Building normative diversity into algorithmic news recommendations

Sanne Vrijenhoek

Building normative diversity into algorithmic news recommendations

Sanne Vrijenhoek

Amsterdam, 2026

Building normative diversity into algorithmic news recommendations

ISBN: 978-94-6534-486-7

Cover: Sanne Vrijenhoek
Printed by: Optima Grafische Communicatie

Institute for Information Law (IViR), University of Amsterdam
P.O. Box 15514
1001 NA Amsterdam
The Netherlands

Building normative diversity into algorithmic news recommendations

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 1 juli 2026, te 11.00 uur

door

Sanne Vrijenhoek

geboren te Amsterdam

Promotiecommissie

Promotores:	prof. dr. N. Helberger	Universiteit van Amsterdam
Co-promotor:	prof. dr. C.H. de Vreese	Universiteit van Amsterdam
	dr. L. Hollink	Centrum Wiskunde & Informatica
Overige leden:	prof. dr. B. Bodó	Universiteit van Amsterdam
	prof. dr. D.C. Trilling	Vrije Universiteit Amsterdam
	dr. D.P. Graus	Universiteit van Amsterdam
	prof. dr. R. Burke	Colorado Boulder
	prof. dr. S. Verberne	Universiteit Leiden
	dr. M.S. Pera	TU Delft

Faculteit der Rechtsgeleerdheid

The work described in this thesis has been carried out within the AI, Media & Democracy Lab and the Institute for Information Law of the University of Amsterdam.

University of Amsterdam

Contents

1 Introduction

1.1 Introduction

How many news articles do you want to read in a day? Despite your best efforts, it is very unlikely that you read *all* the news you come across, let alone all articles that have been published. The BBC publishes on average around 200 items per day[1]; similar numbers can be found at Danish tabloid Ekstra Bladet [144]. Even if reading an article only takes you one minute, you would have to spend more than three hours per day to read all the articles from just one of these outlets. Then, you still might want to get information from other outlets; 7 on average for an adult in the UK [202], ranging from local newspapers, to large (inter)national brands, to influencers on social media.

To manage the overload of news trying to get our attention, we rely on others to help us filter out the news that we want to read [76, 262]. In the time of physical newspapers and linear television programming, this was mostly done by news editors, who are trained to construct an overview of the news that is reflective of their outlet's brand and mission. They would ensure that the news on offer is balanced, with a mix of topics that are important for everyone to know, that have a big impact on a smaller group's life, or that are simply entertaining to read. In doing so, they influence which issues in society get the public's attention, and that through their increased visibility are considered most important [173]. The shift to online news consumption has changed this dynamic [262]. Instead, recommender sys-

[1] https://pressgazette.co.uk/media-audience-and-business-data/at-1500-stories-per-day-mail-online-is-uks-most-prolific-news-website/, accessed 03/11/2025

tems, which can be found on both legacy news websites[2] and social media platforms, aim to predict what we might want to see based on our past engagement patterns, and those of people like us [22, 217].

In 2024, 37% of media organizations thought that recommendations would play a 'very important' role in the upcoming year [192]. Initial experimentation seems promising; while it is notoriously difficult to measure the business value of a recommender system [126], Kruse et al. [143] noted a substantial increase in clicks after deployment of their recommender systems at the Danish tabloid Ekstra Bladet. Rather than leaving the page immediately after reading an article, readers would stay on the website to read more articles. This is important to a domain that is increasingly dependent on social media platforms to direct traffic to their websites, and as such a news recommender system may both make audiences more satisfied with their news offer, and generate additional income in the process [22]. However, Kruse et al. [143] also noted that different recommendation models could recommend very differently. Different algorithms would, with similar data, similar targets and similar model performance, make very different choices with regards to what type of content is recommended, and at what position in the recommendation. Vrijenhoek [277] noticed similar behavior when training benchmark recommender systems on the public Microsoft News Dataset. For example, one model would prioritize sports content, while another almost always included articles on food and recipes in the top 3 recommended articles. Arbitrary design choices imposed a restriction on the type of content users would see, making the recommendation not reflective of the breadth of the organization's news offer.

This is a direct consequence of the way that the technology is designed. In their current form, news recommender systems primarily aim at *engagement* [12]. Their goal is to keep users on their platform as long as possible, and have them engage with many pieces of content. On social media platforms sensationalist and extreme content tends to get a lot of engagement and thus gets disproportionally more distributed, whereas 'boring' news content usually does not [98]. News organizations may start to disproportionately recommend soft news [70]. As a result, recommender systems get a bad reputation; from the Filter Bubble described by Eli Pariser [185, 211], to the 'rabbit hole' of the 'manosphere' on TikTok[3].

In light of the societal importance of news and the issues currently reported with recommender systems, it may then be tempting to say that we should stick to manual curation

[2] See for example this article describing the implementation of a recommender system at the New York Times: https://open.nytimes.com/how-the-new-york-times-is-experimenting-with-recommendation-algorithms-562f78624d26, accessed 19/01/2026

[3] https://www.groene.nl/artikel/vrouwen-geven-niets-om-jou, accessed 03/11/2025

only. There is, however, also a clear case to be made in favor of news recommender systems. Despite the wide availability of news from many different perspectives and sources, the Reuters Digital News Report [193–195] consistently reports a downwards trend in the amount of news that people read. Some people avoid the news altogether, while others only avoid selected topics; almost 40% of those surveyed for the 2024 Reuters Digital News Report [194]. People report feeling overwhelmed by the amount of news, and state that it 'brings down their mood'. On the other hand, people would read the news to get perspective, inspiration, updates, education, help, connection or diversion [85]. While human editors are trained to know what content can give perspective, inspiration, updates etc. to large parts of the population, they cannot account for the needs and interests of individuals. People are aware of and have different perspectives on important topics, are inspired by different things, want to be engaged with or educated on different things. News recommender systems, that account for the individual needs and preferences of audiences, could play a vital role in filling this gap [22, 188].

Building on this, Helberger et al. [112] note that exactly by tailoring to individual readers a *diverse* recommender system could also be used to strengthen the public sphere. After all, if an algorithm could limit what a user sees and engages with, it could also be used to do the opposite, namely to expand their worldview, to help people be more informed, or to expose them to events and ideas they were not aware of before [112]. Doing so requires shifting the dominant recommendation paradigm, which prioritizes engagement, towards ensuring that the diverse information needs of individuals and society are met [17]. Under this view diversity is not a goal of itself, but a means to "reach democratic goals such as an informed citizenry or an inclusive public discourse" [158].

At first glance, this may seem like a fairly simple and straightforward thing to do. If we know what news a person has read in the past, we should recommend them something different. But what does that mean, different? Different in terms of what? What does it mean to satisfy information needs? Do we need to show readers all topics and opinions? In what quantity?

There is no clear and universally true answer to this; it is a *normative* question, the answer to which is based on values, principles, beliefs and moral considerations, and eventually grounded in the role that news plays in society [111]. A commercial entity would answer differently than a public service organization, an organization in the Netherlands different from one in Brazil. To capitalize on the potential of news recommender systems, it is important that diversity-focused news recommender systems are able to reflect the values and nuances of the organization for which they are deployed. However, building technology that explicitly reflects such norms and values has thus far remained largely neglected

in the technical design of news recommender systems [17, 180].

1

The goal of this thesis is to expand on the theoretical foundations laid out by Helberger to work towards news recommender systems that can be evaluated on normative notions of media diversity. Following Helberger [111] and Bernstein et al. [17] I consider a recommendation to be diverse if it "succeeds in informing the user and supports them in fulfilling their role in democratic society" [279]. This leads us to the central question of this dissertation: *How can we evaluate news recommender systems on their normative diversity?*

1.2 Key concepts

This dissertation primarily revolves around the concepts of *news recommendation* and *diversity*. The paragraphs below provide a short introduction on their definition and current state of the art.

1.2.1 News recommender systems

Recommendation as a field is about ranking a set of candidate items in such a way that the users will find something they are interested in as high up in the ranking as possible [27]. News recommender systems then come in many shapes and sizes. First, we can split them into two types: *personalized* and *non-personalized* recommender systems. Non-personalized recommendations are the same for all the users of the system. Think of a 'related stories' section below a news article, listing articles on the same topic or event. These recommendations are often based on similarity, easy for a newsroom to control, and comparatively cheap to run [103]. Personalized news recommendations attempt to tailor to a users' specific preferences, and as a result every user receives a different recommendation. Traditionally, personalized news recommender systems can then be divided into *content-based recommendation* and *collaborative filtering*. Content-based recommendation aims to predict whether an item is relevant to a user based on the content of the article. For example, an item might be relevant because it is similar to an article the user has engaged with in the past. On the other hand, collaborative filtering assigns relevance based on engagement patterns of other readers, with the rationale that if an article is popular among people that are similar to you, you might like this article too. Recent approaches towards neural news recommendation often incorporate elements of both article content and popularity among other users [124, 132]. They may for example differentiate between readers' short-term and long term interests [219, 292] or use knowledge graphs to more accurately model the text of a news article [123].

From a technical perspective, news recommendation has challenges that other recommendation domains do not [132]. While a movie can be watched years after it was initially released, a news article by its definition is only relevant for a very short amount of time before being replaced by other articles. Furthermore, most people on a news website are not logged in; they enter the system through some external referral, like a search engine or social media platform, and exit the system immediately afterwards. Because of this, historic data to base a recommendation on is often limited [144]. Related to this is the data sparsity problem; there are many more articles a user has not seen or interacted with than articles that they have. This all contributes to the difficulty of predicting what makes an article relevant to a reader.

Social science studies identify an additional range of challenges with regards to news recommendations. They note, for example, the lack of journalistic values embedded in the systems [11, 188, 229], and the difficulty of modeling and prioritizing such values [185, 186]. Academic work that aims to tackle this remains primarily theoretical, with few attempts to build practical solutions [180]. At the industry level, technical- and editorial teams often work in isolation, and little knowledge exchange happens between the two [60, 186, 189]. Technical teams working on news recommender systems are often under pressure from the business department to produce quick results [270], and in the process the newsroom is more often than not hardly involved in development [229]. To resolve these issues there is a general agreement on the need for both more interdisciplinary work and more industry-academia collaborations [17, 158, 180].

A note on news recommendation versus social media feeds

The boundaries between recommender systems and social media feeds are often blurred. Academic research on both technologies can be found in similar journals and conferences, while regulatory efforts, such as the European Digital Services Act, focus primarily on social media recommender systems. However in this thesis I explicitly focus on news recommender systems built for and by news organizations, rather than on social media feeds. While the 2025 Reuters Digital News Report [195] reported that people primarily found their news through social media, social media platforms have a much broader purpose than just sharing news. This makes the recommendation task substantively different. Social media platforms have extensive information on their users and their interests, but do not produce content themselves. Conversely, most news organizations have limited information on their users, but fully control the data that is produced and its quality. Most news organizations have the journalistic mission to *inform, update and entertain*, or some variation thereof; the purpose of social media platforms is primarily to entertain, with news

1

as a byproduct and a means to more entertainment[4]. Furthermore, news organizations typically aim to give a concise overview of what is currently important, which means that the number of items that are shown to the readers is limited. In comparison, a social media feed aims to keep the user on the platform as long as possible, and feeds are potentially endless. While some of the findings outlined in this thesis may be relevant for social media platforms, the main focus is on creating technology that benefits 'traditional' news organizations.

1.2.2 Diversity

Early recommender systems were primarily based on similarity: if you are interested in A, you might also be interested in similar item B; or, since user A who is similar to you liked B, you might also like B. But as early as 2001, Smyth and McClave [242] noted that often recommendations were *too* similar, and as a result became monotonous. They claimed that a good recommendation should find the right balance between similarity and diversity. In 2011, Vargas and Castells [273] wrote down what became closest to a standard for evaluating diversity [148]:

$$div(R|u) = \frac{2}{|R|(|R|-1)} \sum_{i_k \in R, l<k} d(i_k, i_l)$$

Here, $R|u$ denotes the recommendation R issued to user u, $i \in R$ the items i in the recommendation, and $d(i_k, i_l)$ the distance between items i_k and i_l. This formula is also known as the *intra-list distance* (ILD), or the average dissimilarity of the items in the recommendation. One of the appeals of ILD is that it is generic, and applicable to all recommendation domains. Whether you talk about movies, music, job candidates or news, defining diversity comes down to defining what makes two items in your recommendation (dis)similar from each other. In news recommendation, this similarity scoring would often be implemented through a cosine similarity, measuring the distance between word vectors representing the news articles [274], or other statistical distance metrics such as the Gini coefficient or Shannon entropy [161, 217]. Bauer et al. [12] note that 62 of the in total 183 academic papers on news recommendation published before the end 2022 mention some notion of diversity[5]. Importantly, this formalization does not take position on what makes two items in a recommendation different from each other. It also does not impose an expectation of the 'right' amount of diversity, though implicitly we may expect that a higher value is considered better. Lastly, the formalization only includes items from the current recommen-

[4] Though there is also evidence of social media platforms directly furthering political agendas. See for example Leerssen [150]

[5] Note that these 183 papers also include theoretical contributions from the social sciences, including the works of Helberger

dation, and does not consider the general news offer, current events, or other items that a user has seen in the past.

Here a clear difference can be observed with how other domains, such as media studies or information law, conceptualize diversity. Media studies are typically concerned with notions of *source* diversity (does the media system in general have enough sources and perspectives to choose from) and *exposure* diversity (what did audiences eventually consume) [180, 190]. Building on that, diversity in the normative sense is about helping people play a role in democratic society; a diverse news offer helps people be informed about important issues, gain a deeper understanding of other groups in society, or helps them to let their voices be heard [76]. It thus ties into larger societal concepts, such as democracy, freedom of expression, participation or tolerance [111]. A recommendation that would score high on diversity according to the intra-list distance may not be considered to be diverse according to the criteria maintained by other fields. Loecherbach et al. [158] note that there is "little to no overlap between concepts and operationalizations [of diversity] used in the different fields interested in media diversity." In short, additional work is necessary to bridge the gap between normative conceptions on diversity, and technical implementations.

1.3 Outline

In order to capitalize on the potential of diverse news recommender systems, additional work is necessary to bridge the gap between normative conceptualizations of diversity and what is technically implementable. To accomplish this, this thesis focuses on the following questions:

> **RQ1:** How is normative diversity defined and understood in both theory and practice?
>
> **RQ2:** How can normative notions of diversity be translated into concrete and informative evaluation metrics for news recommender systems?
>
> **RQ3:** What conditions need to be in place in order for diversity metrics to be adopted in practice?

Each of these research questions is a multifaceted problem, where different disciplines hold different pieces of the puzzle. Furthermore, solutions that need to be adopted in practice also need to be grounded in the reality of industry practitioners. As such, answering the research questions cannot be done in isolation, and need the combination of different perspectives.

In **Chapter 2**, I scope the current state of the art of public service media recommender systems in the Netherlands through semi-structured interviews with practitioners. The

1

chapter describes a think-aloud exercise to identify how the practitioners would conceptualize diversity in their recommender systems. The results of the exercise are split up in *aspects*, or what is being measured, and *tactics*, or how the recommendation is diversified. In doing so, I study whether there are commonalities in how practitioners operationalize diversity, either because they come from the same organization, or because they have similar roles. As a result, this chapter contributes to an understanding of the requirements of industry practitioners at different public service media institutions for their recommender systems and the diversity incorporated therein. By answering how practitioners operationalize diversity, Chapter 2 contributes to answering **RQ1**.

Besides current industry practices, I study what role personalized recommender systems *could* play. In **Chapter 3** I first look at theoretical notions of the different roles that news can play in democratic society, and how news recommender systems could support those roles. Based on four models of democratic theory (the Liberal, Participatory, Deliberative and Critical models), I propose new evaluation metrics for news recommender systems that aim to reflect the nuances of each of the different models: Calibration (to what extent is a recommendation tailored to a readers' preferences), Fragmentation (to what extent are readers aware of the same events), Activation (the amount of neutral versus emotional content in a recommendation), Representation (which viewpoints are expressed) and Alternative Voices (whose viewpoints are expressed). Through its investigation on how diversity is defined in theory, and making a first step towards translating that theory into evaluation metrics, Chapter 3 contributes towards answering both **RQ1** and **RQ2**.

As stated in Section 1.2.2, the standard technical formalization of diversity in recommender systems based on the intra-list distance is not suitable to express normative notions of diversity. Inspired by the original Calibration metric, as proposed by Steck [247], **Chapter 4** proposes to formalize diversity as a *rank-aware divergence* metric. Through this, diversity metrics can be designed by answering the following three questions: 1) what aspect of the (items in the) recommendation needs to be measured; 2) what context distribution are we comparing the recommendation to; and 3) do we expect the divergence between the recommendation and the context distribution to be high or low. We demonstrate these metrics using the benchmark news recommender systems trained on the Microsoft News dataset [293], and the code used for doing so is available on GitHub[6]. Furthermore, I discuss how seemingly innocuous choices in metric design can have a

[6]The repository for doing this was refactored at a later time, and can be found here: https://github.com/svrijenhoek/RADio-

strong impact on the metrics' behavior, and argue that decisions on metric design should be taken with a good understanding of what the metric aims to express or accomplish. As a result, Chapter 4 contributes to answering **RQ2**.

News organizations typically rely on algorithms developed in academic research. However, researchers often do not have access to real-world systems, and are instead dependent on public datasets. **Chapter 5** analyzes the most commonly used public datasets (MIND, Adressa, Globo and EB-NeRD), and specifically how the datasets can and cannot be used for research on diverse news recommendation. To some degree, each of the datasets has aspects that are suitable to a subset of the diversity problem. However, there are also a number of structural problems. All datasets stem from commercial outlets, and cover a very limited amount of time where from a news perspective nothing significant happened; for example, none of the datasets cover the weeks surrounding national elections. This makes it difficult to study and develop new approaches to diversity-aware news recommendation,and the chapter therefore also describes the role the law could play in acquiring data for research and development. By analyzing the data that is necessary to develop diversity-aware news recommender systems, Chapter 5 contributes to answering **RQ3**.

The previous chapters have frequently emphasized that diversity is a multi-faceted concept, and that its operationalization is strongly dependent on the media organization. Furthermore, different groups within an organization may have different perspectives on its meaning. This is incompatible with the standard way of designing a news recommender system, which is, usually, designed by the technical team in isolation [60, 73]. Editorial teams are only involved after the recommender system is well under development, which contributes to the fact that editorial values are often not incorporated [229, 270]. **Chapter 6** describes a collaboration with the Danish national news organization Ekstra Bladet (EB) on the co-design of evaluation metrics that would be relevant for their news recommender system, with participants from Ekstra Bladet's technical, editorial and business teams. Through four workshop sessions of two hours each, the participants discussed the goal of EB's recommender system, and how they could measure whether that goal has been reached. To this end the participants co-designed metrics with the same structure as the one explained in Chapter 4; by defining what needs to be measured, and in what context. I identify the challenges encountered in the process, and describe the steps that need to be taken by both academia and the organization itself to eventually make the recommender systems reflect their editorial values. Chapter 6 combines the insights of all previous chap-

ters, and through its focus on the conditions that need to be in place to make normative diversity metrics land within a news organization primarily contributes to **RQ3**.

Throughout the thesis, I move from a conceptual analysis of the term 'diversity' down to the challenges encountered in its practical implementation. I conclude the thesis by explaining my core findings, and outlining the direction future work could take.

1.3.1 On "social science" and "computer science"

This thesis relies on insights from many different disciplines, including media studies, journalism studies, philosophy, communication science, information law and political science. Continuously listing them all will become repetitive to the reader, but summarizing them under a single term is not a straightforward task. In this thesis I often group those disciplines that work on technical solutions under "computer science", and those that do not under "social science". To experts this may read as reductive of the nuances and contributions of their discipline. To others, it may imply a false dichotomy where there are two camps in opposition of each other, while in reality this is not the case. Nevertheless, this categorization is used for simplicity and clarity, with the understanding that each discipline's unique contributions are instrumental to the interdisciplinary nature of this research.

1.4 Methodology and Contributions

Section 1.1 argued that diversity is an inherently interdisciplinary problem, which is reflected in the set-up of this dissertation. Rather than centering one type of methodology, I focus on the integration of multiple research methodologies to address the complexity of the problem comprehensively. In this, I reflect the call of among others Bernstein et al. [17], Mitova et al. [180] and Loecherbach et al. [158] for both interdisciplinary research approaches, and transdisciplinary (industry/academia) collaboration. Each chapter has benefited from input from experts in their given domain, either directly as co-authors or through one of the many discussion at the AI, Media and Democracy Lab[7], the Institute for Information Law[8], or the general context of the ACM Conference on Recommender Systems[9]. To acknowledge the contributions of the co-authors of the papers that the chapters are based on, they are written using plural personal pronouns. To facilitate technical adoption, the articles are primarily written with a technical audience in mind.

[7] https://www.aim4dem.nl/
[8] https://www.ivir.nl/
[9] https://recsys.acm.org/

In Chapters 2 and 3 I conduct *conceptual research* to broaden and advance our understanding of diversity as a concept. The chapters follow different methodologies; Chapter 2 contains *semi-structured interviews* to understand the current standards and state of the art in industry. In Chapter 3 I combine different strands of *literature* from media studies, communication science, information law and computer sciences to propose alternative evaluation metrics. The research contributes to a clear understanding of the requirements for a diverse news recommender system.

In Chapter 4 I extend the results of the conceptual research and the initial metrics proposed there with more elaborate *metric design*. The chapter proposes a mathematically sound base formalization for a diversity metric, and the *prototype* of a metadata enrichment pipeline. Both metrics and pipeline are tested on the Microsoft News Dataset, including a sensitivity analysis to test whether the metrics respond to alternative configurations. The contribution of this work is in the mathematical formalization of the metrics, and the code is available in a GitHub repository[10]. Other researchers and practitioners can leverage this code for their own systems.

With the added knowledge of the type of data structures necessary for measuring diversity, Chapter 5 describes a *dataset analysis* of the public news recommendation datasets that are currently available to researchers. We analyze these datasets on their suitability to diversity research, and identify blind spots. It furthermore contains *doctrinal legal analysis*, done in close collaboration with a legal scholar, to investigate how the law can facilitate acquiring additional data.

Finally, Chapter 6 studies how media organizations can formulate their own goals and brand of diversity for their own systems through *participatory co-design*. This study underlines the needs and requirements of practitioners, and how metrics can be made informative. It also provides clear pathways for additional research and development.

1.5 Contributions

This dissertation contributes to the state of the art of recommender system design with a multi-method approach, which through its interdisciplinarity leads to a richer conceptualization and advanced understanding of the notion of diversity in the design of news recommender systems. Furthermore, I zoom out of purely technical considerations by also looking at the conditions that need to be in place for the adoption of the proposed evaluation metrics, namely by considering how academia can and cannot contribute to advancing the state of the art, and by studying the modus operandi and requirements of industry

[10] https://github.com/svrijenhoek/RADio-

practitioners.

1

1.5.1 Theoretical contributions

- a proposal of evaluation metrics for the normative diversity of news recommendations based on democratic theory (Chapter 3)
- a flexible and rank-aware base formalization for a diversity metric, expressed as the divergence between two distributions (Chapter 4)
- a doctrinal legal analysis of how the law can facilitate acquiring additional data for diversity research (Chapter 5)*(in close collaboration with a legal scholar)*

1.5.2 Empirical contributions

- a mapping of the different ways in which industry practitioners of public service media organizations in the Netherlands operationalize diversity (Chapter 2)
- an implementation of the normative diversity metrics on the Microsoft News Dataset, including the effects of different design choices (Chapter 4)
- an analysis of public datasets used for news recommendation and their suitability to diversity research (Chapter 5)
- a study of how media organizations can formulate their own diversity goals and metrics through participatory co-design (Chapter 6)

1.5.3 Resource contributions

- a publicly available Python repository for implementing diversity metrics, including a prototype metadata enrichment pipeline for extracting relevant metadata from news texts (Chapter 4)

1.6 How this research has inspired follow-up research by others

The diversity metrics proposed in this dissertation have been explicitly used as departure point for research conducted by various research groups without my direct participation. For example,

- Hada et al. [97] studies the Representation and Fragmentation in online conversations.

- Heitz et al. [104] incorporated the metrics in their news recommendation framework and app Informfully
- Li et al. [152] proposes a random walk algorithm to incorporate the metrics in recommender optimization
- Mattis et al. [171] tests the effects of alternating amounts of Activation and Alternative Voices on several dimensions of participants' democratic participation
- Dattawad et al. [51] calculated the effect of different news frames on Calibration, Activation and Representation

2

How public service media professionals in the Netherlands conceptualize diversity in their recommendations

Authors: Sanne Vrijenhoek, Savvina Daniil, Jorden Sandel, Laura Hollink

Original title: Diversity of What? On the different conceptualizations of diversity in recommender systems

Published in: FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency

https://doi.org/10.1145/3630106.3658926

Abstract

DIVERSITY is a commonly known principle in the design of recommender systems, but also ambiguous in its conceptualization. Through semi-structured interviews we explore how practitioners at three different public service media organizations in the Netherlands conceptualize diversity within the scope of their recommender systems. We provide an overview of the goals that they have with diversity in their systems, which aspects are relevant, and how recommendations should be diversified. We show that even within this limited domain, conceptualization of diversity greatly varies, and argue that it is unlikely that a standardized conceptualization will be achieved. Instead, we should focus on effective communication of what diversity in this particular system means, thus allowing for operationalizations of diversity that are capable of expressing the nuances and requirements of that particular domain.

2.1 Introduction

The concept of diversity is at the same time omnipresent and very ambiguous [78]. In popular discourse, diversity usually refers to the variation of human characteristics, often related to a notion of identity politics [18]; in biological research, as a qualifier to the health of an ecosystem [267]; and in media studies, as a concept adjacent to pluralism, expressing whether a news selection contains a plurality of sources, voices and opinions [135]. While all interpretations are valid and intuitively seem to reflect similar concepts, they differ in their operationalization in a way that is unique to their domain. At the same time, diversity is consistently noted as a social value that is beneficial to pursue, though in different capacities.

In the recommender systems domain accounting for the diversity of a recommendation can help avoid monotony [29, 303]. This follows from the assumption that a recommender system that is purely optimized on predicted relevance will result in a feedback loop and thus prioritize similar content, leading to 'more of the same'. There is as such also a business case to be made for diversity. While it has been challenging to show empirically, diversity may lead to higher user satisfaction and retention, increasing revenue in the process [71, 126]. This argument can be extended for the news domain, where worries over filter bubbles that reinforce existing beliefs by only showing content that align with a user's preferences are especially prevalent [178, 306]. Having a news recommender system that pursues diversity could help expose users to things different than they are used to or expect seeing, and tailor to their specific information needs [111, 112]. For machine learning in general, it is also important to account for diversity in a social context. Since many algorithms are trained on datasets that are not representative of all groups in society, not accounting for diversity may further "amplify stereotypes, alienate users, and further entrench rigid social expectations" [179].

The many different interpretations of diversity are a fundamental challenge to the practical development of recommender systems. Evaluating the performance of a recommender system requires objectively measurable properties, and the field strives to do so in a standardized manner [304]. Standardization allows for comparability and reproducibility of results and algorithms[1], and is achieved through, among others, the use of benchmark datasets, sampling techniques and evaluation metrics such as MRR or

[1] Recent strands of research have noted that this perceived notion of standardization is often false. Even with commonly accepted principles there are often significant differences in implementation, and thus output [233, 255]

NDCG [302], but has not yet been achieved for diversity. Loecherbach et al. [158], who did a literature review of existing measures of diversity in media studies, note that "there is little to no overlap between concepts and operationalizations used in the different fields interested in media diversity". Lawrence [149], executing a conceptual analysis of the term 'diverse books', notes that the topic is inherently political, and that "we have yet to arrive at a clear consensus on what the modifier 'diverse' means in this and other instances".

The conceptual unclarity around diversity makes that not only may we implement it differently, we may also *mean* something completely different when we use the term. One could argue that diversity is, in fact, an essentially contested concept [39, 89], meaning that it is open for discussion and debate and that we are unlikely to reach consensus on its meaning. As such, striving for agreement or a clear definition may lead to a standstill and hinder progress, as it is unclear what a good operationalization or implementation would look like. It may cause the conceptualization of diversity to be driven by the information that is available or more easily suitable for quantification, rather than what is desirable; furthermore, blindly trusting 'objective' measures "may obscure the fact that the conceptualization of diversity is, eventually, a normative choice" [136]. Instead, what we need may be a more flexible operationalization that is capable of reflecting the nuances and requirements of the domain it is deployed in.

The aim of this paper is to explore the dimensions in which industry practitioners within a limited domain conceptualize diversity. To this end we conduct a series of semi-structured interviews with practitioners from three different public service media organizations in the Netherlands: a broadcaster, a news organization, and a library. The choice of these organizations is deliberate; though the medium through which they do it differs, they all play an active and important role in the dissemination and curation of information and ideas. As such, we expect there to be a fairly established conceptualization of diversity within the organization, and comparatively more overlap between them than with, for example, a commercial music recommendation service.

Through analysis of the interviews, and guided by the 'components' of diversity formalization defined in Vrijenhoek et al. [279], we aim to answer the following questions: what *goals* do the organizations aim to achieve with the recommendations, which *aspects* do they consider relevant for diversity, and through which *tactics* are the recommendations diversified. We find that even within the limited domain of public service media recommendation, there is a wide variety of conceptualizations of diversity. With this, we underline that a standardized definition of diversity in recommender systems is likely not achievable. However, this empirical categorisation, albeit non-exhaustive, can assist in the pro-

cess of conceptualizing and implementing diversity in a particular concrete context.

2.2 Related work

Recommender systems have long moved from evaluating recommendations solely based on accuracy-related metrics. Other metrics such as novelty, serendipity and diversity have become a common part of evaluation practices [29]. Diversity in recommendation is a current topic, and multiple diversification methods have been proposed in recent years [87, 148]. In recommender systems research, diversity is often viewed as the opposite of similarity; a list of recommended items is diverse if the items are sufficiently different between them along a set of axes [305]. However, Jesse et al. [129] point to a gap between human perceptions on diversity and intra-list similarity (ILS) commonly used to assess diversity in offline recommender systems experiments. Through a user study, they find that while ILS can be a good proxy, the details of the implementation matter and require validation in a given domain and application.

When diversity is seen as a social value that needs to be incorporated in a system, standardizing its definition and operationalization becomes even more challenging. Mitchell et al. [179] define diversity in a subset selection task (e.g., the construction of a list of items to recommend) as "variety in the representation of individuals in an instance or set of instances, with respect to sociopolitical power differentials (gender, race, etc.)". As such, they view diversity as a concept with inherently sociopolitical connotations in contrast to the more general term 'heterogeneity'.

In the context of media recommendations, diversity is often closely associated with pluralism. According to Karppinen [136], the interpretation of pluralism and diversity in media is dependent on the political and normative understanding of the role of media in society. Additionally, defining media diversity is further complicated by its often contradictory aspects and interpretations. Van Cuilenburg [266] point out an antithesis between the normative media diversity frameworks of *reflection* and *openness*; one requires content distribution to proportionally reflect societal distributions of relevance, while the other corresponds to perfectly equal representation and attention to all people and ideas in society. Similar contradictions have been laid out in the domain of library and information studies [149]. Based on a systematic literature review on media diversity, Loecherbach et al. [158], also referenced in the introduction on the little overlap between different fields' conceptualization of diversity, conclude that relevant research should be guided by an interdisciplinary effort to define and benchmark the concept of media diversity, along with its different sub-dimensions. Other work studying diversity in recommender systems would

typically acknowledge the complexity of the concept, yet model it following existing technical standards; either as a distance to a user's reading history [101], political leaning [107] or as pair-wise distance between the items in a recommendation, for example in topics, categories and/or tone [185]. In this work, we aim to contribute to this effort by attempting to map the dimensions of diversity in media recommender systems.

2.2.1 Diversity in Public Service Media

Diversity in recommendation is especially relevant for public service media (PSM), whose role and societal responsibilities call for a careful consideration of the content they produce and give exposure to. Many media organizations underline the potential of recommender systems and the importance of reflecting editorial values such as diversity therein [23, 93, 143, 187]. PSM are often required to offer diverse content, which might be at odds with the primarily commercial use of recommender systems that mainly intend to increase media consumption, and therefore, profit [118]. Even if we assume societal agreement on the importance of providing diverse content for PSM, the practical implications are harder to lay out. Helberger [111] outlines four models for the normative conceptualization of diversity in recommender systems that serve different purposes: the liberal, the deliberative, the participatory, and the critical models. While each of the models is in its own way relevant for PSM, the work suggests that the critical perspective, which focuses on the visibility of minority voices that are often disadvantaged in public platforms, is less likely to be encountered in commercial applications, and therefore could be partly served by PSM. Regardless, few have succeeded in concrete implementations that reflect diversity as a normative value, in part due to the gap between journalistic values and recommender system evaluation metrics [278]. Møller [182] suggests that "the abstract nature of journalistic values make them hard to account for computationally", and that "[h]uman journalists have an important role to play in these processes not only to help conceptualise the values themselves but also as part of new algorithmic news practices." Translating values into concrete algorithmic practices is thus as challenging as it is necessary.

To bridge the gap between theory and practice, the perspectives of practitioners in a given domain can assist, orient, and ground research. On that note, researchers have conducted interviews with practitioners on the interaction between emerging algorithmic systems and norms and values. Sørensen [254] interviewed developers, data scientists, and project leaders from nine European PSM organizations on the topic of implementing recommender systems. They report that, while interviewees believe diversity to be an essential aspect of their catalogue, they are reluctant to depend on an algorithmic implementa-

tion instead of the traditional editorial control. Additionally, they attribute the general lack of formal definition of diversity to the different understanding between politicians, practitioners, and users. Bastian et al. [11] interviewed practitioners from different departments of two newspapers regarding the impact of algorithmic news recommenders on their organization's norms and mission, as well as how to integrate them in the design of news recommenders. They found that, while the interviewees attach varying degrees of importance to different values, diversity is perceived as one of the core values for news recommender design and implementation. During the interviews we conducted, we specifically focused on the conception and implementation of diversity, which allowed for an in-depth outline of the perspectives and practices of the interviewed professionals.

2.3 Method

For this study we conducted a series of semi-structured interviews with three public service media organizations in the Netherlands: a broadcaster, a news organization, and a library. Interviews were carried out in-person and on site in the offices of the interviewees, and took place over a span of four months, between December 2022 and March 2023. At this time each of these organizations were, at different stages of completion, (exploring the possibility of) developing a recommender system to effectively serve content to their users, which also means that decisions about incorporating diversity in the recommendations were actively being made.

2.3.1 Candidate Selection

Potential candidates were approached through a snowball sampling technique: after initial contacts with the organization were established, interviewees were asked for recommendations of colleagues to interview in the next round. We actively tried to find a set of participants that reflected the composition of the organisations and the relevant figures in the design, deployment and use of the recommender system. This process yielded twelve participants in total: six at the broadcaster, four at the library, and two at the news organization. Following Smets et al. [241], our goal was to have a good spread of different types of stakeholders, and sought participants with roles related to Business (four participants), Curation (two participants), Product owner (three participants) and Technology developer (three participants). During the snowball sampling we would explicitly ask participants whether they knew of potential interviewees with a role we had not covered yet. Finding all different roles did not always succeed, and we acknowledge this as a limitation of our work. This is also why we are adamant about not making normative claims about the definition of

diversity per domain, but rather present it as an exploratory study. Participants were predominantly male (9 out of 12) and white (11 out of 12). This is reflective of the workforce but a caveat for generalization, further outlined in Section 2.3.4. Disclosing too many details about individual participants and their roles would undermine the promised anonymity, and thus each participant is assigned a code reflecting only their organization.

2.3.2 Interview structure

The interviews consisted of four parts. In the first part, interviewees introduced themselves and their role within the organization. In the second part interviewees were asked about their general conceptions of diversity, and how these conceptions related to their organization. Here, the goal was to understand the mission of the organization in question, and the role diversity plays in that mission. In the third part interviewees described their knowledge of the recommender systems that were in use by their organization, either in planning or in production. This served as a check that participants were sufficiently informed to be included in the analysis, and to prime the interviewees for the fourth part, in which they were specifically asked about the role of diversity in the recommender systems of their respective organization. The aim here was to go beyond what currently exists, and instead focus on what the recommender system should look like in an ideal situation. This part of the interview also contained a small experiment, in which the interviewees were asked to take on the role of a recommender system and rank a set of items while keeping diversity in mind. During the experiment participants were asked to voice their thought process by thinking out loud [33]. Our primary interest was not the final recommendation generated, but which aspects of the items participants considered before (not) including an item. For each organization we prepared a set of 15 candidate items based on the organization's (potential) catalogue, and included the metadata a user would see when interacting with the system. We attempted to ensure (to the best of our abilities) that there would be enough diversity in this candidate list present. To cover our own blind spots we would ask participants at the end of the experiment whether there was a type of content that they were missing. We acknowledge that our own conception here steers the type of diversity that could potentially be found (see also Section 2.3.4).

For each of the organizations, the metadata always included the items' title, summary and a cover photo. For the library, we also included the authors' name and a set of keywords about the book; for the broadcaster, the title and description of the series the item was part of (if applicable); and for the news organization, the time of publication and the first few paragraphs of the news item. Based on this candidate set, we asked participants to make a diverse selection of 5 items. In order to not influence people's thought processes

in what could potentially be relevant aspects of diversity, we deliberately left instructions vague, and did not include a profile of a user to create a recommendation for in the instructions. Instead, we considered the participants' questions for clarification as part of what they deemed relevant for diverse recommendation.

2.3.3 Coding and analysis

We executed an inductive thematic analysis on each of the interviews, guided by the research questions posed in Section 2.1: what the organizations aimed to achieve with a (diverse) recommendation, which aspects of the items they would consider during recommendation, and which tactics they would employ to diversify the recommendation. The first two authors created a coding scheme based on half of the interviews, which after completion were discussed and merged. With the resulting coding scheme, the authors annotated the half of the interviews they had not previously seen, and extended the coding scheme when gaps were identified. Additionally, after the respective coding schemes were merged, the two authors independently coded the same interview and proceeded to compare their outputs. This step helped ensure practical consensus, as the authors were able to align on the specific nuanced interpretation of each of the codes, as well as how it applies in the context of the interviews. Based on the resulting coding scheme, we created a diversity 'map' of relevant aspects and how they relate to each other, which will be discussed in the next section. The coding scheme is included in Appendix 2.6. After the first draft of the paper was finished, we reached out to all participants. We shared with them which quotes had been attributed to them, and asked them to reach out in case of misunderstandings.

2.3.4 Limitations

There are a number of important caveats to consider in light of this method and experiment. By presenting the participants with a list of items to recommend, we inadvertently steer the type of diversity they are likely to mention. For example, if our sample did not contain any mention of politics, the participants might also be less likely to mention this as a relevant aspect. Simultaneously, participants may be influenced by what is commonly considered a relevant aspect of diversity, or what interpretations are currently feasible rather than ideal. As such, we cannot draw conclusions on the importance of a certain aspect.

Another difficulty when running this experiment was accounting for the recency of items, which is extremely relevant for the news organization and the broadcaster, where the first will never want to recommend old news, but the latter may sometimes want to include older content. This was especially an issue given that interviews took place on

different days, sometimes weeks apart. For the broadcaster we mixed a stable set of older content with content that was at the day of the interview popular in the system. This has the important caveat that the results between broadcaster interviewees are not fully comparable. For the news organization, we opted for a fixed date a few months in the past. While the interviews are comparable in this case, there is a risk that important contexts are forgotten or mixed up with current events.

Lastly, while diversity in and of itself is already a complicated and multi-faceted concept, the concepts related to it are too. People may talk about 'different ethnicities', and can mean, among others, different skin colors, nationalities or cultures. By extension, our findings are largely influenced by the people that participated in the interviews. Organizations are not a monolith, and many of the interviewees were very explicit about not representing the opinion of every member of the organization they belong to. Furthermore, many of the responses are influenced by the background of the participants themselves, and the fact that this research was conducted in the Netherlands. The Netherlands has a strong public service media system with a focus on representing different groups in society [49]. That being said, participants were overwhelmingly white (11 out of 12) and male (9 out of 12). While this is representative of the general demographic working on recommender systems in the Netherlands, people are not as acutely aware of the needs of groups they are not, themselves, a part of [20, 174]. As such, this is a suitable group for a descriptive study ('what is') into the conceptualizations of diverse recommendations in public service media organisations; however, a complete normative formalization of diversity ('what should be') would require the participation of people from different backgrounds and perspectives in order to provide a complete overview. All limitations considered, the results of this study cannot be used as a definition of diversity. Rather, we aim to show that, even within this limited domain, 'diversity' indeed may refer to a wide variety of things.

2.4 Results

We expect that how organizations operationalize diversity is dependent on what they aim to achieve with their recommendation. We first identify this goal in the following section (Section 2.4.1). Then, we analyze both the different aspects (Section 2.4.2) and tactics (Section 2.4.3) mentioned by the participants.

2.4.1 Goal of recommendation

Being a public organization means that the primary goal of the organization is to provide services that benefit society [200]. As such, each organization's goals with building a rec-

ommender system extend beyond selling ads and optimizing for clicks, as is the norm for most commercial organizations. This difference between being a public or a commercial organization is explicitly mentioned by most interviewees (L2,L3,L4,B2,B3,B4,B5,N1,N2), and awareness of this distinction can be considered central in day-to-day operations. In the absence of optimizing for monetary gains, the goal of recommendation is different for each organization, and is strongly linked to its mission.

News

Both the News organization and the Broadcaster highlight that they are "for and from everyone". For the News organization, this is fairly straightforward: to *"reach the widest possible audience, [and] enable people to be well-informed"* (N2). This means that the news should be accessible to anyone, regardless of skill level, and that journalistic quality takes precedence over personal interest (N1,N2). Each article has a non-personalized 'more like this' section which is automatically populated by a recommender system, but can be supplemented or turned off by the editorial team. Furthermore, there is a personalized recommender system under development that aims to connect readers to "important news they have missed" (N1). For this, the organization is experimenting with how behavioral data and editorial selections can complement each other, in order to compete against the commercial players (N1).

The News organization notes a very strong collaboration between the technical and editorial teams. When opening the app, the users will always land on the editorially curated front page listing the most important news of that moment. At any time, the editorial team has the power to turn off the recommender systems. While diversity is an important concept in the news organization, it is not something that is currently explicitly built into the recommender system. Rather, it is seen as a procedural thing, to be considered at every step of content creation. This goes from choosing which stories and events to cover, to which people are doing the reporting, to writing about events in a neutral way covering multiple perspectives. This control results in a set of items to recommend from that *"have to be told from a diverse standpoint in the first [place]"* (N1). The organization sees the UX design of the recommender system, including where it is placed within the app and accompanied by which headers and explanations as more impactful (N1). They do see potential in personalization of style, telling the news through the user's desired medium (text, video, audio) or in a language complexity level suitable to the reader (N2).

Broadcaster

For the Broadcaster, the mission to be 'from and for everyone' translates a little differently. The organization is effectively an umbrella for multiple smaller broadcasters, each representing a different section of Dutch society. They are the ones pitching ideas and creating the content, though the public broadcaster may request a certain type of program if they feel a particular perspective is missing (B3). The broadcaster then tries to balance on the one hand helping their users recognize themselves in the content they have on offer, while also fostering understanding and knowledge of people, ideas and groups that are 'different' (B1,B2,B4,B5). For this, they see a clear purpose for personalized and diverse recommendation: *"[I]f someone believes or thinks or feels in X, and they look at our platform, that person is also confronted with Y. And that Y is slightly different from what the person had in mind with X."* (B2). Diversity plays a large role in achieving this, but the split in goals between recognition and broadening causes a great deal of conceptual unclarity. As one of the interviewees notes, *"[i]f you are looking at personalization, then we don't want people to all get the same thing recommended. That's what I tend to see as diversity. [...] [P]luralism in recommendations would be [that] we'd like to look into [a] topic from different perspectives. And I think if you talk to different people within this organisation, those two things kind of mix."* (B6).

The current platform is largely manually curated, with separate sections (displayed relatively low on the landing page) for algorithmic recommendations. These algorithmic recommendations balance personal relevance, based on a user's past viewing behavior, with a so-called 'public value score'. This score is an aggregate, obtained through a daily survey sent to users, in which they are asked to score the programs they watched recently on things such as the presence of certain population groups or multiple perspectives. The broadcaster is working on the development of a new platform, which should have a much stronger algorithmic focus. They hold a uniformly strong position on user control: while they, as the broadcaster, should provide a diverse offering, the user is eventually always in charge of what they do and do not watch.

Library

The Library hosts a vast collection containing every book published in or about the Netherlands (L2). They hold a unique position among the cultural institutes, as they have the added role of coordinating 138 other public libraries. There is high interest in digitization and innovation in general, which manifests in creation of proof-of-concept applications that the other libraries can choose to adopt or not (L3). Among other services and initiatives, the Library has the dual goal of having *more people read* and *people read more* (L4).

The Library suspects that the decline of reading among its users is in part caused by a general lack of available time. Currently, there is a recommender system feature in the mobile application hosted by the Library, but its functionality is not satisfactory as *"people cannot find something they might be interested in"* (L4), an issue partly caused by the organization of metadata. Improved personalization can assist with the goal of attracting more patrons, as it increases accessibility to the collection for different types of people. At the same time, the Library is conducting research on the ethics around recommender systems, with topics like bias and privacy being central (L3). In this sense, personalization is not sufficient. There are active efforts to reduce bias that historically existed in the collections, in order to facilitate *"every citizen to be able to participate in society" and "make society [as a whole] better, smarter and more creative"* (L2).

Overall, the Library aims to deploy a well-functioning recommender system to satisfy their readers while remaining inclusive and enacting bias correction. However, building such a system is not trivial, and there are many questions about how development should be approached.

2.4.2 Aspects of diversity

During the recommendation exercise, and the questions leading up to it, participants would mention aspects of the content or the user beyond personal relevance that would lead them to include that item in the diverse recommendation or not. Figure 2.1 provides a schematic overview of these aspects, which can be divided into Item-, Human- and World aspects.

Item aspects

The first dimension considers aspects of the items that are to be recommended. For the Library these would be books, news articles for the news agency, and videos for the Broadcaster. The first group of aspects revolves around *topicality*, or what is actually discussed in the items. The dimensions mentioned by most interviewees are **category**, or sometimes genre, and **topic**. All interviewees but one mentioned balancing different categories and topics in the recommendation, to avoid saturation and to keep a user engaged. This is reflective of how diversity is currently most often conceptualized in technical recommender system literature; see [29].

Related but still separate from the topic are the **geographic location** the item centers on, and the **time period** it discusses. The Library may want to recommend books that discuss a topic through the lens of different time periods (L1), whereas the News organization

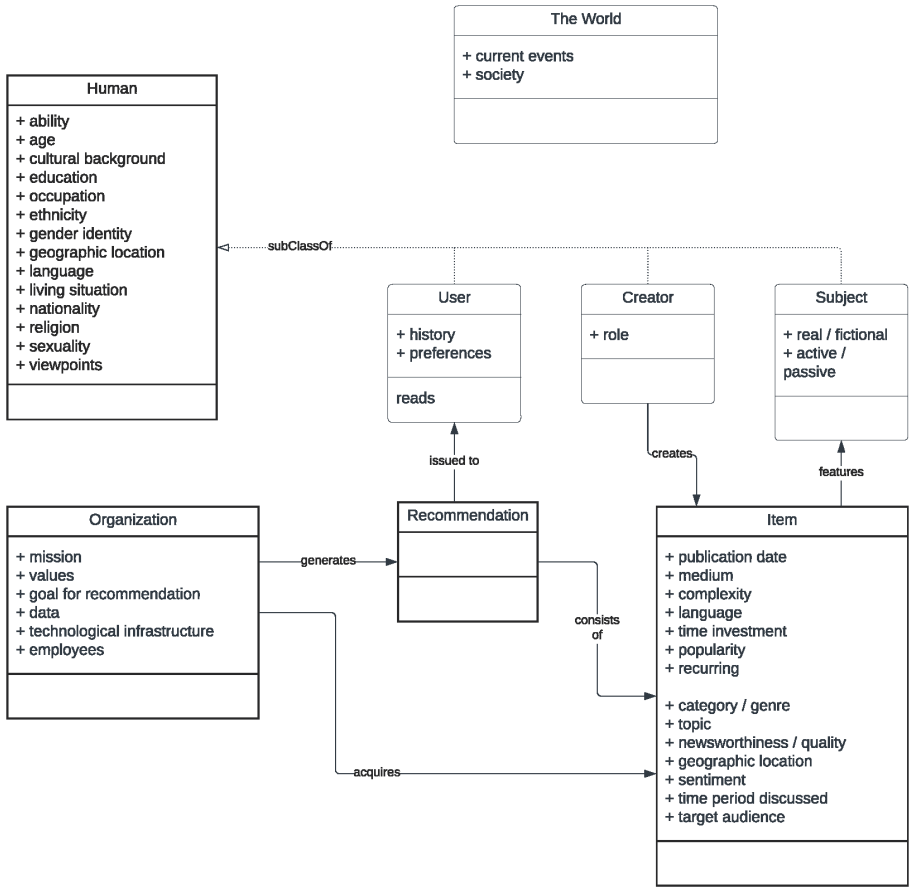


Figure 2.1: Schematic overview of the identified aspects of diversity and how they interact with each other. The 'World' class is unconnected in the graph, but in truth encapsulates and informs everything: from the content that is being produced, to what user wants to read, to what constitutes a 'viewpoint' or a 'minority'.

aims to ensure that they not only cover news from densely populated urban areas, but also from rural areas (N2). Secondly, recommendations may vary on accounts of *stylistic* properties. This relates to the way in which a message is communicated, and whether it is easily accessible for the user. Examples are the **complexity** of the content, the amount of **time investment** a user is required (and willing) to make, the **language** it is written in, the **target audience** of the item, and its **medium** (text, video, audio). Many of these properties are symmetric with characteristics of the user: a certain item complexity level or language requires a certain amount of skill from the user.

Thirdly, there are a range of other item properties which may be prioritized. An important one is an item's **publication date**. While this is of primary importance to the news agency (one would not want to recommend a news item from two months ago), it is much less so for the broadcaster and the library: while there may be some preference for new publications, they also want to help their users discover less popular content. Another important dimension is the perceived **quality** of the item. While the news organization would explicitly optimize for (editorial) quality, opinions differ between participants from the library: while some would find it acceptable if readers only read comic books and manga, others wanted them to be pushed towards more high-brow literature (L4). More differences could be observed in terms of an item's **popularity**: some said that popular items are likely to be items that readers are looking for and should therefore be recommended (L4,B1), while others explicitly distanced themselves from it: *"I do know that out of this selection, [sensationalist article] would have been clicked on more than [serious article]. [...] That's why what we're doing with those recommendations is designed the way we designed it, which is that the journalistic selection we make takes precedence."* (N2). Additionally, items that had a high production **cost** may be prioritized (B3,B4).

Last, but definitely not least, interviewees referred to the people involved in the content. These can be split into the **Creators**, such as producers or authors, and the people that were **Subject** in the items, either through active participation or being discussed by others. Think of guests at a talkshow, politicians discussed in the news or a novel's protagonist. These are further discussed in Section 2.4.2.

Human aspects

While item category is the aspect of diversity mentioned by most interviewees, it is closely followed by notions of the diversity in the context of *people*. While there are a number of attributes that all humans share, there are three 'subclasses' that can be derived with specific roles within the system. These are Users, Creators and Subjects. Users are the people that receive the recommendations, and have a set of unique properties that the other types do

not have: their **history** and **preferences**. Creators are, as the name implies, in some way or another involved in the creation of the items; the authors of books, journalists or producers of shows. Lastly, there are the people that are the Subject of the item's content, such as protagonists of books or people appearing on talk shows. These subjects can be fictional or real, and be active or passive agents within the Item by either speaking on their own behalf or by being described or mentioned by others.

The general Human aspects, such as **age**, **gender identity**, etc, are applicable to each type. A frequently mentioned aspect of diversity is relating to a person's background. **Culture**, **ethnicity**, **nationality** and sometimes **religion** (for example 'Muslims') are often used interchangeably, and the distinction or relation between them is not always clear. For example, B5 notes: *"I think about [...] representing different societal groups. So ethnic minorities. More black and white. Maybe a bit more foreign language [...]"*, while L2 says: *" I don't know if this persona is from a specific country or has a specific background but [author] is an American author."* Identifying the presence of different ethnic and/or cultural groups is notoriously difficult, partly because the data required to do so is often lacking. In each of the domains described above, textual data was the only type of data available, and often no additional information was present beside a person's name[2]. In some cases, data enrichment might be possible (B3). Yet, this requires building a dataset on attributes that are often considered sensitive, which gets even more problematic when considering the fact that, to 'expand horizons' or 'represent', a similar type of dataset would be necessary about the users, which would then lead to issues of privacy (L2); see also Papageorgiou and Mougan [208]. Similar safety concerns can be raised for aspects like **gender identity** and **sexuality**; see Pinney et al. [212].

Culture- and **viewpoint** diversity are sometimes mentioned in the same breath: *"for me, diversity means having multiple colors, multiple opinions, multiple cultures, multiple points of view about something."* (B2). We make a distinction between the two, and denote viewpoint diversity as a person's perspective on events. Viewpoint diversity then often becomes linked to politics, or opinions about current events. A rich body of work exists around so-called viewpoint or stance detection, aiming to algorithmically extract these from text [2, 64, 223]. Even when successful, it is often unclear what a diversity of viewpoints would look effectively like, which is further discussed in Section 2.4.3.

[2]Some approaches have attempted to predict a person's ethnicity based on their name, but this process is not generally accurate and can suffer from misrecognition bias [156]

World aspects

Aspects related to the World are beyond the direct control of a recommender system or user. They influence which content is created, and as such the pool of items that a recommender system can make its selection from. They also interact with a user's preferences to determine what type of content a user is interested in, looking for, or should be reading. **Current events** determine what is newsworthy, and what news a user needs to consume to fulfill their information needs. Naturally this aspect is of high importance to the News organization (N1,N2) and to some extent the Broadcaster (B3, B5), but much less so to the Library. In contrast, **society** represents the world as it currently is, including its biases and existing power distributions. An organization may choose to either reflect or counter these existing structures (see Section 2.4.3). For example, the Library acknowledges that much of their catalogue consists of white, American male authors, and wants to account for this in their recommendations (L1).

2.4.3 Tactics for diversification

During the experiment, the participants considered and deployed different tactics to produce a diverse recommendation. Some participants articulated the tactic they deployed explicitly, while others did not clarify it in as much detail.

Diversity between items

Diversity between items is perhaps the most intuitive interpretation of diversity, and refers to ensuring that all items within a list of recommendations are sufficiently different between them. Following this tactic requires choosing one or more appropriate axes to diversify over. In that sense, constructing a diverse list also entails exclusion, since some aspects have to be deemed less relevant. This issue is particularly important for public organizations; selecting meaningful aspects for which diversity needs to be safeguarded is a crucial decision that must comply with the values and mission of the organization. Multiple interviewees constructed a recommended list by considering diversity between items in the process (B2,B3,B4,B5,B6,N2,L1,L2,L3,L4), especially to recommend items that are diverse between them in terms of category. In particular, the Broadcaster may aim for a *"balance between [e]ntertainment and information"* (B2), given their wide range of offerings and mission *"to inform, educate, and entertain"* (B2). For the Library, item category translates to book genre, and is commonly taken into account when creating a diverse list (L2,L3,L4).

Instead of generally diversifying between items, a somewhat adapted tactic is to recommend a set of items that are different between them in some perspective, but engage with

the same topic or theme. For example, one interviewee opted for selecting *"five books that give [...] an insight on [...] LGBTQ [...] [,h]ow that works or is around the world in different cultures and different times"* (L1). Furthermore, in case of a socially relevant emerging topic, the Broadcaster can create *"a highlight lane and then offer a lot of different opinions that people can scroll through"* (B5).

Finally, interviewees noted as a result of the experiment that *"to combine several aspects [is] really hard"* (L2). It might be that a set of items is diverse over one important axis, but not over a different equally important one. In this context the act of creating a diverse recommendation can be seen as an optimization process that an algorithm can contribute to. Regardless, ensuring diversity between items requires some sort of aspect prioritization, as well as an appropriate justification for it.

Diversity as a within-item measure

A different tactic for ensuring diversity is recommending items that consist of diverse perspectives or types of people within the item itself. For example, according to a participant from the Broadcaster, a travel series that allows the viewer to *"see multiple worlds [...] fits very well into diversity"* (B2). This also pertains to episodes of political programs where *"people from a lot of the major parties [appear], which captures a big part of the political spectrum"* (B4). The News organization notes that their inventory consists of *"very good stories which have to be told from a diverse standpoint in the first [place]"* (N1). From this perspective, guaranteeing item diversity may be a part of the production/acquisition process or an explicit step of the recommendation. During the exercise, one interviewee (B4) suggested combining within-item and between items, by composing a list of items on different topics where each item contains multiple perspectives on the respective topic. According to the interviewee, pursuing diversity in recommendation can also be seen as a process of achieving maximum diversity. This tactic requires that the media organization's catalogue consists of items that can facilitate the maximization process.

Diversity considering the user

User-specific diversity is a personalized form of diversity, where the user's history and/or preferences are taken into account when composing a diverse list of items to recommend. The first way to operationalize this tactic is to cater to a user's specific needs that potentially diverge from the norm. This can manifest in assisting the user with finding uniquely niche content. Based on this outlook, diversity is about *"ensuring that [...] everyone feels that there is something for [them]"* (B1) in the catalogue. The Library and News organiza-

tion also mentioned literacy in this context (L4,N1), as *"how [to] help those with difficulty reading [...] [is] also a diversity topic in a way"* (N1). Additionally, for the Library diversity in accordance with the user can help them *"recognize themselves in the author or in the main characters or the topics [such as] age and physical ability and sexual orientation"* (L2). The second way to apply user-specific diversity is to recommend to a user content that extends further from their current interests, and that they might not be aware of. In other words, an organization can also recommend items so as to *"give people [...] a broader perspective of what is available"* (L3) and to *"broaden the user's horizons"* (B5). This tactic can be seen as a way to ensure that users do not *"stay in their own bubble"* (L2) by *"educating people [that] there's more than [their] bubble"* (B5).

Diversity considering the world

Diversity considering the world entails recommending items that in some way reflect or diverge from the norms that exist in the world, leading to 'similar to world' and 'different from world' diversification tactics. When reflecting the world, one could think of having a good spread of topics that are representative of the important news of that day (N2), or, by representing political parties in a way similar to their distribution in government (B4). With a 'different from world' diversification tactic, the aim is to counter existing power structures. With this interpretation, an item can be considered diverse on its own merit. This can be because the content represents a minority culture, an "outsider [...] in a political sense" (B4), a *"very different part of the world that [the society] knows too little about"* (B2), or even a theme or topic that *"you do not find that many [items] in the catalogue"* (L4). In this case, very popular items may be deemed inherently not diverse, and might not need to receive further exposure: *"[I]t's going to be on the website probably anyway"* (B6).

What constitutes a minority was often implicitly assumed by the interviewees, potentially due to the social context that their organization operates in. For example, multiple interviewees referred to LGBTQ-themed media as inherently diverse (B4,B5,B6,L2), which also prompted their inclusion in a diverse recommendation as part of the experiment. The same can be said for media that gives exposure to people with a minority cultural background or engages with topics not related to the Randstad, a dominant urban area in the Netherlands (N2). From this perspective, diversity considering the world and user-based diversity (in a broadening-horizons way) can overlap when a user is assumed to be the typical or default representation of the majority culture in the given context.

2.5 Discussion

The wide range of interpretations highlighted in Section 2.4 show that, even within this limited domain, conceptualizations of diversity among participants vary greatly. It is therefore unlikely that a standardized definition, encompassing all potential goals, aspects and tactics, can be attained; however, that does not mean that it is impossible to implement a meaningful form of diversity into a recommender system. In this section, we outline the implications the findings of this study have for the implementation of diversity-aware recommender systems.

2.5.1 Implementing diversity: a normative process

Diversity has traditionally been seen as something we always want *more* of: more viewpoints, more topics, more different nationalities. However, during the interview participants often mentioned instances in which they would actually want *less*. An interviewee from the Library wanted to recommend items that fit the amount of time a user was willing to invest, while the News organization only wanted to recommend items of a certain quality. In these cases, meaningful diversity could be achieved by controlling for one aspect and diversifying over another; for example, when recommending different viewpoints for one controlled topic, or vice versa, by showing the opinions of one particular group over a wide range of topics. The same dynamic also occurs in the way interviewees spoke about diversification tactics. In some cases, a recommendation similar to a user's preferences or history would be desirable, in order for them to recognize themselves in the content on offer. In others, it should be different, so as to expose them to new perspectives. This shows that different applications may not only prioritize different aspects of diversity, they may also have different expectations on whether a recommendation should have high or low diversity.

Participants would often mention that recommendations should be 'diverse enough.' This requires an underlying model that not only determines which items are similar and different, but also informs the system whether sufficient diversity has been attained. This is non-trivial, especially in a domain such as news recommendation, where contexts change rapidly. One could imagine using an external source as a reference point: for example, the presence and size of political parties in government, or a country's composition in terms of cultural groups as determined by a national statistics agency. However, this yet again relies on a conscious decision of what is the 'right' model to use and reflect through the recommendations, and each choice will have pros and cons. There is therefore a normative choice to be made about the type of diversity that is desired, with different

aspects to consider, different diversification tactics, and different levels of diversity. The wide variety in the answers given by participants is also an indication that this normative choice is not one that can be taken lightly, and requires a good deal of internal discussion and alignment with all the relevant stakeholders involved before it can be satisfactorily modeled and implemented.

2.5.2 Generalizing to domains beyond public service media

While the results are not directly generalizable to other domains, there are likely elements of the aspects and tactics identified here that would also be applicable. For example, while music recommendation might put less stock in the diversity of people discussed in an item's content, they would be interested in the diversity of Creators [81]; similarly, while in-item diversity would not be applicable, they could be interested in countering existing biases through different-from-world recommendation tactics [59]. We believe that the goals, aspects and tactics of this paper can still serve as a starting point for discussion in other domains, which may make it easier to identify gaps and unique challenges.

2.5.3 Exploring versus defining diversity

Participants varied greatly in which diversity aspects and diversification tactics they mentioned. Some of these differences can be traced to the different ways participants speak: some people are more verbose, more inclined to stick to a specific set of examples, or already have a more developed concept in mind than others. Furthermore, the three organizations were at different stages of developing strategies towards diversity, which may have an impact on said differences between participants. However, they are all working on or considering the implementation of a recommender system. Our participants are thus also actively making decisions about how diversity would be conceptualized and implemented. They are as such representative of the 'real' world, rather than the 'ideal' world, and therefore suitable candidates for an exploration of the dimensions along which diversity is currently conceptualized. For these reasons, we refrain from making claims about whether certain aspects or tactics are 'more important' than others, nor do we make claims about the 'correct' definition of diversity. Our hope is that the overviews of goals, aspects and tactics presented in this study will facilitate building a common understanding and vocabulary between stakeholders with a different background, making it easier to find common ground and establish priorities that reflect the requirements of that particular implementation.

2.6 Conclusion

Through interviews with participants from public service media organizations, we identified a range of different interpretations of diversity within a recommender system. We grouped these into different goals (e.g. broadly informing the public or allowing a user to recognize themselves), aspects (e.g. the topic of the content, or the cultural background of an author), and diversification tactics (e.g. ensuring diversity within a single item or countering biases in society). Given the great variety found in the conceptualization of diversity in recommender systems, even within a limited domain, we find that it is unlikely that a standardized definition can be attained. Instead, rather than striving for this standardization, we argue that it should be conceptualized on a case-by-case basis. We hope that our mapping of goals, aspects, and tactics can contribute in effectively communicating what diversity entails within a specific application.

Appendix: Coding scheme and mentions

Table 2.1: Final coding scheme, containing both aspects and tactics. Aspects are divided into Item, Person and World aspects.

Aspects		B1	B2	B3	B4	B5	B6	L1	L2	L3	L4	N1	N2
Item	category / genre	x	x	x	x	x	x	x	x	x	x	x	
	complexity		x	x	x						x	x	
	cost				x								
	creator (person)	x		x	x	x		x	x	x	x		
	geographic location	x	x	x	x	x	x	x		x		x	x
	language		x	x				x					
	newsworthiness / quality	x	x	x	x	x		x			x	x	x
	medium												x
	subject (person)	x	x	x	x	x	x		x		x		x
	popularity	x	x	x	x	x	x	x		x		x	x
	publication date	x		x	x			x				x	x
	recurring			x								x	
	relevance	x	x	x	x	x		x		x		x	x
	sentiment					x							
	target audience		x	x									x
	time investment										x	x	
	time period discussed				x	x		x					
	topic	x	x	x	x	x	x	x	x	x	x	x	x
Person	ability								x	x			
	age	x	x	x	x			x	x	x			
	cultural background		x		x	x	x	x	x				
	education				x							x	x
	ethnicity		x	x	x				x				
	gender identity		x	x	x	x		x	x	x			
	geographic location	x	x	x	x		x	x		x		x	
	living situation			x	x	x	x		x			x	
	nationality		x	x	x	x			x	x	x	x	
	religion	x	x	x	x	x	x			x			
	sexuality	x			x	x	x	x	x	x	x		
	viewpoint	x	x	x	x	x	x	x	x	x		x	
	user; history	x	x	x	x				x	x		x	
	user; preferences	x		x	x	x			x	x		x	
World	current events			x		x						x	x
	society				x	x				x		x	
Tactics													
	different in item		x		x	x	x					x	
	different in list	x	x	x	x	x	x	x	x	x	x		x
	similar to user	x		x		x	x	x	x	x	x	x	x
	different from user	x	x			x	x		x	x			
	different/similar to world		x		x	x	x	x			x		

3

Democratic theory to inform how news recommender systems should be evaluated

Authors: **Sanne Vrijenhoek**, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk and Natali Helberger

Original title: Recommenders with a mission: assessing diversity in news recommendations

Published in: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR '21). Association for Computing Machinery, New York, NY, USA, 173–183.

https://doi.org/10.1145/3406522.3446019

Abstract

NEWS RECOMMENDERS help users to find relevant online content and have the potential to fulfill a crucial role in a democratic society, directing the scarce attention of citizens towards the information that is most important to them. Simultaneously, recent concerns about so-called filter bubbles, misinformation and selective exposure are symptomatic of the disruptive potential of these digital news recommenders. Recommender systems can make or break filter bubbles, and as such can be instrumental in creating either a more closed or a more open internet. Current approaches to evaluating recommender systems are often focused on measuring an increase in user clicks and short-term engagement, rather than measuring the user's longer term interest in diverse and important information.

This chapter aims to bridge the gap between normative notions of diversity, rooted in democratic theory, and quantitative metrics necessary for evaluating the recommender system. We propose a set of metrics grounded in social science interpretations of diversity and suggest ways for practical implementations.

3.1 Introduction

News recommender algorithms have the potential to fulfill a crucial role in democratic society. By filtering and sorting information and news, recommenders can help users to overcome maybe the greatest challenge of the online information environment: finding and selecting relevant online content - content they need to be informed citizens, be on top of relevant developments, and have their say [76]. Informed by data on what a user likes to read, what people similar to him or her like to read, what content sells best, etc., recommenders use machine learning and AI techniques to make ever smarter suggestions to their users [62, 148, 151, 260]. For the news media, algorithmic recommendations offer a way to remain relevant on the global competition for attention, create higher levels of engagement with content, develop ways of informing citizens and offer services that people are actually willing to pay for [22]. With this comes the power to channel attention and shape individual reading agendas and thus new risks and responsibilities. Recommender systems can be pivotal in deciding what kind of news the public does and does not see. Depending on their design, recommenders can either unlock the diversity of online information [111, 196] for their users, or lock them into routines of "more of the same", or in the most extreme case into so-called filter bubbles [211] and information sphericules.

The most frequently used key performance indicators, or KPIs, for optimizing recommender systems, assess and aim to maximize short-term user engagement, such as click-through rate or time spent on a page [126]. Often, these KPIs are defined by data limitations, and by technological and business demands rather than the societal and democratic mission of the media. More recently however a process of re-thinking algorithmic recommender design has begun, in response to concerns from users [261], regulators (e.g., EU HLEG [203]), academics, and news organizations themselves [22, 161]. Finding ways to develop new metrics and models of more "diverse" recommendations has developed into a vibrant field of experimentation - in academia as well as in the data science and R&D departments of a growing number of media corporations.

But what exactly does diverse mean, and how much diversity is 'enough'? As central as diversity (or pluralism, a notion that is often used interchangeably) is to many debates about the optimal design of news recommenders, as unclear it is what diverse recommender design actually entails [158]. In the growing literature that tries to conceptualise and translate diversity into specific design requirements, a gap between the computer science and the normative literature can be observed. While diversity in the computer science literature is often defined as concrete technical metrics, such as the intra-list distance of recommended items [29, 273] (see also Section 3.2), diversity in the normative

sense is about larger societal concepts: democracy, freedom of expressions, cultural inclusion, mutual respect and tolerance [111]. There is a mismatch between different theoretical understandings of the construct of diversity, similar to the one observed in Fairness research [125]. For news recommenders to be truly able to unlock the abundance of information online and inform citizens better, it is imperative to find ways to overcome the fundamental differences in approaching diversity. There is a need to reconceptualise this central but also elusive concept in a way that both does justice to the goals and values that diversity must promote, as well as facilitates the translation of diversity into metrics that are concrete enough to inform algorithmic design.

This paper describes the efforts of a team from computer science, communication science, and media law and policy experts, to bridge this gap between normative and computational approaches towards diversity, and translate diversity, as a normative concept, to a concrete set of metrics that can be used to evaluate and/or compare different news recommender designs.

We first conceptualise diversity from a technical point of view (Section 3.2) and from a social science interpretation, including its role in democratic models (Section 3.3). In Section 3.4 we expand upon the social science notion of diversity, and propose five metrics grounded in Information Retrieval that reflect our normative approach. We cover the limitations of the proposed metrics and this approach in Section 3.5. We conclude with detailing our implementation of the metrics and the steps to undertake as a media company when intending to adopt this normative notion of diversity in practice.

3.2 A technical perspective

Typically, generating a recommendation is seen as a reranking problem. Given a set of candidate items, the goal is to present these items in such a way that the user finds the item he or she is most interested in at the top, followed by the second-most interesting one, and so on. How well this recommendation reflects the actual interest of the user is called the *accuracy* of the recommendation. *Content-based* approaches aim to maximize this accuracy by looking at the type of items that the user has interacted with before and recommend similar ones. In the context of news recommendations, one could think of finding topics or overall texts that are similar to what is in the user's reading history. On the other hand, in *collaborative filtering* approaches, the algorithm considers what other users similar to the user in question have liked, and recommends those. Most state-of-the-art systems are hybrids of these approaches. Evaluation of the system can be done in both an online and offline fashion; offline often includes testing the system on a piece of held-out

data on its accuracy, whereas online evaluation monitors for increases or decreases of user interactions and click-through rates following the issued recommendations [15].
However, this approach by its definition unduly promotes the items similar to what a user has seen before, locking the user in a feedback loop of "more of the same" [175]. It also introduces a so-called "confounding bias" [32], which happens when an algorithm attempts to model user behavior when the algorithm itself influences that behavior. To tackle this in many currently operational systems "beyond-accuracy" metrics diversity, novelty, serendipity and coverage are introduced. *Diversity* reflects how different the items within the recommendation set are from each other. One intuitive usecase can be found in the context of ambiguous search queries. A user searching for "orange" should receive results about the color, the fruit, the telecom company, the Dutch royal family, and the river in Namibia, and not just about the one the system thinks he or she is most likely to be interested in. The challenge then lies in how to define this difference or distance. In the context of news recommendations many different approaches exist, such as using a cosine similarity on a bag of words model or by calculating the distance between the article's topics [305].
The concepts of *novelty* and *serendipity* are strongly linked. *Novelty* reflects the likeliness that the user has never seen this item before, whereas *serendipity* reflects whether a user was positively surprised by the item in question. However, an item can be novel without being serendipitous (such as the weather forecast), and an item may also be serendipitous without being novel (such as an item that has been seen a long time ago, but becomes relevant again in light of recent events). A common approach to improving novelty and serendipity is by unlocking the "long tail" content of the system, while still optimizing for user accuracy. The long tail refers to the "lesser known" content in the system, that is less popular and therefore seen by less users. By recommending less popular content the recommender systems increase the chance that an item is actually novel to a user.
Lastly, *coverage* reflects to what extent all the items available in the system have been recommended to at least a certain number of users. This metric is naturally strongly influenced by the novelty of the recommendations, as increasing the visibility of lesser-seen items increases the overall coverage of all items.

3.3 A democrative perspective

What becomes apparent from the overview in Section 3.2 is that although there are various attempts to conceptualize evaluation metrics beyond accuracy in the computer science literature, these metrics are constructed for the broad field of recommendation systems,

and are therefore not only relevant in the context of news, but also for music, movies, web search queries and even online dating. However, what they win in generalizability, they lose in specificity. They are not grounded in, and do not refer back to the normative understanding of diversity in the media law, fundamental rights law, democratic theory and media studies/communication science literature, as is also demonstrated in Loecherbach et al. [158].

Before we define more quantitative metrics to assess diversity in news recommendation, we first offer a conceptualization of diversity. Following the definition of the Council of Europe, diversity is not a goal in itself, it is a concept with a mission, and it has a pivotal role in promoting the values that define us as a democratic society. These values may differ according to different democratic approaches. This article builds on a conceptualisation of diversity in recommendations that has been developed by Helberger [111]. Here, Helberger combines the normative understanding of diversity, meaning what should diverse recommendations look like, with more empirical conceptions, meaning what is the impact of diverse exposure on users. There are many theories of democracy, but the paper by Helberger focuses on 4 of the most commonly used theories when talking about the democratic role of the media: *Liberal*, *Participatory*, *Deliberative* and *Critical* theories of democracy (see also [37, 50, 133, 252]).

It is important to note that no model is inherently better or worse than another. Which model is followed is something that should be decided by the media companies themselves, following their mission and dependent on the role they want to play in a democratic society.

3.3.1 The Liberal model

In liberal democratic theory, individual freedom, including fundamental rights such as the right to privacy and freedom of expression, dispersion of power but also personal development and autonomy of citizens stands central. The liberal model is in principal sympathetic to the idea of algorithmic recommendations and considers recommenders as tools to enable citizens to further their autonomy and find relevant content. The underlying premise is that citizens know for themselves best what they need in terms of self-fulfillment and exercising their fundamental rights to freedom of expression and freedom to hold opinions, and even if they do not, this is only to a limited extent a problem for democracy. This is because the normative expectations of what it means to be a good citizen are comparatively low and there is a strict division of tasks, in which *"political elites [...] act, whereas citizens react"* [252].

Under such liberal perspective, diversity would entail a user-driven approach to diver-

sity that reflects citizens interests and preferences not only in terms of content, but also in terms of for example style, language and complexity. The *liberal recommender* is required to inform citizens about prominent issues, especially during key democratic moments such as election time, but else it is expected to take little distance from personal preferences. It is perfectly acceptable for citizens to be consuming primarily cat videos and celebrity news, as long as doing so is an expression of their autonomy.

Summary.

The liberal model of democracy promotes self-development and autonomous decision making. As such, a news recommender following a liberal approach should focus on the following criteria:

- Facilitating the specialization of a user in an area of his/her choosing
- Tailored to a user's preferences, both in terms of content and in terms of style

3.3.2 The Participatory model

An important difference between the liberal and the participatory model of democracy is what it means to be a good citizen. Under participatory conceptions, the role of (personal) freedom and autonomy is to further the common good, rather than personal self-development [115]. Citizens cannot afford to be uninterested in politics because they have an active role to play in helping the community to thrive [252]. Accordingly, the media, and by extension news recommenders must do more than to give citizens 'what they want', and instead provide citizens with the information they need to play their role as active and engaged citizens [9, 83, 130, 134], and to further the participatory values, such as inclusiveness, equality, participation, tolerance. *Participatory recommenders* must also proactively address the fear of missing out on important information and depth, and the concerns about being left out. Here the challenge is to make a selection that gives a fair representation of different ideas and opinions in society, while also helping a user to gain a deeper understanding, and feeling engaged, rather than confused. This also involves that recommenders are able to respond to the different needs of users in which information is being presented. The form of presentation is an aspect that is often neglected in discussions around news recommender diversity, ignoring the fact that different people have different preferences and cognitive abilities to process information. Accordingly, the media should 'frame politics in a way that mobilizes people's interests and participation in politics'. Strömbäck [252] and Ferree et al. [82] speak of 'empowerment': to be truly empowering, media content needs to be presented in different forms and styles [36, 82, 301]. By

extension, this means that diversity is not only a matter of the diversity of content, but also of communicative styles. What would then characterize diversity in a participatory recommender are, on the one hand, active editorial curation in the form of drawing attention to items that citizens 'should know', taking into account inclusive and proportional representation of main political/ideological viewpoints in society; a focus on political content/news, but also: non-news content that speaks to broader public and, on the other hand, a heterogeneity of styles and tones, possibly also emotional, empathetic, galvanizing, reconciliatory.

Summary.

The participatory model of democracy aims to enable people to play an active role in society. It values the idea of the 'common good' over that of the individual. Therefore, a participatory recommender should follow the following principles:

- Different users do not necessarily see the same articles, but they do see the same topics.
- Article's complexity is tailored to a user's preference and capability
- Reflects the prevalent voices in society
- Empathetic writing style

3.3.3 The Deliberative model

The participatory and the deliberative models of democracy have much in common (compare Ferree et al. [82]). Also in the deliberative or discursive conceptions of democracy, community and active participation of virtuous citizens stands central. One of the major differences is that the deliberative model operates on the premise that ideas and preferences are not a given, but that instead we must focus more on the process of identifying and negotiating and, ultimately, agreeing on different values and issues [82, 133]. Political and public will formation is not simply the result of who has the most votes or 'buyers', but it is the result of a process of public scrutiny and intensive reflection [115]. This involves a process of actively comparing and engaging with other also contrary and opposing ideas [167]. The epistemological shift from information to deliberation has important implications for the way the role of news recommenders can be conceptualised. Under a deliberative perspective, it is not enough to 'simply' inform people. The media need to do more, and has an important role in "promoting and indeed improving the quality of public life -

and not merely reporting on and complaining about it" [37]. Strömbäck [252] goes even further and demands that the media should also "actively foster political discussions that are characterised by impartiality, rationality, intellectual honesty and equality among the participants". Diversity in the deliberative conception has the important task of confronting the audience with different and challenging viewpoints that they did not consider before, or not in this way [167]. Concretely, this means that a deliberative recommender should include a higher share of articles presenting various perspectives, diversity of emotions, range of different sources; it should strive for equal representation, as well as on recommending items of balanced content, commentary, discussion formats, background information; potentially some prominence for public service media content (as the mission of many public service media includes the creation of a deliberative public sphere), as well as a preference for rational tone, consensus seeking, inviting commentary and reflection.

Summary.

The focus of the deliberative recommender is on presenting different opinions and values in society, with the goal of coming to a common consensus or agreeing on different values.

- Focus on topics that are currently at the center of public debate
- Within those topics, present a plurality of voices and opinions
- Impartial and rational writing style

3.3.4 The Critical model

A main thrust of criticism of the deliberative model is that it is too much focused on rational choice, on drawing an artificial line between public and private, on overvaluing agreement and disregarding the importance of conflict and disagreement as a form of democratic exercise [134]. The focus on reason and tolerance muffles away the stark, sometimes shrill contrasts and hidden inequalities that are present in society, or even discourage them from developing their identity in the first place. Accordingly, under more radical or critical perspectives, citizens should look beyond the paint of civil and rational deliberation. They should discover and experience the many marginalised voices of those "who are 'outsiders within' the system"[82], and when doing so critically reflect on reigning elites and their ability to give these voices their rightful place in society. Diverse critical recommenders hence do not simply give people what they want. Instead, they actively nudge readers to experience otherness, and draw attention to the marginalised, invisible or less powerful ideas and opinions in society. And again, it is not only the question of what kinds

of content are presented but also the how: whereas in the deliberative and also the participatory model, much focus is on a rational, reconciliatory and measured tone, critical recommenders would also offer room for alternative forms of presentations: narratives that appeal to the 'normal' citizen because they tell an everyday life story, emotional and provocative content, even figurative and shrill tones - all with the objective to escape the standard of civility and the language of the stereotypical "middle-aged, educated, blank white man" [299].

Summary.

The critical recommender aims to provide a platform to those voices and opinions that would otherwise go unheard. From a critical democracy perspective on diversity, recommenders should be optimized on the following principles:

- Emphasis on voices from marginalized groups
- Emotional writing style

3.4 Diversity metrics

The democratic models described in Section 3.3 lead to different conceptualizations of diversity as a value, which again translate into different diversity expectations for recommender systems. In this section, we propose five metrics that follow directly from these expectations, grounded in democratic theory and adapted from existing Information Retrieval metrics: *Calibration*, *Fragmentation*, *Activation*, *Representation* and *Alternative Voices*. For each of these metrics, we explain the concept and link to democratic theory. Furthermore we make a suggestion for operationalization, but note that this work is an initial outline and that much work still needs to be done. Future work should include more work on the validity of the metrics, for example by following the measurement models specified in Jacobs and Wallach [125]. Lastly we mention a number of the limitations of the currently proposed metrics and their operationalizations.

Table 3.1 provides an overview of the different models, metrics and their expected value ranges. Note that not all metrics are relevant to all models.

Before explaining the metrics, we define the following variables that are relevant to multiple metrics:

- p: The list of articles the recommender system could make its selection from, also referred to as the 'pool'

- q: The unordered list of articles in the recommendation set
- Q: The ordered list of articles in the recommendation set
- r: The list of articles in a user's reading history

3.4.1 Calibration

The *Calibration* metric reflects to what extent the issued recommendations reflect the user's preferences. A score of 0 indicates a perfect Calibration, whereas a higher score indicates a larger divergence from the user's preferences.

Explanation.

Calibration is a well-known metric in traditional recommender system literature [247]. It is calculated by measuring the difference in distributions of categorical information, such as topics in the news domain or genres in the movie domain, between what is currently recommended to the user and what the user has consumed in the past. However, we extend our notion of calibration beyond topicality or genre. News recommendations can also be tailored to the user in terms of article style and complexity, allowing the reader to receive content that is attuned to their information needs and processing preferences. This may be split up within different topics; a user may be an expert in the field of politics but less so in the field of medicine, and may want to receive more complex articles in case of the first, and less in case of the second.

In the context of democratic recommenders.

The Calibration metric is most significant for recommenders following the Liberal and Participatory model. The aim of the Liberal model is to facilitate user specialization, and assumes that the user eventually knows best what they want to read. In these models, we expect the Calibration scores to be closer to 0. On the other hand, the Participatory model favors the common good over the individual. We therefore expect a higher degree of divergence in Calibration, at least when considered in light of topicality. Both models, but especially the Participatory model, require that the user receives content that is tailored to their needs in terms of article complexity, and in this context we expect a Calibration score that is closer to zero.

Operationalization.

For the operationalization of a recommender's Calibration score it is important to have information on not only an article's topic and complexity, which can potentially be automatically extracted from an article's body (see for example Feng et al. [79] and Kim and Oh [140]), but also on the user's preferences regarding this matter. Note that topicality can be both generic (politics, entertainment, sports, etc) and more specific (climate change, Arsenal). In light of democratic theory more fine-grained information is preferable, but this is not always available. Steck [247] uses the Kullback-Leibler divergence between two probability distributions as Calibration metric, as follows:

$$Calibration_{(r,q)} = \sum_{c} r(c|u) log \frac{r(c|u)}{\tilde{q}(c|u)}$$

where $r(c|u)$ is the distribution of categorical information c across the articles consumed by the user in the past, and $\tilde{q}(c|u)$ is an approximation of $q(c|u)$ (necessary since KL divergence diverges if $q(c|u) = 0$), which is the distribution of the categories c across the current recommendation set. As mentioned before, a score of 0 indicates that there is no divergence between the two distributions, meaning they are identical. The higher the Calibration score, the larger the divergence. As KL divergence can yield very high scores when dividing by numbers close to zero, outliers can greatly influence the average outcome. Therefore, the aggregate Calibration score is calculated by taking the median of all the Calibration scores for individual users.

Limitations.

This approach is tailored to categorical data, but sometimes our data may be numerical rather than categorical, for example in the case of article complexity. In these cases, a simple distance measure may suffice over the more complex Kullback-Leibler divergence.

3.4.2 Fragmentation

The *Fragmentation* metric denotes the amount of overlap between news story chains shown to different users. A Fragmentation score of 0 indicates a perfect overlap between users, whereas a score of 1 indicates no overlap at all.

Explanation.

News recommender systems create a recommendation by filtering from a large pool of available news items. By doing so they may stimulate a common public sphere, or cre-

ate smaller and more specialized 'bubbles'. This may occur both in terms of topics recommended, which is the focus of the Fragmentation metric, and in terms of presented perspectives, which will be later explained in the Representation metric. Fragmentation specifically compares differences in recommended news story chains, or sets of articles describing the same issue or event from different perspectives, writing styles or points in time [198], *between* users; the smaller the difference, the more aware the users are of the same events and issues in society, and the more we can speak of a joint agenda. When the news story chains shown to the users differ significantly, the public sphere becomes more fragmented, hence the term Fragmentation.

In the context of democratic recommenders.

Both the Participatory and Deliberative models favor a common public sphere, and therefore a Fragmentation score that is closer to zero. The Liberal model on the other hand promotes the specialization of the user in their area of interest, which in turn causes a higher Fragmentation score. Finally the Critical model, with its emphasis on drawing attention to power imbalances prevalent in society as a whole, calls for a low Fragmentation score.

Operationalization.

This metric requires that individual articles can be aggregated into higher-level news story chains over time. This can be done through manual annotation or automated extraction process. Two unsupervised learning approaches for doing this automatically can be found in Nicholls and Bright [198] and Trilling and van Hoof [265]. Once the stories are identified, the Fragmentation score can be defined as the aggregate average distance between all sets of recommendations between all users. Dillahunt et al. [56], which aimed to detect filter bubbles in search engine results, defines this distance with the Kendall Tau Rank Distance (KDT), which measures the number of pairwise disagreements between two lists of ranked items. However, Kendall Tau is not suitable when the two lists can be (largely) disjointed. It also penalizes differences at the top of the list equally to those more at the bottom. Instead we base our approach on the Rank Biased Overlap used in Webber et al. [286]:

$$RBO(Q_1, Q_2, s) = (1 - s) \sum_{d=1}^{\infty} s^{d-1} \cdot A_d$$

where Q_1 and Q_2 denote two (potentially) infinite ordered lists, or two recommendations issued to users 1 and 2, and s a parameter that generates a set of weights with a geometric progression starting at 1 and moving towards 0 that ensures the tail of the recommenda-

tion is counted less severely compared to its head. Because of this there is a natural cut-off point where the score stabilizes. We iterate over the ranks d in the recommendation set, and at each rank we calculate the average overlap A_d. Because Rank-Biased Overlap yields a score between 0 and 1, with 0 indicating two completely disjoint lists and 1 a perfect overlap, and the score that is expressed is semantically opposite of what we aim to express with the Fragmentation metric, we obtain the Fragmentation score by calculating 1 minus the Rank-Biased Overlap. Lastly, the aggregate Fragmentation score is calculated by averaging the Fragmentation score between each user and every other user.

Limitations.

Since this approach is computationally expensive (every user is compared to every other user, which is $O(n^2)$ complexity), some additional work is needed on its scalability in practice, for example through sampling methods.

3.4.3 Activation

The *Activation* metric expresses whether the issued recommendations are aimed at inspiring the users to take action. A score close to 1 indicates a high amount of activating content, whereas a score close to 0 indicates more neutral content.

Explanation.

The way in which an article is written may affect the reader in some way. An impartial article may foster understanding for different perspectives, whereas an emotional article may activate them to undertake action. A lot of work has been done on the effect of emotions and affect on the undertaking of collective group action. This holds especially for anger, in combination with a sense of group efficacy [271]. But positive emotions play a role too; for example, "joy" elicits the urge to get involved, and "hope" to dream big [86]. The link between emotions, affect and activation is described well by Papacharissi [206]: *"...for it is affect that provides the intensity with which we experience emotions like pain, joy, and love, and more important, the urgency to act upon those feelings"*. The Activation metric aims to capture this by measuring the strength of emotions expressed in an article.

In the context of democratic recommenders.

The Activation metric is relevant in three of the four different models. The Deliberative model aims for a common consensus and debate, and therefore would give a certain

measure of prominence to impartial articles with low Activation scores. The Participatory model fosters the common good and understanding, and aims to facilitate users in fulfilling their roles as citizens, undertaking action when necessary. This leads to a slightly wider value range; some activating content is desirable, but nothing too extreme. The Critical model however leaves more room for emotional and provocative content to challenge the status quo. Here high values of Activation should be expected.

Operationalization.

The Circumplex Model of Affect [224] describes a dimensional model where all types of emotions are expressed using the terms *valence* and *arousal*. *Valence* indicates whether the emotion is positive or negative, while *arousal* refers to the strength of the emotion and to what extent it expresses action. Following this, for example, 'excitement' has a positive valence and arousal, whereas 'bored' is negative for both. Based on the theory described above a number of "sentiment analysis" tools have been developed, which typically have the goal of identifying whether people have a positive or negative sentiment regarding a certain product or issue. For example, Hutto and Gilbert [121] provides a lexicon-based tool that for each input piece of text outputs a compound score ranging from -1 (very negative) to 1 (very positive). The absolute values of these scores can be used as an approximation of the arousal and therefore be used to determine the Activation score of a single article. Then, the total Activation score of the recommender system should be calculated two-fold. The average Activation score of the items recommended to each user provides a baseline score for whether the articles overall tend to be activating or neutral. Next, the issued recommendations are compared to the available pool of data as follows:

$$Activation(p, q) = (|polarity(q)| - |polarity(p)|)/2$$

Here p denotes the set of all available articles in the pool, and q those in the recommendation. For both sets we take the mean of the absolute polarity value of each article, which we use as an approximation for Activation. We subtract the mean from the available pool of articles from the mean of the recommendation set, which maps to a range of $[-1, 1]$. A value lower than zero indicates that the recommender system shows less activating content than was available in the pool of data, and therefore favors more neutral articles. Values higher than zero show the opposite; the recommendation sets contained proportionally more activating content than was available in the pool.

Limitations.

Of principle importance is the impact that the article's text has on the reader. However, as we have no direct way of measuring this, we hold to the assumption that a strongly emotional article will also cause similarly strong emotions in a reader, which again translates into higher willingness to act. It must also be noted that people may respond differently to different emotions (for example, anger may incite either approach (action) or avoidance (inaction) tendencies) [230]. We therefore see this approach as an approximation of the concept of activation, affect and emotion in articles, until such a time when more research in the topic allows us to be more nuanced in our perceptions.

3.4.4 Representation

The *Representation* metric expresses whether the issued recommendations provide a balance of different opinions and perspectives, where one is not unduly more or less represented than others. A score close to zero indicates a balance, where the model of democracy that is chosen determines what this balance entails, whereas a higher score indicates larger discrepancies.

Explanation.

Representation is one of the more intuitive interpretations of diversity. Depending on which model of democracy is chosen, news recommendations should contain a plurality of different opinions. Here we care more about *what* is being said than *who* says it, which is the goal of the final metric, Alternative Voices. In order to define what it means to provide a balance of opinions, one needs to refer back to the different models and their goals.

In the context of democratic recommenders.

The Participatory model aims to be *reflective* of "the real political world". Power relations that are therefore present in society should also be present in the news recommendations, with a larger share in the Representation for the more prevalent opinions. On the other hand, the Deliberative model aims to provide an *equal* overview of all opinions without one being more prevalent than the other. The Critical model has a large focus on shifting power balances, and it does so by giving a platform to underrepresented opinions, thereby promoting an *inverse* point of view. In doing this, the Critical model also strongly considers

the characteristics of the opinion holder, specifically whether they are part of a minority group or not, though this is the goal of the last metric, Alternative Voices.

Operationalization.

Representation, and Alternative Voices as well, rely strongly on the correct and complete identification of the opinions and opinion holders mentioned in the news. Though there is research available on the usage of Natural Language patterns to extract opinion data from an article's text [210], additional work is necessary on its applicability in this context. For example, it is of significant importance that not one type of opinion or opinion holder is systematically missed. Once the quality of the extraction is relatively certain, additional work is also necessary on the placement of opinions relative to each other; for example, which opinions are in favor, against or neutral on a statement, and how are these represented in the recommendations. This task is extremely complex, even for humans. In the meantime approximations can be used, for example by considering (spokespersons of) political parties and their position on the political spectrum. This can be done through manual annotations, with hardcoded lists of politicians and their parties, or automatically by for example querying Wikidata for information on persons identified through Named Entity Recognition. To calculate the Representation score, we once again use the Kullback-Leibler Divergence, but this time on the different opinion categories in the recommendations versus the available pool of data:

$$Representation_{(p,q)} = \sum_{o} p(o) log \frac{p(o)}{\tilde{q}(o|u)}$$

This calculation is similar to the one in Section 3.4.1. However, o indicates the different opinions in the data; $p(o)$ represents the proportion of the times this opinion was present in the overall pool of data, whereas $\tilde{q}(o|u)$ represents the proportion of times user u has seen this opinion in their recommendations. A score of 0 means a perfect match between the two, which means that the opinions shown in the recommendations are perfectly representative of those in society. When following the Participatory model *reflective* point of view we want this value to be as close to zero as possible, as being representative of society is its main goal. However, when following one of the other models, we have to make some alterations on the distributions expressed by p. The Critical model's *inverse* point of view aims for the recommendations to diverge as much from the power relations in society as possible. However, since very small differences in distributions can result in a very large KL divergence, simply maximizing the KL divergence is not sufficient. Instead, we inverse the distribution of opinions present in p. Similarly, when choosing the Deliberative

model, we want all opinions in the recommendations to be equally represented, and therefore we choose p as a uniform distribution of opinions. This way, for each of the different approaches holds that the closer the divergence is to zero, the better the recommendations reflect the desired representation of different opinions. For each of the reflective, inverse and equal approaches, the aggregated Representation score is obtained by averaging the Representation score over all recommendations issued to all users.

Limitations.

Kullback-Leibler divergence treats each category as being independent, and does not account for opinions and standpoints that may be more or less similar to other categories.

3.4.5 Alternative Voices

The *Alternative Voices* metric measures the relative presence of people from a minority or marginalised group. A higher score indicates a proportionally larger presence.

Explanation.

Where Representation is largely focused on the explicit content of a perspective (the *what*), Alternative Voices is more concerned with the person holding it (the *who*), and specifically whether this person or organisation is one of a minority or an otherwise marginalised group that is more likely to be underrepresented in the mainstream media. What exactly entails a minority is rather vaguely defined. Article 1 from the 1992 United Nations Minorities Declaration refers to minorities "a non-dominant group of individuals who share certain national, ethnic, religious or linguistic characteristics which are different from those of the majority population", though there is no internationally agreed-upon definition. In practice, this interpretation is often extended with gender identity, disability and sexual orientation. A major challenge of the Alternative Voices metric lies in the actual identification of a minority voice. Though there are a number of studies that aim to detect certain characteristics of minorities from textual data, such as predicting a person's ethnicity and gender based on their first and last name [244], there are no approaches that 1) model all minority characteristics or 2) perform well consistently. This process needs significant additional and most importantly multidisciplinary research, with a large focus on ensuring that doing this type of analysis does not lead to unintended stereotyping, exclusion or misrepresentation. For example, Keyes [137] shows that current studies typically treat gender classification as a purely binary problem, thereby systematically leaving out and wrongly classifying transgender people. Similarly, Hanna et al. [99] argue that race and ethnicity

are strongly social constructs that should not be treated as objective differences between groups. This topic, typically referred to as (algorithmic) Fairness, is an active research field that aims to counter bias and discrimination in data-driven computer systems. One thing is for certain: any recommender system that actively promotes one type of voice over another should make very explicit on what criteria and following which methods it does this. Following this both the identification and the way its algorithms use this information must be fully transparent and auditable. However, for the remainder of this section we will assume that we do have a proper way of identifying people from a minority group, either through manual annotation or automatic extraction.

In the context of democratic recommenders.

The Alternative Voices metric is naturally most significant in the Critical model, which aims to provide a platform to voices that would otherwise go unheard, and therefore has a large focus on the opinions and perspectives from minority groups. To a lesser extent, the same holds for the Participatory and Deliberative models, where the first aims to foster tolerance and empathy, and the second that they should be equally represented.

Operationalization.

The discussion around Fairness in machine learning systems has lead, among others, to a number of definitions of the concept. For the operationalization of Alternative Voices we adapt Equation 10 of Burke et al. [28] for our purposes:

$$AlternativeVoices = \frac{q^+/p^+}{q^-/p^-}$$

Here q^+ denotes the number of mentions of people belonging to a protected group in the recommendations, whereas p^+ denotes the number of mentions of people belonging to a protected group in all the available articles. q^- and p^- denote similar mentions, but for people belonging to the unprotected group. Though the example given in Burke et al. [28] describes the equation being used to identify whether loans from protected and unprotected regions appear equally often, it is also directly applicable to our notion of Alternative Voices; however, rather than counting regions being recommended, we count the number of times that people from minority (protected) versus majority (unprotected) groups are being mentioned in the news. This function maps to 1 when there is a complete balance between people from the protected and the unprotected groups. When the value is larger than 1 more people from unprotected groups appear in the recommendation set, whereas lower than 1 means they appear less.

Again, the aggregate score consists of the average Alternative Voices score over all recommendations issued to all users.

Limitations.

3

A major caveat of this approach is that it assumes that the mere mentioning of minority people is enough to serve the goals of the Alternative Voices metric. This disregards the fact that these people may be mentioned but from another person's perspective, or in a negative light. Further research should focus on not only identifying a person from a minority group, but also whether they are mentioned as an active or passive agent.

Table 3.1: Overview of the different models and expected value ranges for each metric. Note that for the metrics reflecting distance of a distribution (Calibration and Representation), a "High" target value actually means that the resulting value should be close to zero.

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

3.5 General limitations

Though all of the metrics described in Section 3.4 already mention the limitations of that metric specifically, this section describes a number of the limitations of this method as a whole.

ORDERING Of the currently specified metrics, only Fragmentation takes the ordering of the items in the recommendation into account. However, the top result in a recommendation is of significantly more importance than the result in place 10. In future work, the other metrics should be extended in such a way that they reflect this.

FORMALISM TRAP Many of the concepts described here are susceptible to the Formalism Trap described in [232], which is defined as the *"[f]ailure to account for the full meaning of social concepts [...], which can be procedural, contextual and contestable, and cannot be resolved through mathematical formalisms"*. Though our approach aims to model concepts founded in social science and democratic theories, they are merely approximations and to

a large extent simplifications of very complex and nuanced concepts that have been contested and debated in the social sciences and humanities for decades. To claim our approach comes close to covering these subtleties would be presumptuous - however, we do believe it is necessary to provide a starting point in the modeling of concepts that have so far largely been neglected or oversimplified in the evaluation of news recommendations. The pitfalls of this trap should be mitigated by always providing full transparency on how these concepts are implemented, on what kind of data they are based, and most importantly on how they should (and should not) be interpreted.

Bias in the dataset The metrics presented in Section 3.4 typically rely on measuring a difference between the set of recommended items and the full set of articles that were available, the reading history of the user in question or among users. What it does not do is account for inherent bias in the overall dataset, though the possibility of exposure diversity depends on the availability of content in the pool. If the quality and diversity of the pool is low, recommenders have insufficient options to provide good recommendations. That means exposure diversity ultimately is dependent on external diversity. Detecting such a bias in the dataset rather than in the produced recommendations and undertaking steps to remedy this needs additional work.

Nudging for more diverse news consumption The metrics discussed here do not reflect on the process of getting users to actually consume more diverse content. Different users may have different 'tolerance' for diversity, depending on the topic and even on things such as the time of day. Whether or not news recommenders can successfully motivate users to consume more diverse can also depend on the (user-friendly) design of the recommender and the way the recommendations are presented [166]. Designing for more diverse news consumption also gives rise to a different discussion: is it ethical to nudge news consumption, even if it is for a commendable goal such as "more diversity" or "countering filter bubbles", and where do we draw the line between offering more diverse recommendations and manipulating the reader? The complexity and breadth of this topic are out of scope for this paper, but should be considered in future work.

Broader institutional context Efforts to develop more diverse and inclusive news recommendation metrics and models do not, on their own, mean that users will receive more diverse recommendations; that requires a combination of editorial judgement, the availability of internal workflows that translate this judgement into technology design, the room to implement alternative diversity metrics in third party software (which again depends on the degree of professional autonomy and negotiating power between the media

and software providers), and users who engage with the algorithm when presented with a particular recommendation. The design approach must thus additionally consider how values are re-negotiated between stakeholders (e.g. editors, data scientists, regulators, external technology providers), how values are embedded in organizational practices of a news room, and how professional users, citizens, and society create control mechanisms and governance frameworks to realize public values, such as diversity.

Inherent limits to value by design approaches Finally, it is important to be mindful of another lesson from the general diversity by design debate, namely that there are also certain limits to value sensitive design, in our case the extent to which diversity as a normative concept can be operationalized in concrete recommender design. This can have to do with the sheer difficulty of translating certain aspects of diversity, but also with the trade-offs between values that optimizing for exposure diversity can involve. Examples of this are commercial constraints and the need to optimize for profit rather than for diversity, but also the limited effectiveness of recommenders in actually steering user choices.

3.6 Implementation

We are working on the implementation of the concepts and metrics discussed here in an open source tool[1]. The goal of this tool is to implement the metrics described in this paper as evaluation metrics for recommender design, and in doing so enable media companies to evaluate the performance of their own recommendations against those of several baseline recommendations.

Approach By making comparisons between the different recommender approaches, media companies should be able to draw conclusions about which recommender strategy fits their editorial mission best. By also comparing the performance of these algorithms to very simple recommendation approaches, such as a random recommender, the media company can also draw conclusions about where the recommender simply reflects the available data, and where it significantly influences the type of data that is shown. By making these visualizations as intuitive as possible, they should facilitate the discussion between data science teams, editors and upper management around this topic. To make this approach reusable and broadly applicable, it should be implemented and tested on both a benchmark set such as [293] and in a real-life setting. We are in contact with multiple media companies, to inform them about the different models of democracy,

[1] https://www.github.com/svrijenhoek/dart/

facilitate the discussion around this subject, and stimulate and test the implementation of our tool. Simultaneously this topic is continuously being discussed with experts from many different disciplines, as happened for example during a Dagstuhl Workshop[17].

GUIDELINES FOR ADOPTION The ultimate goal of this paper is to propose notions that could be incorporated in recommender system design. In our vision, media companies could approach this in the following steps:

1. Determine which model of democracy to follow
   Following the different models described in Section 3.3, the media company in question should decide which model of democracy the recommender system should reflect. This is something that should be decided in active discussion with the editorial team, and directly in line with the media company's mission.

2. Identify the corresponding metrics
   Use Table 3.1 to determine which metrics are relevant, and what the expected value range for each metric is. For example, when choosing to follow the Deliberative model, the recommender system should optimize for a low Fragmentation, low Activation and equal Representation. Similarly, for the Critical model, it should optimize for high Activation, inverse Representation and high Alternative Voices.

3. Implement into recommender design
   Here it is of key importance to determine the relative importance of each metric, and how to make a trade-off between recommender accuracy and normative diversity. For example, Mehrotra et al. [176] details a number of approaches to combining Relevance and Fairness in Spotify's music recommendation algorithm, and this approach can also be applied in the trade-off between accuracy and the metrics relevant for the chosen model.

We do not consider these metrics to be the final "truth" in the identification of diversity in news recommendations. The metrics and their operationalizations should serve as inspiration and a starting point for discussion, not as restrictions or set requirements for "good" recommender design.

3.7 Discussion

In this paper we have translated normative notions of diversity into five metrics. Each of the metrics proposed here is relevant in the context of democratic news recommenders, and combined they form a picture that aims to be expressive of the nuances in the different models. However there is still a lot of work to be done, both in terms of technical feasibility

and in undertaking steps to make diversity of central importance for recommender system development.

At the basis of our work is that we believe diversity is not a single absolute, but rather an aggregate value with many aspects and a mission in society. In fact, we argue that what constitutes 'good' diversity in a recommender system is largely dependent on its goal, which type of content it aims to promote, and which model of the normative framework of democracy it aims to follow. As none of these models is inherently better or worse than the others, we believe that a media company should take a normative stance and evaluate their recommender systems accordingly.

Different fields and disciplines may have very different notions of the same concept, and navigating these differences is a process of constant negotiation and compromise, but also of expectation management. Abstract concepts such as diversity may never be fully captured by the hard numbers that recommender system practitioners are used to. As recommendation algorithms take on an ever more central role in society, the necessity to bridge this gap and make such concepts more concrete also arises. Social sciences, humanities and computer science will need to meet in the middle between abstract and concrete, and work together to create ethical and interpretable technologies. This work is not a final conclusion on how diversity can be measured in news recommendations, but rather a first step in forming the bridge between the normative notion of diversity and its practical implementation.

4

A divergence-based base formalization for diversity

Authors: **Sanne Vrijenhoek***, Gabriel Bénédict*, Mateo Gutierrez Granada, Daan Odijk and Maarten De Rijke (*equal contribution)

Original title: RADio: rank-aware divergence metrics to measure normative diversity in news recommendations AND RADio* - An Introduction to Measuring Normative Diversity in News Recommendations

Published in: Proceedings of the 16th ACM Conference on Recommender Systems, ACM Transactions on Recommender Systems 3, 1, Article 5 (March 2025)

https://doi.org/10.1145/3523227.3546780, https://doi.org/10.1145/3636465

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 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.

4

4.1 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 [294]. 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 [238] – 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 [22, 185]. 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 [187], or present the reader of a particular news article with other articles about the same topic, but from a different perspective [183]. The use of news recommender systems has a wide range of benefits. They can increase engagement [197] and help raise informed citizens [76]. 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 [83, 252], and counter so-called filter bubbles or echo chambers [185, 211].

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 [132]. 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 [154]. As such, their evaluation has different requirements than those of other types of recommender systems [11, 13, 282, 288]. Recent controversies have shown that merely optimizing for click-through rates and engagement may promote sensationalist content [257], 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 [23, 93]. 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 [209], financial incentives [25], nudging [172] or even identifying user preferences [16, 162] and biases [284]. 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 techni-

[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/

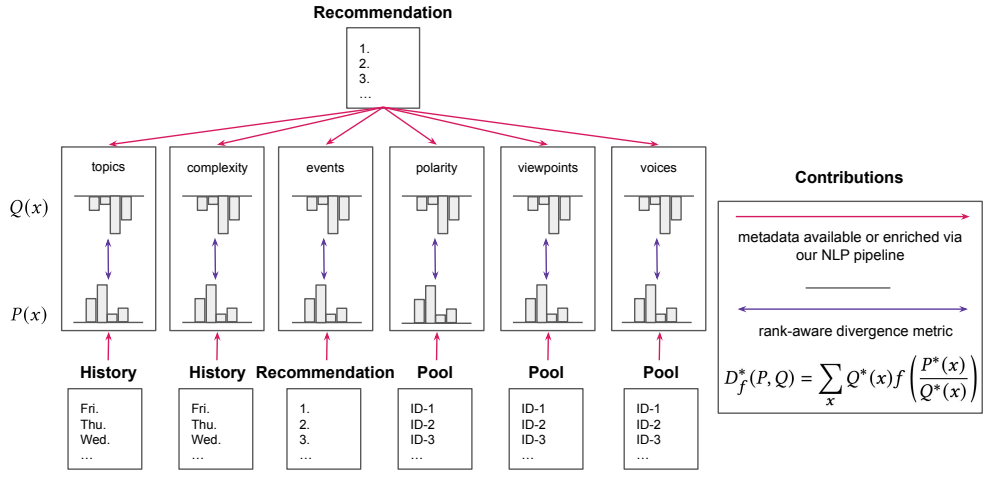


Figure 4.1: Comparing discrete diversity distributions in the context of news recommendations. First, metadata is collected in the news dataset or retrieved via our NLP pipeline (red). Discrete distributions of that metadata are then compared via a rank-aware divergence metric (purple). Recommendation set Q and the context articles P are compared with rank-aware f-Divergence.

cal and editorial purposes. We build on the work of Helberger [111], who provides a theoretical foundation for conceptualizing diversity, and of Vrijenhoek et al. [278], 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 paper, 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 [293].

Figure 4.1 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

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

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. [278]. 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). Lastly, we added an additional section to the Discussion on how practitioners looking to put the metrics in production could approach this.

4.2 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]

4.2.1 Descriptive (General-Purpose) Diversity

Diversity is a central concept in Information Retrieval literature [38, 226], 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 conceptualized as the 'opposite of similarity'. 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) [29, 55, 66, 71, 131, 161, 273]: 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 [72]. 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 [148, 161].

[3] This dichotomy is oftentimes referred to as normative (*what ought to be*) and positive (*what is*) statements [120] 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.

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 [148]. One can optimize for these descriptive notions of diversity (i) before training, by clustering users based on their profile diversity with JS divergence [75], (ii) directly at training time (e.g., for learning-to-rank [24, 29, 273], collaborative filtering [216], graphs [90, 214] or bandits [58, 296]), (iii) by re-ranking a recommendation set and balance diversity vs. relevance [34] or popularity vs. relevance [31], 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 [5, 283]). 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 [161]. Nevertheless, this encouraged recent efforts in training neural-based recommenders that explicitly make a trade-off between accuracy and diversity [218]. Also recently, there have been studies that differentiate between diversity needs of users [295].

4.2.2 Normative Diversity

Diversity is extensively discussed as a normative concept in literature, and has a role in many different areas of science [158, 248], spanning from ecological diversity to diversity as a proxy for fairness in machine learning systems [179]. 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 [111], 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 [111] 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 representation of a plurality of sources, voices and opinions. Lastly 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 [227].

For more details regarding the different models, and what a recommender system following each of these models would look like, we refer to Helberger [111]. 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 [278] take a first step towards normative diversity for recommender systems and reflect the nuances of the different democratic models described above. See Table 4.1 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). On the other hand, **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 4.1. 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. Lastly, 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 [111], 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 [241]. For example, Hada et al. [97] 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.

4.2.3 The Gap Between Normative and Descriptive Diversity

The descriptive diversity metrics described in Section 4.2.1 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 4.2.2. In their comprehensive work on the implementation of media diversity across different domains, Loecherbach et al. [158] note

Table 4.1: 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

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 [38, 226], 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. [23] (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 [93] (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. [158] and Bernstein et al. [17] call for truly interdisciplinary research in bridging this gap, where Bernstein et al. [17] 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.

4.3 Operationalizing Normative Diversity for News Recommendation

With our RADio framework, we further refine the DART metrics that were defined by Vrijenhoek et al. [278] 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 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 4.3.1), then describe how we reformulate the metrics to each use the same divergence-based approach (Section 4.3.2). We then add the rank-aware aspect to the metrics (Section 4.3.3), before applying them to the five concrete DART metrics (Section 4.3.4).

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 [204].

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 [54]) and (iii) allow for discrete rank-based distribution sets, to fit the ranked recommendation setting.

4.3.2 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 [247]), 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 4.1). 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.

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 $x_1, ..., x_n$ and corresponding probabilities $P(x_1), ..., P(x_n)$ 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 [259]. The concept of relative entropy or KL (Kullback–Leibler) divergence [147] between two probability mass functions P and Q (here, a recommendation and its context) is defined as:

$$D_{\mathrm{KL}}(P, Q) = -\sum_{x \in \mathcal{X}} P(x) \log_2 Q(x) + \sum_{x \in \mathcal{X}} P(x) \log_2 P(x). \tag{4.1}$$

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 fulfil the requirements for *symmetry* and *triangle inequality*. This can be resolved by further refining KL Divergence.

Jensen–Shannon Divergence

A succession of steps from KL divergence leads to Jensen-Shannon (JS) divergence. KL divergence was first turned symmetric [128] and then upper bounded [155], to lead to

$$\begin{aligned} D_{\mathrm{JS}}(P, Q) = &-\sum_{x \in \mathcal{X}} \frac{P(x) + Q(x)}{2} \log_2 \left(\frac{P(x) + Q(x)}{2} \right) \\ &+ \frac{1}{2} \sum_{x \in \mathcal{X}} P(x) \log_2 P(x) + \frac{1}{2} \sum_{x \in \mathcal{X}} Q(x) \log_2 Q(x) \end{aligned} \tag{4.2}$$

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 [74] 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.

4

f-Divergence

Liese and Vajda [153] 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

$$D_f(P,Q) = \sum_x Q(x) f\left(\frac{P(x)}{Q(x)}\right), \tag{4.3}$$

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

$$\overline{Q}(x) = (1-\alpha)Q(x) + \alpha P(x) \quad \overline{P}(x) = (1-\alpha)P(x) + \alpha Q(x), \tag{4.4}$$

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 $\sum_x \overline{P}(x) = \sum_x \overline{Q}(x) = 1$. In the following sections P and Q will implicitly refer to $\overline{P}$ and $\overline{Q}$ in order 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 1. *The Kullback-Leibler Divergence is an f-Divergence*

[4] f-Divergence accommodates for other divergence metrics which are out of scope of this research, such as the Hellinger divergence, and the Pearson divergence [153].

Proof. Starting from Equation 4.1,

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

□

Theorem 2. *The Jensen-Shannon Divergence is an f-Divergence*

Proof. Starting from Equation 4.2,

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

□

4.3.3 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 [256, 298], and by extension in conventional descriptive diversity metrics [29] 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 [30, 127], in particular for news recommendation [293]. 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:

$$Q^*(x) = \frac{\sum_i w_{R_i} \mathbb{1}_{i \in x}}{\sum_i w_{R_i}}, \tag{4.5}$$

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., $Q^* = Q$).

4

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 [139]. 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 [30].

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 4.1), 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:

$$D^*_{f_{\mathrm{JS}}}(P, Q) = \sum_x Q^*(x) f_{\mathrm{JS}} \left(\frac{P^*(x)}{Q^*(x)} \right), \tag{4.6}$$

$Q^*(x)$ 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.

4.3.4 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."

[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.

- 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 [278], the original paper contains further normative justifications.

CALIBRATION (Equation 4.7) 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 4.8) 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 4.3.2), 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.

ACTIVATION (Equation 4.9) 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 [278], 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 4.10) 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 4.11) 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 neighborhoods in recommender systems [28]. 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 neces-

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

Table 4.2: 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	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.[7]

sary features from an example news dataset:

$$Calibration = Cal\big(P^*(c|H), Q^*(c|R)\big) = \sum_c Q^*(c|R) f_{\mathrm{JS}}\left(\frac{P^*(c|H)}{Q^*(c|R)}\right) \quad (4.7)$$

$$Fragmentation = Frag\big(P^*(e|R^u), Q^*(e|R^v)\big) = \sum_e Q^*(e|R^v) f_{\mathrm{JS}}\left(\frac{P^*(e|R^u)}{Q^*(e|R^v)}\right) \quad (4.8)$$

$$Activation = Act\big(P(k|S), Q^*(k|R)\big) = \sum_k Q^*(k|R) f_{\mathrm{JS}}\left(\frac{P(k|S)}{Q^*(k|R)}\right) \quad (4.9)$$

$$Representation = Rep\big(P(p|S), Q^*(p|R)\big) = \sum_p Q^*(p|R) f_{\mathrm{JS}}\left(\frac{P(p|S)}{Q^*(p|R)}\right) \quad (4.10)$$

$$AlternativeVoices = AltV\big(P(m|S), Q^*(m|R)\big) = \sum_m Q^*(m|R) f_{\mathrm{JS}}\left(\frac{P(m|S)}{Q^*(m|R)}\right) \quad (4.11)$$

4.4 Experimental Setup

In order to demonstrate RADio's potential effectiveness, we developed an NLP pipeline to retrieve input features to the metrics in Section 4.3.4 and ran them on the MIND dataset, an open source dataset made available by Microsoft News. The MIND dataset [293] 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 [291], NAML [290], LSTUR [4] and NRMS [292]. It was shown that these algorithms outperform general-purpose ones [293] or common collaborative filtering models (such as alternating least squares (ALS)), in particular due to the short lifespan of news items [91]. 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. There are a number of limitations to the usage of the MIND dataset (see Section 4.8 and Vrijenhoek [277]). However, at the moment of 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 paper to improve on the identification and extraction of the relevant parameters proposed in Vrijenhoek et al. [278], 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:

4

1. **Complexity analysis** – Each item is assigned a complexity score based on the Flesch-Kincaid reading ease test [141], implemented in the Python module py-readability-metrics [57]. 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 4.3.4). 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 [265].

3. **Sentiment analysis** – Using the textBlob open source NLP library we assign each article a sentiment polarity score [160]. 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

4. **Named entity recognition** – Using spaCy, we identify the people, organizations and locations mentioned in the text [119], 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]

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 [141]. 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 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 4.2). 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 4.2 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 and TF-IDF of the article's body (text) as article representations.[12]

[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] To make calculating ILD on all pairs of articles computationally feasible we used Spark [300] to load the entire matrix of all intra-list article pairs and parallelize the distance calculation. As this approach was incompatible

4.5 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 4.8. 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 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 4.3[13] displays results for RADio with rank-aware JS divergence, NDCG and ILD. NDCG scores are comparable to the results obtained in [293]. 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 4.2.2.

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 in Section 4.8. 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*

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.

[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.

Table 4.3: 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 4.1.

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	-

in terms of the NDCG score.

Figure 4.2 is the pendant of Table 4.3. 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 4.4, 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. [293]). 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.

4.6 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 ($@n$).

4.6.1 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 4.3.2). Figure 4.5 substantiates that claim empirically, visualizing the sensitivity of RADio to the two described

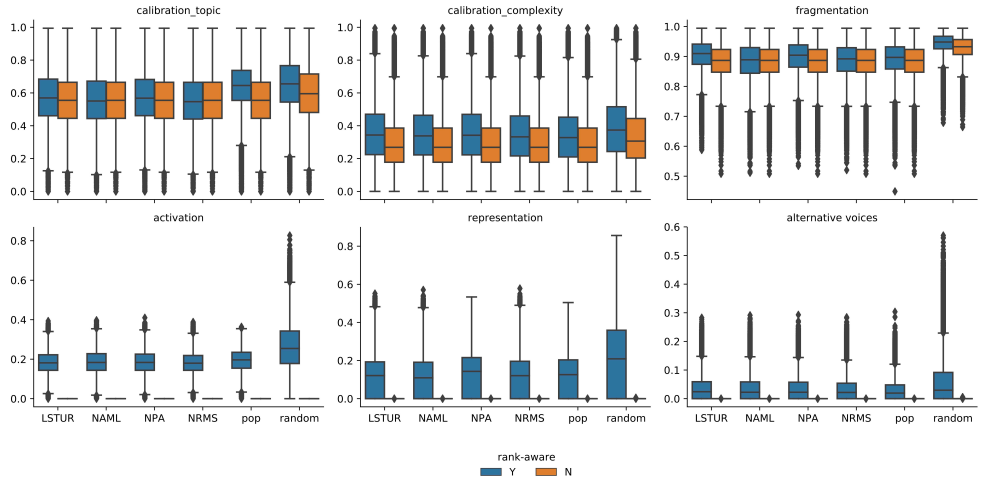


Figure 4.2: Boxplot over each P, Q pair for all algorithms and all metrics for the JS divergence with rank-awareness. This is the pendant of Table 4.3. The horizontal line and the box boundaries show median and Inter Quartile Range (IQR) respectively. Single dots indicate outliers, defined as being further outside the 0.05 to 0.95 quantiles.

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 4.4. This time, metrics are not bounded by $[0, 1]$. If one were to rank the values for each metric, Table 4.3 and 4.4 would score the algorithms in the same order [153]. 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.

4.6.2 Sensitivity to Rank-awareness

In the original formulation of DART metrics [278], rank-awareness was not considered for most of the defined metrics. In our formalization, rank-awareness is the default. In Figure 4.6, 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

Table 4.4: Results for our RADio framework for recommendation algorithms on the MIND dataset. 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 4.1. These metrics are executed on a random sample of 35.000 users.

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

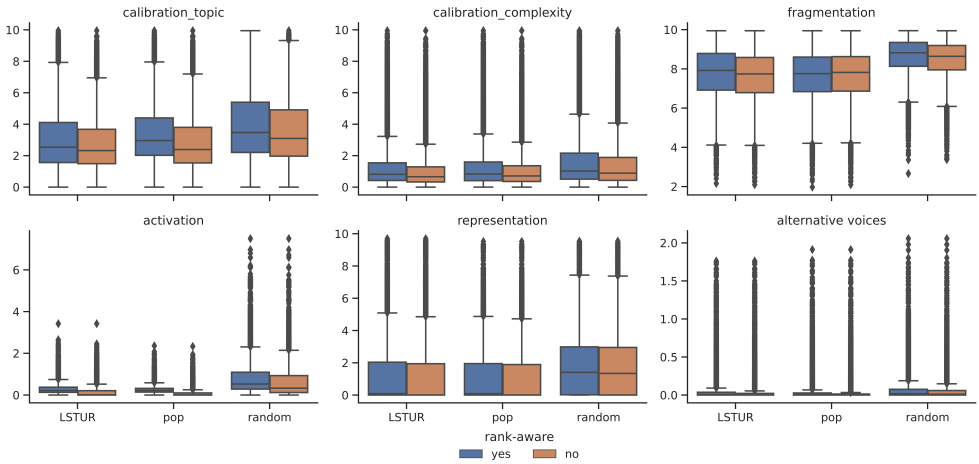


Figure 4.3: Boxplot over each P, Q pair for all metrics for the KL divergence with and without rank-awareness. The horizontal line and the box boundaries show median and Inter Quartile Range (IQR) respectively. Single dots indicate outliers, defined as being further outside the 0.5 to 0.95 quantiles.

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 4.3 and 4.4 show a detailed analysis of rank-awareness of robust estimates (median and quantiles) for KL and the JS divergence.

4.6.3 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 4.7. The figure shows that the effect of rank-awareness becomes stronger with a higher cut-off point, and causes

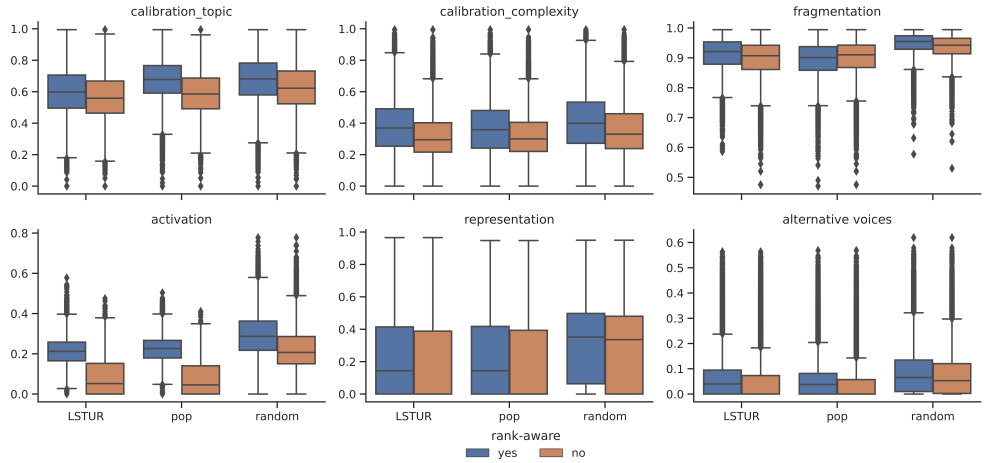


Figure 4.4: Boxplot over each P, Q pair for all metrics for the JS divergence with and without rank-awareness. The horizontal line and the box boundaries show median and Inter Quartile Range (IQR) respectively. Single dots indicate outliers, defined as being further outside the 0.5 to 0.95 quantiles.

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.

4.6.4 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 4.1), 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 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."

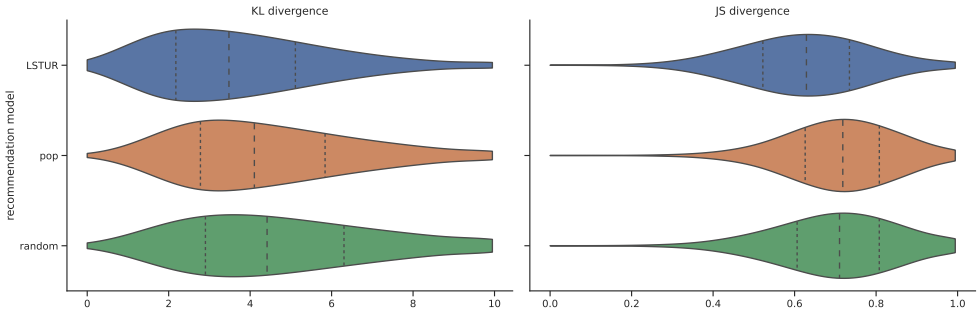


Figure 4.5: Violin kernel density functions [289] over each P, Q pair, for the Calibration (topic) rank-aware metric, rank cutoff @N (no cuttof), using KL (left) and the JS divergence (right). Thick and thin dashes show median and Inter Quartile Range (IQR) respectively.

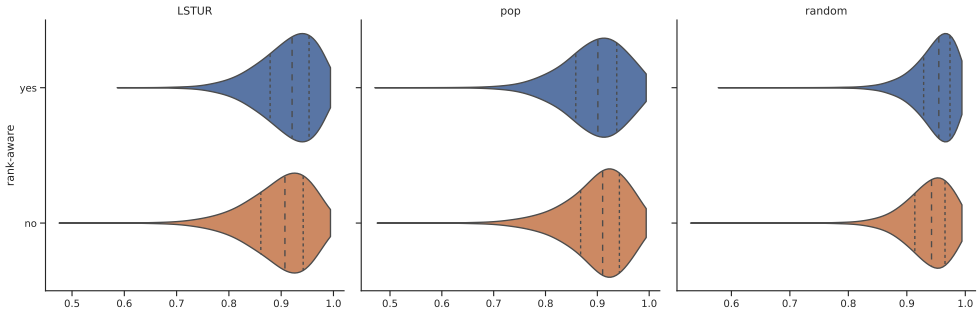


Figure 4.6: Violin kernel density functions [289] over each P, Q pair, for the Fragmentation metric with JS divergence and rank cutoff @N (no cutoff), on three different recommender approaches with (blue) and without (orange) rank-awareness. Thick and thin dashes show median and IQR respectively.

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 trade-off between different target values in the recommendation, or even explicitly optimize on these metrics.

4.7 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 aims to measure the degree in which a recommendation reflects a

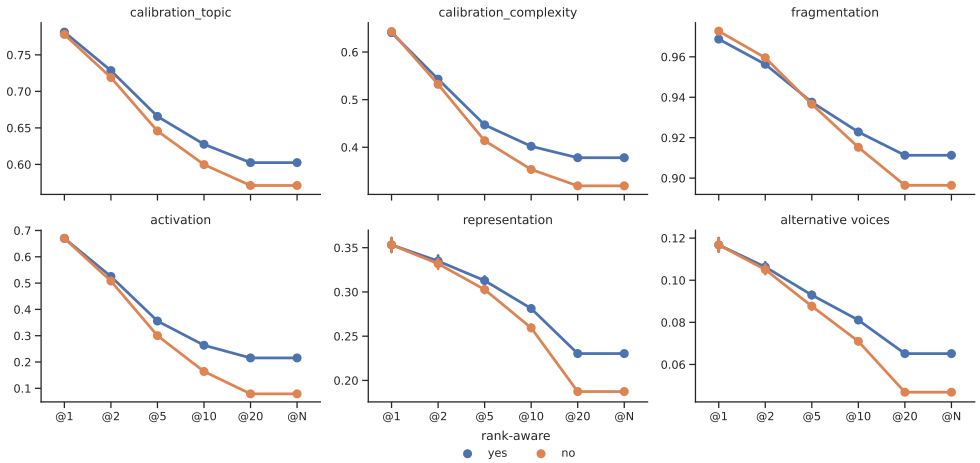


Figure 4.7: Mean and 95% confidence interval for each DART metric implemented with JS divergence for the LSTUR recommender. Sensitivity analysis of RADio on rank-awareness (blue and orange) and rank cutoff.

	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

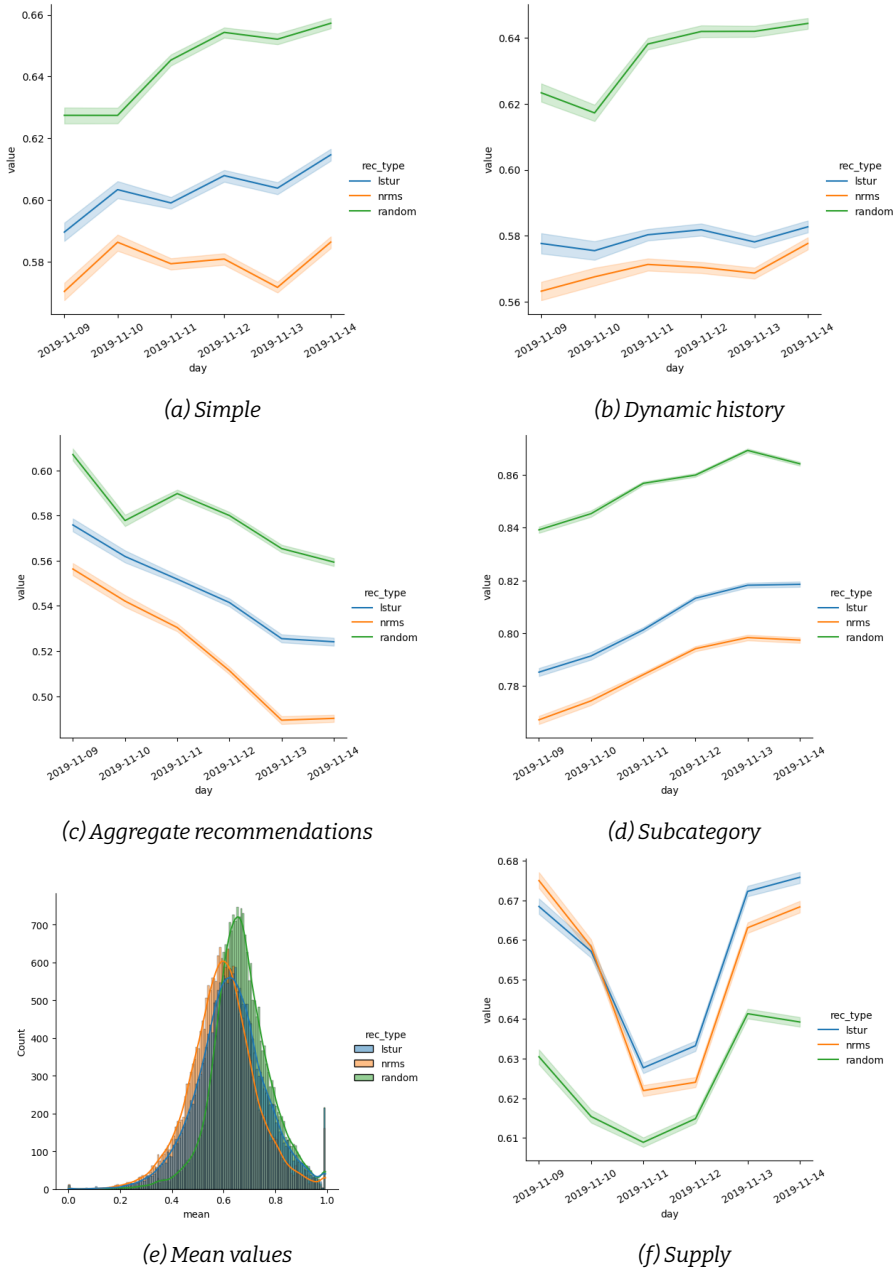
Table 4.5: Distribution of article categories in both the training and validation set of MIND. This includes the full dataset ('all'), the content people have seen and which the recommender ranks ('impr'), what was in the user's history ('hist'), the items the users clicked ('click'), and the top 5 items ranked by the neural recommenders and random baseline.

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 4.5 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 5 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, 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 4.8a, we visualize Calibration in the same way as in Section 4.4: 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 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 4.8b. 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.



(a) Simple

(b) Dynamic history

(c) Aggregate recommendations

(d) Subcategory

(e) Mean values

(f) Supply

Figure 4.8: The Calibration metric with different approaches to metric design

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 4.8c, we conceptualize our metric to not only consider the current recommendation, but also the top 5 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 [247]. 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 4.8d displays the Calibration score when considering subcategory. It shows roughly the same pattern as in Figure 4.8a, although in a higher range: the average value here is centered around 0.8, whereas in Figure 4.8a 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 4.8e, 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 4.8f 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 4.8 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.

4.8 Discussion

Choosing an f-Divergence score as the base for our metrics allows us to construct a single base formalization with a clear interpretation 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. 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 paper, 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, 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 bins. Future work may propose a different approach for calculating divergence between continuous variables. Lastly, 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 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 [8, 63]. 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 [54].

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, 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. 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 [6]. Besides technical and editorial teams, this would also include management and business teams that eventually have to make an informed decision on the acceptable trade-off 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 4.2, 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 [280]. 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 in order to use the metrics for optimization.

Lastly, if metrics such as these normative diversity metrics are to be incorporated in a production system, 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.

4.9 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 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.

5

The public datasets that are available, and how they can(not) be used for diversity

Authors: Max van Drunen* and **Sanne Vrijenhoek*** (*equal contribution)

Original title: How public datasets constrain the development of diversity-aware news recommender systems, and what law could do about it.

https://doi.org/10.48550/arXiv.2510.05952

Abstract

News recommender systems increasingly determine what news individuals see online. Over the past decade, researchers have extensively critiqued recommender systems that prioritise news based on user engagement. To offer an alternative, researchers have analysed how recommender systems could support the media's ability to fulfil its role in democratic society by recommending news based on editorial values, particularly diversity. However, there continues to be a large gap between normative theory on how news recommender systems should incorporate diversity, and technical literature that designs such systems. We argue that to realise diversity-aware recommender systems in practice, it is crucial to pay attention to the datasets that are needed to train modern news recommenders. We aim to make two main contributions. First, we identify the information a dataset must include to enable the development of the diversity-aware news recommender systems proposed in normative literature. Based on this analysis, we assess the limitations of currently available public datasets, and show what potential they do have to expand research into diversity-aware recommender systems. Second, we analyse why and how European law and policy can be used to provide researchers with structural access to the data they need to develop diversity-aware news recommender systems.

5.1 Introduction

News recommender systems play an important role in determining what news (if any) individuals see online by ordering information from most to least relevant [110, 158, 222]. Researchers have extensively criticised the common industry practice of measuring relevance through clicks, arguing this leads media organisations to recommend news articles based on what is most engaging rather than based on the editorial values that should guide the way the way the media informs the public [17, 80]. Over the past decade, researchers have therefore proposed alternative approaches to news recommendation that would support the media's democratic function of informing the public, particularly by allowing media organisations to make more diverse recommendations that expose each reader to perspectives and topics that are new to them. Researchers have analysed extensively how diversity-aware recommender systems should be conceptualised [17, 111, 227, 275, 276], and what technical tools are needed to measure and embed diversity in recommender systems [12, 107, 161, 273, 279].

However, a significant gap remains between the diversity-aware news recommender systems proposed in normative literature, and the kinds that have been realized in technical literature. For example, scholars in law and journalism studies call for recommender systems that allow readers to become experts in their chosen field by showing them new perspectives on a few topics that interest them, or that foster tolerance by showing readers political viewpoints with which they disagree but which do not lead them to see other societal groups as enemies [111, 154, 227]. In contrast, technical literature most often implements diversity as an intra-list similarity function, to check whether the items in the recommendation are not too similar to each other [29, 217]. While work that designs news recommender systems based on more normative conceptualisations of diversity exists, it remains limited and regularly emphasises it is not yet suitable for deployment. Much work has yet to be done before industry-level recommender systems can truly account for normative interpretations of diversity [279].

In this paper, we argue that the public datasets on which researchers rely to create and evaluate news recommender systems are a key factor limiting the development of diversity-aware recommender systems. We show that while these datasets do have some untapped potential to further research into diversity-aware recommender systems, they lack much of the data needed to create the kinds of recommender systems proposed based on normative theory. Instead, they primarily facilitate the development of recommender systems that optimise for readers' short-term preferences. Further, we argue that this constraint cannot be fully addressed with the publication of another single dataset, as the

data needed for diverse recommendations is highly contextual and varies significantly across time and regions. We therefore also explore the European law and policy tools available to secure structural access to the data needed to develop diversity-aware news recommender systems.

With this, we aim to answer the following research question: how do public datasets constrain the development of diversity-aware news recommender systems, and how can European law and policy be used to overcome this constraint? We focus on European law because of its recent advances in data access law and policy, its role for active governmental intervention to support media pluralism, and its wide geographic scope (allowing it to, potentially, support the creation of public datasets in multiple countries). We include the four most-used public datasets in our analysis, but focus in particular on the Microsoft News Dataset (MIND). Due to its size, its publication in 2022 was a significant development for news recommendation research, and our research indicates that it is by far the most used dataset to develop new recommender systems.

Section 2 analyses how public datasets are used to develop news recommender systems, and how they can impact the recommender systems used by media organisations by influencing the research that develops and evaluates the performance of new recommender systems. Section 3 shows what data would have to be included in a dataset to develop the kinds of diversity-aware recommender systems imagined in legal and journalism studies literature. Section 4 evaluates how the data included in the public datasets that are currently most used in news recommendation research limits the development of diversity-aware recommender systems. Section 5 analyses why European law and policy should play a role to address the lack of datasets needed to build better news recommender systems, and how existing European laws and policies intended to expand access to data can do so. Throughout, we aim to make the paper accessible for computer science, legal, and journalism scholars researching how more diverse news recommender systems can be realised. The analysis may therefore at times be overly simplistic for one specific audience.

5.2 From datasets to news recommendations

Training a recommender system involves optimising it to predict a quantitatively measurable target; most often, whether a user will click on an item or not [302]. A recommender system is trained to make these predictions by identifying patterns between item characteristics and user characteristics that are associated with the click. These patterns are subtle. For example, a click on an article about the Olympic Games may indicate the user

is interested in the Olympic Games in general, a specific team, sport, or athlete, or even the politics surrounding it. The broader the scope of content a system needs to cover, the harder it is to find the patterns that predict user behaviour. Furthermore, the way readers interact with a system is highly connected to how content is shown to them, meaning that slight changes in the user interface can lead to very different engagement patterns [14]. To capture these nuances, recommender systems must be trained on large datasets.

However, since behavioural patterns are so dependent on contextual factors one needs to, in a manner of speaking, have a recommender system running in order to generate the data needed to train a recommender system. Organisations may resolve this chicken-and-egg situation by starting out with simpler recommender systems that involve little or no personalization. The interactions generated with these systems will then, over time, serve as input to more complex and sophisticated models. Very few organisations will develop the architecture for said sophisticated models from scratch; rather, they will train and test multiple models to see which is most likely to cover their needs best, and continue to fine-tune that model. This choice is heavily influenced by which models are considered state-of-the-art in academic research. New models are usually published in academic papers and shared benchmarks, accompanied by an analysis of the model's strengths, weaknesses, and typical use cases, which helps industry practitioners decide whether the model will be a good fit for their application. Yet contrary to industry professionals, researchers do most of their work on public datasets. For one, they rarely have access to datasets at the scale that is necessary for training recommender systems, and even if they do, relying on a proprietary dataset limits the reproducibility of their results [80]. This implies that, although media organisations do not directly use public datasets to train their recommender systems, the research that informs their design choices and algorithm selection depends heavily on the few public datasets that are available, and what can be done with them.

One concrete example of how this works in practice can be seen in work by Einarsson et al. [70], which studied the effect of news recommender systems deployed by Ekstra Bladet during the 2022 Danish election [70]. Ekstra Bladet had two recommender systems in production, both of which were trained on Ekstra Bladet's own data. However, the algorithms Ekstra Bladet used to train its recommender systems (NRMS [292] and Matrix Factorization [217]) were developed and evaluated in the research literature, using the Microsoft News Dataset [293].

Given the required complexity and size of news recommendation datasets, only a few datasets are available for recommender systems researchers to work with. As a result, what is in those datasets also strongly influences the type of research that *can* be done. The data included in a dataset constrains what recommender systems developed with the

dataset can be designed to do. 'Simply' training a recommender system that predicts what news users will click will already require a large dataset of news articles and users' interactions with those articles from which these patterns can be derived. Developing a recommender system that balances user preferences with normative values such as diversity requires a richer dataset, which also includes data necessary to assess how the recommendation relates to those values [35, 251]. If such data is not present, executing the research becomes significantly more complicated, which we will further elaborate on in Section 5.4.

Beyond determining the research that can be done, datasets that become a benchmark also influence the type of research that is considered valuable and encouraged within the research community. Indeed, the MIND dataset, mentioned earlier in the context of the Ekstra Bladet algorithms, was created with the explicit purpose "to serve as a benchmark dataset for news recommendation and to facilitate the research in news recommendation"[1]. As a result, researchers are more likely to execute their research on the targets and optimization goals as set out by Microsoft. Another good example is the MovieLens dataset, published in 2005 and still a widely used dataset for recommender system research. MovieLens contained information on users rating a movie on a scale of 1 to 5, which means that to this day, a lot of research is carried out on predicting scores on an ordinal scale [102]. Benchmark datasets thus fulfil an important function in research: they serve as a common reference point by allowing the performance of new recommendation algorithms to be compared to that of existing algorithms tested on the same dataset. However, they also make it more likely that research is done on the one hand on the data structures present in the benchmark dataset, rather than on the structures relevant to news organisations, and on the other hand on the performance goals for which that dataset was constructed, rather than on normative goals.

5.3 Conceptualising diversity in news recommender systems

5.3.1 Data as a constraint on diversity-aware news recommender systems

Scholars from law and journalism studies have proposed many different ways in which recommendations can be made more diverse, including not only accounting for the viewpoints but also the topics, styles, and formats of the recommended content, as well as the characteristics of the users to whom content is recommended. There arguably is not one right way to approach diversity. As the need for diverse news is often ultimately grounded in democratic theory, the type of diversity a recommender system should promote de-

[1] https://learn.microsoft.com/en-us/azure/open-datasets/dataset-microsoft-news?tabs=azureml-opendatasets

pends on the kind of democratic system (e.g., liberal, deliberative, or agonistic democracy) it is used to support [111]. Our goal here is not to advocate for one specific approach to diversity. Nor do we argue a dataset should necessarily be suitable to realise all types of diversity we will outline below. For example, the need for diversity may be outweighed by the privacy concerns of collecting detailed data on the news consumption of marginalised groups. At this stage, we only aim to identify the characteristics of recommendations that are important from normative perspectives on diversity, so we can evaluate which aspects of diversity can and cannot be incorporated in news recommender system research with the data included in currently popular public datasets.

5.3.2 What data is needed to make diverse recommendations?

The diversity of the topics of recommended articles plays a dominant role in the literature on diversity-aware news recommender systems. Many authors highlight that by accounting for what users have seen in the past, recommender systems could expose them to topics that are new to them, and thereby broaden their horizons or alert them to topics affecting other societal groups [100, 253, 276]. Alternatively, recommender systems can show a user extensive information on a few topics to create deeply informed expert citizens, or show individuals news on the topics they prefer to engage with to support their autonomy [111, 122]. Between these extremes sit views that argue for a more mixed approach. A mix of entertaining and hard news could make diversity-aware recommender systems pleasurable to use, while a mix of political and non-political content could give individuals a fuller picture of society [22, 100, 111].

Viewpoint diversity is often addressed in conjunction with topic diversity, and concerns the diversity of perspectives on a particular topic to which users should be exposed [97]. In general, authors advocate for showing users a wide diversity of perspectives within the same topic, either to better inform them, or increase tolerance and social cohesion [17, 111, 165, 276]. The different viewpoints to which users should be exposed are usually ideological in nature; several authors explore how users should be exposed to perspectives from across the political spectrum [17, 107, 276] or to the perspectives of other societal (e.g., ethnic, linguistic, or marginalised) groups to increase tolerance [10, 165, 227]. An important but rarely explicit assumption in this context, is that the diversity of perspectives to which a recommender system should expose users may differ depending on the political perspectives or societal groups that exist in the society in which the system is deployed [281].

The style of news plays an important role in discussions on the type of public debate diversity-aware recommenders should foster. Deliberative approaches that see media as creating space for an informed and rational debate imply the recommended content

should be impartial and promote active discourse [111]. Recommender systems that seek to foster tolerance may recommend positive news about ideological opponents, or respectful discussions between opponents [100]. Finally, from an agonistic perspective the goal of a recommender should be to facilitate conflict while ensuring the different sides do not perceive each other as enemies [227].

The format in which news is recommended matters to recommender systems that respond to users' consumption preferences. If a recommender system is used to serve a wide cross-section of society, it is important to ensure news is recommended in a format (such as video, audio, or text) with which different users can engage [46, 111, 251]. The need for different formats is also implied in some of the approaches to diversity mentioned above. For example, a recommender system that aims to enable users to become experts in a specific topic requires larger background articles and deep dives for users that have a lot of pre-existing knowledge on the topic, and simple explainers for users that do not.

Finally, few authors pay explicit attention to the characteristics of the users to whom content is recommended. One exception is Vermeulen, who emphasises the need to balance diverse exposure with user choice [276]. Additionally, detailed information about the user is required to provide the diverse recommendations described above. For example, a recommender system must be able to assess what news a user knows already in order to give them a new perspective or more deeply inform them. From a fundamental rights perspective, moreover, the need for diversity is also grounded in the right to receive information, and more specifically the need to ensure that "everyone has access to a diverse range of journalistic content" [47, 76]. In that context the Council of Europe has for example referred to the need to ensure diverse news can be accessed despite socio-economic barriers and income level and that minorities, disabled persons, and disadvantaged and local communities are able to access news [46, 47]. Similarly, writing on recommender systems' impact on epistemic welfare, Hyzen et al. [122] emphasise the importance of ensuring a recommender system is able to "ensure that many and diverse users can encounter and engage with epistemically valuable content". Overall, if the goal of a recommender system is to meet the information needs of all members of society, it is also important for the recommender system to be able to effectively recommend news to different societal groups, even if they have different consumption patterns.

5.3.3 Subconclusion

Many of the features that would be needed to make diverse recommendations are hard to identify, and require additional data processing [278, 279]. Think, for example, of identifying the opinions reflected in an article text so a recommender system can show a user

different perspectives on the same topic. There are extensive lines of research into identifying different viewpoints algorithmically, usually referred to as opinion mining or viewpoint detection. However, neither currently have off-the-shelf solutions [220]. Alternatively, researchers or journalists could manually annotate the viewpoints included in an article [169]. However, manual annotation at the scale needed for training data for recommender systems is a time-consuming and costly endeavour. This problem is exacerbated by the conceptual unclarity of diversity, as labelling by non-experts is likely to yield inconsistent results. Despite, and arguably especially given the difficulty of identifying the features relevant to diversity, it is important that these aspects are included in public datasets, if they are to facilitate research developing diversity-aware recommender systems.

5.4 Data as a constraint on diversity-aware news recommender systems

In this section we analyse which datasets are used in recent research developing and testing news recommendation models, the characteristics of those datasets, and how they can(not) facilitate the different aspects of diversity outlined on page 102.

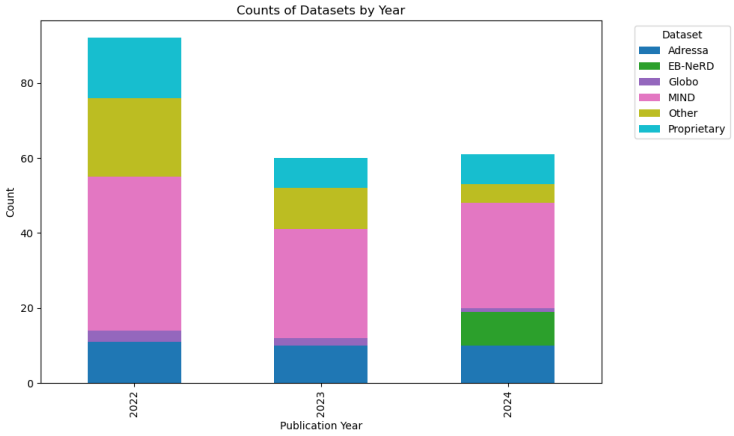


Figure 5.1: Counts of the datasets used in papers on developing news recommender systems between January 1st, 2022 and December 31, 2024. Excludes datasets cited 5 times or fewer.

5.4.1 Datasets used in current news recommendation research

To assess which public datasets are used for news recommendation research, we queried the Scopus search engine for papers with the term “news recommend*” in the title, ab-

stract or keywords, that were published between January 2022 and December 2024. This returned 383 results. We filtered out conference proceedings, papers that only mention news recommendation as a potential field of application, that are behind a paywall, or not written in English. This left us with 314 papers. Out of these 314 papers, 46 are written from a social science perspective, whereas the vast majority (268) is technical. The majority of technical papers (171) implemented and tested a news recommendation model.

Figure 5.1 provides an overview of the datasets used in these papers. It lists all datasets that are cited more than 5 times, and groups the others into the 'Other' and 'Proprietary' categories. The 'Other' category (35 papers) mostly consists of datasets such as MovieLens, which is a standard benchmark dataset for (non-news specific) recommendation tasks, and task-specific datasets such as FakeNewsNet for fake news detection. In addition, the 'Proprietary' category (32 papers) covers datasets that are not shared with the academic community. Such datasets are always used for online experiments that test recommender systems on real users, but without publication of the data these studies are not reproducible.

There are four news-specific public datasets that are frequently used to implement and test news recommendation models: MIND [293](used 98 times), Adressa [95](used 31 times), EB-NeRD [144](published for the 2024 RecSys challenge, used 9 times), and Globo[53] (used 6 times). All four datasets contain (the IDs of) the articles people clicked during a certain time period on a news website: the fully algorithmic aggregator MSN News for MIND, Norwegian Adresseavisen for Adressa, the Brazilian G1 news portal for Globo, and Danish tabloid Ekstra Bladet for EB-NeRD. Table 5.1 provides an overview of the data contained in each of the datasets.

MIND is the most prominent dataset, accounting for 57% of all papers published on developing news recommendation models. The publication of MIND in 2020 was a significant development for news recommender system research. While at first glance its scale seems comparable to that of Adressa, the interactions in MIND take place in a much shorter time period. This means that there are more interactions per published news article. The articles in MIND are written in English, which increases result interpretability for researchers. Most importantly, and contrary to Adressa and Globo, MIND also contains an 'impression log' with articles that users saw but did not click. In addition to this raw data, Microsoft published a package of news recommender algorithms that could be trained on the dataset through open source code. Microsoft also launched a competition, challenging developers to create algorithms that would outperform their own. Performance was determined by the recommender system's ability to correctly rank a list of candidate items based on the likelihood a user clicks on them.

	MIND	Adressa	Globo	EB-NeRD
#interactions	15 million	23 million	3 million	37 million
#users	1 million	3 million	314.000	1 million
#articles	160.000	48.000	46.000	125.000
time period covered	Oct 12 - Nov 22, 2019	Jan 1 - Mar 31, 2017	Oct 1 - Oct 16, 2017	Apr 27 - June 8, 2023
language	English	Norwegian	Portuguese	Danish
organization type	aggregator	local news	national news	national tabloid
article metadata	category, subCategory, title, abstract, URL, named entities in title, named entities in abstract, embeddings	title, category, URL, word count, time published, keywords, author, entities	category, time published, publisher, word count, embeddings	title, subtitle, body, category, subcategory, time published, premium or open, IDs of corresponding images, article type, topic, views, reads, sentiment, entities, embeddings
click metadata	user ID, timestamp, clicked articles, non-clicked articles	timestamp, session start/stop, read time, referrer	user ID, article ID, timestamp, session, referrer	user ID, article ID, session ID, articles viewed, articles clicked, time stamp, read time, scroll percentage, device type
user metadata	history	city, region, country, os, device	region, country, os, device	history, subscription, gender, post code, age

Table 5.1: The context and data included in the four primary datasets used for training news recommender systems. Most of the users on news websites are not logged in - unless IP address is considered it is difficult to determine the number of unique users a dataset encompasses. The URLs to the full articles in MIND and Adressa are expired (https://github.com/msnews/msnews.github.io/issues/22).

5.4.2 The constraints of current datasets from a diversity perspective

The datasets described above are important resources in news recommender system research and development - it would be very challenging for researchers to develop (reproducible) recommendation models without them. However, as we argue below, the characteristics of these datasets (specifically the kind of metadata, content and audiences they contain) pose significant limitations to the development of diversity-aware recommender systems, and make it comparatively easier to develop recommender systems optimised for engagement. Our analysis includes all datasets but focuses in more detail on MIND, as our research indicates this is the most frequently used public dataset in news recommender system research.

Metadata

Most datasets contain only very limited information about the articles that have been recommended. None of the datasets contain readily available information on the styles and formats of the articles, or the perspectives expressed in them. This makes it considerably more difficult to use these datasets to develop recommender systems that incorporate these aspects of diversity. As mentioned in the previous section, it is possible to obtain information on for example the viewpoint expressed in an article by mining the relevant content characteristics from the text of article. However, this process (already difficult in itself) is complicated by the fact that, with the exception of EB-NeRD, the datasets do not contain the full text of the articles. Because of licensing issues, Globo has not published any additional information about their articles; content is only represented through an ID, category code (e.g,.'614') and embeddings created based on all the texts. MIND and Adressa only contain an URL, not the actual content of news articles. Before researchers begin the difficult work of mining characteristics from articles' content, they must therefore verify whether they are legally allowed to scrape the content, build the tools to do so, and ascertain which URLs still function (which, in May 2025, is not the case in either of the datasets).

Adressa, EB-NeRD and MIND all do contain metadata on an article's topic, category, and title; Adressa also has information on the author of an article. This information could be used to develop recommender systems that recommend diverse topics or categories of articles to users. However, the generic nature of the metadata that is available on the articles' topics challenges the development of such recommender systems. A recommender system can only identify patterns in data that is available; as such, the granularity with which article categories are chosen significantly influences the types of patterns that can be identified. For example, all political news in MIND is put under the umbrella category 'politics'. As such, a recommender system developed with MIND could not learn to balance news covering the Democratic versus the Republican Party, let alone non-American politics, unless other metadata such as the title contains sufficient data to infer such specific information.

Content

The content in the datasets is another significant limitation on the datasets' suitability for the development of diversity-aware recommender systems. To train a recommender system to recommend diverse topics or perspectives on matters of public interest (whether to promote social cohesion, create expert citizens, or another objective of diversity-aware news recommenders), the recommended articles must actually include diverse content and perspectives. However, only 30% of the MIND's dataset consists of

articles categorised as news, with the rest fitting into categories such as 'lifestyle', 'sports', and 'food' [277]. Similar distributions can be found in EB-NeRD, where only 28.000 of the 125.000 articles belong to the 'news' category. Adressa contains roughly 60% news, but its categories are mostly focused on the geographic area the article covers, rather than its content ('nyheter|nordtrondelag'). Furthermore, the subcategories MIND uses for news indicate it is focused on the US context, with 'newsus' representing 47% of all news items, 'newspolitics' 17%, and 'newsworld' 8% [277].

This is potentially problematic; while the datasets contain large amounts of articles, they contain relatively few news articles, which are likely to include the kinds of diverse ideological perspectives and address the kinds of public interest issues that the diversity-aware recommender systems proposed in normative literature take into account. This issue is exacerbated by the popularity of MIND and its focus on US news, as a recommender system trained to recommend diverse political perspectives and topics in the US two-party system cannot be directly applied to countries with different political systems and societal groups [234]. This would in turn challenge the ability of recommender systems to facilitate the access of individuals outside the US to content that reflects the diversity of political outlooks or societal groups in the society in which they live. This issue can be partially addressed by use of the other datasets, which reflect the Brazilian, Norwegian and Danish political systems. However, the popularity of MIND indicates this solution has not yet been widely adopted.

Finally, it should be noted that EB-NeRD has some unique potential for the development of style- and format-aware recommender systems. Unlike the other datasets it contains additional information like content type (video, gallery, etc.), accompanying images, and estimated sentiment score. It could therefore be used to research recommender systems that respond to the format consumption preferences of different users (e.g., deep-dives or simple explainers) and affective styles (e.g., deliberative discussion or humanisation of political opponents). Conversely, the lack of such data in MIND, as well as the fact that MIND only contains text articles, make it more difficult to use this dataset to take the style and format of news into account when developing diversity-aware news recommenders.

Audience

There are distinct differences in the information each of the datasets includes about its users. The only information available on MIND's users is when they accessed the system, which items they have interacted with in the past, which items were presented to them in a specific session, and on which items they have clicked. This is commendable from a privacy perspective. However, it makes it impossible to assess for which audiences recommender

systems trained on MIND work, and for whom they do not. Different societal groups (e.g., age, ethnic, or socio-economic background) are likely to have different needs and preferences [272]. If a particular user's or group's consumption pattern falls outside the patterns for which a recommender system is optimised, the recommender system will be less able to determine what news they should be recommended. It is likely that MSN users are not fully representative of the different audiences for whom recommender systems are developed, if only because of MIND's focus on English-language US news. However, as MIND does not contain even aggregate data on the characteristics of its users, it is impossible to know exactly in what ways and for which groups MIND's audience is unrepresentative. The other datasets do contain information on the geographic location a click happened: Globo provides the region, Adressa the city, and EB-NeRD the postcode. This data makes it possible to assess how a recommender system performs for local communities in each country. However, it is not possible to assess how a recommender system performs for different socio-economic, disadvantaged, or minority groups. EB-NeRD is a partial exception to this, as postcode could be combined with other data to infer (for example) socio-economic status and the dataset sometimes contains information on users' gender (7%) and age group (3%).

The limited number of days in each of the datasets poses another limitation. MIND and EB-NeRD only contain interactions of a period of 6 weeks, versus two weeks for Globo and three months for Adressa. Moreover, as users of news websites are often not logged in (only 6.7% of interactions in EB-NeRD are from logged in users), it is difficult or impossible to track changing behavior over time. For example, in MIND only 20% of visits in the dataset are from users that have accessed the system five or more times, which can be explained by the fact that user IDs reset after 24 hours. As such, little of the data that is necessary to reliably observe long term changes in people's reading behavior is available. This limits research on diversity-aware recommender systems that aim to foster long term changes in the audience, for example to promote tolerance, create expert citizens, or stimulate self-development.

Furthermore, all datasets encompass a time period where, electorally, nothing significant happened. It is therefore still impossible to gauge how a model would behave during democratically important moments such as elections. Most of the time, this is a deliberate design choice. On the challenge website EB explicitly states that *"[t]his timeframe was selected to avoid major events, e.g., holidays or elections, that could trigger atypical behavior at Ekstra Bladet."* However Einarsson, Helles & Lomborg, in a prior direct collaboration project with Ekstra Bladet, noted that during the 2019 Danish election the use of news recommender system increased the amount of soft news readers consumed [70]. If diverse

recommendation algorithms are to be researched and developed, they also (perhaps even especially) need to be robust towards exceptional behavior patterns.

5.4.3 Implications for research

Taken together, the data included in the currently most-used datasets are most suitable to develop recommender systems that prioritise news based on users' immediate, short term content preferences. Much of the data needed to develop the kinds of diversity-aware recommender systems proposed in legal and journalism studies literature based on democratic principles is lacking from public datasets. In particular, the recommendation of diverse perspectives, formats, and styles is highly challenging with current datasets, as the relevant content characteristics are either not included as metadata or lacking from the recommended articles themselves. The development of recommender systems that aim to foster long term changes (for example by promoting tolerance or deeply informing individuals about specific topics they select) or are robust enough to perform well during different societally significant events (such as elections or wars) is similarly challenging, due to the short time period covered by the datasets. To a lesser extent, optimising recommender systems to perform well for different societal groups is difficult with current public datasets, as data on the individuals to whom news is recommended is often lacking from the datasets.

The datasets' suitability for creating recommender systems that optimise for engagement is exacerbated by the competitions set out by the companies that have published them. Competitions are a common way for companies to increase researchers' use of their dataset, and allow them to set out the specific performance indicator they would like researchers to optimise for [163]. In the case of MIND, Microsoft set out a competition tasking researchers to create a recommender system that ranks *"articles according to the personal interest of [the] user"* [215]. The 2024 RecSys challenge, which focused on news recommendation and relied on EB-NeRD, did aim to address *"both the technical and normative challenges inherent in the design of effective and responsible recommender systems"*. However, the main evaluation metric used to determine the winner of the competition focused on engagement: the odds that a randomly chosen clicked article is ranked higher than a randomly chosen non-clicked article (referred to in computer science literature as the 'area under the curve' [292]). While participants were encouraged to also include beyond-accuracy objectives in their challenge submission, participants were not evaluated on these metrics, and challenge submissions exclusively maximised accuracy. One of the challenge participants ([105]) noted that, through the design of the challenge the potential to optimise on diversity was limited, and that therefore the

challenge was *"a missed opportunity to move beyond 'leaderboard-chasing culture'"*.

5.5 Ensuring access to training data: the role of European law and policy

5.5.1 The legal case for public support for public datasets

If one were to construct a dataset that is better able to enable the development of diverse news recommender systems proposed in legal and journalism studies literature based on democratic principles, it would include 1) a focus on news relevant to public debate, 2) granular metadata on the content characteristics (topic, viewpoint, format, and style) that a diversity-aware recommender may need to take into account, 3) a large number of frequently returning users over a longer period of time, and 4) with sufficient metadata on these users and the societal groups they represent. Table 1 provides a more complete overview of the data needed to develop diverse recommender systems, based on the analysis in sections 3 and 4.

Aspect of diversity	Data to be included
Topics	News on the political topics about which citizens need to be informed in their democratic society; topics affecting other societal groups; topics (e.g., entertainment, fashion, climate) with which individuals want to engage.
Perspectives	The range of ideological and political perspectives in society; perspectives of societal groups (e.g., ethnic, linguistic, or marginalised) groups.
Style	Impartial and promoting active discourse; Positive and respectful about societal opponents; critical about opponents while not framing them as enemies.
Format	Video/audio/text to meet the consumption preferences of different news users; deep dives and simple explainers depending on users' pre-existing knowledge about a topic.
Users	Different groups (e.g., socio-economic groups, minorities, disabled persons, and disadvantaged and local communities) that need to be able to be informed in democratic society; data on user characteristics relevant to the above, such as the formats they prefer to consume, political leaning, topics they have engaged with.

Table 5.2: Overview of data to be included in datasets to facilitate the development of diversity-aware recommender systems proposed in legal and journalism studies literature

It is unlikely any single dataset will contain all of this data. As noted in section 3, different data is required depending on the specific kind of diversity a recommender system is expected to promote. More importantly, however, are the viewpoints and topics present in the public debate, as well as the societal groups whose perspectives are recommended and for whom recommender systems have to be optimised are different in each society and at different times. While the publication of another dataset with the characteristics above

would already be a significant improvement, it would only facilitate the development of diversity-aware recommender systems that optimise their recommendations for the specific society and the specific time in which the dataset was created.

There is little incentive for any particular media organisation to invest in the creation of a high quality dataset others could use to research diversity-aware recommender systems, much less the creation of multiple datasets that reflect the consumption patterns of different societies at different times. Nor is facilitating the development of diversity-aware recommender systems arguably the societal role of a commercial media or technology organisation. However, states do have an interest in ensuring that the different media organisations in their society can provide diverse perspectives. Under European fundamental rights law, this is moreover a legal obligation. Article 10 of the European Convention on Human Rights has long imposed a positive obligation on states to be the "ultimate guarantor" of diversity, meaning states must not only refrain from censorship, but also actively ensure that media organisations can provide (and the public can receive) diverse perspectives representing different societal groups [168, 181, 245].

States have traditionally fulfilled this obligation by ensuring the existence of different media organisations through subsidies or media concentration laws, or by creating public service media organisations. However, as scholars have argued extensively over the past decade, individuals' access to diverse information not only relies on the existence of different media organisations, but also on the design of technologies such as recommender systems that determine to what information individuals are exposed [26, 46, 109]. So far, legal researchers and policymakers have primarily addressed the media's use of technology by emphasising the media's responsibility to use technologies in line with editorial values, as well as the limits freedom of expression imposes on any regulation of the media's use of technology [48, 113, 201, 275]. However, both to ensure media organisations can actually live up to these responsibilities and to fulfill their own obligation to guarantee diversity in the media system, it is also important to consider how states can create the conditions media organisations need to use technologies like recommender systems responsibly. In particular, as we have argued in this article, the existence of a high quality dataset is an important condition for organisations to ultimately be able to recommend diverse news to their audiences.

Law and policy that supports the creation of public datasets also serves another purpose, namely to safeguard the media's independence from large technology companies. A large strand of journalism studies research has assessed how media organisations rely on large technology companies for technologies used to gather, produce, and distribute news, including news recommender systems [52, 77, 142, 207, 236], as well as the comput-

ing power, research, and indeed training data needed to develop new technologies [235–237]. This reliance challenges media organisations' ability to independently determine how the technologies they use inside newsrooms serve editorial values. Instead, they increasingly adopt the logics of the large technology companies on which they rely [52, 164, 207, 235, 269]. As our research shows, over half the research into the development of new news recommender systems since 2022 relies on a dataset published by Microsoft (MIND), and the datasets currently available are primarily suitable for developing recommender systems that optimise for engagement. As such, public support for the creation of datasets that facilitate the creation of recommender systems that promote different conceptualisations of diversity would lessen MIND's influence on the way media organisations can recommend news to their audiences.

5.5.2 Access to public datasets through European data access law and policy

EU law has significantly expanded its data access obligations in recent years. However, none of the new obligations ensure access to data that could be used to develop diversity-aware recommender systems. The Data Governance Act requires public sector bodies to facilitate the re-use of data, but article 3(2) exempts public service broadcasters and cultural institutions from this obligation. The Data Act's data access obligations are limited to the Internet of Things (Chapter II) or public authorities (Chapter V). Finally, article 40 of the Digital Services Act does grant researchers access to data necessary to assess the effects of large platforms' recommender systems on diversity. However, this data access right is limited to the data that is necessary to understand platforms' impact and risk mitigation measures. While it could be used to evaluate the diversity of the datasets used to train platforms' recommender systems along the criteria laid out on page 103, it therefore likely does not provide researchers with the data necessary to develop alternative techniques and approaches towards building new recommender systems.

EU law and policy is also beginning to facilitate voluntary data sharing. Data spaces, which allow organisations to share data in accordance with common standards, play a particularly important role in the EU's push to strengthen digital media and the European technology industry more broadly [40, 41]. The Commission has in 2022 set out a €8.000.000 grant for the development of a 'media data space' that would provide organisations with content, metadata, and audience data to develop *"data services matching European values, in particular ethics, equality and diversity"*, according to [42]. A pilot project and the European Broadcasting Union (an organisation representing Europe's public service media) both foresee that this media data space could provide access to data necessary to develop recommender systems [43, 67, 68].

In large part, however, the focus on data spaces in European (media) data policy aims to address a different issue than the one posed by the lack of data needed to develop diversity-aware recommender systems. Data spaces are useful to increase access to data currently held by separate organisations. The same is true of other mechanisms in the Data and Data Governance Act facilitating data sharing, such as data intermediaries or data altruism. However, the differences in audience behaviour caused by (for example) different recommender systems' user interfaces, algorithms, and goals, make it highly challenging to identify meaningful patterns between recommendations and user behaviour in such an aggregated dataset. Data spaces may therefore be most suitable as an overarching framework through which datasets for recommender system development can be made available. For example, they can make it easier to test recommender systems on multiple datasets by setting common standards on data structure, and provide means for other actors to enrich datasets by adding metadata concerning, for example, viewpoint diversity [43, 67, 170]. They can also expand access to the data necessary to understand audience preferences and behaviours on a more general level [68]. However, all this would not address the core issue identified in this article, namely the lack of a dataset (or datasets) with which different types of recommender systems promoting different kinds of diversity can be developed.

Though EU law does not currently ensure access to such a dataset, it also does not preempt either governmental or media initiatives to create it. In particular, public service media organisations are arguably well positioned to do so. Public service media organisations already have access to a diverse library of content that can be recommended, and several already operate recommender systems that can be used to generate datasets. Moreover, in contrast to Microsoft and the commercial media organisations that have so published the datasets used in this article, public service media organisations have a specific societal role and sometimes even legal obligation to promote citizens' ability to access diverse information [61, 275]. Public service media organisations have traditionally fulfilled this role by providing diverse content themselves. Increasingly, they have also experimented with using recommender systems to better serve their own audiences, and in this process been forced to navigate the difficulties of incorporating diversity in their recommender systems [118, 254].

However, what remains lacking, and what public service media organisations could begin to provide, is support for the more foundational research that enables media organisations throughout the media system to automate editorial decision-making in line with their editorial values. European law can only play a limited role in facilitating such support; the regulation of public service media remains primarily a national competence. How-

ever, organisations such as the European Broadcasting Union do already operate projects that aim to facilitate cooperation between public service media developing diverse recommender systems [92, 118]. Such approaches could be expanded to also generate publicly accessible datasets needed for researchers to create new kinds of diverse recommender systems. Alternatively, individual public service media organisations could publish datasets on their own initiative. In both cases, the data outlined in Table 2 may serve as a useful starting point for public service media organisations looking to publish the data needed to enable the creation of a wide variety of diversity-aware news recommender systems [251].

5.6 Conclusion

In this paper we have analysed how the lack of suitable public datasets constrains the development of diversity-aware recommender systems. We have analysed what data is needed to develop the kinds of diversity-aware recommender systems proposed in legal and journalism studies literature, and to what extent this data is available in the currently most used public datasets. We have argued current datasets are primarily useful to optimise for short-term engagement, and to a much lesser extent for the recommendation of diverse topics. The development of diverse recommender systems that take the formats, styles, and viewpoints of articles or the characteristics of different audiences into account is particularly limited by current datasets. Furthermore, we argued why European law should address this limitation, how existing data access measures fail to do so, and how public service media playing a more active role in technological development could provide a way forward.

In the short term, our analysis indicates there is potential to better use the currently available datasets to develop diversity-aware recommender systems. The majority of research in the past 3 years has relied on MIND; complementing the use of this dataset with EB-NeRD, Adressa, and Globo would allow research to optimise the performance of recommender systems not only for audiences that resemble 2019 US MSN news users, but also audiences from (specific regions in) Brazil, Denmark, and Norway. In addition, due to its inclusion of article text and wider variety of content, EB-NeRD partially allows for the development of diversity aware recommender systems that take the formats, styles, and viewpoints of articles into account by inferring the relevant characteristics from the text. However, the lack of metadata on (for example) the viewpoints represented in the articles still makes this challenging. Moreover, the limited spectrum of news content, formats, and timeframes in all datasets continue to limit the development of diversity aware recommender systems.

In the long term legal and technical research should therefore also expand its focus from assessing how the media should use technologies such as recommender systems responsibly, to analysing the conditions that need to be in place for researchers to develop and media organisations to deploy technologies that align with editorial values. At present, the public datasets used to develop diversity-aware news recommender systems are supplied by commercial technology companies (Microsoft) or commercial media companies (Adresseavisen, Ekstra Bladet, and G1). Providing the datasets needed to develop the kinds of diversity-aware recommender systems proposed in normative theory is not these companies' societal role, nor do they (given their financial goals) have a strong incentive to do so. Instead, to enable media organisations to enact the responsible use they call for, as well as to live up to their own obligation under European fundamental rights law to safeguard a diverse media system, states should take a more active role in providing the data needed to develop diversity-aware news recommender systems. Collaboration between researchers and public service media organisations has an important role to play here, as these organisations have the diverse content and domain knowledge needed to generate the necessary data, as well as the societal role to promote access to diverse news.

6

Building evaluation metrics for news recommender systems through participatory co-design sessions

Authors: Sanne Vrijenhoek, Natali Helberger, Laura Hollink and Claes de Vreese

Original title: Values to Metrics: Co-designing evaluation metrics for a personalized news recommender system

Abstract

THIS STUDY explores the co-design process of evaluation metrics for a personalized news recommender system at Ekstra Bladet, a Danish national news organization. The research addresses the gap between journalistic values as held by the newsroom and the technical performance metrics that technical teams are familiar with by involving various stakeholders from the news organization in a co-designing process. Through a series of workshops, participants from technical, editorial, business and strategy backgrounds collaboratively defined recommender system objectives, identified risks and opportunities, and designed concrete evaluation metrics that should not only be interpretable to the technical team, but also to the newsroom. These metrics were then implemented and tested on Ekstra Bladet's public dataset, EB-NeRD. The study highlights the challenges of operationalizing abstract concepts underpinning journalistic and editorial values and emphasizes the importance of iterative refinement and stakeholder collaboration in developing a recommender system that aligns with the organization's journalistic mission. The findings suggest that while technical solutions can be developed, ongoing dialogue and appropriate internal workflows are necessary to ensure the system both meets editorial standards and is technically feasible.

6.1 Introduction

The development of artificial intelligence technologies has had a large impact on journalism. Algorithms of varying levels of sophistication are increasingly supporting, and sometimes even replacing, key tasks surrounding the gathering, evaluation, composition, presentation, and distribution of news [44, 261]. Personalized news recommender systems specifically are expected to increasingly influence the news people see online by predicting what readers would like to read next. According to the 2024 'Journalism, Media, and Technology Trends and Predictions' report, 37% of media organizations thought that recommendation would become "very important" to their organization in the upcoming year [192]. However, while news recommender systems are, in essence, taking over the role of human editors, who carefully balance the interests of among others readers, journalists and the business department, it remains unclear how to make news recommendation algorithms uphold those same values [21]. Seaver [231] describes how most state-of-the-art recommender systems are merely 'traps' drawing people in, with the sole purpose of making users return to the system and clicking more. And while value-sensitive design of news recommender systems incorporating journalistic values has been garnering increased attention in recent years (see for example [107, 111, 177]), it remains relatively small and ad-hoc [12].

To achieve an 'editorially aligned' news recommender system, one needs to translate editorial principles into technical artifacts. This is, perhaps unsurprisingly, easier said than done. Møller and Thylstrup [186] note that news values are typically seen as part of a 'gut feeling', meaning that "journalists and editors find it difficult to explain why some stories are deemed important, while others are not". Technologists, on the other hand, need very clear instructions and definitions. To them, the journalistic gut feeling is a problem, also referred to as the 'journalist-technologist gap': different domains of expertise are talking about the same problem in different language [251]. Schjøtt Hansen and Hartley [229] found that "[tech teams] found it difficult to know how and when to include the journalists", and, as a result, journalists were only included in the design of the algorithms as "outsiders looking in"; similarly, Møller [189] notes "a lack of a sustained collaboration between journalists and data scientists". The conceptual unclarity, the lack of a shared vocabulary and understanding between the groups, and the fact that the involvement of journalists and editors in algorithmic design often slows down development, may cause the editorial perspective to be excluded or only included sporadically [228]. As a result, when recommender systems fail to reflect editorial norms and values, many news editors respond with hesitation and may actively resist their deployment in production [84, 94, 205, 297].

For news recommender system to reach their full potential, their design must therefore shift from being driven exclusively by technical teams to a process of *co-design* that reflects how different actors with different roles renegotiate and operationalize norms such as relevance and editorial importance [117, 189, 199]. Following from this, it is impossible to develop one 'gold standard' news recommender system that serves all media organizations' editorial goals and all audience's needs [111, 279]. The purpose of this study is to explore what the process of operationalizing the norms and values underpinning a news recommender system could look like in the context of one specific news organization. We study how different stakeholder groups can formalize together what a news recommender system needs to accomplish, which journalistic values it needs to embody, and how those can be measured in a technical system. To this end, we conducted a case study with the Danish national news organization *Ekstra Bladet*, where we encouraged participants to co-design evaluation metrics for their news recommender systems. Through the design of these metrics, our aim is to translate high-level and abstract concepts as they are understood by the newsroom into concrete code that could be implemented by the technical team.

We describe previous research into the co-design of recommender systems in Section 6.2, and justify why Ekstra Bladet is a suitable candidate for a case study in Section 6.3. Next, we describe the overall structure of the co-design process in Section 6.4. Section 6.5 describes the process and results of our case study with Ekstra Bladet, including defining the goals and values embedded in the news recommender system, the process of turning those goals and values into concrete evaluation metrics, the prototype implementation of the metrics using Ekstra Bladet's own (public) dataset, and the feedback the participants had on that prototype. We end with a critical analysis of the outcomes in Section 6.6.

6.2 Background

To formulate the steps our co-design process follows we will first provide an overview of the development of a typical news recommender system, including how this status quo fails with regards to incorporating journalistic values. Next, we discuss existing literature on co-designing recommender systems, to what extent those methods also apply in our use case, and what this study adds to the existing body of work.

6.2.1 Building and evaluating news recommender systems

The evaluation of news recommender systems has received a considerable amount of attention, but has been found to be very performance-focused. A typical technical recommender systems paper will take a static, historic dataset, comprising of articles and clicks

(for example Microsoft News in 2019 [293]) and try to predict the clicks that happened in that dataset as accurately as possible. This practice has been extensively criticized, though Said et al. [225] note that in the 15 years since the issue has been first raised, the field's status quo has hardly changed. For one, correctly predicting a click on news data from 2019 says very little about a system's ability to successfully predict a click today [80, 225]. Furthermore, there are fundamental issues in the use of a click as an indication of a successful recommendation. A reader may be tempted to click an article with an exaggerated title, but feel disappointed or cheated afterwards. A click-focused recommender system would still interpret this interaction as a success, and not account for the long-term consequences of a reader deciding to leave the system altogether [1, 285]. The newsroom, and by extension also news recommenders, thus needs to find a balance between short-term user engagement, driving clicks and revenue, and long-term 'valuable journalism' [45, 88]. Defining this balance from a journalistic perspective has received considerably less attention in the design of news recommender systems, and the research that exists is fragmented. Bauer et al. [12] identified more than 40 different 'values' present in news recommender systems research. They noted that diversity was studied comparatively much, and questioned whether that was because it was a relatively 'easy' thing to measure. However, when Vrijenhoek et al. [281] investigated the way public service media employees conceptualized diversity for their recommender systems, they also identified 35 different dimensions of the term, many of which overlap with the other values listed by Bauer et al. [12]. This is an inherent challenge to socio-technical systems; the abstract criteria underpinning them are very difficult to model, and doing so is inextricably linked to the perspective of the person doing the modeling [149, 232]. Rather than striving for standardization, Vrijenhoek et al. [281] argue that values such as diversity should be conceptualized on a case-by-case basis, allowing for news recommender systems that are capable of expressing the nuances and requirements of that particular context.

6.2.2 Co-design as an alternative design methodology for recommender systems

According to Ekstrand et al. [73], recommender systems are usually designed and implemented by engineers, researchers and designers, who will maximize financial gains without accounting for the needs of the users of the system. Instead, Ekstrand et al. [73] argue for participatory recommendation, "where designers co-design the system with the people it will affect", and state that "the field so far has few examples of how to implement this in practice". While not explicitly addressed in Ekstrand et al. [73], this line of reasoning does not only concern end users of the system but also applies to professional users whose work is shaped by the operation and presence of recommender systems, such as the newsroom

of a media organization.

Yet, while Ekstrand et al. [73] outline *where* in the co-design process stakeholders should be involved, they do not make it explicit *how* to do so in a meaningful way. According to Avila-Garzon and Bacca-Acosta [7], who studied thirty years of research in and on co-design, there is no clear methodology or even terminology to follow; terms like 'co-design' and 'co-creation' are often used interchangeably, or used as part of the broader 'participatory design' framework. Furthermore, Avila-Garzon and Bacca-Acosta [7] note that "each team establishes their own methodology based on their needs".

More structure can be found in the literature on 'Design Thinking', which outlines a cyclical process with five phases towards designing solutions: Empathize, Define, Ideate, Prototype and Test [65]. Co-design factors into this structure by going through the steps together with the people you are designing for [249]. In the 'Empathize' phase, experiences between participants are shared to create an understanding of each others' background. In 'Define' the exact problem that needs to be solved is specified, followed by the design of potential solutions in 'Ideate'. The solutions are implemented in the 'Prototype' phase, after which the participants once again give their feedback in 'Test'. The feedback acquired in this phase can be used as input to start a next design cycle, which starts again with 'Empathize'.

In news recommendation, there are a few examples of co-design approaches. Van den Bogaert et al. [268] conducted sessions with readers to identify their primary reading needs, which led to the construction of 'reading personas'. Storms et al. [250] worked with readers to design interface elements that would allow them to understand why a particular recommendation was made to them. Both studies focused on news readers, rather than the different teams within the newsroom, and led to qualitative prototypes, but no technical solutions. Einarsson and Ormen [69] built prototypes for news recommender systems dashboards, but did so based on literature, not in direct collaboration with news organizations. A true multi-stakeholder approach is found in Van Es and Nguyen [270], who studied the design of a news recommender system ex-post at the Dutch media organization DPG Media. Van Es and Nguyen [270] organized two workshop sessions with participants from the newsroom and the tech team separately. During the sessions they analyzed how the groups perceived the role and value of news recommender systems differently, and where tensions between them arose. While the participants of the study thought that news recommendation could be used to give readers a more balanced overview of the news, they also found that "technological and market influences confine journalistic values in the design of recommender systems"; not because developers don't *want* the system to uphold journalistic values, but because "internal organizational

pressures promote conformity to business logic". As a potential way forward, they suggest more structural cross-collaboration between teams. This study extends on the findings by Van Es and Nguyen [270] by investigating what such cross-collaboration could look like, through a) involving different stakeholder groups from the earliest design phase and b) encouraging participants to design concrete evaluation metrics that can be translated into technical implementations.

6.3 Setting the scene: News Recommender Systems at Ekstra Bladet

Ekstra Bladet is a Danish tabloid news organization part of the JP/Politiken group. According to the 2025 Reuters Report, it reaches almost 30% of the Danish population online on a weekly basis [195]. Despite its popularity there is a relatively low trust in the brand; only 36% of those surveyed said they trusted the outlet, compared to the 85% for public service media outlet DR News. Ekstra Bladet began experimenting with AI and news recommender systems in late 2019, with the goal of offering readers a "wider, deeper and richer news experience" [96]. Early on, they also saw the need to align these systems with their editorial values, and launched the Platform Intelligence in News (PIN) project, a collaboration with three Danish universities. The project has won multiple innovation awards.

Initial experimentation with recommender systems showed promise. Kruse et al. [143] found that personalization increased click-through rates by 80%. However, they also found that two recommender systems with comparable performance scores could show completely different recommendation patterns, and significantly reshaped the distribution of article categories presented to users. This was studied further by Einarsson et al. [70] who, in collaboration with Ekstra Bladet, investigated the effects of a production recommender system during the 2022 Danish election. This study compared users exposed to algorithmically curated content with those presented only with an editorially selected front page, assessing variation in exposure to hard versus soft news, political topics, and political actors. They found that users receiving personalized recommendations consumed more soft news than those with the editorial selection, and to a lesser extent that they saw more content reporting on right-wing parties. While Einarsson et al. [70] described the effect as "significant but small", and noted the difficulty in attributing the differences in behavior directly to the news recommender systems' logic, the study showed that, when left unchecked, there may be significant divergences between algorithmic and editorial curation.

In 2024 Ekstra Bladet hosted the RecSys Challenge, a yearly competition part of the ACM Conference on Recommender Systems where participants are challenged to build the best

recommender systems for a particular use case. As part of the challenge they published the Ekstra Bladet News Recommendation Dataset (EB-NeRD)[1]. The top challenge submissions all had comparably high predictive accuracy scores, but the challenge organizers noted that they, too, had very different recommendation patterns: for example, in one participants' submission 42% of the recommended articles were from the news category, while for another this was only 4%. The organizers noted that these effects should be studied in more detail [145].

There are multiple reasons to conduct this co-design study with Ekstra Bladet. As far as news personalization goes, they are among if not the most advanced actor among European media organizations. Furthermore, through prior collaborations such as those with Møller and Thylstrup [186] and Einarsson et al. [70] they have built a good track record of freely allowing academics to study their organization. The experiences at Ekstra Bladet offer important insights into the practical challenges of designing a news recommender system, exactly because of their unique blend of both editorial and business considerations.

6.4 Designing the workshops

The co-design workshop series consisted of four two-hour sessions, with the first three sessions taking place two weeks apart and the fourth a few months later. This setup was chosen to be able to work with the participants for an extended amount of time without overburdening their agendas. By having the sessions two weeks apart, the insights from previous weeks would still be relatively fresh in memory. Each session built on the output produced in the previous session.

6.4.1 Recruiting participants

Participants were recruited through our primary contact at Ekstra Bladet. Our goal was to get a good representation of different disciplines involved, grouped into Technical, Editorial, Business and Strategy [241]. Given the difficulty of planning longer sessions with a large group of people, we prioritized the availability of Technical and Editorial participants who had the most stakes in the development of the recommender system. Business and Strategy participants were only required for the first session, but were free to join the three later sessions as well. The participation breakdown per stakeholder group can be found in Table 6.1.

As Ekstra Bladet is part of the media group JP/Politiken, the technical team also develops solutions for news outlets Jyllands Posten and Politiken. However, we explicitly chose

[1]https://recsys.eb.dk/

	Editorial	Technical	Business	Strategy	Total
Session 1	3	2	3	2	10
Session 2	1	4	1	1	7
Session 3	3	4	1	1	9
Session 4	2	4	0	2	8

Table 6.1: Participation of the different stakeholder groups during the workshop sessions

to focus on Ekstra Bladet only, in order to increase understanding between participants. Involvement of the other organizations, which both have different journalistic missions, would have undoubtedly yielded different outcomes.

6.4.2 Workshop design

The workshops followed the Design Thinking structure explained in Section 6.2, where one workshop session roughly corresponded to one phase in the design thinking process. Throughout the sessions, we did 'sensitizing activities' [3] with the participants, which serve to prime participants towards the existence of algorithms and the norms and values that underpin them without the moderator steering the outcome. Participants were asked about existing processes and values within the organization, and to relate those to news recommender systems. In the first session, we asked the participants to think about the different goals of their recommender systems, potential risks and opportunities, and the different stakeholders that should be accounted for in recommendation ('Empathize'). During the second session, we asked them to think concretely about what specific objectives recommender systems could achieve ('Define'). Based on these, the participants thought about what should be measured in the recommendations to establish that a particular goal had or had not been achieved, which were formalized into metrics during the third session ('Ideate'). These metrics were implemented by the research team between the third and fourth sessions using EB-NeRD, Ekstra Bladet's public dataset ('Prototype'). Lastly, the participants provided their input on the constructed evaluation metrics in the fourth session ('Test'), based on which they could decide what steps need to be taken in the future.

In the first two sessions, structured input was gathered through an interactive presentation tool. As such, most quotes from these sessions cannot be attributed to specific individuals. However, we observed during the sessions that participants were not always inclined to write down their thoughts, and that most interesting observations were made during group discussions. As such, to facilitate a more in-depth examination of the discussions and proceedings, we recorded the third and fourth workshop sessions and analyzed them using inductive thematic analysis. To preserve anonymity, participants are referred

to only by numbers.

6.5 Conducting the workshops

The following sections describe the process and results of conducting the workshop.

6.5.1 Empathize

Most news organizations operate in 'knowledge silos' [60], where different departments do not talk much to each other. The main purpose of the Empathize session was for the stakeholder groups to become acquainted with the others' professional goals and expertise. As part of the 'sensitizing activities' [3], the participants discussed the journalistic mission of the organization as a whole, and how personalized news recommender systems can(not) contribute to that mission [278]: what opportunities for the organization arise with a well-developed recommender system, and what risks come with a badly developed one.

Mission

Like most news organizations, it is Ekstra Bladet's journalistic mission to *"inform and entertain"*. Given that EB is a tabloid news outlet, however, the emphasis is likely to be more on the entertainment aspect than it is for EB's sister organizations Politiken and Jyllands-Posten. Besides this, they *"challenge authority"*, *"control those in power"*, and *"give a voice to the little guy"*. As one participant put it, *"we are fast, understandable, critical and sometimes annoying"*. Ekstra Bladet is not for nuanced and detailed reporting; *"they need to hear it from us first, and if they want more, they should go to [other news outlet]."*

Opportunities and risks

As a result of being fast, Ekstra Bladet publishes a lot of content: *"Even our most frequent readers cannot read everything"*. Human editors decide which articles get priority on the front page. Recommender systems can tie into this by connecting readers to articles they have missed, either because they are important to everyone, or because they are specifically relevant to them as individuals. Doing so right is expected to have multiple positive downstream effects; a personalized newsflow can help reach new audiences, as everyone can find something for them on the website; it can increase the time users spend on the platform, and increase revenue through ad sales; and it can help stories find their intended audiences. On the flip side, recommendations, which are *"outside of editorial control"*, may

lead to a news experience that is *"not reflective of the EB brand"* or *"not aligned with strategic decisions"*, as prioritizing only content that people engaged with in the past may *"lead to an increase of soft news consumed"*, and eventually to *"bias and echo chambers"*. During the discussion, one participant noted that *"[W]e don't mind being biased as long as it's [the newsroom] who decides what way we're biased. We just want to make sure that the recommenders are aligned with what the editorial's decision would be." (P5)*[2] Another participant questioned whether any recommender system could ever measure up to the quality of the selections made by the human editor, while others noted the cost, resources and maintenance involved in building a recommender system.

6.5.2 Define

The goal of the *Define* phase is to narrow down which problems the participants are going to solve. The participants were asked to define recommender system *objectives*: high-level goals specifying what the recommender is trying to accomplish, which can later be used to inform what the recommender system should be evaluated on. The participants then voted on which objective they found most important. Each objective consisted of five elements: *what* the recommender is doing, *why* it is doing this, *where* in the news environment it can be found, and *who* interacts with it *when*. By taking this approach, we encouraged participants to not think of a recommender system as a monolith, but rather as a collection of algorithms that can be evaluated separately.

Recommender objectives

The objectives participants defined could be found on different parts of the platform. For the front page, they designed an 'Aktualitets bundle', showcasing articles covering a good mix of currently important topics. As an extension of this they designed a recommender for low-frequent readers, showing the articles that had been featured prominently on the front page, but that had since then been replaced by other articles. Another objective was to provide content to users in the form they wanted, as there was a general intuition that younger audiences preferred audiovisual content while older audiences preferred text. The 'next click' recommender, located either below an article or further down the front page, should purely maximize the number of articles seen by a reader, and tie into their specific interests and preferences. Then, there were specific recommender objectives for content that would be relevant for a very specific group of users, but that others prefer not to see. Examples of these are local news stories, crime, and the so-called 'page 9 girl',

[2] This statement was made in the context of prioritizing one type of content over another, and was not implying that it would be permissible to algorithmically discriminate against a protected group.

a glamour shot of a woman in lingerie, which is somewhat hidden on the platform but still attracts a group of loyal viewers.

Since EB wants there to be 'something for everyone', it is more likely that there are pieces of content that are generally off-putting to other parts of the audience, such as the 'page 9 girl'. There is a higher allowance for entertainment content, and a stronger focus on the commercial aspects of the organization - however, when completely left to predict clicks from historical behavior, it would also be more likely to lose touch with the 'news' part of tabloid news, as is also noted by the workshop participants. As such, much of the discussion focused on making recommendations that struck the right balance between serious news and catering to a users' interests.

6.5.3 Ideate

The goal of the 'Ideate' phase is to move from the high-level problem statements constructed in the 'Define' phase towards designing potential solutions. According to Nishal and Diakopoulos [199] the field of personalized news recommendation is "in the need for robust oversight mechanisms that allow journalists to control and verify algorithmic outputs". Differently put, the newsroom needs to be able to understand what a recommender system does, for example through a dashboard displaying recommender evaluation metrics [69], and be able to take action based on its performance. As such, we structured the 'Ideate' phase around the construction of evaluation metrics that would help participants to determine whether the objectives specified in the 'Define' phase were achieved or not.

To increase understanding of what a metric is and how it can be operationalized, the participants were introduced to standard recommender systems evaluation metrics as well as divergence-based diversity metrics [279], which use an external target distributions to calculate the quality of the recommendation. These could be used to measure a range of different aspects, related to the articles, authors, external world or even the readers themselves [281]. For each metric, the participants answered three questions: what do you intend to measure, in what context do you care about the measurement, and what value do you expect. The participants designed four different metrics: *topic distribution*, *niche interests*, *newsworthiness* and *surprise*, and based on the following discussion we add *valuable content* and *active and expired stories*.

Topic distribution aims to ensure that each recommendation is, as a list, sufficiently diverse. One participant noted that it is not a problem if on the front page considered as a whole a user sees more of one topic, such as soccer, but that each 'bundle' of recommendations should contain a good mix of the Ekstra Bladet brand. *"You should never have a page view where you only see sports, or you only see one sort of thing [...]. You just need to physically*

be in different places on the site" (P11). This was considered more important than optimizing clicks, since it might lose readers in the long term; *"They would end up with [person] having only [relationship] stuff and others having only politics stuff. [...] Then for [person] [it would] feel like Ekstra Bladet wasn't that... having a large mix, even if it could be that [person] might have more page views" (P10).*

Niche interests intends to reflect whether users can find the content that matches their specific interests; and vice versa, whether stories that are written for a very specific audience can find that audience. *"If you know that a reader is interested in a certain niche, then they might not be exposed to what is there, because it disappears from the main page quickly because it doesn't have a broad interest but to the one user. [...] I know that I have readers, but it's hard for me always to get in contact with them" (P6).* Conversely, to the participants it also meant not showing content that would drive users away. Like in the 'relationship' example above, some content may be appealing to a particular audience, but off-putting to another. *"[W]e should not recommend content that makes people leave, right, such as pushing entertainment content to a sports interested, or stuff about crime to someone not wanting crime." (P5)*

However, the participants also note the need to go beyond the users' interest, to challenge them, but also to help them discover new interests. They call this **surprise**, linked closely to the concept of serendipity, which expresses whether the user receives content they did not expect, but are positively surprised by anyway. Like diversity, serendipity is hard to define and the topic of much prior research [19, 240]. Participants used topics the user has not seen before as a proxy: *"you also want to surprise the reader, right? That's sort of a different, different concern. [...] [T]here should both be sort of a low fit with the reader history, [...] but at the same time there should also be some sort of signal that the reader would actually be interested in the story" (P5).*

Newsworthiness is the last metric explicitly designed by the participants, and expresses the journalistic quality of the items in a recommendation. This value should be high on the top of the homepage, but is permitted to go lower the further down a user scrolls. *"We talked about having some sort of metric for newsworthiness. So like this hard news, soft news, because even if it clicks very well, we don't want the soft news in the top. Even if the dog is really cute." (P11)* Related is the preference for **valuable content** (*"it was a commercial aspect here, saying content that we make more money from." (P5)*) and articles produced by Ekstra Bladet itself, rather than those purchased from a central news agency. *"If I were to choose, I would have an algorithm push content that has been created by us. And I, I wouldn't mind making [news agency] into zero and then [...] the only [news agency] would be the one we pushed as breaking news. Because people will have seen them in another place. And I think we need to*

	feature	feature type	context	rank-aware	desired value
topic distribution	topic	multicategorical	all articles published	only recommendation	low divergence
surprise	topic	multicategorical	reading history	recommendation and history	high divergence
niche interests	popularity	int	none	recommendation	more niche content (low popularity) is better
news content	hard or soft news	categorical	none	not applicable	high share of hard news at top of page
premium content	premium label	boolean	not applicable	recommendation	included but not too much
age	publication date	date	none	not applicable	lower difference between recommendation and publication date is better
page9	article type	categorical	none	not applicable	only recommend page9 content to those that have read page9 in the past

Table 6.2: Overview of the metrics designed by workshop participants

have our own DNA." (P6) "And it's boring" (P3).

Another lively discussion, which was not necessarily translated into a metric, is the notion of **active and expired stories**. During high-profile events it is possible that a lot of consecutive, short articles are published on that specific topic. A day later most of those stories will be superseded by either new events or a summary article. This distinction is natural for a human editor that is aware of current events, but not for an algorithmic system, and participants noted that in the past the recommender systems had recommended old or expired articles. However, rather than translating this into a metric, participants felt that this should be controlled by limiting what goes into a recommender system. *"We might publish 10, 20 stories on the same topic. That makes a lot of sense within those 4, 5, 8, 12 hours that that's going on. And then, the following day then, suddenly 1 little corner of it might not be as relevant as some of the others" (P6).*

When discussing political representation, one participant noted *"It's not our job, either, to sort of make that public service, oh we have to have an equal amount of time for each party" (P6).* This ties into the anti-authority and watch-dog mission of the organization[3] As such, the participants did not place a lot of importance on equal political representation, contrary to what was studied in Einarsson et al. [70].

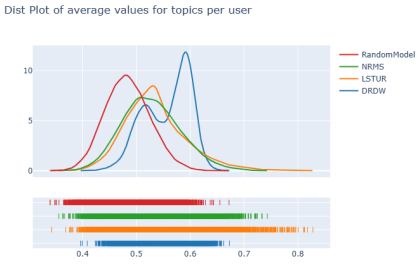
6.5.4 Prototype

After the workshop series ended, the research team took the participants' input and prototyped the evaluation metrics using Ekstra Bladet's public dataset EB-NeRD and a set of already-existing recommendation algorithms developed by other researchers.[4] The following sections explain how this translation from metrics to code was done. All formalizations are done with data that is currently available in the EB-NeRD dataset[5], with no further data processing necessary. The metric designs made by the workshop participants were

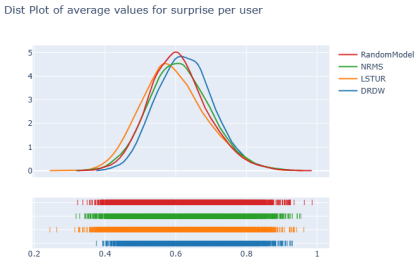
[3] Ekstra Bladet describes its core values as 'anti-authority', 'insistent', 'humorous', 'unimpressed', and 'short and accessible' (translations by the paper author), see also https://ekstrabladet.dk/om_ekstra_bladet/; it is not Ekstra Bladet's job to report on all political actors equally, but rather to report on the relevant news of the day.

[4] Ideally, the prototyping process should be done in collaboration with the organization's technical team.

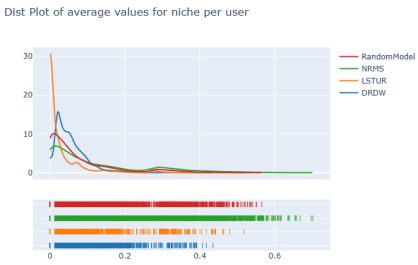
[5] EB-NeRD contains information on article title, subtitle, body (text), category, subcategory, time published, whether the article is premium or open, IDs of corresponding images, article type (text, video, etc), topic, views, reads, sentiment, entities, embeddings



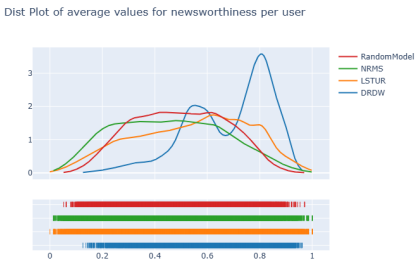
(a) Topics



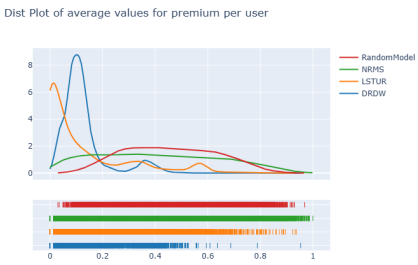
(b) Surprise



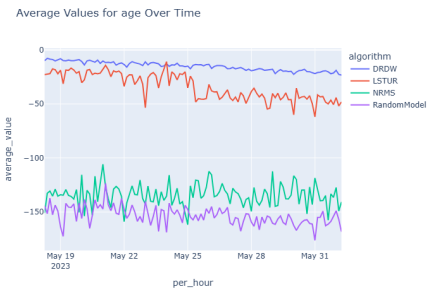
(c) Niche



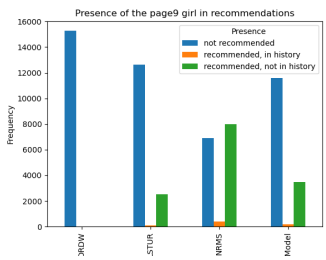
(d) Newsworthiness



(e) Premium



(f) Age



(g) Page 9 articles

Figure 6.1: Implementation of metrics inspired by the input provided by workshop participants, run on algorithms supplied in Li et al. [152] and with the EB-NeRD dataset.

often not a direct match with the data available, and in those cases the research team had to simplify concepts, or approximate them using the data that was available. We do not consider this a limitation of the study, but rather a reflection of what real-life implementation and decision-making would entail. As example algorithms we used a subset of the algorithms (D-RDW [152], LSTUR [4], NRMS [292] and Random) included in Li et al. [152][6]. By comparing the patterns of multiple recommender systems we aim to visualize to the workshop participants how the choice of recommendation algorithm influences the type of content that is shown to readers. Furthermore, the purpose of this prototype is not to draw conclusions about the performance of the different algorithms; rather, it is to test with the workshop participants whether the metrics they designed are informative and actionable. The following sections describe the implementation procedure in detail, visualized in Figure 6.1 and summarized in Table 6.2.

Metric implementation

The metric *Topic distribution* should reflect whether the recommendation is a good reflection of the Ekstra Bladet brand. While participants called the metric 'topic distribution', they talked about it in terms like 'sports' and 'entertainment'; therefore we used 'article_category' as the relevant data point as it is incorporated in EB-NeRD. We account for the rank of an article in the recommendation, which means that we count articles recommended first more strongly than those at the end. The way the participants designed the topic distribution metric is suitable to the rank-aware divergence metrics outlined in Vrijenhoek et al. [279]: the distribution of topics in the recommendation is compared to the distribution of topics on Ekstra Bladet as a whole. The lower this divergence score is, the better the recommendations represent the brand.

Surprise calls for a similar approach as 'topic distribution' described above. As mentioned in Section 6.5.3 the metric is similar to serendipity, which is difficult to operationalize. For the purposes of this paper we use a simplification of the term, and calculate the divergence of the recommendation to a users' reading history; meaning, the more different a recommendation is to what a user has seen in the part, the higher the value will be. Here we assume that the larger the divergence is, the more the user will be surprised by the content. This metric calls for more specific information than article category, and instead use EB-NeRD's 'article_topic', which is multicategorical and contains values such as 'politics', 'education' or 'city life'.

[6]https://github.com/Informfully/Experiments

Measuring the amount of *niche content* in the recommendations could be done with the exact same implementation as surprise, but with an opposite value expectations (lower divergence is better); however, we also modeled it using the popularity scores ('total_pageviews'), assigning a binary 'niche content' label to an article if it had less than the 25th percentile of all pageviews. This implementation was consciously overly simplistic, and intended to spark discussion among participants on what exactly made something a niche article.

Participants also noted that they should not recommend content that makes people leave. We created an extra metric called *page9*, where we are mostly concerned with who receives the recommendation, and the assumption that we only want to recommend page 9 content to those with an explicit interest in it. There are three scenarios: when no page 9 content is recommended; when it is recommended, but the user has no prior history of interacting with page 9 content; and when it is recommended and they do have prior history.

For *newsworthiness*, which the participants designed as the share of hard versus soft news, we mapped the topics present in 'article_category' to either soft or hard news, similar to the approach used in Einarsson et al. [70]. For example, 'celebrity', 'lifestyle' and 'entertainment' are considered soft news, while 'business', 'politics' and 'war' are considered hard news. The metric measures the percentage of hard news articles in the recommendation, counting articles recommended first more strongly than those recommended last.

Premium (valuable content) The participants explicitly mentioned they would prioritize content produced by EB itself, rather than content produced by an external agent. However, this information is not available in the dataset. To approximate what this metric could look like, we use the flag 'isPremium', meaning content that is behind a paywall. Similar to newsworthiness, 'premium' measures the share of premium articles in the recommendation.

Age (active and expired stories) The dataset does not contain explicit information on whether an article is part of a story chain or not. To approximate the notion, we calculate the average difference in days between the date of recommendation and the publication date of the articles. Since the dataset is static and no new articles are added, this number can by definition only go lower for later recommendations recorded in the dataset. This also underlines a general limitation of offline datasets; the visualization shows that often

articles from many years ago are recommended, which is in a real, online setting clearly undesirable.

6.5.5 Test

We met the participants for a fourth time to evaluate the informativeness and usability of the implemented metrics and visualizations. The participants first received a plenary introduction to the implementation process, and the rationale for each of the metrics. Then, they divided into two subgroups to first interpret the individual metrics, followed by the interpretation of algorithms over multiple metrics.

Interpreting metrics

The participants divided into groups that roughly corresponded to the groups they worked in at the end of the Ideation phase, which ensured that there was already some more shared understanding of what the metrics their groups designed were intended to accomplish. The participants were asked to pick the metrics they designed, and discuss a) whether they understood the graphs, b) whether the implementation corresponded to what they had in mind when they designed the metric, and c) whether they would be able to take action based on what was shown in the graph.

Using the provided explanations, the participants were able to interpret the graphs correctly most of the time, though they seemed to feel insecure doing so (*"I think I'm not clever enough to interpret this." (P4) "I say, I don't think I am either." (P5)*). While the divergence-based visualizations gave a more detailed overview of the behaviors of the different algorithms than for example averages or point estimates would, they were also very difficult to interpret, especially to those not used to reading graphs and evaluating algorithms. While such a metric may be something to strive towards in future iterations of the design process, it may not be suitable to facilitate initial discussion and understanding between editorial and technical teams.

In the process of doing the interpretation, a lot of additional knowledge exchange happened between the participants. For example, a participant from the technical team asked a participant from the newsroom how they chose which articles went on the front page, and how long those articles typically stay there, if at all. Most importantly, going through the metrics helped the participants formulate how the operationalization of the metrics could have been better. First of all, in order to interpret the metric, they wanted to know the full details of the operationalizations. When mapping article topic to hard and soft news, was the weather hard news? What about crime? Going further, the participants did

not think that dividing the news offer into hard and soft news was actually a good approximation of 'newsworthiness', despite them designing it as such in the previous session. Instead, a 'newsworthy' article would more likely be one of a set of topics that are important in that particular moment. The same held for 'niche interests', which was consciously implemented in an overly simplistic way using article clicks, but would also have made more sense using article topics (*"maybe the article wasn't clicked a lot because it was just bad." (P1)*). However, both what makes an article topic 'newsworthy' or 'niche' is very time and context dependent. For example, for a long time 'drones' were considered a niche topic, but with the spotting of Russian drones over Copenhagen any news regarding drones was now considered a 'breaking news' article. A recommender system would need to have this information in structured form, for example through a curated list with important versus niche topics. When asked who would construct and maintain such a list, the participants were very adamant in saying that *"in Ekstra Bladet, nobody will do that. Maybe in [other brand]" (P1)*, suggesting that the type of manual labor required for such a task did not fit the organization's way of working. The participants suggested experimenting with AI solutions instead.

This points to an interesting tension. On the one hand, the participants wanted a full understanding of the metrics, but on the other hand, they wanted AI solutions to reduce the amount of manual labor; a solution that by definition will not give them the interpretability they are looking for. In general, it was noteworthy that both groups continuously returned to article topic as the relevant unit of analysis. This is, perhaps, for journalists and editors the most natural way of discussing an article in an abstract sense, and aligns most with their daily practices. It does however point back to the 'algorithmic gut feeling' described in the introduction; intuitively the journalists and editors may know what makes an article topic niche or newsworthy, but they also cannot fully translate that in clear and implementable rules and code. We may have a first intuition on what a news recommender system could be evaluated on, but this process may need to be repeated continuously to operationalize the next abstract concept.

Interpreting algorithms

During the second part of the session, the participants related the metrics back to the recommendation objectives constructed in the 'Define' phase. They were asked to choose which of the metrics were relevant for a particular objective, and to decide which algorithm was most suited for that objective. The goal of this was not necessarily to choose an algorithm that was to be implemented, but to investigate whether the metrics and visualizations were informative enough to make such a decision.

Again, the participants made sound decisions, and seemed to interpret the visualizations correctly. However, they also again did not feel as if they were able to use them in their daily life. For example, P2 noted that *"[I]f I was to understand these and be able to take actions on them, I'd need to learn these by observing them over a long period of time, along with something next to it that I understand, which it would be like looking at the news flow or looking at the more normal site statistics."* In essence, the participants would need to be able to relate what was shown in the metrics to real effects monitored on the front page. It was suggested that the metrics could become more interpretable by splitting them out over user demographics; 'women of this age group' see primarily this, this and this topic. Furthermore, editorial needed more clarity on potential actions they could take; *"Yeah, but then I wouldn't be able to change anything if it was underperforming number-wise. But what could I do then?" (P2).* Ideally, they would want to see how different choices to the recommendation algorithm would affect what was shown to readers, perhaps if they could see changes to traffic on a small group of readers first (*"Oh, you mean like an AB-test". (P11)).* This aligns with the findings of Storms et al. [250], who found that understanding how a news recommender system works is meaningless without the option to subsequently influence it.

With regards to the design and implementation of evaluation metrics for Ekstra Bladet's news recommender systems, we can draw the following conclusions: additional technical work needs to be done on both identifying newsworthy and niche topics and active versus expired stories; proposed solutions need to be aligned to the newsroom's daily practices; and metric interpretability needs to be prioritized.

6.6 Discussion

Van Es and Nguyen [270] concluded that due to business concerns, and the need to produce results quickly, editorial norms were not incorporated in news recommender systems. In our study there was no direct need to produce results, yet it still proved difficult to properly operationalize the newsrooms' values. We identified multiple underlying reasons for this, which can in part be tackled by academic research, and in part need to be resolved by organizations themselves.

6.6.1 From Values to Metrics

We found that the operationalization of one abstract concept, such as what makes a good news recommendation, is often dependent on the operationalization of another abstract concept, such as what makes a news article 'niche' or 'newsworthy'. This may lead to con-

fusion, where it is difficult to know where to start. We found that presenting participants with something tangible to criticize, even if it was knowingly oversimplified, helped them formulate how an operationalization could be better. Repeating this process may eventually lead to a first operationalization that all parties can agree on, on the premise of it being refined again in the future.

Furthermore, we saw that even if the participants seemed to be able to interpret metrics and visualizations correctly, they often felt uncomfortable doing so. This was true for both participants from the newsroom and from the technical teams, as neither seemed to feel like they truly 'owned' the solution, or understood how it would align with their daily practices. While eventually the solutions should be owned by all teams, it may be beneficial to appoint a single person in the organization that fully understands what is there, and can function as a bridge between teams. Here, there is a clear limit to how much academia can accomplish; organizations themselves need to build the expertise to make editorially aligned recommender systems land in the organization.

6.6.2 Three initiatives, three approaches

The metrics designed by the workshop participants are quite different from the ones considered in Einarsson et al. [70], and indeed also the ones considered by Ekstra Bladet themselves during the RecSys challenge [145]. Though both the workshop participants and Einarsson et al. [70] considered the distribution of hard and soft news, Einarsson compared the recommendations to the selection made by human editors, while the workshop participants differentiated between recommendations made high on the homepage and those lower or on article pages. Einarsson considered exposure to political actors, and the tone used when talking about those actors, while some of the workshop participants explicitly stated that this was not their responsibility. On the other hand of the spectrum, the EB RecSys challenge, described in more detail in Section 6.3, only considered the predictive power of the submitted models. In comparison, the workshop participants considered reading time to be of as much value as the actual clicks, and were acutely aware of the tradeoff between short-term and long-term satisfaction. In essence, three attempts at evaluating a recommender system within the same organization ended up with completely different approaches and outcomes. Multiple participants noted that this workshop series was for them the first time they spoke to colleagues from 'the other' department. Perhaps when initiatives like these are organized more frequently, opinions and perspectives would also converge.

6.6.3 Incorporating editorial values vs developing a recommender system that benefits society

Throughout the sessions, participants largely stuck to familiar concepts, such as giving users what is relevant to them, and ensuring that they do not only receive recommendations from one particular topic. This underlines the fact that designing a recommender system that reflects editorial choices does not necessarily make it a recommender system that benefits society. There might be a limit to what we can expect organizations to come up with themselves with regards to monitoring the societal consequences of their technologies, especially in a commercial setting that is very dependent on ad revenues for survival. Yet, allowing the newsrooms to incorporate their values into a news recommender system is an essential first step in deciding what needs to be monitored, and how.

6.6.4 Limitations and Future Work

Ekstra Bladet is one organization, operating in a very specific context. Results of the study can serve as inspiration, but cannot be generalized to other organizations and contexts. While we took care to invite as many different perspectives as we could, the invite-based nature of participant selection may have caused us to miss alternative views, even within the focused Ekstra Bladet context.

Furthermore, the session focused exclusively on professionals as expert users, and did not involve citizen participation. News editors are trained to know what their users want to know at the aggregated level, but cannot accommodate the needs of individuals [187]. Furthermore, Ekstrand et al. [73] explicitly calls to involve the end-users of the recommender systems, whose goals and requirements may differ from those of the people that have built the system. This workshop has yielded initial pointers of what recommender systems *could* be developed, which in future sessions can be used to have targeted sessions with readers. Similar to what happened with the workshop participants, we expect that being presented with concrete and tangible examples and visualizations will support readers in verbalizing how their needs are similar or different from what has been developed thus far.

While the plenary parts of the sessions took place in English, participants typically discussed among themselves in their native tongue, Danish, which the session moderator does not speak. It is possible (and very likely) that important nuances and interesting thoughts have been lost, either in translation or between private and plenary discussion. While participants were asked to write down their thoughts, this often did not happen. One potential remedy for future iterations of the workshop would be to ensure that a few

native speaker moderators join the sessions.

6.7 Conclusion

The co-designing sessions underlined that designing news recommender systems that center editorial values is not a straightforward process. It does not only require the formalization of the goal of the recommender system, but also of what needs to be measured in order to determine whether that goal has been reached; often, this requires formalizing abstract concepts, which are again dependent on the operationalization of other abstract concepts. Truly making evaluation metrics actionable also requires additional effort to align what is visualized with the daily practices of the newsroom. A successful implementation thus also requires realistic expectations of what can be achieved in the short- and long term, and an ongoing collaboration between tech teams and the newsroom, where solutions are proposed, implemented, and refined over time.

Acknowledgments

We are truly grateful to Ekstra Bladet for agreeing to work with us; to the workshop participants, who put in a lot of work and energy during the sessions; and especially to Kasper Lindskow, for his efforts in making this research possible.

7 Conclusion

The following sections describe the main findings of the thesis, structured by the research questions as laid out in Chapter 1.

7.1 How is normative diversity defined and understood in both theory and practice? (RQ1)

Computer scientists often aim for standardization: for example through shared libraries, public datasets, or benchmark algorithms. Through standardization, models and algorithms become comparable, and we can say with more certainty that one algorithm performs better than another. Standardization also makes adoption into production environments more straightforward, as developers can import well-established implementations directly from shared libraries, rather than building systems from scratch. Throughout this thesis, however, I argue that the practice of standardization is inherently at odds with diversity as a concept. Evaluating a news recommender system on its normative diversity requires that we leave room for the specific context and requirements of the organization that builds the system. Rather than aiming for a one-size-fits-all solution, we need to learn how to effectively communicate what it is the system is doing, and what it should be doing in the future.

The interviews conducted with practitioners from public service media organizations in the Netherlands in Chapter 2 identified very diverse interpretations of diversity. In response to the first research question, we created a mapping of the way the interview par-

Table 7.1: Overview of the different models and expected value ranges for each metric, as included in Chapter 3.

	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

ticipants talked about diversity when constructing a recommendation based on a set of candidate items from their own catalogue. The mapping consists of two components: the *aspects* that were considered for diversification, and the *tactics* that were employed to diversify the recommendation. The wide range of different aspects and tactics illustrates that there was no consensus on a definition among the interview participants; each of the participants considered different aspects and tactics while constructing a diverse recommendation.

The lack of a singular definition of diversity is in line with the normative theory described in Chapter 3. Building on Helberger [111], this chapter explains how recommender systems pursuing different democratic models would implement diversity differently. None of the models described in the chapter is inherently better or worse than another. They can, however, be at odds with each other, which means that they cannot be satisfied simultaneously. For example, one cannot give a fully neutral overview of the news, while also giving a platform to underrepresented voices. As a result, choices and tradeoffs need to be made on what behavior a news recommender system should exhibit. How this choice is made is dependent on a news organization's mission and goals, and thus a decision that can only be made internally. To build a bridge between the abstract concepts underpinning the democratic models and technology design, the chapter proposes a set of diversity evaluation metrics: Calibration (to what extent is the recommendation tailored to a reader's preferences), Fragmentation (to what extent are readers aware of the same issues and events), Activation (is the tone of the recommendation affective or neutral), Representation (which viewpoints are included in the recommendation) and Alternative Voices (to what extent are minority groups included in the recommendation). Different combinations and expected value ranges of metrics reflect the different democratic models, which are summarized in Table 7.1. Depending on the recommender system's goal, different metrics would apply. Through their foundation in democratic theory the metrics give direction to the conversation: by answering *why* we care about diversity, we define *what* the recommender system should do.

7.1.1 Combining the metrics and the mapping

Rather than being prescriptive or set in stone, the metrics and models can serve as conversation starters within an organization, and give inspiration for the type of diversity that could be implemented. The mapping can be used to adapt the metrics to the unique context of the organization, or to identify blind spots that are not covered yet. For example, the interview participants frequently mentioned that they wanted the reader to recognize themselves in the content recommended, or in contrast, to help them encounter perspectives they were not yet familiar with. This is related to the Calibration metric: allowing a reader to recognize themselves corresponds to a high degree of Calibration, whereas broadening their views would correspond to a low degree. However, whereas Calibration as it was proposed in Chapter 3 only focuses on the dimensions of topic preference and language complexity, the interviews showed a much wider range of user aspects along which a reader could potentially want to recognize themselves, such as ethnicity, age, political opinion, or other aspects related to social identity. In this case, the mapping can be used to finetune the metrics, or in case structural blind spots are identified, to formulate new ones. The process of adapting a metric to a specific context is described in more detail as part of answering RQ2.

7.1.2 Technical solutions to non-technical problems

Many of the aspects of diversity mentioned in both Chapter 2 and 3 pertain to very sensitive personal data, such as cultural background, gender identity or sexual orientation. In such cases allowing the reader to recognize themselves in the recommendation would require that a news recommendation algorithm somehow predicts or processes this type of data about its readers, which is undesirable at best and unlawful at worst [287]. While there are privacy-preserving methods towards news recommendation [239], employing these would also make it impossible to do well-intentioned recommendation based on sensitive attributes. However, perhaps this problem is much better solved through non-technical solutions. For example, Naudts et al. [191] note that Inclusivity and Political Equality should be warranted by, among others, giving communities meaningful choice between different algorithms, and that communities should be "included, represented, and have one's voice heard during the ideation, design, deployment, and evaluation of recommender systems", with specific attention to marginalized and otherwise vulnerable communities [191]. I reflect more on the notion of meaningful choice and control over recommendations in the Limitations section.

7.2 How can normative notions of diversity be translated into concrete and informative evaluation metrics for news recommender systems? (RQ2)

The diversity mapping from Chapter 2 and the metrics proposed in Chapter 3 can be reconciled into a base mathematical formalization as a rank-aware divergence score [247]. Chapter 4 specifies this formalization as follows:

$$D^*_{f_{\mathrm{JS}}}(P, Q) = \sum_x Q^*(x) f_{\mathrm{JS}} \left(\frac{P^*(x)}{Q^*(x)} \right), \tag{7.1}$$

Here, $Q^*(x)$ denotes the distribution of aspect x in the recommendation, potentially accounting for the rank of an item, and $P^*(x)$ the same in the context distribution the recommendation is compared to. A score of 0 implies that the recommendation and the context exhibit the same distribution, while a score of 1 implies they are fully disjoint. We argue that the Jensen-Shannon divergence (f_{JS}) has desirable properties, such as it being symmetrical and bound between 0 and 1.

With this base formalization, operationalizing a diversity metric comes down to answering the following four questions:

- What aspect (x) should be measured?
- What target distribution (P) should the recommendation be compared to?
- What value ($D^*_{f_{\mathrm{JS}}}(P, Q)$) do we expect from the divergence between the recommendation and the target distribution?
- Should the divergence score account for the rank (*) of an article in the recommendation or not?

Using the Representation metric from Chapter 3 as an example, the aspect that needs to be measured is the mentions of political actors and parties. We aim for the recommendation to be representative of society, so we choose to compare the recommendation to the distribution of political parties in government. We also want to account for the fact that articles that are recommended earlier in the recommendation are more likely to be seen by the reader, so we make the metric rank-aware. And since we want our recommendation to be similar to the external distribution, we aim for a low divergence score.

Using a base formalization makes the metrics easier to interpret: however the metric is set up, a score 0 means that the recommendation and the target distribution are the same,

whereas 1 means they are fully disjoint. The flexibility of the metric also means that it can accommodate alternative diversity aspects and diversification tactics as described in RQ1. The operationalization thus provides a common frame of reference, without imposing restrictions on the type of diversity that can be expressed.

7.2.1 Feasibility in practice

We did a first test of the usability of the metrics with practitioners from Danish tabloid Ekstra Bladet (EB), which is described in Chapter 6. Through participatory co-designing sessions involving EB's technical, editorial and business teams the participants designed a set of evaluation metrics for their news recommender systems. Sometimes simpler metrics sufficed, for example when the participants wanted to know the percentage of news content in a recommendation. Other times the divergence based metrics proved a good fit, for example when the participants wanted to ensure that the recommendations contained a good balance of the different topics that Ekstra Bladet as a whole reported on, or when they wanted to know to what extent the recommendation catered to a user's niche interests. However, many of the participants found the metrics hard to read and interpret, especially participants from editorial teams that were not used to doing so. Additional work is necessary to understand whether the metrics are too complex to be usable, or whether this is also a matter of building the right skills. Furthermore, during the workshop sessions it became clear that diversity as a concept is often again reliant on the definition of other abstract concepts. For example, one of the metrics the participants designed would measure the 'newsworthiness' of the articles in the recommendation, another whether they were 'niche' articles', and another whether the recommendation was 'surprising' to the reader. These are again abstract concepts, and therefore not straightforward to define, and even less straightforward to implement. This shows that defining the values of a recommender system cannot be a one-time effort, but rather needs to be an ongoing cycle of ideation and refinement, which is further reflected on in RQ3.

	Aspect (x)	**Context (P)**
Calibration (topics)	Article topic	Reading history
Calibration (complexity)	Article complexity	Reading history
Fragmentation	Recommended news stories	Other users
Activation	Activation scores, e.g. approximated by the absolute value of a sentiment score	Available articles
Representation	The presence of political actors	Available articles
Alternative Voices	The presence of minority voices versus majority voices	Available articles

Table 7.2: Operationalization of Chapter 2's diversity metrics using Chapter 3's divergence-based formalization. Note that the expected value ranges are dependent on which democratic model is followed; see Table 7.1.

7.2.2 On meaningful principles and implementations

The code that was used to demonstrate the diversity metrics on the Microsoft News Dataset (MIND) is made available on GitHub. Furthermore, Heitz et al. [108] implemented this code as part of their recommendation framework Informfully, which should also facilitate implementation. However, with this comes the risk that recommender system developers will import and run the code blindly, without understanding what the metrics are doing, without building institutional support, and without establishing concrete actions and consequences based on metric behavior. This ties into general concerns regarding 'AI ethics', which the normative diversity metrics are related to. Munn [184] argues that as AI ethics can be interpreted in various ways, the field allows organizations to define their ethics in a manner that suits their interests. As a result, the ethics principles become 'meaningless': the organizations can claim that they act ethically, but in practice the ethics do not effectuate any change to existing practices and underlying issues. Ethics become at best a checkbox to tick, or at worst a distraction that companies can use to avoid regulation. It is apparent that this criticism also applies to the divergence-based diversity metrics: their flexibility and adaptability risk being gamified to exhibit behavior favorable to the organization. Further research is necessary to create guardrails and proper procedures to mitigate this. However, I would also argue that, when done well, the divergence-based metrics have the potential to accomplish the opposite of what Munn [184] cautions for. The steps towards defining a metric force an organization to be explicit about what they care about, and specific in how they operationalize and measure it. As a result, organizations can no longer hide behind vague and abstract concepts. A description of how the metrics are operationalized becomes information that can be scrutinized by external parties, even without direct access to the platform's code.

7.3 What conditions need to be in place in order for diversity metrics to be adopted in practice? (RQ3)

Our interviews with practitioners showed that news organizations usually build their news recommender systems based on concepts and algorithms developed in academia and by other, larger media organizations. However, not many news organizations share their code, and there is a large gap between technical academic research on recommender systems and the real challenges as experienced by news organizations. Predominantly, technical recommender systems research focuses on maximizing the predictive power of recommender systems on static, offline datasets [105], which says little about how

the systems will perform in practice [12, 145]. Throughout this thesis I found, however, that news organizations do not solely need bigger and more complex models, but also lightweight and adaptable ones, that could be explained to editorial teams. On their end, editorial teams need to be able to exert control over the recommender systems, so that they can ensure that the recommendations align with their principles and guidelines [263]. Finally, business and strategy teams need to know whether it is worth freeing up time and resources to work on diverse news recommendation; either because there will be a sufficient return on investment, or because outside pressure compels them to do so. In short, technical academic research can facilitate news recommender systems reaching their full potential by broadening the scope of research that is executed, away from accuracy maximization, and towards research that allows for incorporating editorial values and editorial control.

7.3.1 Academic research to fill knowledge gaps

There is a reason that much of the technical academic research prioritizes optimizing clicks; it is one of the things that is straightforward to do. In contrast, research into diversity-aware recommendation is only possible to a limited extent. Researchers that do not have access to proprietary data from industry partners have to work with public datasets. For news recommendation, there are only a few public dataset suitable. In Chapter 5 we analyzed the ones that are most frequently used: MIND, Adressa, Globo and EB-NeRD. All these datasets are produced by commercial organizations, with a strong focus on entertainment and soft news. As such they are also primarily constructed for the purpose of creating technology that sells more clicks. The timespan they cover is limited to a few weeks, and never includes significant evens such as national elections. These factors make it difficult to conduct research on news recommendation technology that incorporates editorial values, or even to monitor for potential societal consequences. At the same time, it is possible to do more with the existing datasets than is currently done, as each of them individually has unique characteristics that are suitable to solving a subset of the diversification problem. For example, the MIND dataset contains data from different outlets, Adressa covers a local context, and EB-NeRD includes comparatively much additional information on both users and articles. However, if we really want to make progress on developing diversity-aware news recommender systems, a paradigm shift is necessary, away from static, one-time datasets and towards close and continuous collaboration with news organizations, including data-sharing.

7.3.2 The broader institutional context

The lack of a straightforward definition of diversity means that it is not possible to develop a context-agnostic off-the-shelf and open source recommendation algorithm that 'optimizes for diversity'. Because of this, a lot more work is required from news organizations themselves, both in terms of formalizing diversity, and on the actual implementation of the algorithm. It also makes it especially important that the different departments of the organization are involved in development from the very beginning. Neglecting to do so will lead to the development of a product that only covers a subset of perspectives, and make it more likely that, eventually, the newsroom will block the technology from being taken into production. However, involving multiple departments in development is easier said than done. For some of the workshop participants in Chapter 6, participating in said workshop was the first time they spoke to their colleagues from 'other' departments. Furthermore, the fact that defining diversity was dependent on the definition of other abstract concepts raised perhaps more questions than the workshop solved. When this happens, there is a risk that the confusion will make it tempting to hide behind the 'black box' of 'the algorithm' or to abandon the effort altogether, rather than making explicit design choices. This is especially likely to happen given the high-profile nature of both the news and diversity: doing it 'wrong' may lead to criticism and reputation damage.

For diversity-focused news recommender systems to land within a news organization, it is therefore crucial that appropriate internal workflows and procedures are put in place. Experimentation should be possible and encouraged, while being realistic about where the technology falls short, and without the need to oversell what it accomplishes. Preferably, the newsroom needs to do their development internally, as it is unlikely that third-party software allows for the nuances of their specific context. Most of all, there needs to be sustained collaboration between the newsroom, tech teams and academia; not as a one-time effort, but as an ongoing process, where solutions are continuously conceived, tested, and improved. Over time, the different domains will build a shared vocabulary and understanding of the problem, making it possible to leverage each other's skill sets and creativity.

7.4 Limitations

There are a few limitations to this thesis that should be addressed in future work.

7.4.1 Editorial versus user perspectives

While this thesis focuses on increasing the diversity of news recommendations to inform readers, it includes only to a limited extent the perspective of readers themselves. This is not in line with the call of Ekstrand et al. [73] that the design of recommender systems should be done with, and not only for, the people eventually affected by their use. It is of crucial importance that future work does involve readers themselves. However, in my perspective, it is necessary to first model the values as held by the editorial room, before we can work with readers in a meaningful way.

When answering RQ1, I argued that there are many different dimensions to diversity, and that it can potentially mean something else to different people. While this thesis provides a first step towards resolving this, current state-of-the-art recommender systems cannot yet meaningfully incorporate one, let alone all different dimensions of diversity. This means that choices need to be made, and priorities need to be set on what needs to be built into a recommendation algorithm first. News editors, who are professionally trained to balance the news in a way that reflects their outlet's brand, are much better equipped to oversee the potential consequences of prioritizing such editorial decisions than readers. For example, multiple studies have noted that it was difficult for readers to distinguish between diversified and non-diversified recommendations [107, 159, 221]. They also tend to overestimate the amount of news they consume [213]. Furthermore, even when given the option to control or influence a recommender system, readers hardly used the feature [157, 221]. As such, we cannot put the responsibility for constructing a diverse news diet primarily with readers themselves. News organizations first need to take on the responsibility of defining the space in which news recommender systems can operate. When that is achieved, and the methods and frameworks to adapting recommendations are put in place, we can use those same methods and frameworks to conduct those discussions with readers. I explain one potential approach to this in the Future research directions, under section 5.3.

7.4.2 The Western-European context

This thesis builds on Western European values and perspectives on notions such as public service media and the role of news in society. These notions may not correspond to those held in other areas, including the Global South, the United States, or even other parts of Europe. The thesis aims to facilitate the process of translating normative values into technology design; even if those values are not necessarily the same, it is my hope that the process can be replicated in other contexts.

7.4.3 Effectiveness of recommender systems and diversity metrics

The metrics as currently proposed are theoretical. As of yet, it is unknown how effective they would actually be in influencing user choice and reading behavior. Mattis et al. [171] experienced this first-hand when testing different variations of the Alternative Voices metric, finding at best small effects on study participants' support towards pro-democratic goals such as tolerance and political participation. However, they also noted the difficulty of appropriately modeling the construct of an Alternative Voice in their study. Future work needs to pay explicit attention to construct reliability- and validity [125], how the metrics align with human perceptions, and their real-world effects and consequences. Furthermore, the proposed evaluation metrics measure whether the recommendation algorithm skews the distribution of a recommendation in an undesirable way compared to some chosen context distribution. A recommender system can, however, never be more diverse than the pool of items it makes it selection from: if the news organization does not produce diverse content, the recommender will not recommend diverse content either. Most news organizations have editorial standards and guidelines in place to monitor the diversity of content produced; however, it also means that implementing the diversity metrics without accounting for contextual factors may provide a distorted view of reality.

7.5 Future research directions

This thesis has primarily focused on the theoretical foundation of normative diversity metrics. Before the metrics can truly be implemented in a production environment, there are at least three strands of work that need to be expanded upon.

7.5.1 Reranking for diversity

The diversity metrics developed in the Chapter 4 are designed for the *evaluation* of a recommender system, and have not yet been implemented in the recommendation algorithm itself. Besides the diversity of a recommendation, such an algorithm should also be able to account for the predicted relevance score of an article, and other editorial values such as novelty or breaking news. Furthermore, such an algorithm needs to provide some guarantee of optimal results over local minima, and be easy to adapt: changes should be instantaneous and not require a full recommendation model to be retrained [106]. One approach would be to follow up on the post-processing method proposed by Steck [247], who introduced the Calibration metric that the divergence-based diversity metrics are inspired by. This post-processing method involves a greedy reranking algorithm that, by consecutively

selecting the next "best" item, ensures that a recommendation approximates a given distribution. However, additional work is necessary to carefully scrutinize the effects of such an optimization algorithm in the context of news recommendation specifically; for example, it is undesirable that a reranker always puts an item from the same category in the first spot of the recommendation, which is likely to happen with a greedy reranker. Following from this, more thought needs to be put in what the ideal recommendation distribution would be; if every recommendation follows the same distribution, they may quickly become monotonous or counter what they aim to achieve. For example, imagine a reranker that aims to make the recommendation proportionally reflective of the political voices in society, but as a consequence always puts an article about the largest political party in the first place of the recommendation. By extension, more thought needs to be put into the desirability of nudging readers towards more diverse reading behavior or not [172], and Goodhart's bias, which states that "when a measure becomes a target, it ceases to be a good measure" [116]. Kruse et al. [146] and Li et al. [152] have already taken first steps towards diversity-aware optimization, the first through a method built around internal label shift, and the second through a random walk algorithm.

7.5.2 (Generative) AI for metadata augmentation

Many researchers have started looking into the use of Large Language Models (LLMs) to facilitate diverse news recommendation. Based on the findings of this thesis, and research on LLMs in news recommendation and other domains, I conclude that while LLMs could be useful in resolving some of the issues currently hindering progress on diverse news recommendation, using them should be done with care. One of the major problems in incorporating editorial values in a news recommender system is that it usually requires nuanced and sophisticated metadata. Often, these metadata are equally as difficult to define as diversity itself; for example, how to identify minority voices [171, 258], viewpoints [64], or neutrality [246] are all active areas of research on their own. In an ideal world, this metadata would be supplied by the journalist writing an article. However, most journalists simply do not have the time for this. Here, language models could potentially facilitate acquiring this data. Simultaneously, we should be aware of the shortcomings and pitfalls of these models. LLMs have been shown not to consistently uphold cultural values [138] and to pose a risk of propagating stereotypes [243]. Carefully executed research could investigate the suitability of both simple and sophisticated models to identify relevant metadata. This process should provide an honest overview of where (generative) AI models can be of value, and where the gaps in their performance lie.

7.5.3 A dashboard for newsrooms

The workshops conducted with practitioners (Chapter 6) showed that the newsroom participants needed more tangible examples of what the impacts were of different recommender configurations. This could be accomplished by building a dashboard [69] or sandbox environment where practitioners, and later also readers themselves, can experiment and see the effects of their actions instantly. In this environment it should for example be possible to adjust the importance of the different components of the recommender system; what happens when we assign more or less weight to personal relevance, or to a diversity metric, or to the amount of breaking news that is shown. The dashboard should show the recommendations that are issued to a selection of example users (when conducted with professional users), or to a user themselves (when conducted with readers). Central to the study should be how the participants perceive the quality of the recommendations, the control they have over them, and most importantly, whether they understand why the algorithm behaves the way it does [221]. In an ideal situation, building a system like this should allow the newsroom control over the recommendations without the need to go through technical teams, while it allows readers to tune the recommendation algorithm towards their specific preferences.

7.6 Conclusion

Throughout the thesis, a theme that consistently comes up is that incorporating normative diversity into news recommendation is not merely a technical problem, but needs the involvement of multiple academic disciplines and industry practitioners. Doing this in an academic context, however, is often not straightforward. Reviewers and funding bodies often prioritize furthering the state-of-the-art in already-established disciplines, and do not always see the added value another discipline brings to the discussion. Furthermore, it takes time for two people with different backgrounds, be that in terms of discipline or academia and industry, to understand each other, to align on their terminology, and to hone in on which parts of the problem they can solve. This time is often not available to practitioners, who need to dedicate most of their time to the daily operations of their organizations, but also not for academics who need to show results within a set timeline. It is my impression thus far, however, that there are many individuals, venues and institutions working on recommender systems that welcome research bridging the gap between technology and other disciplines. The ACM Conference on Recommender Systems (RecSys), for example, has a long tradition of industry participation, and has also accepted less technical work into the main track. Many of its satellite workshops focus on the societal im-

pact of recommender systems. However, bar a few exceptions, it is unlikely that technical and non-technical academics will attend the same conferences. This limits the possibility to organically meet, or serendipitously find inspiration in approaches common in other disciplines. To facilitate interdisciplinary research, we need to create more space for researchers to meet, collaborate, share their work, and have it recognized as an academic contribution focused on a specific domain rather than a specific discipline. As Trilling [264] puts it, "first the question, then the method" (translated).

In that regard this thesis heavily relies on the work of Helberger [111], who concisely laid out how democratic theory would apply to news recommendation; not to criticize the field as it currently is, but to paint a picture on how to move it forward. However, moving from theory to implementation can be a rather painful process. It requires that theories that have been refined, criticized, and refined again over decades are simplified and rid of nuance to the point where they can be translated into numbers. Lin and Lewis [154] note the discrepancy between 'ideal theory' and 'actual performance': it is fundamentally unlikely that theoretical concepts can be perfectly captured in technical applications. However, in the context of news recommendation specifically Helberger et al. [114] note that "practical implementation may stay far behind the normative ideal [...]. Even so, the result would be a recommendation algorithm that at least aspires to be more inclusive and diverse." Rather than being paralyzed by the notion of building the perfect recommender system (or technology in general), we should aim for consciously imperfect solutions, the limitations of which are understood and accepted by everyone, and that can be improved upon in the future. All in all, we need to look for workable simplifications, rather than reductive generalizations.

Bibliography

[1] Arpit Agarwal, Nicolas Usunier, Alessandro Lazaric, and Maximilian Nickel. System-2 Recommenders: Disentangling Utility and Engagement in Recommendation Systems via Temporal Point-Processes. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1763–1773, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3659004. URL https://dl.acm.org/doi/10.1145/3630106.3659004.

[2] Abeer AlDayel and Walid Magdy. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597, 2021.

[3] Oscar Alvarado, Elias Storms, David Geerts, and Katrien Verbert. Foregrounding Algorithms: Preparing Users for Co-design with Sensitizing Activities. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, NordiCHI '20, pages 1–7, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-7579-5. doi: 10.1145/3419249.3421237. URL https://dl.acm.org/doi/10.1145/3419249.3421237.

[4] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. Neural News Recommendation with Long- and Short-term User Representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 336–345, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1033. URL https://www.aclweb.org/anthology/P19-1033.

[5] Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of The Web Conference 2020*, WWW '20, page 2155–2165, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380281. URL https://doi.org/10.1145/3366423.3380281.

[6] Jonathan Hendrickx Annelien Smets and Pieter Ballon. We're in this together: A multi-stakeholder approach for news recommenders. *Digital Journalism*, 10(10):1813–1831, 2022. doi: 10.1080/21670811.2021.2024079. URL https://doi.org/10.1080/21670811.2021.2024079.

[7] Cecilia Avila-Garzon and Jorge Bacca-Acosta. Thirty Years of Research and Methodologies in Value Co-Creation and Co-Design. *Sustainability*, 16(6):2360, January 2024. ISSN 2071-1050. doi: 10.3390/su16062360. URL https://www.mdpi.com/2071-1050/16/6/2360.

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

[9] Edwin C. Baker. *Media Concentration and Democracy: Why Ownership Matters*. New York: Cambridge University Press, 1998.

[10] Mariella Bastian, Mykola Makhortykh, and Tom Dobber. News personalization for peace: how algorithmic recommendations can impact conflict coverage. *International Journal of Conflict Management*, 30(3):309–328, June 2019. ISSN 1044-4068, 1044-4068. doi: 10.1108/IJCMA-02-2019-0032. URL http://www.emerald.com/ijcma/article/30/3/309-328/117967.

[11] Mariella Bastian, Natali Helberger, and Mykola Makhortykh. Safeguarding the Journalistic DNA: Attitudes towards the Role of Professional Values in Algorithmic News Recommender Designs. *Digital Journalism*, 9 (6):835–863, July 2021. ISSN 2167-0811. doi: 10.1080/21670811.2021.1912622. URL https://doi.org/10.1080/21670811.2021.1912622. _eprint: https://doi.org/10.1080/21670811.2021.1912622.

[12] Christine Bauer, Chandni Bagchi, Olusanmi A. Hundogan, and Karin Van Es. Where Are the Values? A Systematic Literature Review on News Recommender Systems. *ACM Transactions on Recommender Systems*,

2(3):1–40, September 2024. ISSN 2770-6699. doi: 10.1145/3654805. URL https://dl.acm.org/doi/10.1145/3654805.

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

[14] Joeran Beel and Haley Dixon. The 'Unreasonable' Effectiveness of Graphical User Interfaces for Recommender Systems. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '21, pages 22–28, New York, NY, USA, June 2021. Association for Computing Machinery. ISBN 978-1-4503-8367-7. doi: 10.1145/3450614.3461682. URL https://dl.acm.org/doi/10.1145/3450614.3461682.

[15] Joeran Beel, Marcel Genzmehr, Stefan Langer, Andreas Nürnberger, and Bela Gipp. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, RepSys '13, page 7–14, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450324656. doi: 10.1145/2532508.2532511. URL https://doi.org/10.1145/2532508.2532511.

[16] B Douglas Bernheim, Luca Braghieri, Alejandro Martínez-Marquina, and David Zuckerman. A theory of chosen preferences. *American Economic Review*, 111(2):720–54, 2021.

[17] 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. Diversity in News Recommendations. Technical report, 2021. URL http://arxiv.org/abs/2005.09495. arXiv:2005.09495 [cs].

[18] Mary Bernstein. Identity politics. *Annu. Rev. Sociol.*, 31:47–74, 2005.

[19] Brett Binst, Lien Michiels, and Annelien Smets. What Is Serendipity? An Interview Study to Conceptualize Experienced Serendipity in Recommender Systems. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pages 243–252, New York City USA, June 2025. ACM. ISBN 979-8-4007-1313-2. doi: 10.1145/3699682.3728325. URL https://dl.acm.org/doi/10.1145/3699682.3728325.

[20] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. The forgotten margins of ai ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 948–958, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533157. URL https://doi.org/10.1145/3531146.3533157.

[21] S. Blassnig, E. Strikovic, E. Mitova, A. Urman, A. Hannák, C. de Vreese, and F. Esser. A Balancing Act: How Media Professionals Perceive the Implementation of News Recommender Systems. *Digital Journalism*, 2023. doi: 10.1080/21670811.2023.2293933. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85181904971&doi=10.1080%2f21670811.2023.2293933&partnerID=40&md5=9c84525c852b01b4f8d09e0121ecab5d.

[22] Balázs Bodó. Selling News to Audiences – A Qualitative Inquiry into the Emerging Logics of Algorithmic News Personalization in European Quality News Media. *Digital Journalism*, 7(8):1054–1075, September 2019. ISSN 2167-0811. doi: 10.1080/21670811.2019.1624185. URL https://doi.org/10.1080/21670811.2019.1624185. _eprint: https://doi.org/10.1080/21670811.2019.1624185.

[23] Christina Boididou, Di Sheng, Felix J Mercer Moss, and Alessandro Piscopo. Building public service recommenders: Logbook of a journey. In *Fifteenth ACM Conference on Recommender Systems*, RecSys '21, pages 538–540, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582. doi: 10.1145/3460231.3474614. URL https://doi.org/10.1145/3460231.3474614.

[24] Allan Borodin, Hyun Chul Lee, and Yuli Ye. 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, pages 155–166, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312486. doi: 10.1145/2213556.2213580. URL https://doi.org/10.1145/2213556.2213580.

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

[26] Elda Brogi, Roberta Carlini, Iva Nenadi\' c, Pier Luigi Parcu, and Mario Viola de Azevedo Cunha. EU and media policy: conceptualising media pluralism in the era of online platforms. The experience of the Media Pluralism Monitor. In *Research Handbook on EU Media Law and Policy*, pages 16–31. Edward Elgar Publishing, September 2021. ISBN 978-1-78643-933-8. URL https://www.elgaronline.com/display/edcoll/9781786439321/9781786439321.00007.xml.

[27] Robin Burke, Alexander Felfernig, and Mehmet H. Göker. Recommender Systems: An Overview. *AI Magazine*, 32(3):13–18, June 2011. ISSN 2371-9621. doi: 10.1609/aimag.v32i3.2361. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2361.

[28] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. Balanced neighborhoods for multi-sided fairness in recommendation. In *Conference on Fairness, Accountability and Transparency*, pages 202–214, 2018.

[29] Pablo Castells, Neil Hurley, and Saúl Vargas. Novelty and Diversity in Recommender Systems. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 603–646. Springer US, New York, NY, 2022. ISBN 978-1-0716-2197-4. doi: 10.1007/978-1-0716-2197-4_16. URL https://doi.org/10.1007/978-1-0716-2197-4_16.

[30] Soumen Chakrabarti, Rajiv Khanna, Uma Sawant, and Chiru Bhattacharyya. Structured learning for non-smooth ranking losses. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 88–96, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401906. URL https://doi.org/10.1145/1401890.1401906.

[31] Abhijnan Chakraborty, Gourab K. Patro, Niloy Ganguly, Krishna P. Gummadi, and Patrick Loiseau. Equality of voice: Towards fair representation in crowdsourced top-k recommendations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 129–138, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287570. URL https://doi.org/10.1145/3287560.3287570.

[32] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 224–232, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359016. doi: 10.1145/3240323.3240370. URL https://doi.org/10.1145/3240323.3240370.

[33] Elizabeth Charters. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education Journal*, 12(2), 2003.

[34] Laming Chen, Guoxin Zhang, and Hanning Zhou. 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, pages 5627–5638, Red Hook, NY, USA, 2018. Curran Associates Inc.

[35] Jin Yao Chin, Yile Chen, and Gao Cong. The Datasets Dilemma: How Much Do We Really Know About Recommendation Datasets? In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 141–149, Virtual Event AZ USA, February 2022. ACM. ISBN 978-1-4503-9132-0. doi: 10.1145/3488560.3498519. URL https://dl.acm.org/doi/10.1145/3488560.3498519.

[36] Clifford Christians. The media and moral literacy. page 62, 2006.

[37] Clifford Christians, Theodore L. Glasser, Denis McQuail, Kaarle Nordenstreng, and Robert A. White. *Normative theories of the media: Journalism in democratic societies*. University of Illinois Press, 12 2009. ISBN 9780252034237.

[38] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 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, pages 659–666, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581644. doi: 10.1145/1390334.1390446. URL https://doi.org/10.1145/1390334.1390446.

[39] David Collier, Fernando Daniel Hidalgo, and Andra Olivia Maciuceanu. Essentially contested concepts: Debates and applications. *Journal of political ideologies*, 11(3):211–246, 2006.

[40] Commission. Europe's Media in the Digital Decade: An Action Plan to Support Recovery and Transformation. Technical Report COM(2020) 784 final, Commission, December 2020. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0784.

[41] Commission. Commission Staff Working Document on Common European Data Spaces. Technical Report SWD(2022) 45 final, Brussels, February 2022. URL https://digital-strategy.ec.europa.eu/en/library/staff-working-document-data-spaces.

[42] Commission. Digital Europe Programme (DIGITAL) Call for proposals Cloud Data and TEF (DIGITAL-2022-CLOUD-AI-03). Technical report, September 2022. URL https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/digital/wp-call/2022/call-fiche_digital-2022-cloud-ai-03_en.pdf.

[43] Commission. Pilot Project - Digital European Platform of Quality Content Providers (2nd Phase). Technical Report CNECT/2021/OP/0002, 2022. URL https://europemedialab.eu/wp-content/uploads/2022/10/Media-Data-Space_Interim-Report_Summary_final.pdf.

[44] Hannes Cools and Claes H. De Vreese. From Automation to Transformation with AI-Tools: Exploring the Professional Norms and the Perceptions of Responsible AI in a News Organization. *Digital Journalism*, pages 1–20, May 2025. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2025.2505982. URL https://www.tandfonline.com/doi/full/10.1080/21670811.2025.2505982.

[45] Irene Costera Meijer and Hildebrand P. Bijleveld. Valuable Journalism: Measuring news quality from a user's perspective. *Journalism Studies*, 17(7):827–839, October 2016. ISSN 1461-670X. doi: 10.1080/1461670X.2016.1175963. URL https://doi.org/10.1080/1461670X.2016.1175963. _eprint: https://doi.org/10.1080/1461670X.2016.1175963.

[46] Council of Europe. Recommendation of the Committee of Ministers to member States on media pluralism and transparency of media ownership. Technical Report CM/Rec(2018)1, Strasbourg, 2018. URL https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680790e13.

[47] Council of Europe. Recommendation CM/Rec(2022)4 of the Committee of Ministers to member States on promoting a favourable environment for quality journalism in the digital age. Technical report, Strasbourg, 2022. URL https://www.coe.int/en/web/freedom-expression/committee-of-ministers-adopted-texts/-/asset_publisher/aDXmrol0vvsU/content/recommendation-cm-rec-2022-4-of-the-committee-of-ministers-to-member-states-on-promoting-a-favourable-environment-for-quality-journalism-in-the-digita.

[48] Council of Europe. Guidelines on the Responsible Implementation of Artificial Intelligence Systems in Journalism. Technical report, Steering Committee on Media and Information Society, Strasbourg, 2023. URL https://rm.coe.int/cdmsi-2023-014-guidelines-on-the-responsible-implementation-of-artific/1680adb4c6.

[49] Joop Daalmeijer. Public service broadcasting in the netherlands. *Trends in Communication*, 12(1):33–45, 2004.

[50] Lincoln Dahlberg. Re-constructing digital democracy: An outline of four 'positions'. *New Media & Society*, 13(6):855–872, 2011. doi: 10.1177/1461444810389569.

[51] Sourabh Dattawad, Agnese Daffara, and Tanise Ceron. Leveraging Media Frames to Improve Normative Diversity in News Recommendations, September 2025. URL http://arxiv.org/abs/2509.02266. arXiv:2509.02266 [cs].

[52] Mathias-Felipe de Lima-Santos, Allen Munoriyarwa, Adeola Abdulateef Elega, and Charis Papaevangelou. Google News Initiative's Influence on Technological Media Innovation in Africa and the Middle East. *Media and Communication*, 11(2):330–343, June 2023. ISSN 2183-2439. doi: 10.17645/mac.v11i2.6400. URL https://www.cogitatiopress.com/mediaandcommunication/article/view/6400.

[53] Gabriel de Souza Pereira Moreira, Felipe Ferreira, and Adilson Marques da Cunha. News Session-Based Recommendations using Deep Neural Networks. In *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*, DLRS 2018, pages 15–23, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 978-1-4503-6617-5. doi: 10.1145/3270323.3270328. URL https://dl.acm.org/doi/10.1145/3270323.3270328.

[54] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 862–872, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445924. URL https://doi.org/10.1145/3442188.3445924.

[55] Tommaso Di Noia, Vito Claudio Ostuni, Jessica Rosati, Paolo Tomeo, and Eugenio Di Sciascio. An analysis of users' propensity toward diversity in recommendations. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 285–288, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450326681. doi: 10.1145/2645710.2645774. URL https://doi.org/10.1145/2645710.2645774.

[56] Tawanna R Dillahunt, Christopher A Brooks, and Samarth Gulati. Detecting and visualizing filter bubbles in google and bing. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1851–1856, 2015.

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

[58] Qinxu Ding, Yong Liu, Chunyan Miao, Fei Cheng, and Haihong Tang. A hybrid bandit framework for diversified recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4036–4044, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16524.

[59] Karlijn Dinnissen and Christine Bauer. Amplifying artists' voices: Item provider perspectives on influence and fairness of music streaming platforms. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 238–249, 2023.

[60] Tomás Dodds, Astrid Vandendaele, Felix M. Simon, Natali Helberger, Valeria Resendez, and Wang Ngai Yeung. Knowledge Silos as a Barrier to Responsible AI Practices in Journalism? Exploratory Evidence from Four Dutch News Organisations. *Journalism Studies*, 26(6):740–758, April 2025. ISSN 1461-670X, 1469-9699. doi: 10.1080/1461670x.2025.2463589. URL https://www.tandfonline.com/doi/full/10.1080/1461670X.2025.2463589.

[61] Karen Donders. *Public Service Media in Europe: Law, Theory and Practice*. Routledge, London, June 2021. ISBN 978-1-351-10556-9.

[62] Konstantin Nicholas Dörr. Mapping the field of algorithmic journalism. *Digital Journalism*, 4(6):700–722, 2016. doi: 10.1080/21670811.2015.1096748.

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

[64] Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions With Debated Topics. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 135–145, Regensburg Germany, March 2022. ACM. ISBN 978-1-4503-9186-3. doi: 10.1145/3498366.3505812. URL https://dl.acm.org/doi/10.1145/3498366.3505812.

[65] d.school at Stanford. An Introduction to Design Thinking PROCESS GUIDE. Technical report. URL https://dschool-old.stanford.edu/sandbox/groups/designresources/wiki/36873/attachments/74b3d/ModeGuideBOOTCAMP2010L.pdf.

[66] Yu Du, Sylvie Ranwez, Nicolas Sutton-Charani, and Vincent Ranwez. Is diversity optimization always suitable? toward a better understanding of diversity within recommendation approaches. *Information Processing & Management*, 58(6):102721, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102721. URL https://www.sciencedirect.com/science/article/pii/S0306457321002053.

[67] EBU. *Dataspace for Cultural and Creative Industries*. October 2022. URL https://tech.ebu.ch/publications/dataspace-for-cultural-and-creative-industries.

[68] EBU. Personalisation and Recommendation Ecosystem developed by Broadcasters for Broadcasters. Technical report, EBU, Geneva, 2023. URL https://docs.peach.ebu.io/.

[69] Arni Mar Einarsson and Jacob Ormen. From static to dynamic evaluation: The potentials and challenges of dashboards to account for NRSs. *Big Data & Society*, 12(2):20539517251340635, June 2025. ISSN 2053-9517. doi: 10.1177/20539517251340635. URL https://doi.org/10.1177/20539517251340635.

[70] Arni Mar Einarsson, Helles , R., , and S. Lomborg. Algorithmic agenda-setting: the subtle effects of news recommender systems on political agendas in the Danish 2022 general election. *Information, Communication & Society*, 28(2):218–238, January 2025. ISSN 1369-118X. doi: 10.1080/1369118X.2024.2334411. URL https://doi.org/10.1080/1369118X.2024.2334411. _eprint: https://doi.org/10.1080/1369118X.2024.2334411.

[71] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 161–168, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450326681. doi: 10.1145/2645710.2645737. URL https://doi.org/10.1145/2645710.2645737.

[72] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 172–186. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/ekstrand18b.html.

[73] Michael D. Ekstrand, Afsaneh Razi, Aleksandra Sarcevic, Maria Soledad Pera, Robin Burke, and Katherine Landau Wright. Recommending With, Not For: Co-Designing Recommender Systems for Social Good. *ACM Transactions on Recommender Systems*, page 3759261, August 2025. ISSN 2770-6699. doi: 10.1145/3759261. URL http://arxiv.org/abs/2508.03792. arXiv:2508.03792 [cs].

[74] Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

[75] Farzad Eskandanian, Bamshad Mobasher, and Robin Burke. A clustering approach for personalizing diversity in collaborative recommender systems. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, UMAP '17, pages 280–284, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346351. doi: 10.1145/3079628.3079699. URL https://doi.org/10.1145/3079628.3079699.

[76] Sarah Eskens, Natali Helberger, and Judith Moeller. Challenged by news personalisation: five perspectives on the right to receive information. *Journal of Media Law*, 9(2):259–284, 2017. doi: 10.1080/17577632.2017.1387353.

[77] Alexander Fanta and Ingo Dachwitz. Google, the media patron. How the digital giant ensnares journalism., November 2020. URL https://osf.io/preprints/socarxiv/3qbp9_v1/.

[78] Sina Fazelpour and Maria De-Arteaga. Diversity in sociotechnical machine learning systems. *Big Data & Society*, 9(1):20539517221082027, 2022.

[79] Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd international conference on computational linguistics: Posters*, pages 276–284. Association for computational linguistics, 2010.

[80] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 101–109, Copenhagen Denmark, September 2019. ACM. ISBN 978-1-4503-6243-6. doi: 10.1145/3298689.3347058. URL https://dl.acm.org/doi/10.1145/3298689.3347058.

[81] Andres Ferraro, Xavier Serra, and Christine Bauer. Break the loop: Gender imbalance in music recommenders. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 249–254, 2021.

[82] Myra Marx Ferree, William A Gamson, Jürgen Gerhards, and Dieter Rucht. Four models of the public sphere in modern democracies. *Theory and society*, 31(3):289–324, 2002.

[83] Raul Ferrer-Conill and Edson C. Tandoc Jr. The audience-oriented editor. *Digital Journalism*, 6(4):436–453, 2018. doi: 10.1080/21670811.2018.1440972.

[84] César Fieiras-Ceide, Martí-n Vaz-Álvarez, and Miguel Túñez-López. Designing personalisation of European public service media (PSM): trends on algorithms and artificial intelligence for content distribution. *Profesional de la información*, 32(3), May 2023. ISSN 1699-2407. doi: 10.3145/epi.2023.may.11. URL https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/87298. Number: 3.

[85] Richard Fletcher. More than 'just the facts': How news audiences think about 'user needs' | Reuters Institute for the Study of Journalism, June 2024. URL https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024/more-just-facts-how-news-audiences-think-about-user-needs.

[86] Barbara L Fredrickson. Positive emotions broaden and build. In *Advances in experimental social psychology*, volume 47, pages 1–53. Elsevier, 2013.

[87] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

[88] Silke Fürst. In the Service of Good Journalism and Audience Interests? How Audience Metrics Affect News Quality. *Media and Communication*, 8(3):270–280, August 2020. ISSN 2183-2439. doi: 10.17645/mac.v8i3.3228. URL https://www.cogitatiopress.com/mediaandcommunication/article/view/3228.

[89] Walter Bryce Gallie. Essentially contested concepts. In *Proceedings of the Aristotelian society*, volume 56, pages 167–198. JSTOR, 1955.

[90] Lu Gan, Diana Nurbakova, Léa Laporte, and Sylvie Calabretto. 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*, pages 2001–2004, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. URL https://doi.org/10.1145/3397271.3401213.

[91] Florent Garcin, Christos Dimitrakakis, and Boi Faltings. Personalized news recommendation with context trees. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 105–112, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450324090. doi: 10.1145/2507157.2507166. URL https://doi.org/10.1145/2507157.2507166.

[92] Graziadio, Marco and Weström, Pär. The European Data Market Study 2024-2026: Story on the Common European Media Data Space. Technical report, European Commission, Brussels, 2024. URL https://ec.europa.eu/newsroom/repository/document/2024-31/EDM_20242026_StoryMedia_Data_Space_final_Qfq92zEU7HzdPZzSQFFBnxX14vI_107670.pdf.

[93] Andreas Grün and Xenija Neufeld. Challenges experienced in public service media recommendation systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 541–544, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582. doi: 10.1145/3460231.3474618. URL https://doi.org/10.1145/3460231.3474618.

[94] Andreas Grün and Xenija Neufeld. Transparently Serving the Public: Enhancing Public Service Media Values through Exploration. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1045–1048, Singapore Singapore, September 2023. ACM. ISBN 979-8-4007-0241-9. doi: 10.1145/3604915.3610243. URL https://dl.acm.org/doi/10.1145/3604915.3610243.

[95] Jon Atle Gulla, Lemei Zhang, Peng Liu, Özlem Özgöbek, and Xiaomeng Su. The Adressa dataset for news recommendation. In *Proceedings of the International Conference on Web Intelligence*, pages 1042–1048, Leipzig Germany, August 2017. ACM. ISBN 978-1-4503-4951-2. doi: 10.1145/3106426.3109436. URL https://dl.acm.org/doi/10.1145/3106426.3109436.

[96] Neha Gupta. How Denmark's Ekstra Bladet used AI to boost subscriptions by 35%, February 2025. URL https://wan-ifra.org/2025/02/how-denmarks-ekstra-bladet-used-ai-to-boost-subscriptions-by-35/.

[97] Rishav Hada, Amir Ebrahimi Fard, Sarah Shugars, Federico Bianchi, Patricia Rossini, Dirk Hovy, Rebekah Tromble, and Nava Tintarev. 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*, pages 33–41, Singapore Singapore, February 2023. ACM. ISBN 978-1-4503-9407-9. doi: 10.1145/3539597.3570487. URL https://dl.acm.org/doi/10.1145/3539597.3570487.

[98] N. Hagar and N. Diakopoulos. Algorithmic indifference: The dearth of news recommendations on TikTok. *New Media and Society*, 2023. doi: 10.1177/14614448231192964. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85170026581&doi=10.1177%2f14614448231192964&partnerID=40&md5=1a1660b46075d416201fa47c306c483b.

[99] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 501–512, 2020.

[100] Jaron Harambam, Natali Helberger, and Joris van Hoboken. Democratizing algorithmic news recommenders: how to materialize voice in a technologically saturated media ecosystem. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180088, October 2018. doi: 10.1098/rsta.2018.0088. URL https://royalsocietypublishing.org/doi/full/10.1098/rsta.2018.0088.

[101] Jaron Harambam, Dimitrios Bountouridis, Mykola Makhortykh, and Joris Van Hoboken. Designing for the better by taking users into account: A qualitative evaluation of user control mechanisms in (news) recommender systems. In *Proceedings of the 13th ACM conference on recommender systems*, pages 69–77, 2019.

[102] F. Maxwell Harper and Joseph A. Konstan. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, January 2016. ISSN 2160-6455, 2160-6463. doi: 10.1145/2827872. URL https://dl.acm.org/doi/10.1145/2827872.

[103] Marcel Hauck, Michael Huber, Juri Diels, David Wittenberg, and Dietmar Jannach. Balanced Public Service Media Recommendation Trade-offs with a Light Carbon Footprint. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, RecSys '25, pages 915–918, New York, NY, USA, September 2025. Association for Computing Machinery. ISBN 979-8-4007-1364-4. doi: 10.1145/3705328.3748106. URL https://dl.acm.org/doi/10.1145/3705328.3748106.

[104] L. Heitz, J.A. Croci, M. Sachdeva, and A. Bernstein. Informfully – Research Platform for Reproducible User Studies. pages 660–669, 2024. doi: 10.1145/3640457.3688066. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85208782918&doi=10.1145%2f3640457.3688066&partnerID=40&md5=0c6ca691e96a29a76ba84fa844947e1f.

[105] L. Heitz, S. Vrijenhoek, and O. Inel. Recommendations for the Recommenders: Reflections on Prioritizing Diversity in the RecSys Challenge. pages 22–26, 2024. doi: 10.1145/3687151.3687155. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85208403426&doi=10.1145%2f3687151.3687155&partnerID=40&md5=fece0791606f82da0c080f32c1cca00b.

[106] Lucien Heitz. Classification of Normative Recommender Systems. *Proceedings of the First Workshop on the Normative Design and Evaluation of Recommender Systems (NORMalize 2023) co-located with the 17th ACM Conference on Recommender Systems (RecSys 2023)*, September 2023.

[107] Lucien Heitz, Juliane A. Lischka, Alena Birrer, Bibek Paudel, Suzanne Tolmeijer, Laura Laugwitz, and Abraham Bernstein. Benefits of Diverse News Recommendations for Democracy: A User Study. *Digital Journalism*, 10(10):1710–1730, November 2022. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2021.2021804. URL https://www.tandfonline.com/doi/full/10.1080/21670811.2021.2021804.

[108] Lucien Heitz, Runze Li, Oana Inel, and Abraham Bernstein. Informfully Recommenders - Reproducibility Framework for Diversity-aware Intra-session Recommendations. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 792–801, Prague Czech Republic, September 2025. ACM. ISBN 979-8-4007-1364-4. doi: 10.1145/3705328.3748148. URL https://dl.acm.org/doi/10.1145/3705328.3748148.

[109] Natali Helberger. Exposure Diversity as a Policy Goal. *Journal of Media Law*, 4(1):65–92, July 2012. ISSN 1757-7632, 1757-7640. doi: 10.5235/175776312802483880. URL https://www.tandfonline.com/doi/full/10.5235/175776312802483880.

[110] Natali Helberger. Policy Implications From Algorithmic Profiling and the Changing Relationship Between Newsreaders and the Media. *Javnost - The Public*, 23(2):188–203, April 2016. ISSN 1318-3222, 1854-8377. doi: 10.1080/13183222.2016.1162989. URL https://www.tandfonline.com/doi/full/10.1080/13183222.2016.1162989.

[111] Natali Helberger. On the Democratic Role of News Recommenders. *Digital Journalism*, 7(8):993–1012, September 2019. ISSN 2167-0811. doi: 10.1080/21670811.2019.1623700. URL https://doi.org/10.1080/21670811.2019.1623700. _eprint: https://doi.org/10.1080/21670811.2019.1623700.

[112] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, 21(2):191–207, February 2018. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2016.1271900. URL https://www.tandfonline.com/doi/full/10.1080/1369118X.2016.1271900.

[113] Natali Helberger, Max Van Drunen, Sarah Eskens, Mariella Bastian, and Judith Möller. A freedom of expression perspective on AI in the media – with a special focus on editorial decision making on social media platforms and in the news media. *European Journal of Law and Technology*, 11(3), December 2020. ISSN 2042-115X. URL https://ejlt.org/index.php/ejlt/article/view/752.

[114] Natali Helberger, Max Van Drunen, Judith Moeller, Sanne Vrijenhoek, and Sarah Eskens. Towards a Normative Perspective on Journalistic AI: Embracing the Messy Reality of Normative Ideals. *Digital Journalism*, 10(10):1605–1626, November 2022. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2022.2152195. URL https://www.tandfonline.com/doi/full/10.1080/21670811.2022.2152195.

[115] David Held. *Models of democracy*. Stanford University Press, 2006.

[116] Christopher A. Hennessy and Charles A. E. Goodhart. Goodhart's Law and Machine Learning: A Structural Perspective. *International Economic Review*, 64(3):1075–1086, 2023. ISSN 1468-2354. doi: 10.1111/iere.12633. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/iere.12633. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/iere.12633.

[117] Alfred Hermida and Felix M. Simon. AI in the Newsroom: Lessons from the Adoption of The Globe and Mail's Sophi. *Journalism Practice*, 19(10):2323–2340, October 2025. ISSN 1751-2786. doi: 10.1080/17512786.2025.2471781. URL https://doi.org/10.1080/17512786.2025.2471781. _eprint: https://doi.org/10.1080/17512786.2025.2471781.

[118] Jockum Hildén. The Public Service Approach to Recommender Systems: Filtering to Cultivate. *Television & New Media*, 23(7):777–796, November 2022. ISSN 1527-4764. doi: 10.1177/15274764211020106. URL https://doi.org/10.1177/15274764211020106.

[119] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. Technical report, spaCy, 2022. URL https://spacy.io/usage/spacy-101.

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

[121] Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.

[122] Aaron Hyzen, Hilde Van Den Bulck, Manuel Puppis, Michelle Kulig, and Steve Paulussen. Epistemic welfare and algorithmic recommender systems: overcoming the epistemic crisis in the digitalized public sphere. *Communication Theory*, 36(1):46–57, February 2026. ISSN 1050-3293, 1468-2885. doi: 10.1093/ct/qtaf018. URL https://academic.oup.com/ct/article/36/1/46/8240891.

[123] A. Iana, M. Alam, and H. Paulheim. A survey on knowledge-aware news recommender systems. *Semantic Web*, 15(1):21–82, 2024. doi: 10.3233/SW-222991. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85182653927&doi=10.3233%2fSW-222991&partnerID=40&md5=d6f3978470b9bea10edfb669bcbe96fe.

[124] Andreea Iana, Goran Glavas, and Heiko Paulheim. Simplifying Content-Based Neural News Recommendation: On User Modeling and Training Objectives. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2384–2388, Taipei Taiwan, July 2023. ACM. ISBN 978-1-4503-9408-6. doi: 10.1145/3539618.3592062. URL https://dl.acm.org/doi/10.1145/3539618.3592062.

[125] Abigail Z. Jacobs and Hanna Wallach. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, Virtual Event Canada, March 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445901. URL https://dl.acm.org/doi/10.1145/3442188.3445901.

[126] Dietmar Jannach and Michael Jugovac. Measuring the Business Value of Recommender Systems. *ACM Trans. Manage. Inf. Syst.*, 10(4):16:1–16:23, December 2019. ISSN 2158-656X. doi: 10.1145/3370082. URL https://doi.org/10.1145/3370082.

[127] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002. ISSN 1046-8188. doi: 10.1145/582415.582418. URL https://doi.org/10.1145/582415.582418.

[128] Harold Jeffreys. 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):453–461, 1946.

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

[130] Edson C. Tandoc Jr. and Ryan J. Thomas. The ethics of web analytics. *Digital Journalism*, 3(2):243–258, 2015. doi: 10.1080/21670811.2014.909122.

[131] Michael Jugovac, Dietmar Jannach, and Lukas Lerche. Efficient optimization of multiple recommendation quality factors according to individual user tendencies. *Expert Systems with Applications*, 81:321–331, 2017. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2017.03.055. URL https://www.sciencedirect.com/science/article/pii/S0957417417302075.

[132] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. News recommender systems – Survey and roads ahead. *Information Processing & Management*, 54(6):1203–1227, November 2018. ISSN 03064573. doi: 10.1016/j.ipm.2018.04.008. URL https://linkinghub.elsevier.com/retrieve/pii/S030645731730153X.

[133] Kari Karppinen. Conceptions of democracy in media and communications studies.

[134] Kari Karppinen. Uses of democratic theory in media and communication studies. *Observatorio*, 7(3):1–17, 2013. ISSN 1646-5954.

[135] Kari Karppinen. *Rethinking media pluralism*. Fordham Univ Press, 2013.

[136] Kari Karppinen. The limits of empirical indicators: Media pluralism as an essentially contested concept. In *Media pluralism and diversity: Concepts, risks and global trends*, pages 287–296. Springer, 2015.

[137] Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), November 2018. doi: 10.1145/3274357. URL https://doi.org/10.1145/3274357.

[138] Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. How Well Do LLMs Represent Values Across Cultures? Empirical Analysis of LLM Responses Based on Hofstede Cultural Dimensions, August 2025. URL http://arxiv.org/abs/2406.14805. arXiv:2406.14805 [cs].

[139] Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias Heintz. The plista dataset. In *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge*, NRS '13, pages 16–23, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450323024. doi: 10.1145/2516641.2516643. URL https://doi.org/10.1145/2516641.2516643.

[140] Dongwoo Kim and Alice Oh. Topic chains for understanding a news corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 163–176. Springer, 2011.

[141] J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.

[142] Lisa Merete Kristensen and Jannie Møller Hartley. The Infrastructure of News: Negotiating Infrastructural Capture and Autonomy in Data-Driven News Distribution. *Media and Communication*, 11(2), May 2023. ISSN 2183-2439. doi: 10.17645/mac.v11i2.6388. URL https://www.cogitatiopress.com/mediaandcommunication/article/view/6388.

[143] Johannes Kruse, Kasper Lindskow, Michael Riis Andersen, and Jes Frellsen. Creating the next generation of news experience on ekstrabladet.dk with recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1067–1070, Singapore Singapore, September 2023. ACM. doi: 10.1145/3604915.3610248. URL https://dl.acm.org/doi/10.1145/3604915.3610248.

[144] Johannes Kruse, Kasper Lindskow, Saikishore Kalloori, Marco Polignano, Claudio Pomo, Abhishek Srivastava, Anshuk Uppal, Michael Riis Andersen, and Jes Frellsen. EB-NeRD a large-scale dataset for news recommendation. In *Proceedings of the Recommender Systems Challenge 2024*, pages 1–11, Bari Italy, October 2024. ACM. ISBN 979-8-4007-1127-5. doi: 10.1145/3687151.3687152. URL https://dl.acm.org/doi/10.1145/3687151.3687152.

[145] Johannes Kruse, Kasper Lindskow, Saikishore Kalloori, Marco Polignano, Claudio Pomo, Abhishek Srivastava, Anshuk Uppal, Michael Riis Andersen, and Jes Frellsen. RecSys Challenge 2024: Balancing Accuracy and Editorial Values in News Recommendations. In *18th ACM Conference on Recommender Systems*, pages 1195–1199, Bari Italy, October 2024. ACM. ISBN 979-8-4007-0505-2. doi: 10.1145/3640457.3687164. URL https://dl.acm.org/doi/10.1145/3640457.3687164.

[146] Johannes Kruse, Kasper Lindskow, Michael Riis Andersen, Ryotaro Shimizu, Julian McAuley, Pierre-Alexandre Mattei, and Jes Frellsen. Normative Alignment of Recommender Systems via Internal Label Shift. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, RecSys '25, pages 1240–1245, New York, NY, USA, September 2025. Association for Computing Machinery. ISBN 979-8-4007-1364-4. doi: 10.1145/3705328.3759309. URL https://dl.acm.org/doi/10.1145/3705328.3759309.

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

[148] Matevž Kunaver and Tomaž Požrl. Diversity in recommender systems – A survey. *Knowledge-Based Systems*, 123:154–162, May 2017. ISSN 0950-7051. doi: 10.1016/j.knosys.2017.02.009. URL https://www.sciencedirect.com/science/article/pii/S0950705117300680.

[149] E. E. Lawrence. The trouble with diverse books, part I: on the limits of conceptual analysis for political negotiation in Library & Information Science. *Journal of Documentation*, 76(6):1473–1491, July 2020. ISSN 0022-0418. doi: 10.1108/JD-04-2020-0057. URL https://doi.org/10.1108/JD-04-2020-0057.

[150] Paddy Leerssen. From Murdoch to Musk: Platform ownership and the political economy of online content governance. *Platforms & Society*, 2:29768624251386260, October 2025. ISSN 2976-8624. doi: 10.1177/29768624251386260. URL https://doi.org/10.1177/29768624251386260.

[151] Seth C. Lewis and Oscar Westlund. Big data and journalism. *Digital Journalism*, 3(3):447–466, 2015. doi: 10.1080/21670811.2014.976418.

[152] Runze Li, Lucien Heitz, Oana Inel, and Abraham Bernstein. D-RDW: Diversity-Driven Random Walks for News Recommender Systems, August 2025. URL http://arxiv.org/abs/2508.13035. arXiv:2508.13035 [cs].

[153] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006. doi: 10.1109/TIT.2006.881731.

[154] Bibo Lin and Seth C. Lewis. The One Thing Journalistic AI Just Might Do for Democracy. *Digital Journalism*, 10(10):1627–1649, November 2022. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2022.2084131. URL https://www.tandfonline.com/doi/full/10.1080/21670811.2022.2084131.

[155] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[156] Jeffrey W Lockhart, Molly M King, and Christin Munsch. Name-based demographic inference and the unequal distribution of misrecognition. *Nature Human Behaviour*, pages 1–12, 2023.

[157] Felicia Loecherbach and Damian Trilling. 3bij3 – Developing a framework for researching recommender systems and their effects. *Computational Communication Research*, 2(1):53–79, February 2020. URL https://computationalcommunication.org/ccr/article/view/11.

[158] Felicia Loecherbach, Judith Moeller, Damian Trilling, and Wouter Van Atteveldt. The Unified Framework of Media Diversity: A Systematic Literature Review. *Digital Journalism*, 8(5):605–642, May 2020. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2020.1764374. URL https://www.tandfonline.com/doi/full/10.1080/21670811.2020.1764374.

[159] Felicia Loecherbach, Kasper Welbers, Judith Moeller, Damian Trilling, and Wouter Van Atteveldt. Is this a click towards diversity? Explaining when and why news users make diverse choices. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, pages 282–290, New York, NY, USA, June 2021. Association for Computing Machinery. ISBN 978-1-4503-8330-1. doi: 10.1145/3447535.3462506. URL https://dl.acm.org/doi/10.1145/3447535.3462506.

[160] Steven Loria. textblob documentation. Technical report, Textblob, 2021. URL https://textblob.readthedocs.io/en/dev/quickstart.html.

[161] Feng Lu, Anca Dumitrache, and David Graus. Beyond Optimizing for Clicks: Incorporating Editorial Values in News Recommendation. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 145–153, Genoa Italy, July 2020. ACM. ISBN 978-1-4503-6861-2. doi: 10.1145/3340631.3394864. URL https://dl.acm.org/doi/10.1145/3340631.3394864.

[162] Hongyu Lu, Min Zhang, and Shaoping Ma. 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*, SIGIR '18, pages 435–444, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210007. URL https://doi.org/10.1145/3209978.3210007.

[163] Dieuwertje Luitse. Platform power in AI: The evolution of cloud infrastructures in the political economy of artificial intelligence. *Internet Policy Review*, 13(2), June 2024. ISSN 2197-6775. doi: 10.14763/2024.2.1768. URL https://policyreview.info/articles/analysis/platform-power-ai-evolution-cloud-infrastructures.

[164] Dieuwertje Luitse, Tobias Blanke, and Thomas Poell. AI competitions as infrastructures of power in medical imaging. *Information, Communication & Society*, 28(10):1735–1756, July 2025. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2024.2334393. URL https://www.tandfonline.com/doi/full/10.1080/1369118X.2024.2334393.

[165] M. Makhortykh and M. Bastian. Personalizing the war: Perspectives for the adoption of news recommendation algorithms in the media coverage of the conflict in Eastern Ukraine. *Media, War and Conflict*, 15(1): 25–45, 2022. doi: 10.1177/1750635220906254. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85081561386&doi=10.1177%2f1750635220906254&partnerID=40&md5=8e9ea3cf2ef68bf325b9cb1894016c1f.

[166] Mykola Makhortykh, Claes de Vreese, Natali Helberger, Jaron Harambam, and Dimitrios Bountouridis. We are what we click: Understanding time and content-based habits of online news readers. *new media & society*, page 1461444820933221, 2020.

[167] Bernard Manin. On legitimacy and political deliberation. *Political theory*, 15(3):338–368, 1987.

[168] ECtHR Manole and Others v Moldova. Manole and Others v. Moldova, September 2009. URL https://hudoc.echr.coe.int/eng?i=001-94075.

[169] Laura Mascarell, Tatyana Ruzsics, Christian Schneebeli, Philippe Schlattner, Luca Campanella, Severin Klingler, and Cristina Kadar. Stance Detection in German News Articles. In Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos, editors, *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 66–77, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.fever-1.8. URL https://aclanthology.org/2021.fever-1.8/.

[170] Ariadne Matas, Antoine Isaac, Esther Huyer, Fred Saunderson, Julia Fallon, Kerstin Arnold, Marie-V\' eronique Leroi, Paul Keller, and Valentine Charles. Data Governance for the common European data space for cultural heritage: Strategy and Plan. Technical report, Europeana, 2023. URL https://pro.europeana.eu/files/Europeana_Professional/Publications/Data_Governance_Strategy_and_Plan.pdf.

[171] Nicolas Mattis, Philipp K. Masur, Judith Moeller, and Wouter Van Atteveldt. It ain't easy: using normatively motivated news diversification to facilitate policy support, tolerance, and political participation. *Information, Communication & Society*, 28(4):651–668, March 2025. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2024.2423892. URL https://www.tandfonline.com/doi/full/10.1080/1369118X.2024.2423892.

[172] Nicolas Michael Mattis, Philipp K. Masur, Judith Möller, and Wouter van Atteveldt. Nudging towards diversity: A theoretical framework for facilitating diverse news consumption through recommender design. Technical report, SocArXiv, July 2021. URL https://osf.io/preprints/socarxiv/wvxf5/. type: article.

[173] Maxwell McCOMBS. Building Consensus: The News Media's Agenda-Setting Roles. *Political Communication*, 14(4):433–443, October 1997. ISSN 1058-4609, 1091-7675. doi: 10.1080/105846097199236. URL http://www.tandfonline.com/doi/abs/10.1080/105846097199236.

[174] Nora McDonald and Shimei Pan. Intersectional ai: A study of how information science students think about ethics and their impact. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), oct 2020. doi: 10.1145/3415218. URL https://doi.org/10.1145/3415218.

[175] Sean M McNee, John Riedl, and Joseph A Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1097–1101, 2006.

[176] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 2243–2251, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3272027. URL https://doi.org/10.1145/3269206.3272027.

[177] Lien Michiels, Jens Leysen, Annelien Smets, and Bart Goethals. What Are Filter Bubbles Really? A Review of the Conceptual and Empirical Work. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22 Adjunct, pages 274–279, New York, NY, USA, July 2022. Association for Computing Machinery. ISBN 978-1-4503-9232-7. doi: 10.1145/3511047.3538028. URL https://dl.acm.org/doi/10.1145/3511047.3538028.

[178] Lien Michiels, Jorre Vannieuwenhuyze, Jens Leysen, Robin Verachtert, Annelien Smets, and Bart Goethals. How should we measure filter bubbles? a regression model and evidence for online news. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 640–651, 2023.

[179] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 117–123, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375832. URL https://doi.org/10.1145/3375627.3375832.

[180] Eliza Mitova, Sina Blassnig, Edina Strikovic, Aleksandra Urman, Aniko Hannak, Claes H. de Vreese, and Frank Esser. News recommender systems: a programmatic research review. *Annals of the International Communication Association*, 47(1):84–113, January 2023. ISSN 2380-8985. doi: 10.1080/23808985.2022.2142149. URL https://doi.org/10.1080/23808985.2022.2142149. _eprint: https://doi.org/10.1080/23808985.2022.2142149.

[181] ECtHR NIT S. R. L. v the Republic of Moldova. NIT S.R.L. v. the Republic of Moldova, April 2022. URL https://hudoc.echr.coe.int/eng?i=001-216872.

[182] Lynge Asbjørn Møller. Designing algorithmic editors: How newspapers embed and encode journalistic values into news recommender systems. *Digital Journalism*, pages 1–19, 2023.

[183] Mats Mulder, Oana Inel, Jasper Oosterman, and Nava Tintarev. Operationalizing framing to support multiperspective recommendations of opinion pieces. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 478–488, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445911. URL https://doi.org/10.1145/3442188.3445911.

[184] Luke Munn. The uselessness of AI ethics. *AI and Ethics*, 3(3):869–877, August 2023. ISSN 2730-5961. doi: 10.1007/s43681-022-00209-w. URL https://doi.org/10.1007/s43681-022-00209-w.

[185] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 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):959–977, 2018.

[186] Hartley Jannie Møller and Nanna Bonde Thylstrup. The Algorithmic Gut Feeling – Articulating Journalistic Doxa and Emerging Epistemic Frictions in AI-Driven Data Work. *Digital Journalism*, March 2025. ISSN 2167-0811. URL https://www.tandfonline.com/doi/abs/10.1080/21670811.2024.2319641.

[187] L.A. Møller. Recommended for You: How Newspapers Normalise Algorithmic News Recommendation to Fit Their Gatekeeping Role. *Journalism Studies*, 23(7):800–817, 2022. doi: 10.1080/1461670X.2022.2034522. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124352638&doi=10.1080%2f1461670X.2022.2034522&partnerID=40&md5=2a8dc8d0b7bd5ca9b0bab31430f255c4.

[188] Lynge Asbjørn Møller. Between Personal and Public Interest: How Algorithmic News Recommendation Reconciles with Journalism as an Ideology. *Digital Journalism*, 10(10):1794–1812, November 2022. ISSN 2167-0811. doi: 10.1080/21670811.2022.2032782. URL https://doi.org/10.1080/21670811.2022.2032782. _eprint: https://doi.org/10.1080/21670811.2022.2032782.

[189] Lynge Asbjørn Møller. Bridging the Tech-Editorial Gap: Lessons from Two Case Studies of the Development and Integration of Algorithmic Curation in Journalism. *Journalism Studies*, 24(11):1458–1475, August 2023. ISSN 1461-670X, 1469-9699. doi: 10.1080/1461670X.2023.2227283. URL https://www.tandfonline.com/doi/full/10.1080/1461670X.2023.2227283.

[190] Pm Napoli. Deconstructing the diversity principle. *Journal of Communication*, 49(4):7–34, 1999. ISSN 1460-2466. doi: 10.1111/j.1460-2466.1999.tb02815.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.1999.tb02815.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-2466.1999.tb02815.x.

[191] L. Naudts, N. Helberger, M. Veale, and M. Sax. A Right to Constructive Optimization: A Public Interest Approach to Recommender Systems in the Digital Services Act. *Journal of Consumer Policy*, 48(3):269–296, September 2025. ISSN 1573-0700. doi: 10.1007/s10603-025-09586-1. URL https://doi.org/10.1007/s10603-025-09586-1.

[192] Nic Newman. Journalism, media, and technology trends and predictions 2024. Technical report, Reuters Institute for the Study of Journalism, 2024. URL https://reutersinstitute.politics.ox.ac.uk/journalism-media-and-technology-trends-and-predictions-2024.

[193] Nic Newman, Richard Fletcher, Craig T. Robertson, Kirsten Eddy, and Rasmus Kleis Nielsen. Reuters Institute digital news report 2022. Technical report, Reuters Institute for the Study of Journalism, 2022. URL https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022.

[194] Nic Newman, Richard Fletcher, Craig T. Robertson, Amy Ross Arguedas, and Rasmus Kleis Nielsen. Reuters Institute digital news report 2024. Technical report, Reuters Institute for the Study of Journalism, 2024. URL https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024.

[195] Nic Newman, Amy Ross Arguedas, Craig T. Robertson, Rasmus Kleis Nielsen, and Richard Fletcher. Reuters Institute digital news report 2025. Technical report, Reuters Institute for the Study of Journalism, 2025. URL https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025.

[196] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. Exploring the filter bubble: The effect of using recommender systems on content diversity. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 677–686, 2014. doi: 10.1145/2566486.2568012.

[197] Newman Nic, Richard Fletcher, Antonis Kalogeropoulos, David AL Levy, and Rasmus Kleis Nielsen. Reuters institute digital news report 2018. *Reuters Institute for the Study of Journalism*, page 39, 2018.

[198] Tom Nicholls and Jonathan Bright. Understanding news story chains using information retrieval and network clustering techniques. *Communication Methods and Measures*, 13(1):43–59, 2019.

[199] Sachita Nishal and Nicholas Diakopoulos. Values as Problems, Principles, and Tensions in Sociotechnical System Design for Journalism. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, pages 2975–2991, Madeira Portugal, July 2025. ACM. ISBN 979-8-4007-1485-6. doi: 10.1145/3715336.3735717. URL https://dl.acm.org/doi/10.1145/3715336.3735717.

[200] Christian S Nissen et al. Public service media in the information society, 2006.

[201] Ofcom. Ofcom's strategic approach to AI 2024/25. Technical report, London, 2024. URL https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-2-6-weeks/273274—ofcoms-proposed-plan-of-work-for-2024-25/associated-documents/ofcoms-strategic-approach-to-ai.pdf?v=321367.

[202] Ofcom. News Consumption in the UK 2025 Research Findings. 2025.

[203] High-Level Expert Group on Artificial Intelligence. Draft ethics guidelines for trustworthy ai. Technical report, European Commission, 2018.

[204] Mícheál. O'Searcoid. *Metric Spaces*. Springer Undergraduate Mathematics Series. Springer London, London, 1st ed. 2007. edition, 2007. ISBN 1-84628-627-1.

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

[206] Zizi Papacharissi. Affective publics and structures of storytelling: sentiment, events and mediality. *Information, Communication & Society*, 19(3):307–324, 2016. doi: 10.1080/1369118X.2015.1109697. URL https://doi.org/10.1080/1369118X.2015.1109697.

[207] Charis Papaevangelou. Funding Intermediaries: Google and Facebook's Strategy to Capture Journalism. *Digital Journalism*, 12(2):234–255, February 2024. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2022.2155206. URL https://www.tandfonline.com/doi/full/10.1080/21670811.2022.2155206.

[208] Ioanna Papageorgiou and Carlos Mougan. Necessity of processing sensitive data for bias detection and monitoring: A techno-legal exploration. In *NeurIPS 2023 Workshop on Regulatable ML*, 2023.

[209] Javier Parapar and Filip Radlinski. Towards unified metrics for accuracy and diversity for recommender systems. In *Fifteenth ACM Conference on Recommender Systems*, pages 75–84, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582. URL https://doi.org/10.1145/3460231.3474234.

[210] Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R Curran, and Irena Koprinska. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999, 2013.

[211] Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

[212] Christine Pinney, Amifa Raj, Alex Hanna, and Michael D Ekstrand. Much ado about gender: Current practices and future recommendations for appropriate gender-aware information access. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 269–279, 2023.

[213] Markus Prior. The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure. *Public Opinion Quarterly*, 73(1):130–143, 2009. ISSN 1537-5331, 0033-362X. doi: 10.1093/poq/nfp002. URL https://academic.oup.com/poq/article-lookup/doi/10.1093/poq/nfp002.

[214] Shameem A. Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. A coverage-based approach to recommendation diversity on similarity graph. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 15–22, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340359. doi: 10.1145/2959100.2959149. URL https://doi.org/10.1145/2959100.2959149.

[215] Ying Qiao, Jiun-Hung Chen, Winnie Wu, Fangzhao Wu, Chuhan Wu, Tao Qi, Yingwei Yi, Ling Luo, and Xing Xie. MIND News Recommendation Competition, July 2020. URL https://competitions.codalab.org/competitions/24122#learn_the_details-task.

[216] Lijing Qin and Xiaoyan Zhu. Promoting diversity in recommendation by entropy regularizer. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 2698–2704. AAAI Press, 2013. ISBN 9781577356332.

[217] S. Raza and C. Ding. News recommender system: a review of recent progress, challenges, and opportunities. *Artificial Intelligence Review*, 55(1):749–800, 2022. doi: 10.1007/s10462-021-10043-x. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85111138409&doi=10.1007%2fs10462-021-10043-x&partnerID=40&md5=dd829ce34f43bf34661b3217be457d4e.

[218] Shaina Raza and Chen Ding. Deep neural network to tradeoff between accuracy and diversity in a news recommender system. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5246–5256. IEEE, 2021. doi: 10.1109/BigData52589.2021.9671467.

[219] L. Ren, D. Chen, Z. Miao, and H. Zhang. News Recommendation Model Based on Long-Term and Short-Term Interests. pages 183–189, 2022. doi: 10.1145/3565291.3565321. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85145874118&doi=10.1145%2f3565291.3565321&partnerID=40&md5=daace2a022942eff6d404a50485f520b.

[220] Myrthe Reuver, Nicolas Mattis, Marijn Sax, Suzan Verberne, Nava Tintarev, Natali Helberger, Judith Moeller, Sanne Vrijenhoek, Antske Fokkens, and Wouter van Atteveldt. Are we human, or are we users? The role of natural language processing in human-centric news recommenders that nudge users to diverse content. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 47–59, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nlp4posimpact-1.6. URL https://aclanthology.org/2021.nlp4posimpact-1.6.

[221] Anna Marie Rezk, Auste Simkute, Ewa Luger, John Vines, Chris Elsden, Michael Evans, and Rhianne Jones. Agency Aspirations: Understanding Users' Preferences And Perceptions Of Their Role In Personalised News Curation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–16, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642634. URL https://dl.acm.org/doi/10.1145/3613904.3642634.

[222] Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer, 2011.

[223] Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. This item might reinforce your opinion: Obfuscation and labeling of search results to mitigate confirmation bias. In *Proceedings of the 32nd ACM conference on hypertext and social media*, pages 189–199, 2021.

[224] James A Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1): 145, 2003.

[225] Alan Said, Maria Pera, and Michael Ekstrand. *We're Still Doing It (All) Wrong: Recommender Systems, Fifteen Years Later*. September 2025. doi: 10.48550/arXiv.2509.09414.

[226] Tetsuya Sakai and Zhaohao Zeng. Which diversity evaluation measures are "good"? In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 595–604, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729. doi: 10.1145/3331184.3331215. URL https://doi.org/10.1145/3331184.3331215.

[227] Marijn Sax. Algorithmic News Diversity and Democratic Theory: Adding Agonism to the Mix. *Digital Journalism*, 10(10):1650–1670, November 2022. ISSN 2167-0811. doi: 10.1080/21670811.2022.2114919. URL https://doi.org/10.1080/21670811.2022.2114919. _eprint: https://doi.org/10.1080/21670811.2022.2114919.

[228] A. Schjøtt and J. Møller Hartley. *Pursuing the Promises of Personalisation: Fetish and Friction in Futures of Automated News*. BerlinDe Gruyter, 2024. ISBN 978-3-11-079224-9. doi: 10.1515/9783110792256-015. URL https://dare.uva.nl/search?identifier=b130c093-2ba5-460f-85bc-f21783cce661;startDoc=1.

[229] Anna Schjøtt Hansen and Jannie Møller Hartley. Designing What's News: An Ethnography of a Personalization Algorithm and the Data-Driven (Re)Assembling of the News. *Digital Journalism*, 11(6):924–942, July 2023. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2021.1988861. URL https://www.tandfonline.com/doi/full/10.1080/21670811.2021.1988861.

[230] Andreas RT Schuck and Alina Feinholdt. News framing effects and emotions. *Emerging Trends in the Social and Behavioral Sciences: an Interdisciplinary, Searchable, and Linkable Resource*, pages 1–15, 2015.

[231] Nick Seaver. Captivating algorithms: Recommender systems as traps. *Journal of Material Culture*, 24(4): 421–436, December 2019. ISSN 1359-1835. doi: 10.1177/1359183518820366. URL https://doi.org/10.1177/1359183518820366.

[232] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 59–68, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287598. URL https://dl.acm.org/doi/10.1145/3287560.3287598.

[233] Faisal Shehzad and Dietmar Jannach. Everyone'sa winner! on hyperparameter tuning of recommendation models. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 652–657, 2023.

[234] K. Shivaram, P. Liu, M. Shapiro, M. Bilgic, and A. Culotta. Reducing Cross-Topic Political Homogenization in Content-Based News Recommendation. pages 220–228, 2022. doi: 10.1145/3523227.3546782. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85139563808&doi=10.1145%2f3523227.3546782&partnerID=40&md5=562ce76c1a7bfab39c9a8404628085d5.

[235] Felix M. Simon. Uneasy Bedfellows: AI in the News, Platform Companies and the Issue of Journalistic Autonomy. *Digital Journalism*, 10(10):1832–1854, November 2022. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2022.2063150. URL https://www.tandfonline.com/doi/full/10.1080/21670811.2022.2063150.

[236] Felix M. Simon. Artificial Intelligence in the News: How AI Retools, Rationalizes, and Reshapes Journalism and the Public Arena. 2024. doi: 10.7916/ncm5-3v06. URL https://doi.org/10.7916/ncm5-3v06.

[237] Felix M. Simon. Escape Me If You Can: How AI Reshapes News Organisations' Dependency on Platform Companies. *Digital Journalism*, 12(2):149–170, February 2024. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2023.2287464. URL https://www.tandfonline.com/doi/full/10.1080/21670811.2023.2287464.

[238] Jane B Singer. User-generated visibility: Secondary gatekeeping in a shared media space. *New Media & Society*, 16(1):55–73, 2014. doi: 10.1177/1461444813477833. URL https://doi.org/10.1177/1461444813477833.

[239] Manel Slokom, Alan Hanjalic, and Martha Larson. Towards user-oriented privacy for recommender system data: A personalization-based approach to gender obfuscation for user profiles. *Information Processing & Management*, 58(6):102722, November 2021. ISSN 0306-4573. doi: 10.1016/j.ipm.2021.102722. URL https://www.sciencedirect.com/science/article/pii/S0306457321002065.

[240] Annelien Smets. Intended, afforded, and experienced serendipity: overcoming the paradox of artificial serendipity. *Ethics and Information Technology*, 27(3):33, June 2025. ISSN 1572-8439. doi: 10.1007/s10676-025-09841-6. URL https://doi.org/10.1007/s10676-025-09841-6.

[241] Annelien Smets, Jonathan Hendrickx, and Pieter Ballon. We're in This Together: A Multi-Stakeholder Approach for News Recommenders. *Digital Journalism*, November 2022. ISSN 2167-0811. URL https://www.tandfonline.com/doi/abs/10.1080/21670811.2021.2024079.

[242] Barry Smyth and Paul McClave. Similarity vs. Diversity. In David W. Aha and Ian Watson, editors, *Case-Based Reasoning Research and Development*, pages 347–361, Berlin, Heidelberg, 2001. Springer. ISBN 978-3-540-44593-7. doi: 10.1007/3-540-44593-5_25.

[243] Pia Sommerauer, Giulia Rambelli, and Tommaso Caselli. Simulating Identity, Propagating Bias: Abstraction and Stereotypes in LLM-Generated Text. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19812–19831, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1080. URL https://aclanthology.org/2025.findings-emnlp.1080/.

[244] Gaurav Sood and Suriyan Laohaprapanon. Predicting race and ethnicity from the sequence of characters in a name. *arXiv preprint arXiv:1805.02109*, 2018.

[245] Robert Spano. The concept of media pluralism under the European Convention on Human Rights – substantive principles and procedural safeguards. *Journal of Media Law*, 15(2):168–178, July 2023. ISSN 1757-7632, 1757-7640. doi: 10.1080/17577632.2024.2305419. URL https://www.tandfonline.com/doi/full/10.1080/17577632.2024.2305419.

[246] Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, 58(3):102505, May 2021. ISSN 03064573. doi: 10.1016/j.ipm.2021.102505. URL https://linkinghub.elsevier.com/retrieve/pii/S0306457321000157.

[247] Harald Steck. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, pages 154–162, New York, NY, USA, September 2018. Association for Computing Machinery. ISBN 978-1-4503-5901-6. doi: 10.1145/3240323.3240372. URL https://dl.acm.org/doi/10.1145/3240323.3240372.

[248] Daniel Steel, Sina Fazelpour, Kinley Gillette, Bianca Crewe, and Michael Burgess. Multiple diversity concepts and their ethical-epistemic implications. *European Journal for Philosophy of Science*, 8(3):761–780, 2018.

[249] Marc Steen. Co-Design as a Process of Joint Inquiry and Imagination. *Design Issues*, 29(2):16–28, April 2013. ISSN 0747-9360, 1531-4790. doi: 10.1162/DESI_a_00207. URL https://direct.mit.edu/desi/article/29/2/16-28/69111.

[250] Elias Storms, Oscar Alvarado, and Luciana Monteiro-Krebs. 'Transparency is Meant for Control' and Vice Versa: Learning from Co-designing and Evaluating Algorithmic News Recommenders. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2):405:1–405:24, November 2022. doi: 10.1145/3555130. URL https://dl.acm.org/doi/10.1145/3555130.

[251] Jonathan Stray. Editorial values for news recommenders: Translating principles to engineering. In *News Quality in the Digital Age*. Routledge, 2023. ISBN 978-1-003-25799-8. Num Pages: 15.

[252] Jesper Strömbäck. In search of a standard: four models of democracy and their normative implications for journalism. *Journalism Studies*, 6(3):331–345, 2005. doi: 10.1080/14616700500131950.

[253] Emily Sullivan, Dimitrios Bountouridis, Jaron Harambam, Shabnam Najafian, Felicia Loecherbach, Mykola Makhortykh, Domokos Kelen, Daricia Wilkinson, David Graus, and Nava Tintarev. Reading News with a Purpose: Explaining User Profiles for Self-Actualization. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 241–245, Larnaca Cyprus, June 2019. ACM. ISBN 978-1-4503-6711-0. doi: 10.1145/3314183.3323456. URL https://dl.acm.org/doi/10.1145/3314183.3323456.

[254] Jannick Kirk Sørensen. Public Service Media, Diversity and Algorithmic Recommendation: RecSys 2019: 13th ACM Conference on Recommender Systems. *CEUR Workshop Proceedings*, 2554:6–11, 2019. ISSN 1613-0073.

[255] Yan-Martin Tamm, Rinchin Damdinov, and Alexey Vasilev. Quality metrics in recommender systems: Do we calculate metrics consistently? In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 708–713, 2021.

[256] Niek Tax, Sander Bockting, and Djoerd Hiemstra. A cross-benchmark comparison of 87 learning to rank methods. *Information Processing & Management*, 51(6):757–772, 2015. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2015.07.002. URL https://www.sciencedirect.com/science/article/pii/S0306457315000862.

[257] Ori Tenenboim and Akiba A Cohen. What prompts users to click and comment: A longitudinal study of online news. *Journalism*, 16(2):198–217, 2015.

[258] Guusje Thijs, Damian Trilling, and Anne C. Kroon. Contextualized Word Embeddings Expose Ethnic Biases in News. In *Proceedings of the 16th ACM Web Science Conference*, WEBSCI '24, pages 290–295, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 979-8-4007-0334-8. doi: 10.1145/3614419.3643994. URL https://dl.acm.org/doi/10.1145/3614419.3643994.

[259] Joy A. Thomas Thomas M. Cover. *Elements of Information Theory*. Wiley, 1991.

[260] N. Thurman and S. Schifferes. The future of personalisation at news websites: Lessons from a longitudinal study. *Journalism Studies*, 13(5-6), March 2012. doi: 10.1080/1461670X.2012.664341.

[261] Neil Thurman. Computational journalism. In *The handbook of journalism studies*, pages 180–195. Routledge, 2019.

[262] Neil Thurman, Judith Moeller, Natali Helberger, and Damian Trilling. My Friends, Editors, Algorithms, and I. *Digital Journalism*, 7(4):447–469, April 2019. ISSN 2167-0811. doi: 10.1080/21670811.2018.1493936. URL https://doi.org/10.1080/21670811.2018.1493936. _eprint: https://doi.org/10.1080/21670811.2018.1493936.

[263] Nava Tintarev, Bart P. Knijnenburg, and Martijn C. Willemsen. Measuring the benefit of increased transparency and control in news recommendation. *AI Magazine*, 45(2):212–226, 2024. ISSN 2371-9621. doi: 10.1002/aaai.12171. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12171. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.12171.

[264] Damian Trilling. Dit vind je vast ook leuk: Waarom een veranderende journalistiek een creatievere journalistiekwetenschap nodig heeft. 2025. URL https://research.vu.nl/files/434591138/Rede_Trilling_digitaal_compleet.pdf.

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

[266] Jan Van Cuilenburg. On competition, access and diversity in media, old and new: Some remarks for communications policy in the information age. *New media & society*, 1(2):183–207, 1999.

[267] Alje van Dam. Diversity and its decomposition into variety, balance and disparity. *Royal Society open science*, 6(7):190452, 2019.

[268] Lawrence Van den Bogaert, David Geerts, and Jaron Harambam. Putting a Human Face on the Algorithm: Co-Designing Recommender Personae to Democratize News Recommender Systems. *Digital Journalism*, 12(8):1097–1117, September 2024. ISSN 2167-0811. doi: 10.1080/21670811.2022.2097101. URL https://doi.org/10.1080/21670811.2022.2097101. _eprint: https://doi.org/10.1080/21670811.2022.2097101.

[269] Max Van Drunen. Editorial independence in an automated media system. *Internet Policy Review*, 10(3), September 2021. ISSN 2197-6775. doi: 10.14763/2021.3.1569. URL https://policyreview.info/articles/analysis/editorial-independence-automated-media-system.

[270] Karin Van Es and Dennis Nguyen. Balancing Needs and Values: A Multi-Stakeholder Examination of Algorithmic News Recommenders in the Netherlands. *Journalism Practice*, pages 1–19, August 2024. ISSN 1751-2786, 1751-2794. doi: 10.1080/17512786.2024.2392654. URL https://www.tandfonline.com/doi/full/10.1080/17512786.2024.2392654.

[271] Martijn Van Zomeren, Russell Spears, Agneta H Fischer, and Colin Wayne Leach. Put your money where your mouth is! explaining collective action tendencies through group-based anger and group efficacy. *Journal of personality and social psychology*, 87(5):649, 2004.

[272] Hanne Vandenbroucke. *Not One News Recommender To Fit Them All: How Different Rec-ommender Strategies Serve Various User Segments*. September 2025. doi: 10.1145/3705328.3748034.

[273] Saúl Vargas and Pablo Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 109–116, Chicago Illinois USA, October 2011. ACM. ISBN 978-1-4503-0683-6. doi: 10.1145/2043932.2043955. URL https://dl.acm.org/doi/10.1145/2043932.2043955.

[274] S. Vercoutere, G. Joris, T. De Pessemier, and L. Martens. Improving selection diversity using hybrid graph-based news recommenders. *User Modeling and User-Adapted Interaction*, 34(4):955–993, 2024. doi: 10.1007/s11257-024-09399-w. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195680782&doi=10.1007%2fs11257-024-09399-w&partnerID=40&md5=8838848c3ae86b6c6e04c800498453b7.

[275] Judith Vermeulen. Access Diversity Through Online News Media and Public Service Algorithms: An Analysis of News Recommendation in Light of Article 10 ECHR. In James Meese and Sara Bannerman, editors, *The Algorithmic Distribution of News: Policy Responses*, pages 269–287. Springer International Publishing, Cham, 2022. ISBN 978-3-030-87086-7. doi: 10.1007/978-3-030-87086-7_14. URL https://doi.org/10.1007/978-3-030-87086-7_14.

[276] Judith Vermeulen. To Nudge or Not to Nudge: News Recommendation as a Tool to Achieve Online Media Pluralism. *Digital Journalism*, 10(10):1671–1690, November 2022. ISSN 2167-0811. doi: 10.1080/21670811.2022.2026796. URL https://doi.org/10.1080/21670811.2022.2026796. _eprint: https://doi.org/10.1080/21670811.2022.2026796.

[277] Sanne Vrijenhoek. Do You MIND? Reflections on the MIND Dataset for Research on Diversity in News Recommendations. In Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo, editors, *Advances in Bias and Fairness in Information Retrieval*, pages 147–154, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-37249-0.

[278] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. Recommenders with a Mission: Assessing Diversity in News Recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, pages 173–183, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8055-3. doi: 10.1145/3406522.3446019. URL https://dl.acm.org/doi/10.1145/3406522.3446019.

[279] Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten De Rijke. RADio – Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 208–219, Seattle WA USA, September 2022. ACM. ISBN 978-1-4503-9278-5. doi: 10.1145/3523227.3546780. URL https://dl.acm.org/doi/10.1145/3523227.3546780.

[280] Sanne Vrijenhoek, Lien Michiels, Johannes Kruse, Alain Starke, Nava Tintarev, and Jordi Viader Guerrero. Normalize: The first workshop on normative design and evaluation of recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 1252–1254, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702419. doi: 10.1145/3604915.3608757. URL https://doi.org/10.1145/3604915.3608757.

[281] Sanne Vrijenhoek, Savvina Daniil, Jorden Sandel, and Laura Hollink. Diversity of What? On the Different Conceptualizations of Diversity in Recommender Systems. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 573–584, Rio de Janeiro Brazil, June 2024. ACM. doi: 10.1145/3630106.3658926. URL https://dl.acm.org/doi/10.1145/3630106.3658926.

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

[283] Isaac Waller and Ashton Anderson. Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. In *The World Wide Web Conference*, WWW '19, pages 1954–1964, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313729. URL https://doi.org/10.1145/3308558.3313729.

[284] Ningxia Wang and Li Chen. User bias in beyond-accuracy measurement of recommendation algorithms. In *Fifteenth ACM Conference on Recommender Systems*, pages 133–142, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582. URL https://doi.org/10.1145/3460231.3474244.

[285] Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. Clicks can be Cheating: Counterfactual Recommendation for Mitigating Clickbait Issue. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1288–1297, Virtual Event Canada, July 2021. ACM. ISBN 978-1-4503-8037-9. doi: 10.1145/3404835.3462962. URL https://dl.acm.org/doi/10.1145/3404835.3462962.

[286] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.

[287] Hilde Weerts, Raphaële Xenidis, Fabien Tarissan, Henrik Palmer Olsen, and Mykola Pechenizkiy. Algorithmic Unfairness through the Lens of EU Non-Discrimination Law: Or Why the Law is not a Decision Tree. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 805–816, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 979-8-4007-0192-4. doi: 10.1145/3593013.3594044. URL https://dl.acm.org/doi/10.1145/3593013.3594044.

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

[289] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis – Violin plot*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org/reference/geom_violin.html.

[290] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. Neural news recommendation with attentive multi-view learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3863–3869. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/536. URL https://doi.org/10.24963/ijcai.2019/536.

[291] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. Npa: Neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2576–2584, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330665. URL https://doi.org/10.1145/3292500.3330665.

[292] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. Neural News Recommendation with Multi-Head Self-Attention. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6389–6394, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1671. URL https://aclanthology.org/D19-1671/.

[293] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. MIND: A Large-scale Dataset for News Recommendation. *ACL*, 2020.

[294] Shangyuan Wu, Edson C. Tandoc, and Charles T. Salmon. Journalism reconfigured: Assessing human-machine relations and the autonomous power of automation in news production. *Journalism Studies*, 20 (10):1440 – 1457, 2019. ISSN 1461670X. URL https://search-ebscohost-com.proxy.uba.uva.nl/login.aspx?direct=true&db=ufh&AN=137190691&site=ehost-live&scope=site.

[295] Wen Wu, Li Chen, and Yu Zhao. Personalizing recommendation diversity based on user personality. *User Modeling and User-Adapted Interaction*, 28(3):237–276, 2018.

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

[297] Zhen Yang. How The New York Times incorporates editorial judgment in algorithms to curate its home page. URL https://www.niemanlab.org/2024/10/how-the-new-york-times-incorporates-editorial-judgment-in-algorithms-to-curate-its-home-page/.

[298] Emine Yilmaz and Stephen Robertson. On the choice of effectiveness measures for learning to rank. *Information Retrieval*, 13(3):271–290, June 2010. ISSN 1386-4564. doi: 10.1007/s10791-009-9116-x.

[299] Iris Marion Young. Communication and the other: Beyond deliberative democracy. *Democracy and difference: Contesting the boundaries of the political*, 31:120–135, 1996.

[300] 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. Apache spark: A unified engine for big data processing. *Commun. ACM*, 59(11): 56–65, oct 2016. ISSN 0001-0782. doi: 10.1145/2934664. URL https://doi.org/10.1145/2934664.

[301] John Zaller. A new standard of news quality: Burglar alarms for the monitorial citizen. *Political Communication*, 20(2):109–130, 2003.

[302] Eva Zangerle and Christine Bauer. Evaluating Recommender Systems: Survey and Framework. *ACM Comput. Surv.*, 55(8):170:1–170:38, December 2022. ISSN 0360-0300. doi: 10.1145/3556536. URL https://dl.acm.org/doi/10.1145/3556536.

[303] Mi Zhang and Neil Hurley. Avoiding monotony: Improving the diversity of recommendation lists. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, page 123–130, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580937. doi: 10.1145/1454008.1454030. URL https://doi.org/10.1145/1454008.1454030.

[304] Jieming Zhu, Quanyu Dai, Liangcai Su, Rong Ma, Jinyang Liu, Guohao Cai, Xi Xiao, and Rui Zhang. Bars: Towards open benchmarking for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2912–2923, 2022.

[305] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32, 2005.

[306] Frederik Zuiderveen Borgesius, Damian Trilling, Judith Möller, Balázs Bodó, Claes H De Vreese, and Natali Helberger. Should we worry about filter bubbles? *Internet Policy Review. Journal on Internet Regulation*, 5 (1), 2016.

Summary

News recommender systems aim to predict which news items their users would like to read based on their past reading behaviour. However, rather than only catering to a readers' preferences, a *diverse* recommender system could also be used to expand a reader's world view, to help them be more informed, or to expose them to events and ideas they were not aware of before. This thesis aims to answer the question: *How can we evaluate news recommender systems on their normative diversity?*

In **Chapter 2** we show through interviews with public service media professionals that there is not one standardized interpretation of what the term diversity means. We created a mapping of the way the interview participants talked about diversity when constructing a recommendation based on a set of candidate items from their own catalogue. The mapping consists of two components: the *aspects* that were considered for diversification, and the *tactics* that were employed to diversify the recommendation.

The lack of a standardized definition aligns with normative theory: Helberger [111] noted that different democratic models (Liberal, Participatory, Deliberative, Critical) imply different desired forms of diversity, which can conflict and cannot be optimized for simultaneously. To bridge the abstract concepts underpinning the democratic models and technology design, **Chapter 3** proposes a set of diversity evaluation metrics: *Calibration* (to what extent is the recommendation tailored to a reader's preferences), *Fragmentation* (to what extent are readers aware of the same issues and events), *Activation* (is the tone of the recommendation affective or neutral), *Representation* (which viewpoints are included in the recommendation) and *Alternative Voices* (to what extent are minority groups included in the recommendation). Different combinations and expected value ranges of metrics reflect the different democratic models. Through their foundation in democratic theory the metrics give direction to the problem of defining diversity in a specific context: by answering *why* we care about diversity, we define *what* the recommender system should do.

The mapping and proposed metrics can be reconciled into a base mathematical formalization as a rank-aware divergence score, which is explained in **Chapter 4**. With that, defin-

ing a diversity metric comes down to answering the questions *what aspect should be measured*, *what target distribution should the recommendation be compared to*, *what value is intended* and *should the divergence score account for the rank of an article in the recommendation or not*. The operationalization thus provides a common frame of reference for interpretation, without imposing restrictions on the type of diversity that can be expressed.

Despite the large interest in developing diversity-aware news recommender systems, a significant gap remains between theoretical, normative literature and what has been developed in practice. **Chapter 5** argues that the public datasets available to researchers constrain and shape the research that can be executed. While each of the most frequently used public datasets has unique characteristics suitable to working on a sub-problem of diversity-aware recommendation, the time- and context dependent nature of news requires a paradigm shift away from one-off static datasets and toward ongoing collaboration and data-sharing with news organizations.

In order for the metrics to be adopted they need to land within news organizations' operations and practices. **Chapter 6** explains a workshop series that brought together different stakeholder groups within a Danish national news organization to collectively define evaluation metrics for their news recommender systems. The study highlights the challenges of operationalizing abstract concepts underpinning journalistic and editorial values and emphasizes the importance of iterative refinement and stakeholder collaboration, and highlights that ongoing dialogue and appropriate internal workflows are necessary to ensure the system both meets editorial standards and is technically feasible.

The limitations of the thesis include its lack of involvement of news readers themselves and focus on Western-European values. Furthermore, through their theoretical nature the actual behavioural and democratic impact of the metrics remains unclear. Future research should prioritize developing algorithms that put the notions developed in the thesis into practice, and test those with both the newsroom and readers themselves.

Nederlandse samenvatting

Nieuwsaanbevelingssystemen voorspellen welke nieuwsartikelen hun gebruikers willen lezen op basis van met name hun eerdere leesgedrag. Echter zou een nieuwsaanbevelingssysteem dat naast de voorkeuren van lezers ook de *diversiteit* van het aanbod meeneemt ook gebruikt kunnen worden om het wereldbeeld van lezers te verbreden, hen beter te informeren, of hen op de hoogte te brengen van nieuwe gebeurtenissen en ideeën. Dit proefschrift beantwoordt daarom de volgende vraag: *Hoe kunnen we nieuwsaanbevelingssystemen beoordelen op hun normatieve diversiteit?*

In **Hoofdstuk 2** laten we aan de hand van interviews met professionals van publieke mediabedrijven zien dat er geen eenduidige, gestandaardiseerde interpretatie van diversiteit mogelijk is. We brachten in kaart (mapping) hoe de geïnterviewden over diversiteit spraken wanneer zij een aanbeveling samenstelden op basis van een set kandidaatitems uit hun eigen catalogus. We identificeerden twee componenten: de *aspecten* die werden meegenomen voor diversificatie, en de *tactieken* die werden ingezet om de aanbeveling te diversificeren.

Het feit dat er geen gestandaardiseerde definitie van diversiteit mogelijk is sluit aan bij normatieve theorie: Helberger [111] stelt dat verschillende democratische modellen (liberaal, participatief, deliberatief, kritisch) verschillende vormen van diversiteit impliceren, en dat die vormen haaks op elkaar kunnen staan en dus niet gelijktijdig gemaximaliseerd kunnen worden. Om de stap te maken van de abstracte concepten die aan de democratische modellen ten grondslag liggen naar technologisch ontwerp, stelt **Hoofdstuk 3** een set nieuwe evaluatie 'metrics' voor diversiteit voor: *Calibration* (in welke mate is de aanbeveling afgestemd op de voorkeuren van een lezer), *Fragmentation* (in welke mate zijn lezers op de hoogte van dezelfde kwesties en gebeurtenissen), *Activation* (is de toon van de aanbeveling affectief of neutraal), *Representation* (welke standpunten worden in de aanbeveling opgenomen) en *Alternative Voices* (in welke mate worden minderheidsgroepen in de aanbeveling genoemd). Verschillende combinaties en verwachte

waardebereiken van deze metrics reflecteren de verschillende democratische modellen. Door hun basis in democratische theorie geven de metrics richting aan het definiëren van diversiteit in een specifieke context: door te beantwoorden *waarom* we diversiteit belangrijk vinden, bepalen we *wat* het aanbevelingssysteem zou moeten doen.

De mapping en de voorgestelde metrics kunnen worden samengebracht in een wiskundige formalisering als een positie-gevoelige divergentiescore, die wordt uitgewerkt in **Hoofdstuk 4**. Daarmee komt het definiëren van een diversiteitsmetric neer op het beantwoorden van de vragen *welk aspect moet worden gemeten*, *met welke distributie moet de aanbeveling worden vergeleken*, *welke waarde wordt nagestreefd* en *of de divergensiescore al dan niet rekening moet houden met de positie van een artikel in de aanbeveling*. De operationalisering biedt zo een referentiekader voor interpretatie, zonder beperkingen op te leggen aan het type diversiteit dat kan worden uitgedrukt.

Ondanks de grote interesse in het ontwikkelen van diversiteitsbewuste nieuwsaanbevelingssystemen blijft er een groot gat bestaan tussen de theoretische, normatieve literatuur en wat er in de praktijk is ontwikkeld. **Hoofdstuk 5** beargumenteert dat mede door de publieke datasets die voor onderzoekers beschikbaar zijn het onderzoek dat gedaan kan worden beperkt is. Hoewel elk van de meest gebruikte publieke datasets unieke elementen heeft waarmee aan een deelprobleem van diversiteitsbewuste aanbeveling gewerkt kan worden, is door het tijd- en contextafhankelijke karakter van nieuws een verschuiving nodig van statische datasets naar continue samenwerking en datadeling met nieuwsorganisaties.

Voordat de metrics in praktijk kunnen worden gebracht moeten zij opgenomen worden in de bestaande processen binnen nieuwsorganisaties. **Hoofdstuk 6** beschrijft een reeks workshops waarin verschillende stakeholdergroepen binnen een Deense nationale nieuwsorganisatie bij elkaar werden gebracht om gezamenlijk metrics voor hun nieuwsaanbevelingssystemen te definiëren. De studie onderstreept de grote uitdagingen die komen met het operationaliseren van abstracte concepten die ten grondslag liggen aan journalistieke en redactionele waarden. Daarmee beargumenteert het dat het noodzakelijk is om te werken in korte iteraties waarbij er nauwe samenwerking is met de verschillende stakeholders, en laat het zien dat voortdurende dialoog en passende interne werkprocessen cruciaal zijn om ervoor te zorgen dat het systeem zowel aan redactionele normen voldoet als technisch haalbaar is.

De limitaties van dit proefschrift zijn het het feit dat de perspectieven van nieuwslezers zelf geen expliciet onderdeel zijn geweest van het onderzoek en dat de focus ligt op West-Europese waarden. Bovendien blijft door het theoretische karakter van het onderzoek de daadwerkelijke effectiviteit en democratische impact van de metrics onduidelijk. Toekom-

stig onderzoek zou prioriteit moeten geven aan het ontwikkelen van algoritmen die de in dit proefschrift ontwikkelde concepten in de praktijk brengen, en deze vervolgens zowel bij de redactie als bij lezers zelf te testen.

Publications

Papers included in dissertation

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 (CHIIR '21).* Association for Computing Machinery, New York, NY, USA, 173–183. https://doi.org/10.1145/3406522.3446019

Sanne Vrijenhoek*, Gabriel Bénédict*, Mateo Gutierrez Granada, Daan Odijk, and Maarten De Rijke (2022). "RADio – Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations". In *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys '22).* Association for Computing Machinery, New York, NY, USA, 208–219. https://doi.org/10.1145/3523227.3546780

Sanne Vrijenhoek, Savvina Daniil, Jorden Sandel, and Laura Hollink (2024). "Diversity of What? On the Different Conceptualizations of Diversity in Recommender Systems". *In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24).* Association for Computing Machinery, New York, NY, USA, 573–584. https://doi.org/10.1145/3630106.3658926

Max van Drunen* and **Sanne Vrijenhoek*** (2026). "How public datasets constrain the development of diversity-aware news recommender systems, and what law could do about it." *Under review*

Sanne Vrijenhoek, Natali Helberger, Laura Hollink and Claes de Vreese (2026). "Values to Metrics: Co-designing evaluation metrics for a personalized news recommender system." *Under review*

*equal contribution.

Additional publications

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**, Theresa Zueger (2020). "Diversity in news recommendations". In *Dagstuhl Reports, Vol. 9, Issue 1, ISSN 2193-2433*.

Reuver, M., Mattis, N., Sax, M., Verberne, S., Tintarev, N., Helberger, N., Moeller, J., **Vrijenhoek, S.**, Fokkens, A., van Atteveldt, W. (2021). "Are we human, or are we users? The role of natural language processing in human-centric news recommenders that nudge users to diverse content". In *Proceedings of the 1ste Workshop on NLP for Positive Impact (August 2021 ed., pp. 47-59). Association for Computational Linguistics*. https://doi.org/10.18653/v1/2021.nlp4posimpact-1.6

Helberger, N., van Drunen, M., **Vrijenhoek, S.**, Möller, J. (2021, Feb 26). "Regulation of news recommenders in the Digital Services Act: empowering David against the Very Large Online Goliath". In *Internet Policy Review*. https://policyreview.info/articles/news/regulation-news-recommenders-digital-services-act-empowering-david-against-very-large

Helberger, N., van Drunen, M., Moeller, J., **Vrijenhoek, S.**, Eskens, S. (2022). "Towards a Normative Perspective on Journalistic AI: Embracing the Messy Reality of Normative Ideals". In *Digital Journalism, 10(10), 1605–1626*. https://doi.org/10.1080/21670811.2022.2152195

Vrijenhoek, S. (2023, April). "Do you MIND? Reflections on the MIND dataset for research on diversity in news recommendations". In *International Workshop on Algorithmic Bias in Search and Recommendation (pp. 147-154)*. Cham: Springer Nature Switzerland.

Vrijenhoek, S. (2024). "RADio-: a simplified codebase for evaluating normative diversity in recommender systems". In *12th International Workshop on News Recommendation and Analytics (INRA)*.

Lucien Heitz, **Sanne Vrijenhoek**, and Oana Inel. 2024. "Recommendations for the Recommenders: Reflections on Prioritizing Diversity in the RecSys Challenge". In *Proceedings of the Recommender Systems Challenge 2024 (RecSysChallenge '24). Association for Computing Machinery, New York, NY, USA, 22–26*. https://doi.org/10.1145/3687151.3687155

Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Chloe Bakalar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, Sara Johansen, Lianne Kerlin, David Vickrey, Spandana Singh, **Sanne Vrijenhoek**, Amy Zhang, McKane Andrus, Natali Helberger, Polina Proutskova, Tanushree Mitra, Nina Vasan, "Building human values into recommender systems: An interdisciplinary synthesis", *ACM Transactions on Recommender Systems, no. 2, issue 3, pp. 1-57, 2024*.

Workshop reports

Makri, S., McKay, D., Buchanan, G., Chang, S., Lewandowski, D., MacFarlane, A., Cole, L., **Vrijenhoek, S**. and Ferraro, A. (2021), "Search a Great Leveler? Ensuring More Equitable Information Acquisition". In *Proceedings of the Association for Information Science and Technology, 58: 613-618*. https://doi.org/10.1002/pra2.511

Alessandro Piscopo, Oana Inel, **Sanne Vrijenhoek**, Martijn Millecamp, and Krisztian Balog (2022). "QUARE: 1st Workshop on Measuring the Quality of Explanations in Recommender Systems". In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22). Association for Computing Machinery, New York, NY, USA, 3478–3481*. https://doi.org/10.1145/3477495.3531699

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, USA, 1252–1254*. https://doi.org/10.1145/3604915.3608757

Alain Starke, **Sanne Vrijenhoek**, Lien Michiels, Johannes Kruse, and Nava Tintarev. 2024. "NORMalize 2024: The Second Workshop on Normative Design and Evaluation of Recommender Systems". In *Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24). Association for Computing Machinery, New York, NY, USA, 1242–1244*.

https://doi.org/10.1145/3640457.3687103

Lien Michiels, **Sanne Vrijenhoek**, Alain D. Starke, Johannes Kruse, and Savvina Daniil. 2025. "NORMalize 2025: The Third Workshop on Normative Design and Evaluation of Recommender Systems". In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25). Association for Computing Machinery, New York, NY, USA, 1386–1388*. https://doi.org/10.1145/3705328.3748002

Author contributions

The main chapters in this thesis are based on publications, as referenced on the first page of each chapter. The contributions of the authors to these publications are listed below.

CHAPTER 2: How PSM professionals in the Netherlands conceptualize diversity

Author	Contribution[†]
Sanne Vrijenhoek	Conceptualization (lead), Funding Acquisition (equal), Methodology (equal), Investigation (equal), Formal Analysis (equal), Writing – Original Draft Preparation (equal), Writing – Review & Editing (equal)
Savvina Daniil	Conceptualization (support), Funding Acquisition (equal), Methodology (equal), Investigation (equal), Formal Analysis (equal), Writing – Original Draft Preparation (equal), Writing – Review & Editing (equal)
Jorden Sandel	Investigation (support)
Laura Hollink	Writing – Review & Editing (equal), Supervision

CHAPTER 3: Democratic theory to inform evaluation

Author	Contribution[†]
Sanne Vrijenhoek	Conceptualization, Methodology, Investigation, Writing - Original Draft Preparation, Writing - Review & Editing
Mesut Kaya	Methodology (support)
Nadia Metoui	Methodology (support)
Judith Möller	Conceptualization, Writing - Review & Editing
Daan Odijk	Conceptualization, Writing - Review & Editing
Natali Helberger	Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing, Supervision

CHAPTER 4: A divergence-based base formalization for diversity

Author	Contribution†
Sanne Vrijenhoek	Conceptualization, Methodology, Investigation, Software, Formal Analysis, Visualization, Writing - Original Draft Preparation, Writing - Review & Editing
Gabriel Bénédict	Conceptualization, Methodology, Investigation, Software, Formal Analysis, Visualization, Writing - Original Draft Preparation, Writing - Review & Editing
Mateo Gutierrez Granada	Conceptualization, Methodology, Investigation, Software, Writing - Review & Editing
Daan Odijk	Writing - Review & Editing, Supervision
Maarten de Rijke	Writing - Review & Editing, Supervision

CHAPTER 5: Public datasets and how they can(not) be used

Author	Contribution†
Max van Drunen	Conceptualization (lead), Legal Analysis, Writing – Original Draft Preparation (equal)
Sanne Vrijenhoek	Conceptualization (support), Technical Analysis, Writing – Original Draft Preparation (equal)

CHAPTER 6: Building metrics through participatory co-design sessions

Author	Contribution†
Sanne Vrijenhoek	Conceptualization, Methodology, Investigation, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing (equal)
Natali Helberger	Writing - Review & Editing (equal), Supervision
Claes de Vreese	Writing - Review & Editing (equal), Supervision
Laura Hollink	Writing - Review & Editing (equal), Supervision

† Following the Contributor Role Taxonomy (CRediT), see https://credit.niso.org/.

Acknowledgements

When I first walked into the Institute for Information Law I had no idea what I was in for. That this book is here in front of you now can be attributed to a good amount of youthful naïveté, a mysterious ability to seem confident when I have no clue what I am doing, and an amazing environment that gave me absolutely every opportunity to succeed.

First and most importantly, I am very grateful to Natali Helberger. Without your vision and ideas about diversity and news recommender systems this would have been a very boring book about topic distances or something. Equally important, you have been an example that being incredibly smart and successful does not need to come at the cost of being kind and empathetic. From the very beginning you made me feel like what I had to say was worth listening to. Whenever academia or life threw a curve-ball, I felt seen and supported, and it is because of that that the university became a place where I felt like I belonged. It goes without saying that my life would have looked very differently if it hadn't been for you (would I still have been a consultant? Scary thought). With the utmost sincerity: thank you.

I am also very grateful to my co-promotors Claes de Vreese and Laura Hollink. Thank you both very much for your input and efforts over the years, often on methodologies that were not your own (or mine, for that matter). Claes, I look forward to the next time my mother calls me because "your supervisor is on tv!" Laura, I am very excited and grateful to be able to continue this work at CWI, together with you and the great team you have built there.

Of course, I should also thank my PhD committee: Balázs Bodó, Robin Burke, David Graus, Sole Pera, Damian Trilling and Suzan Verberne. I am already excited for our discussions during the defense day, and hope for many more in the future.

This journey started in 2018 with a vacancy text that I have often lovingly referred to as 'the vaguest job description of all time' which, looking back, probably had more to do with my lack of skills than with the text itself. During the interview I choked on a piece of chocolate that Natali offered, and I could not really speak for a good five minutes. So, I suppose I should express real and sincere gratitude (and a bit of wonder) to co-interviewers Daan Odijk and Judith Möller, who somehow still saw some potential amidst the coughs.

As a recently converted consultant it was not easy to find my footing within the Law faculty. The fact that the vocabulary of the average young IVIRian was at least twice the size of mine did not help, either. Yet, over the years people that at first I did not understand at all became people whose company I enjoyed very much.
Theresa, if I am grateful for our friendship, I imagine that the owners of Katoen and Crea are even more so. Thank you for all the days and nights seamlessly switching between any intelligent and un-intelligent topic under the sun. To Laurens, I should first apologize for all the times I kept you off work because of yet another issue I wanted to get your thoughts on. Thank you for being such thoughtful and kind academic, but more importantly for being such a thoughtful and kind friend. To Charis, my fellow cat lover and extreme West dweller; thank you for not making me feel stupid when I had to yet again Google the meaning of 'epistemic harm'. Wherever you end up (I hope it's Amsterdam), I hope I can come visit a lot so that you can remind me to respect the coffee. To Marijn, thank you for the memes, and for always making time for me when I needed to rage about something (which was quite often). Also, thank you for being such an inspirational social media presence. To Plixavra, thank you for being the only person that would willingly listen to the stories about my cats, for making me want to eat Greek food during New Year's, and for being there in a difficult moment despite the Dutch language. A special thank-you to Joran, who, when I asked him to proofread *Recommenders with a mission*, pushed me to step out of my comfort zone and propose real, math-y metrics, which is how this train really started moving. Thanks also to Petros, as I hope we will still get to climb mount Olympos sometime; to Max, for opening the door for me on my very first day; to Jill, Naomi and Jef, who taught me that bravery is a muscle that can be trained; to Midas, whose sighs are a work of art; to Agustin, Anh, Anushka, Bengi, Gabi, Giónata, Isabella, João, John, Linda, Ljubiša (I spent a good few minutes finding out how to do š in latex), Magdalena, Mariella, Nathalie, Paddy and Ronan, for all the laughs, discussions and lunchtime conversations. Even if I can no longer call myself the Lost Computer Scientist of the Law Faculty, at least I now know the meaning of the word 'normative'.

At the AI, Media and Democracy Lab I found a group of people that are kind, supportive, curious about other disciplines, and incredibly generous with their own skills and knowledge. With Sara, who patiently and positively anchors the floaty academics to the real world (without you, the Lab never would have worked quite so well); Hannes, who writes entire paragraphs in the time it takes me to form a single coherent thought (and runs at the same pace); Sophie, who selflessly taught two computer scientists how to conduct interviews; Theresa, Laurens and Max, who have been mentioned before but I really just want to emphasize how great they are; Anna, Aqsa, Guillermo, Kimon, Leona, Manel, Nick, Pooja, Stan, Teresa, Tomás, Valeria and Zilin; we may have frozen in the winter, and melted in the summer, but our Tuesdays in the attic of IAS were always my favorite day of the week.

To my friends at CWI, who make me believe there is hope for our field. Savvina, thank you for being the opposite to all the self-importance that is (often) academia. Thank you also for believing in our paper when I started second-guessing. Without you, I never would have fist-bumped a 2 meter tall guy with a very big gun under a partytent at the bottom of a mountain in a favela in Rio de Janeiro. Andrei, Davide, Dirk, Lynda, Manel and Thomas, and now also Benjamin, Clem, Mae, and Yahui. I am incredibly excited and proud to now call you my colleagues.

Beyond the people in Amsterdam I am also incredibly proud to have my extended academic family. To Lien, my Belgian twin: your friendship and support has meant the absolute world to me throughout these years. Between running around the room during NORMalize and our stellar karaoke performances, conferences are so much more fun (not to mention smarter and more competent) with you around. To Alain, our big brother from Brabant, whose talent for pubquizzes is only exceeded by his talent for social media jokes. Thank you for always being willing to steal me a croissant from the breakfast buffet I was too cheap to pay for. I solemnly promise not to call you a computer scientist again. Thanks also to the Belgians, even if I sometimes understand you better in English: Annelien, Kim (yes you are Belgian now), Olivier and Robin; ex-VUers Myrthe and Felicia; Johannes, the madman; co-authors Gabriël and Mateo; Mesut, for introducing me to the Calibration metric; Alessandro, Dana, Stephann, George, James and Kasper, who made me feel like there were people out there that were interested in my work; IRLab friends Ana, Harrie, Jasmin, Mohamed, Pooya, Sam, Sami, Maria and Mariya, many of whom became my first conference friends in SIGIR Madrid. And of course, to Philipp, who really does not complain as much as the stories would have you believe: between the many conferences

we have visited together and almost carrying me through the Kruidvat when I had my concussion, you have seen me at my best and my worst. Thank you for being my friend.

Frankly, it is amazing how many people you meet at conferences, and thus have hung out with on multiple continents, yet you have never been to their houses. It would be impossible to list all my conference friends, given for example the 135 people in the 'Jeunen's dinner party' Whatsapp group. You make academic conferences something I genuinely enjoy and look forward to.

One would almost forget that there is life beyond academia. To Marjolein and Tamara, my Dwazen, because when life breaks down, I know you will be there to pick up the pieces. To all my former volleyballteam- and clubmates at SVU, VVA and Armixtos, who over the years helped me come out of my shell. To Jolijn and Alissa, from the Nachttoernooi to Werchter to the Wild Wall in Jiankou; Giulia, the boss lady who has more style in her pinky finger than I have in my whole body; Lindsay, because we will forever be de Harde Kern, even if we are more Harde Nerds now; and to Green Army Manôu, Esmée, Malou, and of course my favorite running coach Lotte. And to Elza, thank you for bringing uncomplicated fun to everything that you do. From the countless board dinners and meetings we have had, to running during COVID, to silent disco at the Paaspop afterparty, time spent with you never disappoints.

Als laatste ben ik dankbaar voor mijn familie. Ik was niet wie ik was zonder jullie. Voor mama: jij leerde me op mijn eigen benen te staan, maar ook om te luisteren; om jezelf niet groter te maken dan je bent, maar vooral ook niet kleiner. Ik kon gekke dingen doen, zoals een vast contract opzeggen om een jaar als onderzoeksassistent te gaan werken, omdat ik wist dat ik altijd jou had om op terug te vallen. Als mensen zeggen dat wij zo op elkaar lijken, vind ik dat iets om trots op te zijn.
Voor papa, die me leerde dat als iedereen in de sloot springt, dat niet betekent dat ik er ook achteraan moet springen. Droppo sportivo is (gelukkig) geen neurochirurg geworden, maar ik denk graag dat als je nog bij ons was geweest, je vandaag heel trots zou zijn.
Voor Maud, mijn kleine grote zus, wiens verre reizen en avonturen ik twintig jaar later een beetje na probeer te doen; voor Bart, mijn grote kleine broer, want hoe verschillend we ook zijn, je moet toch wel blind zijn om niet te zien dat we broer en zus zijn. Veel liefs ook voor Jule (mag ik binnenkort nog een keer vioolles?), Gido, Sterre, Fenja en Tuur; voor de familie in Venlo: oma Doortje, Margriet, Jeanne, Ad, Bep, Jan, Marjan, Renée, Imke, Bram, Daan, Bob en Joke. Bedankt voor alle keren dat jullie je best hebben gedaan om te volgen waar ik me nou eindelijk mee bezighoud. Ik hoop dat in na al die jaren eindelijk

een verhaal heb kunnen vertellen waar chocola gemaakt van kan worden.

The saying (cliché, really) says that it takes a village, but I think it is safe to say that this book has taken a metropolis. If you have made it all the way here, to the end of this book, and if I were to wax poetic, to the end of this chapter of my life:

thank you.

Sanne, somewhere between April and June 2026