

Recommenders with a Mission: Assessing Diversity in News Recommendations

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

Mesut Kaya
Independent Researcher[1]
mesutt.kayaa@gmail.com

Nadia Metoui
TU Delft[2]
n.metoui@tudelft.nl

Judith Möller
Department of Communication Science, University of Amsterdam
j.e.moller1@uva.nl

Daan Odijk
RTL Nederland B.V.
daan.odijk@rtl.nl

Natali Helberger
Institute for Information Law, University of Amsterdam
n.helberger@uva.nl

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 paper 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.

ACM Reference Format:
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 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '21), March 14–19, 2021, Canberra, ACT, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3406522.3446019

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 [13]. 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 [12, 29, 30, 50]. 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 [4]. 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 [19, 37] for their users, or lock them into routines of "more of the same", or in the most extreme case into so-called filter bubbles [42] 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 [23]. 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 [49], regulators (e.g., EU HLEG [39]), academics, and news organizations themselves [4, 32]. 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 [31]. 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 [6, 53] (see also Section 2), diversity in the normative sense is about larger societal concepts:

[1]This work was done while the author was at TU Delft.
[2]This work was done while the author was at Amsterdam School of Communication Research.



This work is licensed under a Creative Commons Attribution International 4.0 License.

CHIIR '21, March 14–19, 2021, Canberra, Australia.
© 2021 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-8055-3/21/03. https://doi.org/10.1145/3406522.3446019

democracy, freedom of expressions, cultural inclusion, mutual respect and tolerance [19]. There is a mismatch between different theoretical understandings of the construct of diversity, similar to the one observed in Fairness research [22]. 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 2) and from a social science interpretation, including its role in democratic models (Section 3). In Section 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 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.

2 A technical conception of diversity in news recommenders

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 [2].

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" [35]. It also introduces a so-called "confounding bias" [7], 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 [58].

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 A democrative conception of diversity in news recommenders

What becomes apparent from the overview in Section 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. [31].

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 [19]. 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 [9, 10, 25, 48]).

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.

3.1 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"*[48].

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

3.2 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 [20]. Citizens cannot afford to be uninterested in politics because they have an active role to play in helping the community to thrive [48]. 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 [1, 16, 24, 26], 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 [48] and Ferree et al. [15] speak of 'empowerment': to be truly empowering, media content needs to be presented in different forms and styles [8, 15, 57]. 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 The Deliberative model

The participatory and the deliberative models of democracy have much in common (compare Ferree et al. [15]). 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 [15, 25]. 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 [20]. This involves a process of actively comparing and engaging with other also contrary and opposing ideas [34]. 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" [9]. Strömbäck [48] 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 [34]. 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

3.4 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 [26]. 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"[15], 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, reconciliary 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"[56].

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

4 Diversity metrics

The democratic models described in Section 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 [22]. Lastly we mention a number of the limitations of the currently proposed metrics and their operationalizations.

Table 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

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.

4.1.1 Explanation. Calibration is a well-known metric in traditional recommender system literature [47]. 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.

4.1.2 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.

4.1.3 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. [14] and Kim and Oh [28]), 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 [47] 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.

4.1.4 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.

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.

4.2.1 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 create 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 [38], *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.

4.2.2 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.

4.2.3 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 [38] and Trilling and van Hoof [51]. 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. [11], 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. [54]:

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

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 recommendation 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.

4.2.4 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.

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.

4.3.1 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 [52]. But positive emotions play a role too; for example, "joy" elicits the urge to get involved, and "hope" to dream big [17]. The link between emotions, affect and activation is described well by Papacharissi [40]: *"...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.

4.3.2 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.

4.3.3 Operationalization. The Circumplex Model of Affect [43] 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 [21] 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.

4.3.4 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) [44]. 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.

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.

4.4.1 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.

4.4.2 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.

4.4.3 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 [41], 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 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.

4.4.4 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.

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.

4.5.1 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 [46], 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 [27] 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. [18] 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.

4.5.2 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.

4.5.3 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. [5] 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. [5] 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.

4.5.4 Limitations. 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.

5 General limitations

Though all of the metrics described in Section 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 [45], 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 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 [33]. 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.

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

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

	Calibration (topic)	Calibration (style)	Fragmentation	Affect	Representation	Alternative Voices
Liberal	High	High	High	-	-	-
Participatory	Low	High	Low	Medium	Reflective	Medium
Deliberative	-	-	Low	Low	Equal	Medium
Critical	-	-	-	High	Inverse	High

Table 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.

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 [55] and in a real-life setting. We are in contact with multiple media companies, to inform them about the different models of democracy, 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[3].

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, 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 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. [36] 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.

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.

8 Acknowledgments

This research has been supported by the SIDN Fund and Adessium Foundation in project 'Algorithms for freedom of expression and a well-informed public' and ERC Grant 'Profiling and targeting news readers – implications for the democratic role of the digital media, user rights and public information policy' (no. 638514).

References

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

[2] Joeran Beel, Marcel Genzmehr, Stefan Langer, Andreas Nürnberger, and Bela Gipp. 2013. 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).* Association for Computing Machinery, New York, NY, USA, 7–14. https://doi.org/10.1145/2532508.2532511

[3] Abraham Bernstein, Claes de Vreese, Natali Helberger, Wolfgang Schulz, Katharina Zweig, Christian Baden, Michael A Beam, Marc P Hauer, Lucien Heitz, Pascal Jürgens, et al. 2020. Diversity in News Recommendations. *arXiv preprint arXiv:2005.09495* (2020).

[4] Balazs Bodo. 2019. Selling News to Audiences – A Qualitative Inquiry into the Emerging Logics of Algorithmic News Personalization in European Quality News Media. *Digital Journalism* 0, 0 (2019), 1–22. https://doi.org/10.1080/21670811.2019.1624185

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

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

[7] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18).* Association for Computing Machinery, New York, NY, USA, 224–232. https://doi.org/10.1145/3240323.3240370

[8] Clifford Christians. 2006. The media and moral literacy. (2006), 62.

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

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

[11] Tawanna R Dillahunt, Christopher A Brooks, and Samarth Gulati. 2015. 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.* 1851–1856.

[12] Konstantin Nicholas Dörr. 2016. Mapping the field of Algorithmic Journalism. *Digital Journalism* 4, 6 (2016), 700–722. https://doi.org/10.1080/21670811.2015.1096748

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

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

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

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

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

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

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

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

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

[22] Abigail Z Jacobs and Hanna Wallach. 2019. Measurement and fairness. *arXiv preprint arXiv:1912.05511* (2019).

[23] Dietmar Jannach and Michael Jugovac. 2019. Measuring the Business Value of Recommender Systems. *ACM Trans. Manage. Inf. Syst.* 10, 4, Article 16 (Dec. 2019), 23 pages. https://doi.org/10.1145/3370082

[24] Edson C. Tandoc Jr. and Ryan J. Thomas. 2015. The Ethics of Web Analytics. *Digital Journalism* 3, 2 (2015), 243–258. https://doi.org/10.1080/21670811.2014.909122

[25] Kari Karppinen. [n.d.]. Conceptions of democracy in media and communications studies.

[26] Kari Karppinen. 2013. Uses of Democratic Theory in Media and Communication Studies. *Observatorio* 7, 3 (2013), 1–17.

[27] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 88 (Nov. 2018), 22 pages. https://doi.org/10.1145/3274357

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

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

[30] Seth C. Lewis and Oscar Westlund. 2015. Big Data and Journalism. *Digital Journalism* 3, 3 (2015), 447–466. https://doi.org/10.1080/21670811.2014.976418

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

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

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

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

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

[36] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. 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).* Association for Computing Machinery, New York, NY, USA, 2243–2251. https://doi.org/10.1145/3269206.3272027

[37] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity. In *Proceedings of the 23rd International Conference on World Wide Web.* 677–686. https://doi.org/10.1145/2566486.2568012

[38] Tom Nicholls and Jonathan Bright. 2019. Understanding News Story Chains using Information Retrieval and Network Clustering Techniques. *Communication Methods and Measures* 13, 1 (2019), 43–59.

[39] High-Level Expert Group on Artificial Intelligence. 2018. *Draft Ethics guidelines for trustworthy AI.* Technical Report. European Commission.

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

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

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

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

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

[45] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19).* Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

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

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

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

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

[50] N. Thurman and S. Schifferes. 2012. The Future of Personalisation at News Websites: Lessons from a Longitudinal Study. *Journalism Studies* 13, 5-6 (March 2012). https://doi.org/10.1080/1461670X.2012.664341

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

[52] Martijn Van Zomeren, Russell Spears, Agneta H Fischer, and Colin Wayne Leach. 2004. 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 (2004), 649.

[53] Saúl Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 109–116.

[54] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)* 28, 4 (2010), 1–38.

[55] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. *ACL* (2020).

[56] Iris Marion Young. 1996. Communication and the other: Beyond deliberative democracy. *Democracy and difference: Contesting the boundaries of the political* 31 (1996), 120–135.

[57] John Zaller. 2003. A new standard of news quality: Burglar alarms for the monitorial citizen. *Political Communication* 20, 2 (2003), 109–130.

[58] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*. 22–32.