RADio – Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations

Sanne Vrijenhoek*
Institute for Information Law, University of Amsterdam
Amsterdam, The Netherlands
s.vrijenhoek@uva.nl

Gabriel Bénédict*
University of Amsterdam, RTL Nederland B.V.
Amsterdam, The Netherlands
gabriel.benedict@rtl.nl

Mateo Gutierrez Granada
RTL Nederland B.V.
Hilversum, The Netherlands
mateo.gutierrez.granada@rtl.nl

Daan Odijk
RTL Nederland B.V.
Hilversum, The Netherlands
daan.odijk@rtl.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

ABSTRACT

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.

KEYWORDS

News recommendation, Diversity, Divergence, Normative framework

ACM Reference Format:
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 *Sixteenth ACM Conference on Recommender Systems (RecSys '22), September 18–23, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3523227.3546780

*These authors contributed equally.

RecSys '22, September 18–23, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9278-5/22/09…$15.00
https://doi.org/10.1145/3523227.3546780

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 [82]. 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 to them at the click of a button [63] – 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, 52]. 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 [53], or present the reader of a particular news article with other articles about the same topic, but from a different perspective [54]. The use of news recommender systems has a wide range of benefits. They can increase engagement [55] and help raise informed citizens [28]. A news recommender system may broaden the horizons of their users by presenting diverse recommendations, including items different from what they are used to or expect seeing. They could even foster tolerance and understanding [29, 66], and counter so-called filter bubbles or echo chambers [52, 58].

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 [39]. 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. As such, their evaluation has different requirements than those of other types of recommender systems [4, 5, 72, 75]. Recent controversies have shown that merely optimizing for click-through rates and engagement may promote sensationalist content [68], 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, 32]. News recommender systems exist in a complex space

[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/

consisting of many different areas and disciplines, each with their own goals and challenges; think of balancing diversity and accuracy [57], nudging [50] or even identifying user preferences [6, 49] and biases [74]. In this paper, 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 [33], who provides a theoretical foundation for conceptualizing diversity, and of Vrijenhoek et al. [71], 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 possible shortcomings in these metrics: there could be more consideration for the theory of metrics and distance functions, generalizability to other normative concepts, unification under one framework, and rank-awareness. In this paper, we focus on the mathematical aspects of a rank-aware metric, versatile to different normative concepts and as such addressing these shortcomings. 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 [80]. 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.

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 [17, 62], 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 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, 13, 19, 23, 24, 38, 48, 70]: 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 [25]. 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 [43, 48].

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 [43, 81]. One can optimize for these descriptive notions of diversity (i) before training, by clustering users based on their profile diversity with JS divergence [27], (ii) directly at training time (e.g., for learning-to-rank [10, 13, 70], collaborative filtering [60], graphs [30, 59] or bandits [21, 84]), (iii) by re-ranking a recommendation set and balance diversity vs. relevance [16] or popularity vs. relevance [15], 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, 73]). With any of these four methods, a trade-off 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 [48]. Nevertheless, this encouraged recent efforts in training neural-based recommenders that explicitly make a trade-off between accuracy and diversity [61]. Also recently, there have been studies that differentiate between diversity needs of users [83].

2.2 Normative Diversity

Diversity is extensively discussed as a normative concept in literature, and has a role in many different areas of science [46, 65], spanning from ecological diversity to diversity as a proxy for fairness in machine learning systems [51]. 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 [33], we define a normatively diverse news recommendation as one that succeeds in informing the user and supports them in fulfilling their role in democratic society. Out of the many theoretical models that exist in literature, Helberger [33] describes four different models from the normative framework of democracy, each with a different view on what it means to properly inform citizens: the **Liberal** model, which aims to enable personal development and autonomy, the **Participatory** model, which aims to enable users to fulfill their role as active citizens in a democratic society, the **Deliberative** model, which aims to foster discussion and debate by equally presenting different viewpoints and opinions in a rational and neutral way, and the **Critical** model, which aims to challenge the status quo and to inspire the readers to take action against existing injustices in society.

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

[3] This dichotomy is oftentimes referred to as normative (*what ought to be*) and positive (*what is*) statements [35] but can easily be confused with concepts such as positive / negative examples in Machine Learning. We thus opt for the more explicit normative / descriptive duo.

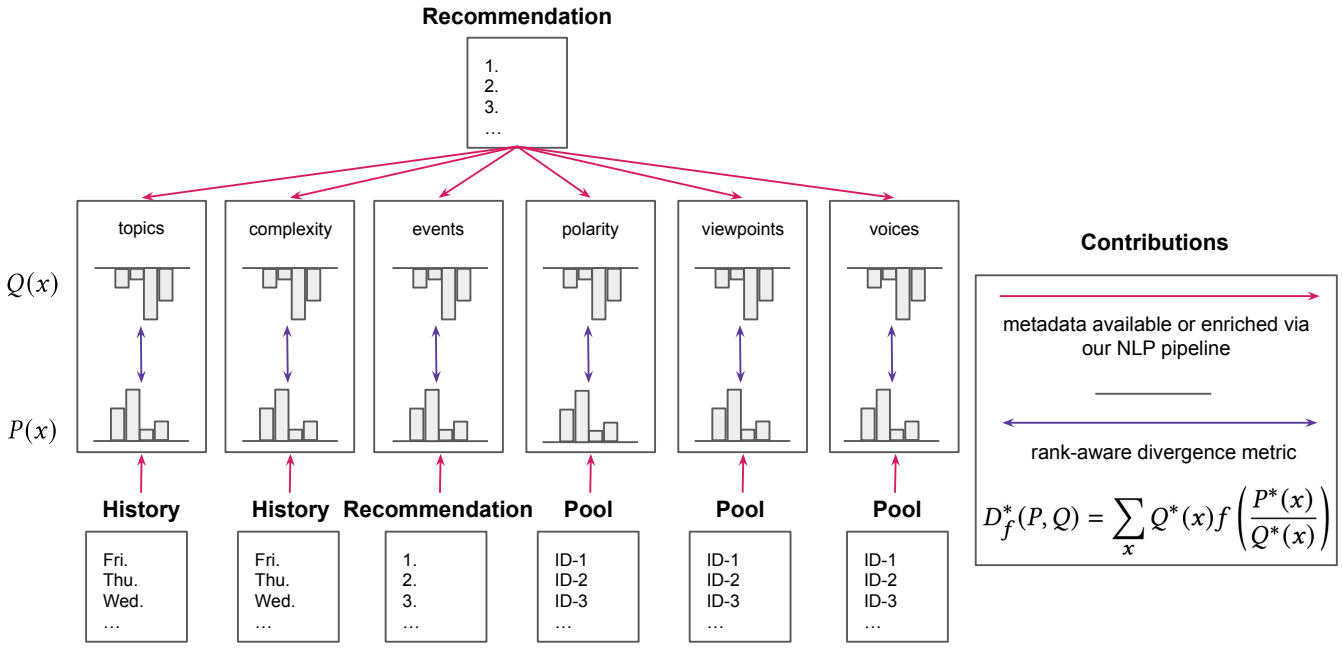

Figure 1: Comparing discrete diversity distributions in the context of news recommendations. First, metadata is collected in the news dataset or retrieved via our NLP pipeline (red). Discrete distributions of that metadata are then compared via a rank-aware divergence metric (purple). Recommendation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

For more details regarding the different models, and what a recommender system following each of these models would look like, we refer to Helberger [33]. 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.

Based on these models, the DART metrics [71] take a first step towards normative diversity for recommender systems and reflect the nuances of the different democratic models described above: **Calibration**, **Fragmentation**, **Activation**, **Representation** and **Alternative Voices**. Table 1 provides an overview of the DART metrics and their expected value ranges for the different models, and will be further elaborated later in the paper.

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. [46] 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 [17, 62], may not be considered to be diverse according to the criteria maintained by newsroom editors. Both Loecherbach et al. [46] 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. [71] in order 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 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

Table 1: Overview of the different models and expected value ranges for each metric. 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.

	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

article in the recommendation ranking, rather than considering the set as a whole (e.g. ILD).

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*, that express intuitions about concepts of distance [56].

We add that our distance measure should (i) be bounded by $[0; 1]$, for comparisons of different recommendation algorithms (ii) be unified, so as to fairly consider different diversity aspects (as opposed to e.g. using weighted averages or maxima in [18]) 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 issued 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 [64]), 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 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 (Jensen-Shannon divergence) as our preferred metric.

3.2.1 Kullback-Leibler Divergence. The concept of relative entropy or KL (Kullback–Leibler) divergence [42] between two probability mass functions P and Q (here, a recommendation and its context) is defined as:

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

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. When $P = Q$, $D_{\mathrm{KL}}(P, Q) = D_{\mathrm{KL}}(P, P) = 0$, that identity property is not guaranteed by cross entropy alone. This is the main reason to prefer KL divergence over cross entropy. Though KL Divergence satisfies the *identity* requirement, the *symmetry* and *triangle inequality* are not fulfilled. This can be resolved by further refining KL Divergence.

3.2.2 Jensen–Shannon Divergence. A succession of steps from KL divergence lead to Jensen-Shannon (JS) divergence. KL divergence was first turned symmetric [37] and then upper bounded [45], to lead to

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

When the base 2 logarithm is used, the JS divergence bounds are $0 \leq D_{\mathrm{JS}}(P, Q) \leq 1$. Additionally, Endres and Schindelin [26] show that $\sqrt{D_{\mathrm{JS}}}$ is a proper distance which 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.

Liese and Vajda [44] 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

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

[4] f-Divergence accommodates for other divergence metrics which are out of scope of this research [44].

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 [64], we write $\overline{P}$ and $\overline{Q}$:

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

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. In order for the entire probabilistic distributions $\overline{P}$ and $\overline{Q}$ to remain proper statistical distributions, we normalize them to ensure $\sum_x \overline{P}(x) = \sum_x \overline{Q}(x) = 1$. To avoid notation congestion, P and Q will implicitly refer to $\overline{P}$ and $\overline{Q}$, in the following sections.

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 [67, 85], and by extension in conventional descriptive diversity metrics [13] 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 [14, 36], in particular for news recommendation [80]. 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:

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

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 [40]. 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 [14].

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:

$$D_f^*(P, Q) = \sum_x Q^*(x) f\left(\frac{P^*(x)}{Q^*(x)}\right), \quad (6)$$

$Q^*(x)$ 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 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.

$R_i^u \in \{1, 2, 3, \ldots\}$ 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 [71], 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 [71], 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, etc.[6] This approach is based on the implementation of balanced neighbourhoods in recommender systems [12]. With m we refer to the distribution of protected vs. non-protected groups, with $m \in \{Minority, Majority\}$. $P(m|S)$ and $Q^*(m|S)$ 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:

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

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

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

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

$$AlternativeVoices = AltV\big(P(m|S), Q^*(m|R)\big) = \sum_m Q^*(m|R) f\left(\frac{P(m|S)}{Q^*(m|R)}\right) \tag{11}$$

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

4 EXPERIMENTAL SETUP

In order 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 a public dataset. It should be noted that this pipeline is an imperfect approximation, and that each metric individually would benefit from more sophisticated methods. The MIND dataset [80] 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 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,[8] including news-specific neural recommendation methods NPA [78], NAML [77], LSTUR [1] and NRMS [79]. It was shown that these algorithms outperform general-purpose ones [80] or common collaborative filtering models (such as alternating least squares (ALS)), in particular due to the short lifespan of news items [31]. These algorithms are trained on the impression logs in order to predict which items the users are most likely to click on. For the purpose of this paper 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.

Since RADio computes the average of all $\{P, Q\}$ pairs, we retrieve confidence intervals over paired distances too, as illustrated in the sensitivity analyses below. 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).

We scrape articles via the URLs provided in the MIND dataset. Each article's metadata is enriched with five methods:

(1) **Complexity analysis** – Each item is assigned a complexity score based on the Flesch-Kincaid reading ease test [41], implemented in the Python module py-readability-metrics [20]. 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 [69].
(3) **Sentiment analysis** – Using the textBlob open source NLP library we assign each article a sentiment polarity score [47]. Our focus is on the relative neutrality of articles, we thus take the absolute value of the negative / positive polarity score.

[8]Code available at https://github.com/microsoft/recommenders/tree/main/examples/00_quick_start

Table 2: Overview of the implementation approach for different methods. Numbers in bold correspond to the corresponding steps in the metadata enrichment pipeline presented above.

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

(4) **Named entity recognition** – Using spaCy, we identify the people, organizations and locations mentioned in the text [34], 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[9] entry through fuzzy name matching, to figure out if they are politicians, or in the case of organizations, political parties.[10]

We implement RADio with the pipeline above. 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.

5 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 Section 5 from the interpretation of normative interpretation of the metrics in Section 6. We choose to implement RADio with rank-awareness and JS divergence with a rank cutoff @N (the entire ranking list) as our default. After commenting on the overall results, we further motivate that 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 displays results for RADio with rank-aware JS divergence.[12] Higher values imply higher divergence scores, but whether high or low divergence is desired depends on the goal of the recommender system, which we will further elaborate in Section 6. 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. Between the neural recommenders, most scores for LSTUR, NPA, NRMS and NAML are in lower ranges. Note that they produce similar recommendations (see NDCG values and Wu et al. [80]). 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 people's reading history, though the most popular baseline performs marginally better in terms of Calibration of complexity. In the following, we further analyse 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.

5.1 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 2 substantiates that claim empirically, visualizing the sensitivity of RADio to the two described f-Divergence metrics: KL and JS Divergence. Clear differences can

[9] https://www.wikidata.org/
[10] 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.
[11] https://github.com/svrijenhoek/RADio/
[12] More visualizations are available on the online repository
[12] 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, since we only computed one inference cycle.

Table 3: Results for our RADio framework for recommendation algorithms on the MIND dataset. 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.

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

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. Due to the large sample in MIND, the random baseline is an approximation of a diverse recommendation set, given the candidate articles. 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. Although KL Divergence is a well-known metric that can be found in many applications and is simpler to express mathematically, we found the JS divergence to be a better fit both theoretically and empirically.

5.2 Sensitivity to Rank-awareness

In the original formulation of DART metrics [71], rank-awareness was not considered for most of the defined metrics. In our formalization, rank-awareness is the default. In Figure 3, 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.

5.3 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 4. 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 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.[13]

5.4 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 pool 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 public service media organization that aims to reflect Participatory norms and values in their recommendations. The Participatory model prescribes low Fragmentation and low Activation, 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." The Critical model 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 trade-off between different target values in the recommendation, or even explicitly optimize on these metrics.

6 DISCUSSION

Choosing an f-Divergence score as the base for our metrics allows us to construct a single base formalization with a clear interpretation amongst 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. However, it also comes with a number of limitations. For one, f-Divergence does not

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

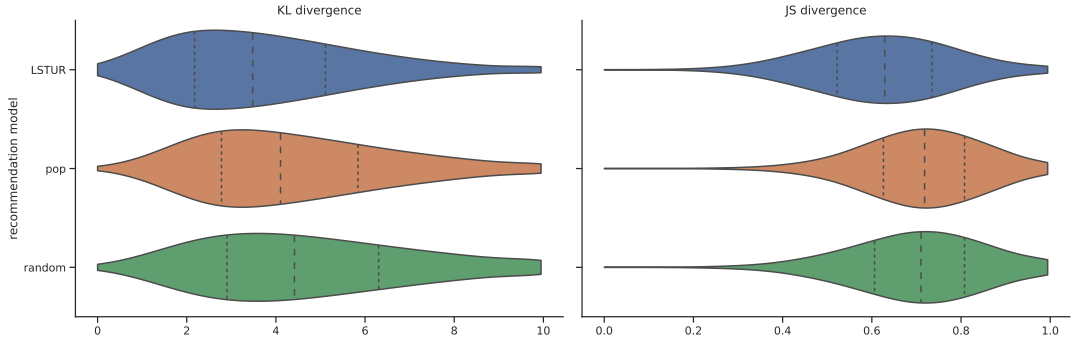

Figure 2: Violin kernel density functions [76] over each P, Q pair, for the Calibration (topic) rank-aware metric, rank cutoff @N (no cuttof), using KL (left) and the JS divergence (right). Thick and thin dashes show median and Inter Quartile Range (IQR) respectively.

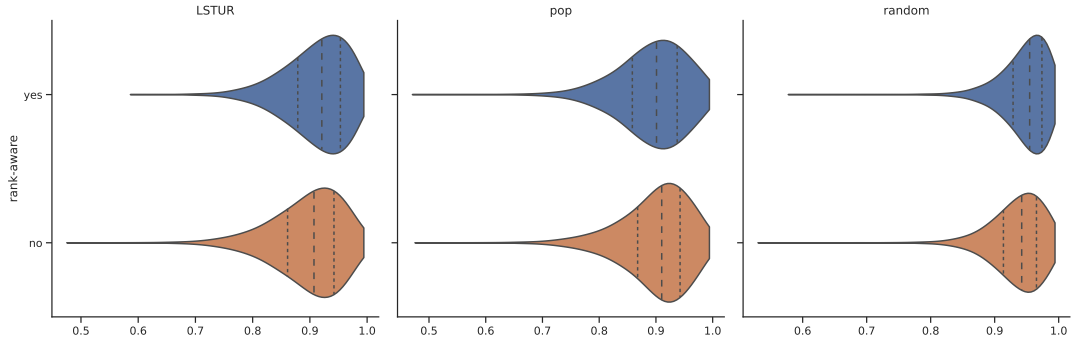

Figure 3: Violin kernel density functions [76] over each P, Q pair, for the Fragmentation metric with JS divergence and rank cutoff @N (no cuttof), on three different recommender approaches with (blue) and without (orange) rank-awareness. Thick and thin dashes show median and IQR respectively.

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, in order 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 binsFuture work may propose a different approach for calculating divergence between continuous variables. 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 paper 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, 22]. 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 [18]. Additionally, 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, sport 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 trade-off between diversity and click-through rate, and what can reasonably be expected of a news recommender system.

7 CONCLUSION

In this paper 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

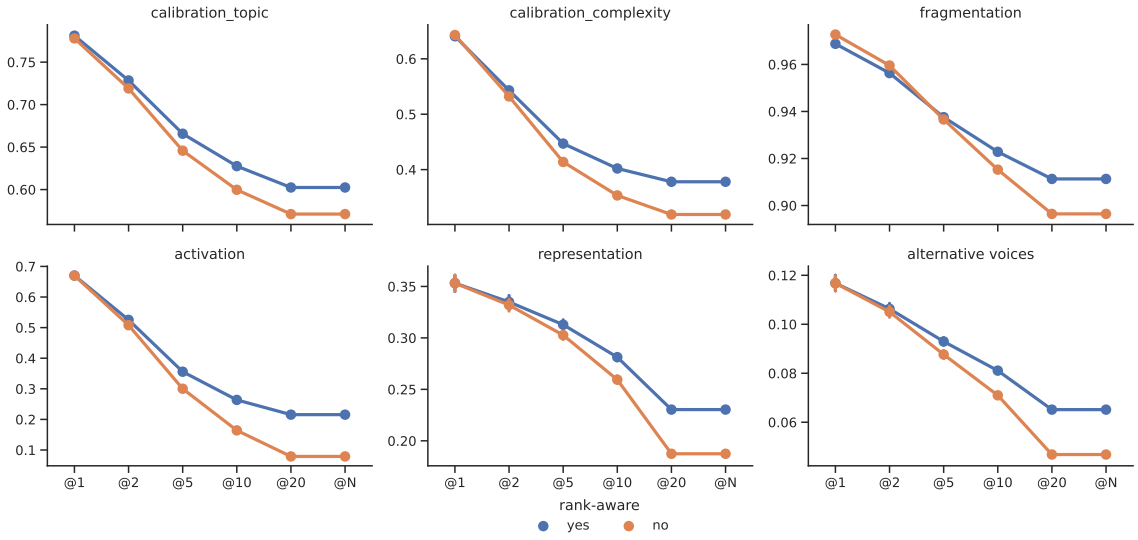

Figure 4: Mean and 95% confidence interval for each DART metric implemented with JS divergence for the LSTUR recommender. Sensitivity analysis of RADio on rank-awareness (blue and orange) and rank cutoff.

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 proof-reading. We also thank Ilias Koutsakis for helping cleaning up the code. 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). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

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*. 336–345.

[2] Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. 2020. Algorithmic Effects on the Diversity of Consumption on Spotify. *Proceedings of The Web Conference 2020* (2020).

[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. *Digital Journalism* 0, 0 (2021), 1–29. https://doi.org/10.1080/21670811.2021.1912622 arXiv: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. *Communication Research* 41, 8 (2014), 1019–1041.

[6] B Douglas Bernheim, Luca Braghieri, Alejandro Martínez-Marquina, and David Zuckerman. 2021. A Theory of Chosen Preferences. *American Economic Review* 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, et al. 2020. Diversity in News Recommendations. *arXiv preprint arXiv:2005.09495* (2020).

[8] Balazs Bodo. 2019. Selling News to Audiences – A Qualitative Inquiry into the Emerging Logics of Algorithmic News Personalization in European Quality News Media. *Digital Journalism* 0, 0 (2019), 1–22. 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 *Fifteenth ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 538–540. 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* (Scottsdale, Arizona, USA) *(PODS '12)*. Association for Computing Machinery, New York, NY, USA, 155–166. 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* (Maynooth, Ireland) *(AICS'01)*. 85–94.

[12] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced Neighborhoods for Multi-sided Fairness in Recommendation. In *Conference on Fairness, Accountability and Transparency*. 202–214.

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

[14] Soumen Chakrabarti, Rajiv Khanna, Uma Sawant, and Chiru Bhattacharyya. 2008. Structured Learning for Non-Smooth Ranking Losses. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, Nevada, USA) *(KDD '08)*. Association for Computing Machinery, New York, NY, USA, 88–96. https://doi.org/10.1145/1401890.1401906

[15] 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* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 129–138. https://doi.org/10.1145/3287560.3287570

[16] 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* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 5627–5638.

[17] 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* (Singapore, Singapore) *(SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 659–666. https://doi.org/10.1145/1390334.1390446

[18] 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 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 862–872. https://doi.org/10.1145/3442188.3445924

[19] 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* (Foster City, Silicon Valley, California, USA) *(RecSys '14)*. Association for Computing Machinery, New York, NY, USA, 285–288. https://doi.org/10.1145/2645710.2645774

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

[21] Qinxu Ding, Yong Liu, Chunyan Miao, Fei Cheng, and Haihong Tang. 2021. A Hybrid Bandit Framework for Diversified Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 5 (May 2021), 4036–4044. https://ojs.aaai.org/index.php/AAAI/article/view/16524

[22] Tim Draws, Nava Tintarev, and Ujwal Gadiraju. 2021. Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics. *ACM SIGKDD Explorations Newsletter* 23, 1 (2021), 50–58.

[23] 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. *Information Processing & Management* 58, 6 (2021), 102721. https://doi.org/10.1016/j.ipm.2021.102721

[24] 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* (Foster City, Silicon Valley, California, USA) *(RecSys '14)*. Association for Computing Machinery, New York, NY, USA, 161–168. https://doi.org/10.1145/2645710.2645737

[25] 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, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 172–186. https://proceedings.mlr.press/v81/ekstrand18b.html

[26] Dominik Maria Endres and Johannes E Schindelin. 2003. A New Metric for Probability Distributions. *IEEE Transactions on Information theory* 49, 7 (2003), 1858–1860.

[27] 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* (Bratislava, Slovakia) *(UMAP '17)*. Association for Computing Machinery, New York, NY, USA, 280–284. https://doi.org/10.1145/3079628.3079699

[28] Sarah Eskens, Natali Helberger, and Judith Moeller. 2017. Challenged by News Personalisation: Five Perspectives on the Right to Receive Information. *Journal of Media Law* 9, 2 (2017), 259–284. https://doi.org/10.1080/17577632.2017.1387353

[29] Raul Ferrer-Conill and Edson C. Tandoc Jr. 2018. The Audience-Oriented Editor. *Digital Journalism* 6, 4 (2018), 436–453. https://doi.org/10.1080/21670811.2018.1440972

[30] 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, USA, 2001–2004. https://doi.org/10.1145/3397271.3401213

[31] Florent Garcin, Christos Dimitrakakis, and Boi Faltings. 2013. Personalized News Recommendation with Context Trees. In *Proceedings of the 7th ACM Conference on Recommender Systems* (Hong Kong, China) *(RecSys '13)*. Association for Computing Machinery, New York, NY, USA, 105–112. https://doi.org/10.1145/2507157.2507166

[32] Andreas Grün and Xenija Neufeld. 2021. Challenges Experienced in Public Service Media Recommendation Systems. In *Fifteenth ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 541–544. https://doi.org/10.1145/3460231.3474618

[33] Natali Helberger. 2019. On the Democratic Role of News Recommenders. *Digital Journalism* 0, 0 (2019), 1–20. https://doi.org/10.1080/21670811.2019.1623700

[34] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2022. spaCy: Industrial-strength Natural Language Processing in Python. *Release 3.2.1* (2022). https://spacy.io/usage/spacy-101

[35] David Hume. 1739. *A Treatise of Human Nature*. London.

[36] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (Oct. 2002), 422–446. https://doi.org/10.1145/582415.582418

[37] Harold Jeffreys. 1946. An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 186, 1007 (1946), 453–461.

[38] Michael Jugovac, Dietmar Jannach, and Lukas Lerche. 2017. Efficient Optimization of Multiple Recommendation Quality Factors According to Individual User Tendencies. *Expert Systems with Applications* 81 (2017), 321–331. https://doi.org/10.1016/j.eswa.2017.03.055

[39] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News Recommender Systems–Survey and Roads Ahead. *Information Processing & Management* 54, 6 (2018), 1203–1227.

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

[41] J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. (1975).

[42] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 1 (1951), 79 – 86. https://doi.org/10.1214/aoms/1177729694

[43] Matevz Kunaver and Tomaz Pozrl. 2017. Diversity in Recommender Systems – A Survey. *Knowledge-Based Systems* 123 (2017), 154 – 162. https://doi.org/10.1016/j.knosys.2017.02.009

[44] F. Liese and I. Vajda. 2006. On Divergences and Informations in Statistics and Information Theory. *IEEE Transactions on Information Theory* 52, 10 (2006), 4394–4412. https://doi.org/10.1109/TIT.2006.881731

[45] Jianhua Lin. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information theory* 37, 1 (1991), 145–151.

[46] Felicia Loecherbach, Judith Moeller, Damian Trilling, and Wouter van Atteveldt. 2020. The Unified Framework of Media Diversity: A Systematic Literature Review. *Digital Journalism* (2020), 1–38.

[47] Steven Loria. 2021. textblob Documentation. *Release 0.7.0* (2021). https://textblob.readthedocs.io/en/dev/quickstart.html

[48] Feng Lu, Anca Dumitrache, and David Graus. 2020. Beyond Optimizing for Clicks: Incorporating Editorial Values in News Recommendation. *arXiv preprint arXiv:2004.09980* (2020).

[49] 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 *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 435–444. https://doi.org/10.1145/3209978.3210007

[50] Nicolas Michael Mattis, Philipp K Masur, Judith Moeller, and Wouter van Atteveldt. 2021. Nudging towards Diversity: A Theoretical Framework for Facilitating Diverse News Consumption through Recommender Design. (2021).

[51] 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*. 117–123.

[52] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 2018. Do Not Blame It on the Algorithm: An Empirical Assessment of Multiple Recommender Systems and their Impact on Content Diversity. *Information, Communication & Society* 21, 7 (2018), 959–977.

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

[54] Mats Mulder, Oana Inel, Jasper Oosterman, and Nava Tintarev. 2021. Operationalizing Framing to Support Multiperspective Recommendations of Opinion Pieces. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 478–488.

[55] Newman Nic, Richard Fletcher, Antonis Kalogeropoulos, David AL Levy, and Rasmus Kleis Nielsen. 2018. Reuters Institute Digital News Report 2018. *Reuters Institute for the Study of Journalism* (2018), 39.

[56] Mícheál. O'Searcoid. 2007. *Metric Spaces* (1st ed. 2007. ed.). Springer London, London.

[57] Javier Parapar and Filip Radlinski. 2021. Towards Unified Metrics for Accuracy and Diversity for Recommender Systems. In *Fifteenth ACM Conference on Recommender Systems*. Association for Computing Machinery, New York, NY, USA, 75–84. https://doi.org/10.1145/3460231.3474234

[58] Eli Pariser. 2011. *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think*. Penguin.

[59] 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* (Boston, Massachusetts, USA) *(RecSys '16)*. Association for Computing Machinery, New York, NY, USA, 15–22. https://doi.org/10.1145/2959100.2959149

[60] Lijing Qin and Xiaoyan Zhu. 2013. Promoting Diversity in Recommendation by Entropy Regularizer. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (Beijing, China) *(IJCAI '13)*. AAAI Press, 2698–2704.

[61] Shaina Raza and Chen Ding. 2021. Deep Dynamic Neural Network to trade-off between Accuracy and Diversity in a News Recommender System. *CoRR* abs/2103.08458 (2021). arXiv:2103.08458 https://arxiv.org/abs/2103.08458

[62] 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* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 595–604. https://doi.org/10.1145/3331184.3331215

[63] Jane B Singer. 2014. User-generated Visibility: Secondary Gatekeeping in a Shared Media Space. *New Media & Society* 16, 1 (2014), 55–73. https://doi.org/10.1177/1461444813477833 arXiv:https://doi.org/10.1177/1461444813477833

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

[65] Daniel Steel, Sina Fazelpour, Kinley Gillette, Bianca Crewe, and Michael Burgess. 2018. Multiple Diversity Concepts and their Ethical-epistemic Implications. *European Journal for Philosophy of Science* 8, 3 (2018), 761–780.

[66] Jesper Strömbäck. 2005. In Search of a Standard: Four Models of Democracy and their Normative Implications for Journalism. *Journalism Studies* 6, 3 (2005), 331–345. https://doi.org/10.1080/14616700500131950

[67] Niek Tax, Sander Bockting, and Djoerd Hiemstra. 2015. A Cross-benchmark Comparison of 87 Learning to Rank Methods. *Information Processing & Management* 51, 6 (2015), 757–772. https://doi.org/10.1016/j.ipm.2015.07.002

[68] 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.

[69] Damian Trilling and Marieke van Hoof. 2020. Between Article and Topic: News Events as Level of Analysis and Their Computational Identification. *Digital Journalism* 8, 10 (2020), 1317–1337. https://doi.org/10.1080/21670811.2020.1839352 arXiv:https://doi.org/10.1080/21670811.2020.1839352

[70] 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*. 109–116.

[71] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a Mission: Assessing Diversity in News Recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 173–183. https://doi.org/10.1145/3406522.3446019

[72] Julian Wallace. 2018. Modelling Contemporary Gatekeeping: The Rise of Individuals, Algorithms and Platforms in Digital News Dissemination. *Digital Journalism* 6, 3 (2018), 274–293.

[73] Isaac Waller and Ashton Anderson. 2019. Generalists and Specialists: Using Community Embeddings to Quantify Activity Diversity in Online Platforms. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1954–1964. https://doi.org/10.1145/3308558.3313729

[74] Ningxia Wang and Li Chen. 2021. User Bias in Beyond-Accuracy Measurement of Recommendation Algorithms. In *Fifteenth ACM Conference on Recommender Systems*. Association for Computing Machinery, New York, NY, USA, 133–142. https://doi.org/10.1145/3460231.3474244

[75] 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. *Journalism Studies* 19, 3 (2018), 315–333.

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

[77] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Attentive Multi-view Learning. *arXiv preprint arXiv:1907.05576* (2019).

[78] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: Neural News Recommendation with Personalized Attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2576–2584.

[79] 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)*. 6389–6394.

[80] 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. *ACL* (2020).

[81] Qiong Wu, Yong Liu, Chunyan Miao, Yin Zhao, Lu Guan, and Haihong Tang. 2019. Recent Advances in Diversified Recommendation. *ArXiv* abs/1905.06589 (2019).

[82] 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. *Journalism Studies* 20, 10 (2019), 1440 – 1457. https://search-ebscohost-com.proxy.uba.uva.nl/login.aspx?direct=true&db=ufh&AN=137190691&site=ehost-live&scope=site

[83] Wen Wu, Li Chen, and Yu Zhao. 2018. Personalizing Recommendation Diversity based on User Personality. *User Modeling and User-Adapted Interaction* 28, 3 (2018), 237–276.

[84] 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. *arXiv preprint arXiv:2102.03787* (2021).

[85] Emine Yilmaz and Stephen Robertson. 2010. On the Choice of Effectiveness Measures for Learning to Rank. *Information Retrieval* 13, 3 (June 2010), 271–290. https://doi.org/{10.1007/s10791-009-9116-x}