)** [11, 15, 21, 25, 26, 41, 52, 77]: mean pairwise dissimilarity between recommended item lists. ILD requires the specification of a distance function between lists and thus leaves it up to interpretation as to what it means for two lists to be distant. In theory, it could still be interpreted with a metric that accounts for the presence of different sources or viewpoints [27]. However, in practice, diversity is most often implemented as a descriptive distance metric such as cosine similarity between two bag-of-words models or word embeddings [46, 52].

Other popular "beyond-accuracy" metrics related to diversity are novelty (how different is this item from what the user has seen in the past), serendipity (is the user positively surprised by this item), and coverage (what percentage of articles are recommended to at least one user). These metrics can be taken into account at different points in the machine learning pipeline [46, 91]. One can optimize for these descriptive notions of diversity (i) before training, by clustering users based on their profile diversity with JS divergence [29], (ii) directly at training time (e.g., for learning-to-rank [10, 15, 77], collaborative filtering [64], graphs [32, 63], or bandits [23, 94]), (iii) by re-ranking a recommendation set and balance diversity vs. relevance [18] or popularity vs. relevance [17], and (iv) by defining a post-recommendation metric to measure diversity for each recommendation set or at user-level (e.g., the generalist-specialist score [2, 83]). With any of these four methods, a tradeoff must be made between the relevance of a recommendation issued to users and the level of descriptive diversity, though there have also been studies indicating that increasing diversity does not necessarily need to negatively affect relevance [52]. Nevertheless, this encouraged recent efforts in training neural-based recommenders that explicitly make a tradeoff between accuracy and diversity [65]. Also recently, there have been studies that differentiate between diversity needs of users [93].

2.2 Normative Diversity

Diversity is extensively discussed as a normative concept in literature and has a role in many different areas of science [50, 71], spanning from ecological diversity to diversity as a proxy for fairness in machine learning systems [55]. While these interpretations of diversity are often related, they do not fully cover the nuances of a diverse *news* recommender system, the work on which stems from democratic theory and the role of media in society. Following Helberger [36], we define a normative diverse news recommendation as one that succeeds in informing the user and supports them in fulfilling their role in democratic society. Out of the many theoretical models that exist in literature, Helberger [36] describes four different models from the normative framework of democracy, each with a different view on what it means to properly inform citizen. The **Liberal** model states that eventually the users themselves know what is best for them and primarily aims to enable personal development and autonomy. A recommender system following the Liberal model would have a strong focus on aligning with users' personal preferences. Most current recommender systems are, inadvertently, in line with this model. In comparison, the **Participatory** model takes on a more paternalistic role, as it values the common good over the individual and endeavors to know what is good for both. A recommender system following the Participatory model aims to enable users to fulfill their role as active citizens in a democratic society, informing them of important events in a way that is suitable to their needs and preferences. The **Deliberative** model fosters discussion and debate by equally presenting different viewpoints and opinions in a rational and neutral way, with the eventual goal of either reaching a consensus or agreeing on different values. Deliberative recommender systems have a strong focus on the

Table 1. Overview of the Different Models and Expected Value Ranges for Each Metric

	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

It should be noted that a high score should be interpreted as high *divergence*; as such, a high score does not necessarily mean a better score.

representation of a plurality of sources, voices, and opinions. Last, the **Critical** model aims to challenge the status quo and to inspire the readers to take action against existing injustices in society. A Critical recommender system provides a platform to voices that would otherwise go unheard and favors emotion-driven and activating content. However, it should *not* promote hate speech or conspiracy theories; those with different opinions should be seen as opponents or adversaries, not enemies [67].

For more details regarding the different models and what a recommender system following each of these models would look like, we refer to Helberger [36]. Since none of these models is inherently better or worse than the other, which model is followed is a decision that needs to be made by the media organization itself and should be in line with their norms and values.

Conceptualized based on these models, the DART metrics [80] take a first step towards normative diversity for recommender systems and reflect the nuances of the different democratic models described above. See Table 1 for an overview of the DART metrics and their interplay with democratic models. **Calibration** aims to reflect to what extent a recommendation is tailored to a user's personal preferences. This can be considered both in terms of an article's content and express, for example, which topics a user is interested in, but also in terms of its complexity. This would then differentiate information needs for different users: A user can consume more complex material on topics they are an expert in than on ones they are unfamiliar with. **Fragmentation** expresses to what extent there is a common public sphere, where users are frequently recommended the same items and therefore have a similar understanding of the content in the system. This metric is conceptually most related to the concept of the filter bubble. Next, **Activation** expresses the tone of articles: Are they written in a neutral and rational tone, or one that is more emotional and activating? While neutrality is a core principle of quality journalism, a softer approach can be more affective and increase empathy. The last two metrics correspond to different aspects of viewpoint diversity. **Representation** aims to reflect the *content* of the different viewpoints in the recommendations. Here, it is important to think about how these viewpoints should be distributed across recommendations: for example, uniformly (a.k.a. equal representation), in proportion to the occurrence of viewpoints in the medium (a.k.a. reflective representation) or in inverse proportion to the occurrence of viewpoints in the medium (a.k.a. inverse representation). However, **Alternative Voices** considers the viewpoint *holder*, and more specifically whether this person belongs to a minority group or not. One could think of Alternative Voices in the context of algorithmic fairness and aim to reflect an equal share of Alternative Voices compared to the overall supply of data.

When considered together, the metrics aim to reflect the characteristics of the different democratic models, summarized in Table 1. The **Liberal** model focuses on personal development and preferences. It is therefore most concerned with aspects of Calibration and tolerant of a high degree of Fragmentation. The **Participatory** model, with its focus on a public sphere and informing audiences, aims for the opposite value in Fragmentation, but low divergence on Calibration in terms of article complexity. Viewpoints should reflect society, with a larger share for more

dominant viewpoints. In comparison, with its focus on promoting rational discussion and debate, the **Deliberative** model favors a system with an equal representation of different viewpoints, presented in a rational tone. Last, the **Critical** model has a strong focus on promoting minority voices, and viewpoints recommended should therefore be the inverse of what is most commonly heard.

There is a lot of discussion to be had on this topic. The four models described are the ones most frequently discussed [36] and not exhaustive in the goals one could have with a news recommender system. The metrics following them may also differ, in their definition or operationalization, depending on the use case. The main take-away is that there is no single Golden Standard for diversity and what makes for a good news recommender system, but rather that this is very much dependent on what you wish to accomplish with the system. This is ultimately a discussion that needs to be had with the different stakeholders involved in recommender system design [69]. For example, Hada et al. [35] did an implementation of Fragmentation and Representation in social media conversations, focusing on Daylight Savings Time and the more polarized topic of immigration. While Fragmentation and Representation are not directly tied in one of the democratic models, they found that the two metrics combined were capable of expressing the nuances of the discourse and that the effects were stronger for the polarized discussion than for the control group.

2.3 The Gap between Normative and Descriptive Diversity

The descriptive diversity metrics described in Section 2.1 are general-purpose and meant to be applicable in all domains of recommendation. However, in their simplicity, a large gap can be observed between this interpretation of diversity and the social sciences' perspective on media diversity that is detailed in Section 2.2. In their comprehensive work on the implementation of media diversity across different domains, Loecherbach et al. [50] note that there is "little to no overlap between concepts and operationalizations (of diversity) used in the different fields interested in media diversity." As such, a recommendation that would score high on diversity according to traditional information retrieval-based metrics [19, 66], may not be considered to be diverse according to the criteria maintained by newsroom editors. This notion is also shared in industry: Boididou et al. [9] (BBC) say that *"We need to devise appropriate metrics to capture the quality criteria and values expressed in our editorial guidelines,"* whereas Grün and Neufeld [34] (ZDF) note that *"the evaluation of KPIs and public values is more complex than the measurement of common machine learning metrics applied to recommendation algorithms."* To move the issue forward, both Loecherbach et al. [50] and Bernstein et al. [7] call for truly interdisciplinary research in bridging this gap, where Bernstein et al. [7] argue for close collaboration between academia and industry and the foundation of joint labs. This work is a step in that direction, as we provide a versatile and mathematically grounded rank-aware metric that can be used by practitioners to monitor their normative goals.

3 OPERATIONALIZING NORMATIVE DIVERSITY FOR NEWS RECOMMENDATION

With our RADio framework, we further refine the DART metrics that were defined by Vrijenhoek et al. [80] to resolve a number of the shortcomings of the metrics' initial formalizations. In their current form, each of the metrics has different value ranges; for example, *Activation* has a value range $[-1, 1]$, where a higher score indicates a higher degree of activating content, and *Calibration* has a range of $[0, \infty]$, where a lower score indicates a better Calibration. These different value ranges reduce the interpretability of the metrics, making them harder to explain and as such less likely to be adopted by news editors. Furthermore, the proposed metrics do not take the position of an article in a recommendation list into account. News recommendations are ranked lists of

articles that are typically presented to users in such a way that the likelihood of a recommended article to be considered by the user decreases further down the ranking. As such, in the evaluation of the diversity of the recommender system, we should also account for the position of an article in the recommendation ranking, rather than considering the set as a whole.

Thus, the two major challenges that we seek to address are that (i) scores should be comparable between the metrics and across recommendation systems, and (ii) scoring of both unranked and ranked sets of recommendations should be possible. In this section, we first detail these requirements (Section 3.1), then describe how we reformulate the metrics to each use the same divergence-based approach (Section 3.2). We then add the rank-aware aspect to the metrics (Section 3.3) before applying them to the five concrete DART metrics (Section 3.4).

3.1 Requirements

We first enunciate the classical definition of a distance metric before specifying three desirable metric criteria for news recommendations. Take a set X of random variables and $x, y, z \in X$, then a metric D is a proper distance measure if $D(x, y) = 0 \Leftrightarrow x = y$, $D(x, y) = D(y, x)$ and $D(x, y) \leq D(x, z) + D(z, y)$. These are, respectively, the axioms of *identity*, *symmetry* and *triangle inequality*, which express intuitions about concepts of distance [60].

We add that our distance measure should (i) be bounded by $[0; 1]$, for comparisons of different recommendation algorithms; (ii) be unified, to fairly consider different diversity aspects (as opposed to, e.g., using weighted averages or maxima in Reference [20]); and (iii) allow for discrete rank-based distribution sets, to fit the ranked recommendation setting.

3.2 f-Divergence

We model the task of measuring diversity as a comparison between probability distributions: *the difference in distribution between the recommendations* (Q) *and its context* (P). Each diversity metric prescribes its own Q and P. The elements in the distribution Q can be recommendation items (cf. Calibrated Recommendations [70]) but can also be higher-level concepts, such as distributions of topics and viewpoints. The context P may refer to either the overall supply of available items, the user profile, such as the reading history or explicitly stated preferences, or the recommendations that were issued to other users (see Figure 1). Intuitively, when P is linked to the same user as Q, we measure within user diversity (e.g., towards preventing getting locked in "filter bubbles"). When P is linked to another user than the one linked to Q, we measure diversity across users (e.g., monitoring diversity of viewpoints represented across personalized homepages). In the following, we formalize the role of P and Q in two different metric settings, starting with the simple and common KL divergence metric, before presenting its refinement, the Jensen-Shannon divergence, as our preferred metric.

3.2.1 Kullback-Leibler Divergence. For every random variable there is an inherent level of "uncertainty" in its possible outcomes. In Information Theory, the concept of entropy is a way to describe the average level of "uncertainty" a random variable carries. The entropy of a discrete random variable X with possible outcomes $x_1, \ldots, x_n$ and corresponding probabilities $P(x_1), \ldots, P(x_n)$ is defined as $H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$. This value describes the average level of "information" required to describe that random variable [75]. The concept of relative entropy or **KL (Kullback–Leibler)** divergence [45] between two probability mass functions P and Q (here, a recommendation and its context) is defined as:

$$D_{\mathrm{KL}}(P, Q) = -\sum_{x \in \mathcal{X}} P(x) \log_2 Q(x) + \sum_{x \in \mathcal{X}} P(x) \log_2 P(x). \tag{1}$$

This is often also expressed as $D_{\mathrm{KL}}(P,Q) = H(P,Q) - H(P)$, with $H(P,Q)$ the cross-entropy of P and Q, and $H(P)$ the entropy of P. Both cross-entropy and KL divergence can be thought of as measurements of how far the probability distribution Q is from the reference probability distribution P. Cross-entropy however, does not guarantee the *identity* property. That is, when $P = Q$, it holds that $H(P,Q) = H(P,P) = H(P) > 0$: The cross-entropy of P with itself (the entropy of P) is never 0. KL Divergence, though, does satisfy the *identity* property, bringing forward a good argument to prefer KL divergence over cross-entropy. However, neither of them are truly proper metrics, as neither fulfill the requirements for *symmetry* and *triangle inequality*. This can be resolved by further refining KL Divergence.

3.2.2 Jensen–Shannon Divergence. A succession of steps from KL divergence leads to **Jensen-Shannon (JS)** divergence. KL divergence was first turned symmetric [40] and then upper bounded [49], to lead to

$$D_{\mathrm{JS}}(P,Q) = -\sum_{x\in\mathcal{X}} \frac{P(x)+Q(x)}{2} \log_2\left(\frac{P(x)+Q(x)}{2}\right) + \frac{1}{2}\sum_{x\in\mathcal{X}} P(x)\log_2 P(x) + \frac{1}{2}\sum_{x\in\mathcal{X}} Q(x)\log_2 Q(x). \tag{2}$$

When the base 2 logarithm is used, the JS divergence bounds are $0 \le D_{\mathrm{JS}}(P,Q) \le 1$. Additionally, Endres and Schindelin [28] show that $\sqrt{D_{\mathrm{JS}}}$ is a proper distance that fulfills the identity, symmetry, and the triangle inequality properties. When we refer to D_{JS} or JS divergence below, we therefore implicitly refer to the square root of the JS formulation with log base 2.

3.2.3 f-Divergence. Liese and Vajda [47] defined *f-Divergence* [D_f]: a generic formulation of several divergence metrics. Among them are the JS and KL divergences.[4] Further along the text, we use D_f as a shorthand notation for KL and JS divergences. D_f in discrete form is

$$D_f(P,Q) = \sum_x Q(x) f\left(\frac{P(x)}{Q(x)}\right), \tag{3}$$

where $f_{\mathrm{KL}}(t) = t\log t$ and $f_{\mathrm{JS}}(t) = \frac{1}{2}\left[(t+1)\log\left(\frac{2}{t+1}\right) + t\log t\right]$. To avoid misspecified metrics (i.e., division by zero) [70], we write $\overline{P}$ and $\overline{Q}$:

$$\overline{Q}(x) = (1-\alpha)Q(x) + \alpha P(x) \quad \overline{P}(x) = (1-\alpha)P(x) + \alpha Q(x), \tag{4}$$

where α is a small number close to zero. $\overline{P}$ prevents artificially setting D_f to zero when a category (e.g., a news topic) is represented in Q and not in P. In the opposite case (when a category is represented in P and not in Q), $\overline{Q}$ avoids zero divisions. For the entire probabilistic distributions $\overline{P}$ and $\overline{Q}$ to remain proper statistical distributions, we normalize them to ensure $\sum_x \overline{P}(x) = \sum_x \overline{Q}(x) = 1$. In the following sections, P and Q will implicitly refer to $\overline{P}$ and $\overline{Q}$ to avoid notation congestion.

Below, we verify by simple plug-in and factorization that both the Kullback-Leibler and the Jensen-Shannon Divergence are f-Divergences.

Theorem 3.1. *The Kullback-Leibler Divergence is an f-Divergence*

[4] f-Divergence accommodates for other divergence metrics that are out of scope of this research, such as the Hellinger divergence and the Pearson divergence [47].

Proof. Starting from Equation (1),

$$\begin{aligned} D_{\mathrm{KL}}(P,Q) &= -\sum_{x\in\mathcal{X}} P(x)\log_2 Q(x) + \sum_{x\in\mathcal{X}} P(x)\log_2 P(x) \\ &= \sum_{x\in\mathcal{X}} Q(x)\left[\frac{P(x)}{Q(x)}\left(\log_2 P(x) - \log_2 Q(x)\right)\right] \\ &= \sum_{x\in\mathcal{X}} Q(x) f_{KL}\left(\frac{P(x)}{Q(x)}\right). \end{aligned}$$

□

Theorem 3.2. *The Jensen-Shannon Divergence is an f-Divergence*

Proof. Starting from Equation (2),

$$\begin{aligned} D_{\mathrm{JS}}(P,Q) &= -\sum_{x\in\mathcal{X}} \frac{P(x)+Q(x)}{2}\log_2\left(\frac{P(x)+Q(x)}{2}\right) + \frac{1}{2}\sum_{x\in\mathcal{X}} P(x)\log_2 P(x) + \frac{1}{2}\sum_{x\in\mathcal{X}} Q(x)\log_2 Q(x) \\ &= \sum_{x\in\mathcal{X}} Q(x)\frac{1}{2}\left[\left(\frac{P(x)}{Q(x)}+1\right)\log_2\left(\frac{2}{\left(\frac{P(x)}{Q(x)}+1\right)}\right) + \frac{P(x)}{Q(x)}\log_2\left(\frac{P(x)}{Q(x)}\right)\right] \\ &= \sum_{x\in\mathcal{X}} Q(x) f_{JS}\left(\frac{P(x)}{Q(x)}\right). \end{aligned}$$

□

3.3 Rank-aware f-Divergence Metrics

Our ranked recommendation setting (characteristic (iii) above) motivates a further reformulation of our f-Divergence metric. It is well entrenched in **Learning To Rank (LTR)** literature [73, 95] and by extension in conventional descriptive diversity metrics [15] that a user is a lot less likely to see items further down a recommended ranked list (i.e., diminishing inspection probabilities). Note that the ranking oftentimes reflects relevance to the user, but it is not always the case for news (e.g., editorial layout of a news homepage).

We extend our metrics with an optional discount factor for P and Q to weigh down the importance of results lower in the ranked recommendation list. The ranking relevancy metrics **Mean Reciprocal Rank (MRR)** and **Normalized Discounted Cumulative Gain (NDCG)** are popular rank-aware metrics for LTR [16, 39], in particular for news recommendation [90]. In line with the LTR literature, we first define the discrete probability distribution of a ranked recommendation set Q^*, given each item i in the recommendation list R:

$$Q^*(x) = \frac{\sum_i w_{R_i}\mathbb{1}_{i\in x}}{\sum_i w_{R_i}}, \tag{5}$$

where w_{R_i}, the weight of a rank for item i, can be different depending on the discount form. For MMR, $w_{R_i} = 1/R_i$, for NDCG, $w_{R_i} = 1/\log_2(R_i+1)$. When $w_{R_i} = 1$, Q^* is not discounted (i.e., $Q^* = Q$).

In news recommendation, the *sparsity bias* plays a predominant role: Users will interact with a small fraction of a large item collection, such as scrollable news recommendation websites [43]. We thus opt for weighing based on MRR rather than NDCG, because it applies a heavier discount along the ranking than NDCG. Note that the latter is said to be more suited for query-related rankings, where the user has a particular information need related to a query and thus higher propensity to scroll down a page [16].

The context distribution P is discounted in the same manner when it is a ranked recommendation list. When P is a user's reading history (see Figure 1), the discount on P increases with time: Articles read recently are weighted higher than articles read longer ago. There are situations when rank-awareness is not applicable, for example, when P is the entire pool of available articles.[5] With rank-aware Q^* and optionally rank-aware P^*, we formulate RADio, our rank-aware f-Divergence metric:

$$D^*_{f_{\text{JS}}}(P, Q) = \sum_x Q^*(x) f_{\text{JS}} \left(\frac{P^*(x)}{Q^*(x)} \right), \tag{6}$$

$Q^*(x)$ and $P^*(x)$ accommodate for multiple situations: For example, $Q^*(c|R)$ is the rank-aware distribution of news categories c over the recommendation set R. In the following, we specify $P^*(x|\cdot)$ and $Q^*(x|\cdot)$ in accordance to each normative concept of interest for our universal metric.

3.4 Normative Diversity metrics as Rank-aware f-Divergences

In this section, we describe the RADio formalization of the general f-Divergence formulation above and apply it to the five DART metrics. We leave the exact implementation of the metrics in practice for a particular open news recommendation dataset to the next section. More formally, we define the following global parameters:

- 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. In this work, metrics are defined for a specific user at a certain point in time, therefore, R implicitly refers to R^u, unless stated otherwise. While this section contains some contextualization of the DART metrics [80], the original paper contains further normative justifications.

Calibration (Equation (7)) measures to what extent the recommendations are tailored to a user's preferences. The user's preferences are deduced from their reading history (H). Calibration can have two aspects: the divergence of the recommended articles' *categories* and *complexity*. The former is expected to be extracted from news metadata and thus categorical by nature, the latter is a binned (categorical) probabilistic measure extracted via a language model. As such, we compare $P^*(c|H)$, the rank-aware distribution of categories or complexity score bins c over the users' reading history, and $Q^*(c|R)$ the same in the recommendations issued to the user.

Fragmentation (Equation (8)) reflects to what extent we can speak of a common public sphere or whether the users exist in their own bubble. We measure Fragmentation as the divergence between every pair of users' recommendations. Here, we consider $P^*(e|R^u)$ as the rank-aware distribution of news events e over the recommendations R for user u, and $Q^*(e|R^v)$ the same but for user v. KL Divergence is asymmetric (see Section 3.2.1), which means that its outcome differs, depending on which user's recommendation is chosen as the target and which as the reference distribution. To avoid this, we compute the Fragmentation score as the average of KL Divergences with switched parameters. JS divergence is already symmetric and is thus implemented as for the other metrics. In theory, Fragmentation requires a user's recommendation to be compared to those of all other users. This is not feasible with a sizeable dataset and the requirement of a reasonable compute time. Instead, we opt to randomly sample user pairs.

[5]There are several features along which such a pool of data could be ranked besides recency, such as the popularity during the last hour, day, or week. As this is an editorial decision, we remain agnostic as to the choice of that feature and refrain from ranking, though it remains possible in theory.

Activation (Equation (9)) Most off-the-shelf sentiment analysis tools analyze a text and return a value $(0, 1]$ when the text expresses a positive emotion, a value $[-1, 0)$ when the expressed sentiment is negative, and 0 if it is completely neutral. The more extreme the value, the stronger the expressed sentiment is. As proposed in Reference [80], we use an article's absolute sentiment score as an approximation to determine the height of the emotion and therefore the level of Activation expressed in a single article. This then yields a continuous value between 0 and 1. $P(k|S)$ denotes the distribution of (binned) article Activation score k within the pool of items that were available at that point (S). $Q^*(k|R)$ expresses the same, but for the binned Activation scores in the rank-aware recommendation distribution.

Representation (Equation (10)) aims to approximate a notion of viewpoint diversity (e.g., mentions of political topics or political parties), where the viewpoints are expressed categorically. Here p refers to the presence of a particular viewpoint, and $P(p|S)$ is the distribution of these viewpoints within the overall pool of articles, while $Q^*(p|R)$ expresses the rank-aware distribution of viewpoints within the recommendation set.

Alternative Voices (Equation (11)) is related to the Representation metric in the sense that it also aims to reflect an aspect of viewpoint diversity. Rather than focusing on the content of the viewpoint, it focuses on the viewpoint holder, and specifically whether they belong to a "protected group" or not. Examples of such protected/unprotected groups could be non-male/male, non-white/white, and so on.[6] This approach is based on the implementation of balanced neighborhoods in recommender systems [14]. With m, we refer to the distribution of protected vs. non-protected groups, with $m \in \{Minority, Majority\}$. $P(m|S)$ and $Q^*(m|R)$ refer to the distribution of these groups in the pool of available articles and rank-aware recommendation distribution, respectively.

Below is a summary of the formalization of DART with the RADio framework, the notation of which is defined in this section. In the next section, we show how to retrieve the necessary features from an example news dataset:

$$Calibration = Cal\Big(P^*(c|H), Q^*(c|R)\Big) = \sum_c Q^*(c|R) f_{\text{JS}}\left(\frac{P^*(c|H)}{Q^*(c|R)}\right), \tag{7}$$

$$Fragmentation = Frag\Big(P^*(e|R^u), Q^*(e|R^v)\Big) = \sum_e Q^*(e|R^v) f_{\text{JS}}\left(\frac{P^*(e|R^u)}{Q^*(e|R^v)}\right), \tag{8}$$

$$Activation = Act\Big(P(k|S), Q^*(k|R)\Big) = \sum_k Q^*(k|R) f_{\text{JS}}\left(\frac{P(k|S)}{Q^*(k|R)}\right), \tag{9}$$

$$Representation = Rep\Big(P(p|S), Q^*(p|R)\Big) = \sum_p Q^*(p|R) f_{\text{JS}}\left(\frac{P(p|S)}{Q^*(p|R)}\right), \tag{10}$$

$$AlternativeVoices = AltV\Big(P(m|S), Q^*(m|R)\Big) = \sum_m Q^*(m|R) f_{\text{JS}}\left(\frac{P(m|S)}{Q^*(m|R)}\right). \tag{11}$$

4 EXPERIMENTAL SETUP

To demonstrate RADio's potential effectiveness, we developed an NLP pipeline to retrieve input features to the metrics in Section 3.4 and ran them on the MIND dataset, an open source dataset made available by Microsoft News. The MIND dataset [90] contains the interactions of 1 million randomly sampled and anonymized users with the news items on MSN News between October 12 and November 22, 2019. Each interaction contains an impression log, listing which articles

[6]For more examples, see the UK 2010 Equality Act: https://www.legislation.gov.uk/ukpga/2010/15/part/2/chapter/1

were presented to the user, which were clicked on, and the user's reading history. The MIND dataset was published accompanied by a performance comparison on news recommender algorithms trained on this dataset,[7] including news-specific neural recommendation methods NPA [88], NAML [87], LSTUR [1], and NRMS [89]. It was shown that these algorithms outperform general-purpose ones [90] or common collaborative filtering models (such as **alternating least squares (ALS)**), in particular due to the short lifespan of news items [33]. These algorithms are trained on the impression logs to predict which items the users are most likely to click on. For the purpose of this article, we will evaluate these neural recommendation methods with the RADio framework (on the DART metrics) and compare their performance with two naive baseline methods, based on a reasonable set of candidates (the original impression log): a random selection and a selection of the most popular items, where the popularity of the item is approximated by the number of recorded clicks in the dataset. There are a number of limitations to the usage of the MIND dataset (see Section 8 and Vrijenhoek [78]). However, at the time of this writing, it is the only large open source news dataset with sufficiently granular logs for such experiments. Future work would benefit from a dataset that is explicitly prepared for normative diversity, bypassing the need for an additional preprocessing pipeline. It is not the goal of this article to improve on the identification and extraction of the relevant parameters proposed in Vrijenhoek et al. [80], but rather to express their outcomes in a meaningful way. We scrape articles via the URLs provided in the MIND dataset. Each article's metadata is enriched in five steps:

(1) **Complexity analysis** – Each item is assigned a complexity score based on the Flesch-Kincaid reading ease test [44], implemented in the Python module py-readability-metrics [22]. Complexity is then discretized into bins to accommodate for the discrete form of D_f^*.
(2) **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. This way, we add a level of analysis between individual news items and higher-level categories (see Section 3.4). We use a TF-IDF-based unsupervised clustering algorithm based on cosine similarity and a three days moving window, following the setup of Trilling and van Hoof [76].
(3) **Sentiment analysis** – Using the textBlob open source NLP library, we assign each article a sentiment polarity score [51]. Our focus is on the relative neutrality of articles; we thus take the absolute value of the negative/positive polarity score.
(4) **Named entity recognition** – Using spaCy, we identify the people, organizations, and locations mentioned in the text [37] and count their frequency.
(5) **Named entity augmentation** – For the entities identified in the text in the previous step, we attempt to link them to their Wikidata[8] entry through fuzzy name matching to figure out if they are politicians, or in the case of organizations, political parties.[9]

Linking back to the DART metrics, this means that for Calibration, we use the article category that is supplied in the dataset and the complexity score calculated with the Flesch-Kincaid reading ease test [44]. We compare those to what was in the users' reading history. For Fragmentation, we compare the different news stories, identified through story clustering, that are recommended to different users. The Activation score of an article is determined by the absolute sentiment polarity score, and for Representation, we look at the distribution of political actors as identified through Named Entity Linking. For Activation, Representation, and Alternative Voices, we compare

[7] Code available at https://github.com/microsoft/recommenders/tree/main/examples/00_quick_start
[8] https://www.wikidata.org/
[9] In the future, one could also use additional data available on Wikidata for further refinement of the metrics, such as gender or place of birth/ethnicity for persons, industry type for organizations, or country code for locations.

Table 2. Overview of the Implementation Approach for Different Methods

	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 Reference [16]
Activation	Available articles	Continuous	Activation 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.[10]

Numbers in bold correspond to the corresponding steps in the metadata enrichment pipeline presented above.

the recommendations to the available pool of data, here interpreted as the set of items the recommender system could make its selection from (see also Table 2). Since RADio computes the average of all $\{P, Q\}$ pairs, we retrieve confidence intervals over paired distances, too, as illustrated in the sensitivity analyses in the next section. In a traditional model evaluation setting, it would be desirable to generate confidence intervals via different model seeds or cross-validation splits. We refrain from doing this for our metric evaluation, as this would introduce a multidimensional confidence interval (e.g., over $\{P, Q\}$ pairs and over model seeds). It should be noted that this pipeline is an imperfect approximation and that each metric individually would benefit from more sophisticated methods. Table 2 links the numbered list above with the DART metrics. It provides an overview of the different metrics and their respective context distribution P over normative concepts. The code for this implementation is available online.[11] We evaluate the outcome of our RADio framework for different recommender strategies (LSTUR, NAML, NPA, NRMS, most popular and random), with both KL Divergence and Jensen-Shannon as divergence metrics, with and without discounting for the position in the recommendation and at different ranking cutoffs. Additionally, we calculate the **intra-list diversity (ILD)** as a typical descriptive diversity metric. We used cosine similarity as a distance metric [11] and TF-IDF [13] of the article's body (text) as article representations.[12]

5 EXPERIMENTAL RESULTS

Having described our methodology and experimental setup around the operationalization of DART metrics, we analyze the results of the experiments on MIND. We separate descriptive analysis of the results in this section from the normative interpretation of the metrics in Section 8. As the default method, we choose to implement RADio with rank-awareness and JS divergence with a rank cutoff @N, which corresponds to the entire ranking list. After commenting on

[10]We acknowledge that this is a largely oversimplified approach towards identifying minority voices. This is a complex question that cannot be resolved to satisfaction within the scope of this article.

[11]https://github.com/svrijenhoek/RADio/

[12]To make calculating ILD on all pairs of articles computationally feasible, we used Spark [96] to load the entire matrix of all intra-list article pairs and parallelize the distance calculation. As this approach was incompatible with the way we did our random selection, which happened later in the pipeline, we have at this time no ILD score to report for the random recommender.

Table 3. Results for Our RADio Framework for Recommendation Algorithms on the MIND Dataset

Algorithm	Calibration (topic)	Calibration (complexity)	Fragmentation	Activation	Representation	Alternative voices	NDCG	ILD
LSTUR	0.5847	0.3632	0.9046	0.1819	0.1261	0.0409	0.4134	0.9837
NAML	0.5709	0.3593	0.8836	0.1842	0.1230	0.0384	0.4091	0.9895
NPA	0.5838	0.3619	0.8979	0.1841	0.1359	0.0390	0.4068	0.9868
NRMS	0.5662	0.3548	0.8872	0.1794	0.1278	0.0362	0.4163	0.9897
Most popular	0.6526	0.3477	0.8923	0.1949	0.1268	0.0342	0.2750	0.9179
Random	0.6636	0.3981	0.9439	0.2715	0.2578	0.0698	0.2949	-

We use our preferred setup: JS divergence with rank-awareness @10. For interpretation of the results, it should be noted that though a *higher* score does imply higher divergence, this does not necessarily mean this is a *better* score. Rather, what it means to be better is dependent on the metric and the model chosen, for which we refer to Table 1.

the overall results, we further motivate this choice with a sensitivity analysis to different hyperparameters. We alter the divergence metric (KL or JS), rank-awareness (with and without a discount), and ranking cutoffs (@n, with $n = 1, 2, 5, 10, 20, N$) for the different recommender models.

Table 3[13] displays results for RADio with rank-aware JS divergence, NDCG, and ILD. NDCG scores are comparable to the results obtained in Reference [90]. Scores for ILD are generally high and show little difference between the neural recommenders. This is in line with the nature of natural language: Without accounting for synonyms or using word embeddings it is not surprising that little similarity would be found between texts. The most popular recommender does differ significantly in terms of ILD compared to the neural recommenders. Explaining this behavior requires a more in-depth look at the articles that have a lot of clicks recorded in the dataset. Given the nature of MSN News, it is likely that these are often sports or lifestyle articles. This would imply a similar topic and thus a common vocabulary and lower diversity. However, the only conclusion that can be drawn from this is that a most popular recommender tends to recommend more items of the same category and is as such not very useful in terms of the requirements for normative diversity, as described in Section 2.2.

For the normative diversity metrics, higher values imply higher divergence scores. Whether high or low divergence is desired depends on the goal of the recommender system, which we will further elaborate on in Section 8. The random recommender scores highest on divergence for all metrics and is also one of the least relevant by definition (see NDCG score). *Most popular* and *random* have comparable NDCG results. Popularity scores for the articles are derived from the clicks recorded in the MIND interaction logs, and many articles have zero or only one click recorded. When the candidate list contains exclusively articles with a similar number of clicks, this forces the *most popular* recommender to a random choice, which explains the artificial similarity between *most popular* and *random* in terms of the NDCG score.

Figure 2 is the pendant of Table 3. It displays robust estimates based on the median and quantiles. Outliers are predominant for the Alternative Voices metric, as most articles do not contain any Alternative Voices. As detailed in Section 4, Alternative Voices are tagged as such if a named entity could not be found on Wikipedia.

Between the neural recommenders, most scores for LSTUR, NPA, NRMS, and NAML are close to zero, indicating a low divergence between the recommendations and the context distribution. Note that they produce similar recommendations (see NDCG values and Wu et al. [90]). Some notable differences can be observed when comparing these neural methods to the baselines. For example, we see that the neural recommenders are more Calibrated to the items present in

[13]We computed NDCG for popular and random and report on the original NDCG of the MIND publication for the neural recommenders, as it is more informative to the reader. We obtained similar results but no exact match.

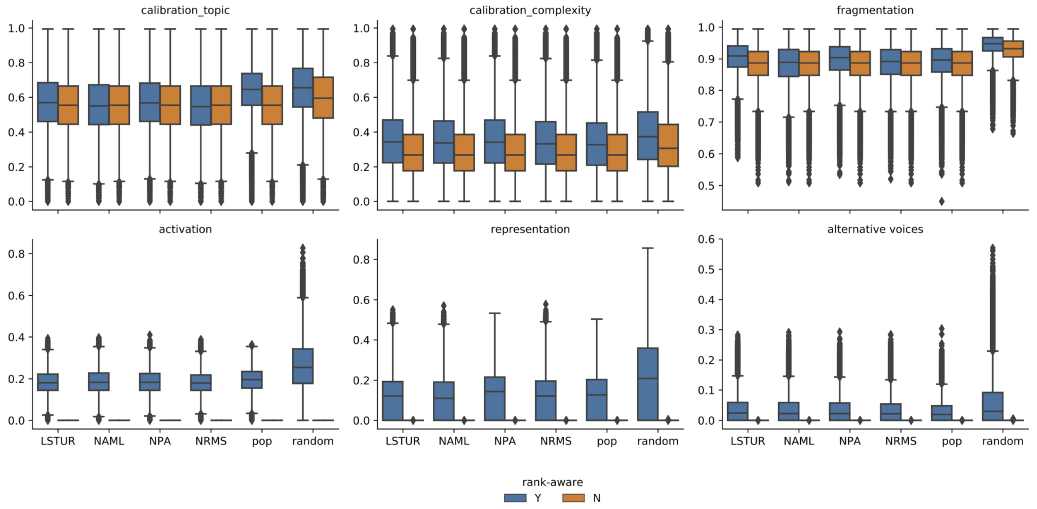


Fig. 2. Boxplot over each P, Q pair for all algorithms and all metrics for the JS divergence with rank-awareness. This is the pendant of Table 3. The horizontal line and the box boundaries show median and Inter Quartile Range (IQR), respectively. Single dots indicate outliers, defined as being further outside the 0.05 to 0.95 quantiles.

people's reading history, though the most popular baseline performs marginally better in terms of Calibration of complexity. In the following, we further analyze the entire distribution of individual recommendation list divergences and test the sensitivity of RADio to different settings. Boxplots for all metrics and all recommender strategies are available in the online repository, where we highlight the importance of rank-awareness.

6 SENSITIVITY ANALYSIS

This section contains a detailed analysis of the sensitivity of RADio to different settings: the divergence metric (JS or KL), rank-awareness (yes or no), and rank cut-off point (@n).

6.1 Sensitivity to the Divergence Metric

JS divergence is our preferred implementation of universal diversity metrics. It is a proper distance metric and bounded between 0 and 1 (see Section 3.2). Figure 5 substantiates that claim empirically, visualizing the sensitivity of RADio to the two described f-Divergence metrics: KL and JS Divergence. Clear differences can be observed in the distributions; KL divergence is skewed towards lower divergence, while JS divergence yields a more centered distribution of values. Additionally, JS divergence applies more contrast between the neural recommender systems and the naive recommendation methods and especially the random baseline. In certain cases, KL introduces consequential skew in the distribution of individual P, Q comparison pairs across recommendation models; this does not occur to that extent with JS.

Nevertheless, for demonstration purposes, we show results for the RADio framework coupled with KL divergence in Table 4. This time, metrics are not bounded by $[0, 1]$. If one were to rank the values for each metric, then Tables 3 and 4 would score the algorithms in the same order [47]. In conclusion, although KL divergence is a well-known metric that can be found in many applications and is simpler to express mathematically, we find JS divergence to be a better fit both theoretically and empirically.

Table 4. Results for Our RADio Framework for Recommendation Algorithms on the MIND Dataset

Algorithm	Calibration (topic)	Calibration (complexity)	Fragmentation	Activation	Representation	Alternative voices
LSTUR	2.6038	1.1432	7.7201	0.1481	0.1078	0.0142
NAML	2.5333	1.1287	7.3926	0.1531	0.1047	0.0127
NPA	2.5945	1.1390	7.6202	0.1521	0.1237	0.0134
NRMS	2.5013	1.1204	7.4519	0.1442	0.1114	0.0113
Most popular	2.9384	1.1082	7.6377	0.1605	0.1028	0.0102
Random	3.6038	1.5985	8.6295	0.8079	1.1248	0.0420

KL divergence with rank-awareness @10. For interpretation of the results, it should be noted that though a *higher* score does imply higher divergence, this does not necessarily mean this is a *better* score. Rather, what it means to be better is dependent on the metric and the model chosen, for which we refer to Table 1. These metrics are executed on a random sample of 35,000 users.

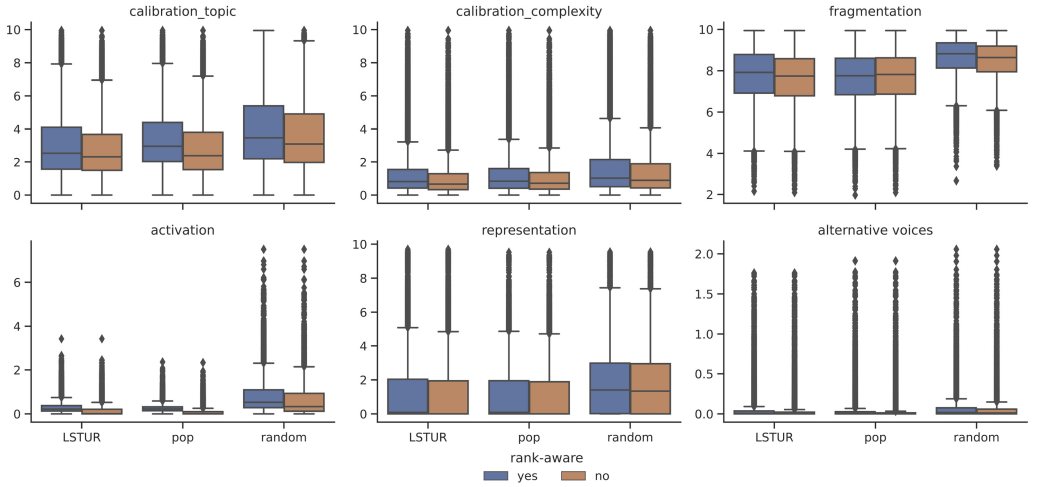


Fig. 3. Boxplot over each P, Q pair for all metrics for the KL divergence with and without rank-awareness. The horizontal line and the box boundaries show median and Inter Quartile Range (IQR), respectively. Single dots indicate outliers, defined as being further outside the 0.5 to 0.95 quantiles.

6.2 Sensitivity to Rank-awareness

In the original formulation of DART metrics [80], rank-awareness was not considered for most of the defined metrics. In our formalization, rank-awareness is the default. In Figure 6, we visualize 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. Figures 3 and 4 show a detailed analysis of rank-awareness of robust estimates (median and quantiles) for KL and the JS divergence.

6.3 Sensitivity @n

One could also consider adding a cut-off point where only the top n recommendations are considered for evaluation, the results of which are shown in Figure 7. The figure shows that the effect of rank-awareness becomes stronger with a higher cut-off point and causes the divergence score to stabilize after roughly 10 recommendations. This is because our MMR rank-awareness strongly

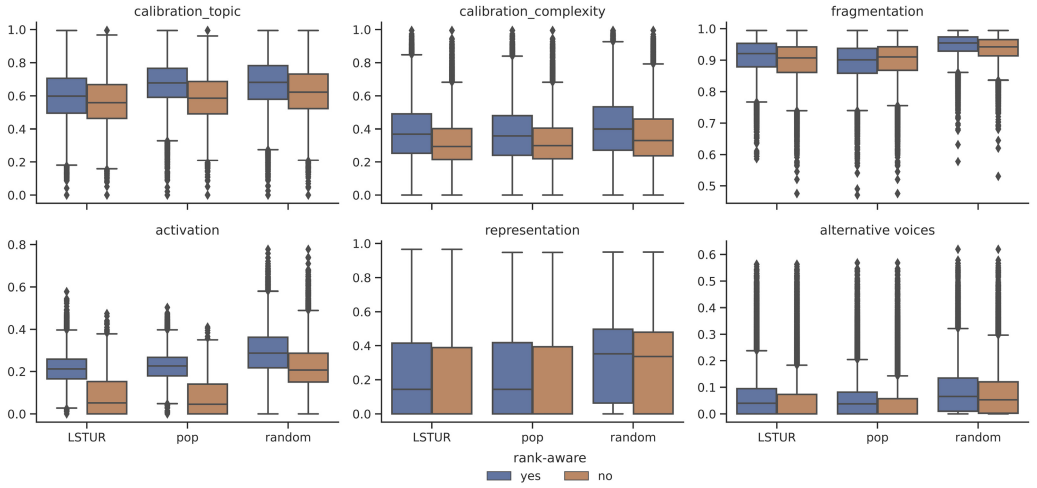


Fig. 4. Boxplot over each P, Q pair for all metrics for the JS divergence with and without rank-awareness. The horizontal line and the box boundaries show median and Inter Quartile Range (IQR), respectively. Single dots indicate outliers, defined as being further outside the 0.5 to 0.95 quantiles.

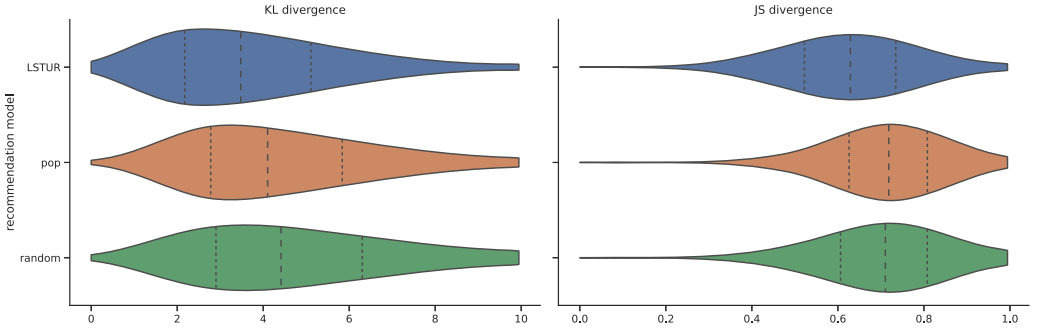


Fig. 5. Violin kernel density functions [86] over each P, Q pair, for the Calibration (topic) rank-aware metric, rank cutoff @N (no cutoff), using KL (left) and the JS divergence (right). Thick and thin dashes show median and Inter Quartile Range (IQR), respectively.

discounts values further down the ranking. @20 and @N (no cutoff) are similar for all metrics, because MIND rarely contains more than 20 recommendation candidates. Note that when calculating the divergence score for Activation, Representation, or Alternative Voices without rank-awareness and without cutoff point, there is no divergence to be reported, as recommendation and target distribution are identical in these cases.

6.4 Normative Evaluation

By comparing divergence scores across different recommender strategies, we can draw conclusions on the way they influence exposure of news to users. This is especially the case when comparing neural methods to the random recommender, which should reflect the characteristics of the overall supply of data. Combining this with DART's different theoretical models of democracy (summarized in Table 1), one can make informed decisions on which recommender system is better suited to one's normative stance than others. Imagine, for example, a media organization that aims to help

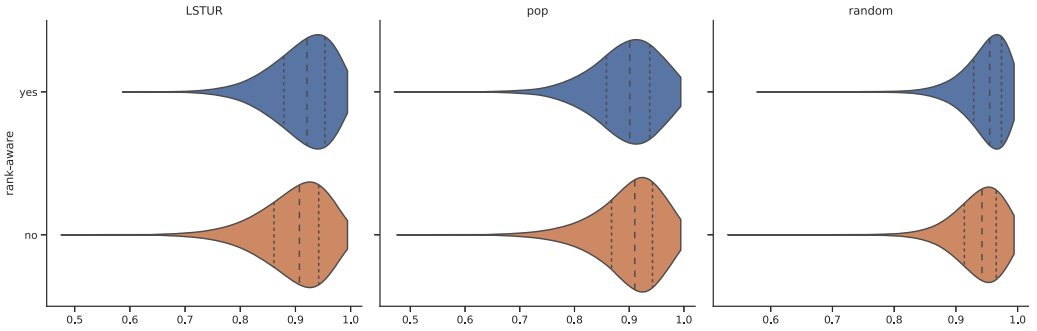


Fig. 6. Violin kernel density functions [86] over each P, Q pair, for the Fragmentation metric with JS divergence and rank cutoff @N (no cutoff), on three different recommender approaches with (blue) and without (orange) rank-awareness. Thick and thin dashes show median and IQR, respectively.

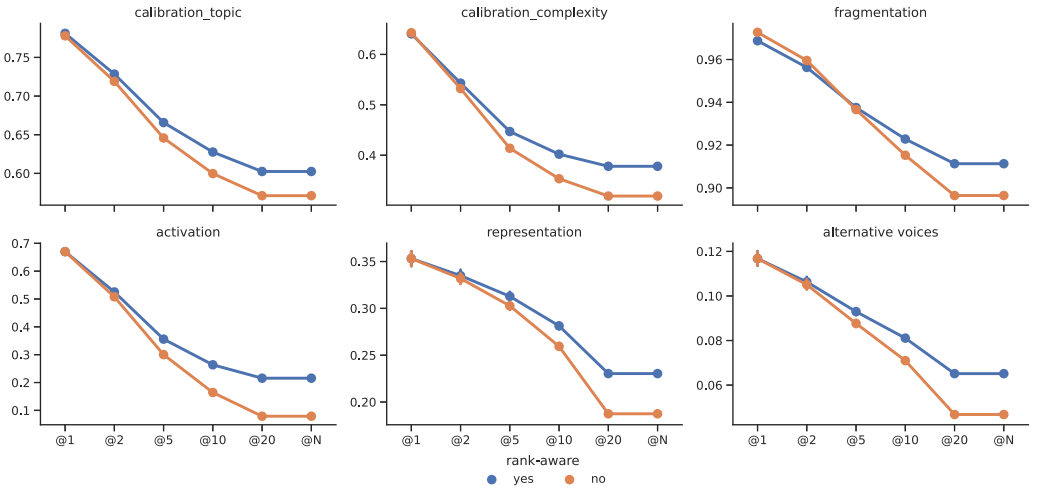


Fig. 7. Mean and 95% confidence interval for each DART metric implemented with JS divergence for the LSTUR recommender. Sensitivity analysis of RADio on rank-awareness (blue and orange) and rank cutoff.

their users find content on topics they are interested in. In this case, they are looking for a low divergence on Calibration, which is shown in the scores of the neural recommenders. This would indicate that those models are more suitable for this organization's goals. In comparison, imagine a large media organization that wants to dedicate a small section of their website to Critical principles, consisting of one element with recommendations called "A different perspective." This calls for a high divergence score in both Representation and Alternative Voices. Given that the random recommender scores best according to these principles, the neural recommenders would not be very suitable for this goal. Nevertheless, the conclusion that a random recommender is suitable for Critical norms and values is moot. Additional steps should be taken to further improve upon these scores: recommendation algorithm developers could tweak the tradeoff between different target values in the recommendation or even explicitly optimize on these metrics.

7 THE EFFECTS OF METRIC DESIGN CHOICES

In the following section, we will take a more detailed look at the effects of different design choices for a normative diversity metric, using the Calibration metric as an example. Calibration, which

Table 5. Distribution of Article Categories in Both the Training and Validation Set of MIND

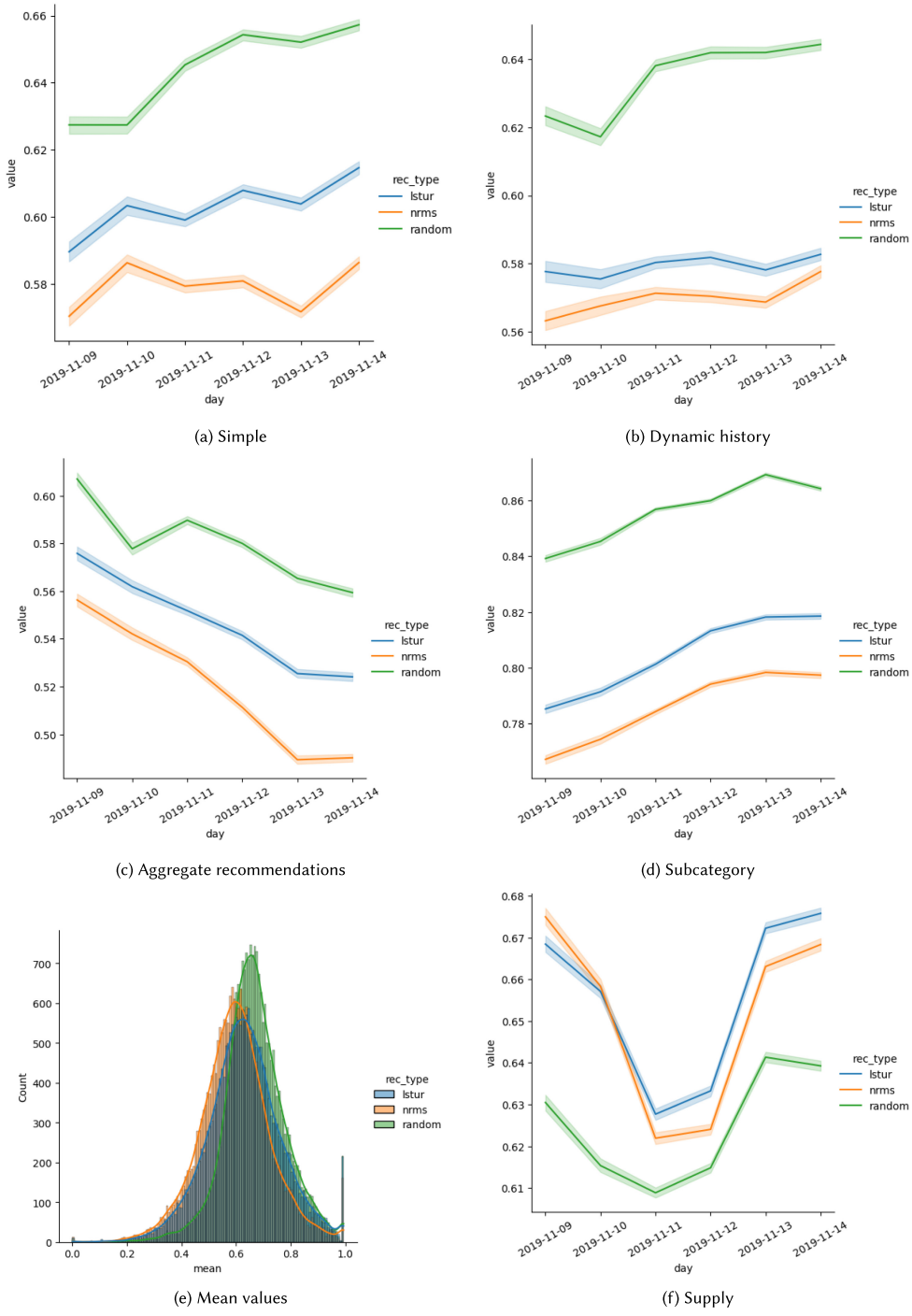
	Training							Validation						
	all	**impr**	**hist**	**click**	**LSTUR**	**NRMS**	**rand**	**all**	**impr**	**hist**	**click**	**LSTUR**	**NRMS**	**rand**
news	0.3	0.31	0.293	0.299	0.313	0.311	0.31	0.303	0.233	0.279	0.233	0.213	0.198	0.233
sports	0.315	0.116	0.138	0.125	0.129	0.136	0.116	0.3	0.163	0.143	0.246	0.275	0.254	0.163
finance	0.058	0.096	0.071	0.087	0.085	0.078	0.096	0.06	0.069	0.069	0.035	0.024	0.015	0.069
foodanddrink	0.044	0.042	0.049	0.04	0.033	0.038	0.042	0.048	0.068	0.05	0.056	0.09	0.092	0.068
lifestyle	0.045	0.104	0.104	0.106	0.114	0.114	0.105	0.048	0.171	0.105	0.178	0.161	0.174	0.171
travel	0.049	0.039	0.03	0.032	0.025	0.022	0.039	0.047	0.027	0.03	0.02	0.027	0.026	0.027
video	0.045	0.016	0.02	0.018	0.019	0.019	0.016	0.041	0.021	0.019	0.022	0.017	0.017	0.021
weather	0.042	0.027	0.013	0.023	0.021	0.022	0.027	0.039	0.028	0.012	0.015	0.017	0.013	0.027
autos	0.03	0.032	0.036	0.03	0.027	0.027	0.031	0.034	0.03	0.036	0.025	0.015	0.016	0.03
health	0.029	0.037	0.046	0.041	0.04	0.044	0.037	0.033	0.034	0.047	0.034	0.044	0.045	0.034

This includes the full dataset ("all"), the content people have seen and which the recommender ranks ("impr"), what was in the user's history ("hist"), the items the users clicked ("click"), and the top five items ranked by the neural recommenders and random baseline.

aims to measure the degree in which a recommendation reflects a user's personal preferences, is in this experimental setting the most suitable for this type of analysis, as it relies on categorical data that is directly available in MIND. Differences in behavior and performance will as such directly be caused by differences in the behavior of the recommender system and the design choices of the metric, rather than as a result of the data extraction pipeline.

Table 5 displays the distribution of article categories in the different steps of the recommendation pipeline. For both the training and validation sets, we analyze the article categories present in the full dataset ("all"), the items that were shown to the user and which the recommender system ranks ("impr"), what was present in the users' reading history ("hist"), and which items they eventually clicked on ("click"). We compare this to the categories present in the top-five items recommended by the neural recommenders LSTUR and NRMS and the random baseline. The categories in the full datasets are very similarly distributed between the training and validation sets, with both the news and sports categories comprising about 30% of the data, and the other categories ranging between 3%–6%. The items in the user histories are also decidedly similar. In both sets, the random recommender reflects the distribution of the impressions, which is expected behavior. Discrepancies between the two sets can be observed in both the candidate list and the content that users eventually clicked. The training data contains many more articles of the category "news," whereas the validation set has a much higher share of sports articles, especially among the clicked content. This is also reflected in the recommendations generated by the neural recommenders. The difference can be explained by the dynamic nature of news itself, which also underlines a caveat of using a dataset that only consists of a single day: If a major world event happened on that day, then it stands to reason that there is both a lot more content in that category and a higher interaction rate of users with it. This availability of content will then also influence the rate in which a recommender system can generate highly Calibrated recommendations. To account for this variability, we carry out the analysis on the predictions generated on the training set, which comprises six days. We emphasize that this analysis is not conducted to make judgments about the *performance* of the algorithms, in which case it would be imprudent to include training data in analysis, but rather to exemplify the multiple design choices that are relevant when constructing a normative diversity metric.

In Figure 8(a), we visualize Calibration in the same way as in Section 4: as the average divergence between the categories in individual recommendations and the user's reading history, though here it is plotted over time. This graph shows clear differences between the different recommender strategies. Furthermore, we find a statistically significant negative correlation between



(a) Simple

(b) Dynamic history

(c) Aggregate recommendations

(d) Subcategory

(e) Mean values

(f) Supply

Fig. 8. The Calibration metric with different approaches to metric design.

the divergence scores and the number of items in a user's history ($p<0.001$): The more items a user has consumed in the past, the lower the Calibration divergence is. In MIND, the user's history is static. Even if a user has accessed the website more than once, there will be no change recorded in the dataset. This makes it impossible to say how the recommender systems would adapt the recommendations over time, as more information about the user becomes available. We simulate a dynamic history in the diversity calculations by including the registered clicks of past interactions in the current history. This is visualized in Figure 8(b). Compared to the simple approach, we see a slightly lower and mostly more consistent divergence score. Some conflation with the fact that this is training data may happen here. The recommender systems are trained to predict the clicked items, which are also added to the history, resulting in a lower divergence score.

With the dynamic availability of content and the user's relative interest in major world events, it may not be possible for every single recommendation to achieve the desired normative diversity score, though they may do so over time. In Figure 8(c), we conceptualize our metric to not only consider the current recommendation, but also the top-five items of each recommendation in the past, weighted by recency. Here, we see that the divergence indeed decreases over time. However, f-Divergence penalizes the score comparatively much if smaller categories are not present in the recommendation [70]. Having more individual items in the calculations thus increases the chance that smaller categories are also represented, leading to a lower divergence score. This is reflected by the fact that the random recommender also decreases in divergence over time, though to a lesser extent than the neural recommenders (on average, a 0.03 versus a 0.06 decrease). Besides the top-level category, MIND also supplies article subcategories, which splits up "news" into, among others, "newsus" and "newspolitics," and sports into "football_nfl" and "baseball_mlb." The more fine-grained categorization could potentially be much more expressive of a user's interest. Figure 8(d) displays the Calibration score when considering subcategory. It shows roughly the same pattern as in Figure 8(a), although in a higher range: The average value here is centered around 0.8, whereas in Figure 8(a) it is around 0.6. When deploying Calibration in production, an informed decision needs to be made about the level of detail that is meaningful for the metric.

One other thing that might be relevant to consider is the number of users for whom the system does not work: those who, even on average, receive recommendations that are very divergent from their preferences. This is visualized in Figure 8(e), which plots the frequency of mean divergence scores per user. The information obtained from such an analysis could be used to investigate who the users are that consistently receive sub-par recommendations, and whether they share characteristics. This could then steer future development and tweaking of the algorithm. In a similar vein, one could investigate whether there are users who exist in their own "bubble" and receive recommendations that are very different from the "brand" of the organization. This is conceptualized in Figure 8(f) by calculating the divergence between the recommendations and the general distribution of articles. This type of analysis could be especially relevant when measuring organizational values and metrics such as Representation or Alternative Voices.

The figures in 8 convey very different types of information, even though they are based on the same feature and on the same set of recommendations. This shows that the design choices made when implementing the metric greatly influence the patterns it shows. Again, how the metric should behave is dependent on what the organization deploying the recommender system aims to measure and wants to achieve with it. As such, it is necessary that each of the choices are explainable to all stakeholders involved and that the process is well-documented.

8 DISCUSSION

Choosing an f-Divergence score as the base for our metrics allows us to construct a single base formalization with a clear interpretation among all metrics; when the value is 0, the distribution

between the recommendations and the chosen context is identical. The larger the measurements, the larger the divergence is. It is also flexible to the use of different target distributions and relevant features and has as such a strong advantage over ILD, which requires a static model of what it means for two articles to be different. Formalizing the diversity of the recommendation is as such brought down to three questions: what feature are we interested in, which distribution do we want to compare to, and should the divergence between these distributions be low, high, or something else. The answers to these questions could be informed by normative diversity metrics described in this article, but could also be interpreted differently. However, f-Divergence also comes with a number of limitations. For one, it does not take the relations between categorical values into account, and the ordering of the categorical values in the input vector is arbitrary. For example, two left-wing political parties in the Representation metric may be more similar than an extremely left-wing and an extremely right-wing party, but this is currently not taken into account. Related to this, to make continuous values suitable for our general discrete definition of f-Divergence, they need to be discretized into arbitrarily defined bins. This means that two very similar values may end up in different bins. Future work may propose a different approach for calculating divergence between continuous variables. Last, even design choices within the metric itself, such as the level of detail expressed in the relevant metric or the number of recommendations included in the calculation of the divergence score, may have large impacts on its behavior. Furthermore, the content that is available to a recommender also logically bounds to which extent it can generate normatively diverse recommendations. Careful justification and documentation of design choices when constructing the metric are thus a necessity.

Regarding the data enrichment pipeline, we identify a number of enhancement points. While some metrics, such as topic Calibration, work with simple data on news topics that is often directly available in a dataset, other metrics require a more sophisticated data enrichment pipeline. The differences in these approaches appear in the results: The metrics with more trivial metadata retrieval setups show clear and distinct patterns for different recommender algorithms, but this is not the case for the more complicated ones. Furthermore, it is not possible to determine the quality of the pipeline, as we do not have a ground truth for evaluation. For future work, we suggest to take the base formalizations as constructed in this article as a starting point and work to improve the extraction of the relevant parameters for metrics such as Representation, Alternative Voices, and Activation. Especially for the first two, there is already a large body of work that can facilitate this process [3, 24]. Human evaluation, including the input from editorial teams, would then be a promising way to evaluate these three normative metrics, similar to the work in the context of language generation bias [20].

At the same time, more insight needs to be gained on the influence of the choice of dataset. The MIND dataset contains a significant amount of so-called soft news, including articles on lifestyle, sports, and entertainment, whereas the DART metrics are mostly applicable to hard news. The influence of the chosen dataset needs to be investigated in more detail, which can then lead to more informed decision-making on the tradeoff between diversity and click-through rate and what can reasonably be expected of a news recommender system. Similarly, while the RADio metrics already provide an improvement over ILD by taking a target distribution into account instead of only what is present in the recommendations, they cannot look "beyond." A dataset or target distribution that is already skewed will not be identified as such. A possible solution could be to externalize the target distribution. For example, one could compare to the content produced by other organizations, such as (a number of) competitors or state-funded media. Alternatively, one could consider official statistics, such as the distribution of political parties in government. Possibly, this is an issue that cannot be solved purely through algorithmic means and should also be approached in a procedural way. Many news organizations already have organization-wide

protocols and procedures in place to guarantee the quality of content produced, and perhaps these protocols should be extended to also account for algorithmic recommendation.

Steps for adoption

Adopting the normative diversity metrics in practice is not a task that can be undertaken by an individual. We envision this process in the following steps:

(1) Identify the relevant stakeholders - As a first step, all the relevant stakeholders should be involved in the process [69]. Besides technical and editorial teams, this would also include management and business teams that eventually have to make an informed decision on the acceptable tradeoff between generating clicks and diversity.
(2) Determine the purpose of the recommender systems - Together, the stakeholders should determine what the purpose of the envisioned recommender system is. This can be inspired by one of the democratic models described in Section 2 but should most of all stem from open discussion. A recommender system can have multiple envisioned purposes, and these purposes can be at odds with each other; for example, finding relevant content and encountering different perspectives [81]. In such cases, the dynamic between the different purposes must be made as explicit and concrete as possible.
(3) Determine what type of diversity reflects this purpose - With Step 2 in mind, the stakeholders should define the type of diversity they are looking for using the three questions mentioned earlier in this section: What feature to measure, which distribution to compare to, and the expected divergence range. When the purpose is to help a user find relevant content, this means a low divergence of recommended topics between the recommendation and the user's history. For new perspectives, this means a high divergence between the perspectives in the recommendation versus the user's reading history. Collaboration between the different stakeholders is still instrumental: While the editorial and business teams may have the clearest vision on what a recommendation needs to accomplish, the technical teams need to communicate what is technically feasible. Many of the aspects related to diversity are technically hard to measure and often even conceptually ambiguous (what is a "perspective"?). Eventually, all stakeholders should agree on chosen simplifications, be aware of what these metrics can and cannot measure, and know where additional work and research is still necessary.

In its current form, RADio is an evaluation framework. We do not believe that the metrics should *replace* existing evaluation metrics, but should rather be used concurrently. Ideally, however, the normative diversity metrics would be incorporated as target for recommender system optimization, rather than post hoc evaluation. For example, they could be used to inform a reranking algorithm that ensures that, over time, the recommendations issued to a user reflect the chosen target distribution. Current diversification methods often rely on a distance measure between individual items. This is not suitable for the normative diversity metrics: As they are based on divergence scores, the recommendation list can only be considered as a whole. As such, additional theoretical work is necessary to use the metrics for optimization.

Last, if metrics such as these normative diversity metrics are to be incorporated in a production system, then they need to be carefully constructed and monitored. Neglecting to do so might lead to so-called "ethics-washing," where empty metrics are in place that convey no meaning but merely serve as a checkbox; and gamification of the metrics, where external parties learn how to utilize the metrics to boost the content that fits their purposes. Once more, this is not an effort that can be accomplished by technical teams alone and is one more reason to close the gap between the technical teams and the newsroom.

9 CONCLUSION

In this article, we have made a first attempt at constructing and implementing new evaluation criteria for news recommender systems, with a foundation in normative theory. Based on the DART metrics, first theoretically conceptualized in earlier work, we propose to look at diversity as a divergence score, observing differences between the issued recommendations and a metric-specific target distribution. We proposed RADio, a unified rank-aware f-Divergence metric framework that is mathematically grounded and that fits several possible use cases within the original DART metrics and we hope beyond in future work. We showed that JS divergence was preferred over other divergence metrics: At first mathematically, as JS is a proper distance metric; and empirically, via a sensitivity analysis to different cutoff, rank-awareness, and divergence metric regimes. When our approach is adopted in practice, it enables the evaluation of news recommender systems on normative principles beyond user relevance. Finally, we wish to emphasize that the metrics proposed are meant to supplement standard recommender system evaluation metrics in the same way that current beyond-accuracy metrics do. Most importantly, they are meant to bridge the gap between different disciplines involved in the process of news recommendation and to support more informed discussion between them. We hope for future research to foster interdisciplinary teams, leveraging each fields' unique skills and specialties.

ACKNOWLEDGMENTS

We thank Max van Drunen, Natali Helberger, Maartje ter Hoeve, Dan Li and Marijn Sax for proofreading. We also thank Ilias Koutsakis for help in cleaning up the code.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REPRODUCIBILITY

The code used for this article is available on http://www.github.com/svrijenhoek/RADio/. This includes both the diversity metrics and the data enrichment pipeline necessary for extracting relevant features from the MIND dataset.

REFERENCES

[1] 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*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, 336–345.

[2] Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. 2020. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of The Web Conference 2020 (WWW'20)*. Association for Computing Machinery, New York, NY, 2155–2165. https://doi.org/10.1145/3366423.3380281

[3] Christian Baden and Nina Springer. 2017. Conceptualizing viewpoint diversity in news discourse. *Journalism* 18, 2 (2017), 176–194.

[4] Mariella Bastian, Natali Helberger, and Mykola Makhortykh. 2021. Safeguarding the journalistic DNA: Attitudes towards the role of professional values in algorithmic news recommender designs. *Digit. Journal.* 9, 6 (2021), 835–863. DOI: https://doi.org/10.1080/21670811.2021.1912622

[5] Michael A. Beam. 2014. Automating the news: How personalized news recommender system design choices impact news reception. *Commun. Res.* 41, 8 (2014), 1019–1041.

[6] B. Douglas Bernheim, Luca Braghieri, Alejandro Martínez-Marquina, and David Zuckerman. 2021. A theory of chosen preferences. *Am. Econ. Rev.* 111, 2 (2021), 720–54.

[7] 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

[8] Balázs Bodó. 2019. Selling news to audiences—A qualitative inquiry into the emerging logics of algorithmic news personalization in European quality news media. *Digit. Journal.* 7, 8 (2019), 1054–1075. DOI: https://doi.org/10.1080/21670811.2019.1624185

[9] Christina Boididou, Di Sheng, Felix J. Mercer Moss, and Alessandro Piscopo. 2021. Building public service recommenders: Logbook of a journey. In *Proceedings of the 15th ACM Conference on Recommender Systems (RecSys'21)*. Association for Computing Machinery, New York, NY, 538–540. DOI: https://doi.org/10.1145/3460231.3474614

[10] Allan Borodin, Hyun Chul Lee, and Yuli Ye. 2012. Max-sum diversification, monotone submodular functions and dynamic updates. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS'12)*. Association for Computing Machinery, New York, NY, 155–166. DOI: https://doi.org/10.1145/2213556.2213580

[11] Keith Bradley and Barry Smyth. 2001. Improving recommendation diversity. In *Proceedings of the 12th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'01)*. 85–94.

[12] Joshua A. Braun and Jessica L. Eklund. 2019. Fake news, real money: Ad tech platforms, profit-driven hoaxes, and the business of journalism. *Digit. Journal.* 7, 1 (2019), 1–21.

[13] Khoo Khyou Bun and M. Ishizuka. 2001. Emerging topic tracking system. In *Proceedings of the 3rd International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems (WECWIS'01)*. 2–11. DOI: https://doi.org/10.1109/WECWIS.2001.933900

[14] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 202–214. https://proceedings.mlr.press/v81/burke18a.html

[15] Pablo Castells, Neil J. Hurley, and Saul Vargas. 2015. Novelty and diversity in recommender systems. In *Recommender Systems Handbook*. Springer, 881–918.

[16] 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 (KDD'08)*. Association for Computing Machinery, New York, NY, 88–96. DOI: https://doi.org/10.1145/1401890.1401906

[17] Abhijnan Chakraborty, Gourab K. Patro, Niloy Ganguly, Krishna P. Gummadi, and Patrick Loiseau. 2019. Equality of voice: Towards fair representation in crowdsourced Top-K recommendations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'19)*. Association for Computing Machinery, New York, NY, 129–138. DOI: https://doi.org/10.1145/3287560.3287570

[18] Laming Chen, Guoxin Zhang, and Hanning Zhou. 2018. Fast greedy MAP inference for determinantal point process to improve recommendation diversity. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Curran Associates Inc., Red Hook, NY, 5627–5638.

[19] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*. Association for Computing Machinery, New York, NY, 659–666. DOI: https://doi.org/10.1145/1390334.1390446

[20] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT'21)*. Association for Computing Machinery, New York, NY, 862–872. DOI: https://doi.org/10.1145/3442188.3445924

[21] Tommaso Di Noia, Vito Claudio Ostuni, Jessica Rosati, Paolo Tomeo, and Eugenio Di Sciascio. 2014. An analysis of users' propensity toward diversity in recommendations. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys'14)*. Association for Computing Machinery, New York, NY, 285–288. DOI: https://doi.org/10.1145/2645710.2645774

[22] Carmine DimAscio. 2020. py-readability-metrics. Retrieved from https://github.com/cdimascio/py-readability-metrics

[23] Qinxu Ding, Yong Liu, Chunyan Miao, Fei Cheng, and Haihong Tang. 2021. A hybrid bandit framework for diversified recommendation. *Proc. AAAI Conf. Artif. Intell.* 35, 5 (May 2021), 4036–4044. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/16524

[24] Tim Draws, Nava Tintarev, and Ujwal Gadiraju. 2021. Assessing viewpoint diversity in search results using ranking fairness metrics. *ACM SIGKDD Explor. Newslett.* 23, 1 (2021), 50–58.

[25] Yu Du, Sylvie Ranwez, Nicolas Sutton-Charani, and Vincent Ranwez. 2021. Is diversity optimization always suitable? Toward a better understanding of diversity within recommendation approaches. *Inf. Process. Manag.* 58, 6 (2021), 102721. DOI: https://doi.org/10.1016/j.ipm.2021.102721

[26] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys'14)*. Association for Computing Machinery, New York, NY, 161–168. DOI: https://doi.org/10.1145/2645710.2645737

[27] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, 172–186. Retrieved from https://proceedings.mlr.press/v81/ekstrand18b.html

[28] Dominik Maria Endres and Johannes E. Schindelin. 2003. A new metric for probability distributions. *IEEE Trans. Inf. Theor.* 49, 7 (2003), 1858–1860.

[29] Farzad Eskandanian, Bamshad Mobasher, and Robin Burke. 2017. A clustering approach for personalizing diversity in collaborative recommender systems. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP'17)*. Association for Computing Machinery, New York, NY, 280–284. DOI: https://doi.org/10.1145/3079628.3079699

[30] Sarah Eskens, Natali Helberger, and Judith Moeller. 2017. Challenged by news personalisation: Five perspectives on the right to receive information. *J. Media Law* 9, 2 (2017), 259–284. DOI: https://doi.org/10.1080/17577632.2017.1387353

[31] Raul Ferrer-Conill and Edson C. Tandoc Jr. 2018. The audience-oriented editor. *Digit. Journal.* 6, 4 (2018), 436–453. DOI: https://doi.org/10.1080/21670811.2018.1440972

[32] Lu Gan, Diana Nurbakova, Léa Laporte, and Sylvie Calabretto. 2020. Enhancing recommendation diversity using determinantal point processes on knowledge graphs. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, 2001–2004. DOI: https://doi.org/10.1145/3397271.3401213

[33] Florent Garcin, Christos Dimitrakakis, and Boi Faltings. 2013. Personalized news recommendation with context trees. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys'13)*. Association for Computing Machinery, New York, NY, 105–112. DOI: https://doi.org/10.1145/2507157.2507166

[34] Andreas Grün and Xenija Neufeld. 2021. Challenges experienced in public service media recommendation systems. In *Proceedings of the 15th ACM Conference on Recommender Systems (RecSys'21)*. Association for Computing Machinery, New York, NY, 541–544. DOI: https://doi.org/10.1145/3460231.3474618

[35] Rishav Hada, Amir Ebrahimi Fard, Sarah Shugars, Federico Bianchi, Patricia Rossini, Dirk Hovy, Rebekah Tromble, and Nava Tintarev. 2023. Beyond digital "Echo Chambers": The role of viewpoint diversity in political discussion. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (Singapore, Singapore) (WSDM'23)*. Association for Computing Machinery, New York, NY, 33–41. https://doi.org/10.1145/3539597.3570487

[36] Natali Helberger. 2019. On the democratic role of news recommenders. *Digit. Journal.* 7, 8 (2019), 993–1012. DOI: https://doi.org/10.1080/21670811.2019.1623700

[37] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2022. *spaCy: Industrial-strength Natural Language Processing in Python*. Technical Report. spaCy. https://spacy.io/usage/spacy-101

[38] David Hume. 1739. *A Treatise of Human Nature*. Clarendon Press, London.

[39] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (Oct. 2002), 422–446. DOI: https://doi.org/10.1145/582415.582418

[40] Harold Jeffreys. 1946. An invariant form for the prior probability in estimation problems. *Proc. R. Societ. London. Series A. Math. Physic. Sci.* 186, 1007 (1946), 453–461.

[41] Michael Jugovac, Dietmar Jannach, and Lukas Lerche. 2017. Efficient optimization of multiple recommendation quality factors according to individual user tendencies. *Expert Syst. Applic.* 81 (2017), 321–331. DOI: https://doi.org/10.1016/j.eswa.2017.03.055

[42] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News recommender systems—Survey and roads ahead. *Inf. Process. Manag.* 54, 6 (2018), 1203–1227.

[43] Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias Heintz. 2013. The Plista dataset. In *Proceedings of the International News Recommender Systems Workshop and Challenge (NRS'13)*. Association for Computing Machinery, New York, NY, 16–23. DOI: https://doi.org/10.1145/2516641.2516643

[44] J. Peter Kincaid and Leroy J. Delionbach. 1973. Validation of the automated readability index: A follow-up. *Human Factors* 15, 1 (1973), 17–20.

[45] S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *Ann. Math. Statist.* 22, 1 (1951), 79–86. DOI: https://doi.org/10.1214/aoms/1177729694

[46] Matevz Kunaver and Tomaz Pozrl. 2017. Diversity in recommender systems—A survey. *Knowl.-based Syst.* 123 (2017), 154–162. DOI: https://doi.org/10.1016/j.knosys.2017.02.009

[47] F. Liese and I. Vajda. 2006. On divergences and informations in statistics and information theory. *IEEE Trans. Inf. Theor.* 52, 10 (2006), 4394–4412. DOI: https://doi.org/10.1109/TIT.2006.881731

[48] Bibo Lin and Seth C. Lewis. 2022. The one thing journalistic AI just might do for democracy. *Digit. Journal.* 10, 10 (2022), 1627–1649. DOI: https://doi.org/10.1080/21670811.2022.2084131
[49] Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theor.* 37, 1 (1991), 145–151.
[50] Felicia Loecherbach, Judith Moeller, Damian Trilling, and Wouter van Atteveldt. 2020. The unified framework of media diversity: A systematic literature review. *Digital Journalism* 8, 5 (2020), 605–642. https://doi.org/10.1080/21670811.2020.1764374
[51] Steven Loria. 2021. *textblob Documentation.* Technical Report. Textblob. https://textblob.readthedocs.io/en/dev/quickstart.html
[52] 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 (UMAP'20).* Association for Computing Machinery, New York, NY, 145–153. https://doi.org/10.1145/3340631.3394864
[53] Hongyu Lu, Min Zhang, and Shaoping Ma. 2018. Between clicks and satisfaction: Study on multi-phase user preferences and satisfaction for online news reading. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18).* Association for Computing Machinery, New York, NY, 435–444. DOI: https://doi.org/10.1145/3209978.3210007
[54] Nicolas Mattis, Philipp Masur, Judith Möller, and Wouter van Atteveldt. 2022. Nudging towards news diversity: A theoretical framework for facilitating diverse news consumption through recommender design. *New Media & Society* Online First (2022), 14614448221104413. https://doi.org/10.1177/14614448221104413
[55] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES'20).* Association for Computing Machinery, New York, NY, 117–123. https://doi.org/10.1145/3375627.3375832
[56] 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. *Inf., Commun. Societ.* 21, 7 (2018), 959–977.
[57] Lynge Asbjørn Møller. 2022. Recommended for you: How newspapers normalise algorithmic news recommendation to fit their gatekeeping role. *Journal. Stud.* 23, 7 (2022), 800–817. DOI: https://doi.org/10.1080/1461670X.2022.2034522
[58] 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 (FAccT'21).* Association for Computing Machinery, New York, NY, 478–488. https://doi.org/10.1145/3442188.3445911
[59] Newman Nic, Richard Fletcher, Antonis Kalogeropoulos, David A. L. Levy, and Rasmus Kleis Nielsen. 2018. Reuters Institute digital news report 2018. *Reuters Instit. Study Journal.* (2018), 39.
[60] Mícheál. O'Searcoid. 2007. *Metric Spaces* (1st ed). Springer London, London.
[61] Javier Parapar and Filip Radlinski. 2021. Towards unified metrics for accuracy and diversity for recommender systems. In *Proceedings of the 15th ACM Conference on Recommender Systems.* Association for Computing Machinery, New York, NY, 75–84. DOI: https://doi.org/10.1145/3460231.3474234
[62] Eli Pariser. 2011. *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think.* Penguin.
[63] Shameem A. Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. 2016. A coverage-based approach to recommendation diversity on similarity graph. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys'16).* Association for Computing Machinery, New York, NY, 15–22. DOI: https://doi.org/10.1145/2959100.2959149
[64] Lijing Qin and Xiaoyan Zhu. 2013. Promoting diversity in recommendation by entropy regularizer. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI'13).* AAAI Press, 2698–2704.
[65] Shaina Raza and Chen Ding. 2021. Deep dynamic neural network to trade-off between accuracy and diversity in a news recommender system. In *2021 IEEE International Conference on Big Data (Big Data).* IEEE, 5246–5256. https://doi.org/10.1109/BigData52589.2021.9671467
[66] Tetsuya Sakai and Zhaohao Zeng. 2019. Which diversity evaluation measures are "Good"? In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19).* Association for Computing Machinery, New York, NY, 595–604. DOI: https://doi.org/10.1145/3331184.3331215
[67] Marijn Sax. 2022. Algorithmic news diversity and democratic theory: Adding agonism to the mix. *Digit. Journal.* 10, 10 (2022), 1650–1670. https://doi.org/10.1080/21670811.2022.2114919
[68] Jane B. Singer. 2014. User-generated visibility: Secondary gatekeeping in a shared media space. *New Media Societ* 16, 1 (2014), 55–73. DOI: https://doi.org/10.1177/1461444813477833
[69] Annelien Smets, Jonathan Hendrickx, and Pieter Ballon. 2022. We're in this together: A multi-stakeholder approach for news recommenders. *Digit. Journal.* 10, 10 (2022), 1813–1831. https://doi.org/10.1080/21670811.2021.2024079

[70] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys'18)*. Association for Computing Machinery, New York, NY, 154–162. DOI: https://doi.org/10.1145/3240323.3240372

[71] Daniel Steel, Sina Fazelpour, Kinley Gillette, Bianca Crewe, and Michael Burgess. 2018. Multiple diversity concepts and their ethical-epistemic implications. *Eur. J. Philos. Sci.* 8, 3 (2018), 761–780.

[72] Jesper Strömbäck. 2005. In search of a standard: Four models of democracy and their normative implications for journalism. *Journal. Stud.* 6, 3 (2005), 331–345. DOI: https://doi.org/10.1080/14616700500131950

[73] Niek Tax, Sander Bockting, and Djoerd Hiemstra. 2015. A cross-benchmark comparison of 87 learning to rank methods. *Inf. Process. Manag.* 51, 6 (2015), 757–772. DOI: https://doi.org/10.1016/j.ipm.2015.07.002

[74] 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.

[75] Joy A. Thomas and Thomas M. Cover. 1991. *Elements of Information Theory*. Wiley.

[76] Damian Trilling and Marieke van Hoof. 2020. Between article and topic: News events as level of analysis and their computational identification. *Digit. Journal.* 8, 10 (2020), 1317–1337. DOI: https://doi.org/10.1080/21670811.2020.1839352

[77] Saúl Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys'11)*. Association for Computing Machinery, New York, NY, 109–116. https://doi.org/10.1145/2043932.2043955

[78] Sanne Vrijenhoek. 2023. Do you MIND? Reflections on the MIND dataset for research on diversity in news recommendations. In *Advances in Bias and Fairness in Information Retrieval*, Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo (Eds.). Springer Nature Switzerland, Cham, 147–154.

[79] 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, 208–219. DOI: https://doi.org/10.1145/3523227.3546780

[80] 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 Conference on Human Information Interaction and Retrieval (CHIIR'21)*. Association for Computing Machinery, New York, NY, 173–183. DOI: https://doi.org/10.1145/3406522.3446019

[81] Sanne Vrijenhoek, Lien Michiels, Johannes Kruse, Alain Starke, Nava Tintarev, and Jordi Viader Guerrero. 2023. NORMalize: The first workshop on normative design and evaluation of recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys'23)*. Association for Computing Machinery, New York, NY, 1252–1254. https://doi.org/10.1145/3604915.3608757

[82] Julian Wallace. 2018. Modelling contemporary gatekeeping: The rise of individuals, algorithms and platforms in digital news dissemination. *Digit. Journal.* 6, 3 (2018), 274–293.

[83] Isaac Waller and Ashton Anderson. 2019. Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. In *Proceedings of the World Wide Web Conference (WWW'19)*. Association for Computing Machinery, New York, NY, 1954–1964. DOI: https://doi.org/10.1145/3308558.3313729

[84] Ningxia Wang and Li Chen. 2021. User bias in beyond-accuracy measurement of recommendation algorithms. In *Proceedings of the 15th ACM Conference on Recommender Systems*. Association for Computing Machinery, New York, NY, 133–142. DOI: https://doi.org/10.1145/3460231.3474244

[85] Kasper Welbers, Wouter van Atteveldt, Jan Kleinnijenhuis, and Nel Ruigrok. 2018. A gatekeeper among gatekeepers: News agency influence in print and online newspapers in the Netherlands. *Journal. Stud.* 19, 3 (2018), 315–333.

[86] Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis – Violin Plot*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org/reference/geom_violin.html

[87] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with attentive multi-view learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI'19)*. International Joint Conferences on Artificial Intelligence Organization, 3863–3869. https://doi.org/10.24963/ijcai.2019/536

[88] 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 (KDD'19)*. Association for Computing Machinery, New York, NY, 2576–2584. https://doi.org/10.1145/3292500.3330665

[89] 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 6389–6394. https://doi.org/10.18653/v1/D19-1671

[90] 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 3597–3606. https://doi.org/10.18653/v1/2020.acl-main.331

[91] Qiong Wu, Yong Liu, Chunyan Miao, Yin Zhao, Lu Guan, and Haihong Tang. 2019. Recent advances in diversified recommendation. *ArXiv* abs/1905.06589 (2019).

[92] Shangyuan Wu, Edson C. Tandoc, and Charles T. Salmon. 2019. Journalism reconfigured: Assessing human-machine relations and the autonomous power of automation in news production. *Journal. Stud.* 20, 10 (2019), 1440–1457. Retrieved from https://search-ebscohost-com.proxy.uba.uva.nl/login.aspx?direct=true&db=ufh&AN=137190691&site=ehost-live&scope=site

[93] Wen Wu, Li Chen, and Yu Zhao. 2018. Personalizing recommendation diversity based on user personality. *User Model. User-adapt. Interact.* 28, 3 (2018), 237–276.

[94] Ruobing Xie, Qi Liu, Shukai Liu, Ziwei Zhang, Peng Cui, Bo Zhang, and Leyu Lin. 2021. Improving accuracy and diversity in matching of recommendation with diversified preference network. *IEEE Transactions on Big Data* 8, 4 (2022), 955–967. https://doi.org/10.1109/TBDATA.2021.3103263

[95] Emine Yilmaz and Stephen Robertson. 2010. On the choice of effectiveness measures for learning to rank. *Inf. Retr.* 13, 3 (June 2010), 271–290. DOI: https://doi.org/{10.1007/s10791-009-9116-x}

[96] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. 2016. Apache Spark: A unified engine for big data processing. *Commun. ACM* 59, 11 (Oct. 2016), 56–65. DOI: https://doi.org/10.1145/2934664

Received 18 January 2023; revised 17 September 2023; accepted 15 November 2023