One Recommender Fits All? An Exploration of User Satisfaction With Text-Based News Recommender Systems

Wieland, Mareike; Nordheim, Gerret von; Kleinen-von Königslöw, Katharina

Article

One Recommender Fits All? An Exploration of User Satisfaction With Text-Based News Recommender Systems

Mareike Wieland *, Gerret von Nordheim and Katharina Kleinen-von Königslöw

Institute of Journalism/Media Research, University of Hamburg, Germany;
E-Mails: mareike.wieland@uni-hamburg.de (M.W.), gerret.vonnordheim@uni-hamburg.de (G.v.N.),
katharina.kleinen@uni-hamburg.de (K.K.-v.K.)

* Corresponding author

Abstract

Journalistic media increasingly address changing user behaviour online by implementing algorithmic recommendations on their pages. While social media extensively rely on user data for personalized recommendations, journalistic media may choose to aim to improve the user experience based on textual features such as thematic similarity. From a societal viewpoint, these recommendations should be as diverse as possible. Users, however, tend to prefer recommendations that enable "serendipity"—the perception of an item as a welcome surprise that strikes just the right balance between more similarly useful but still novel content. By conducting a representative online survey with n = 588 respondents, we investigate how users evaluate algorithmic news recommendations (recommendation satisfaction, as well as perceived novelty and unexpectedness) based on different similarity settings and how individual dispositions (news interest, civic information norm, need for cognitive closure, etc.) may affect these evaluations. The core piece of our survey is a self-programmed recommendation system that accesses a database of vectorized news articles. Respondents search for a personally relevant keyword and select a suitable article, after which another article is recommended automatically, at random, using one of three similarity settings. Our findings show that users prefer recommendations of the most similar articles, which are at the same time perceived as novel, but not necessarily unexpected. However, user evaluations will differ depending on personal characteristics such as formal education, the civic information norm, and the need for cognitive closure.

1. Introduction

News recommendations are widespread, not only on large social media platforms but also in journalistic media (Kunert & Thurman, 2019). In a fragmented and rich information environment, algorithm-based recommender systems help users find relevant content (Bernstein et al., 2020). As social media and news aggregators nowadays have become a common way of accessing news, news organizations face pressure to offer a similar user experience to meet users' expectations (Nielsen, 2016). Implementing news recommendation algorithms on their web pages and mobile applications has thus become an integral part of their revenue strategies (Bodó, 2019; Kunert & Thurman, 2019). At the same time, not all news companies may be able, or want, to employ the "data-hungry" personalization strategies of the recommendation systems developed by large tech platforms. For them, recommendation algorithms based exclusively or primarily on content characteristics might be the more useful

option for satisfying the expectations of their users, their normative goals, and even their economic aims.

Combining a prototype for a text-based news recommender system with an online survey representative of German internet users (n = 588), this article explores how satisfactory a text-based news recommendation algorithm is perceived by news users, and whether certain user dispositions might impact this satisfaction, and make it necessary to optimize the content-based news recommendation engine for specific user groups. Based on our results, we propose optimizing recommender systems by capturing specific user characteristics in a targeted (and explicit) way.

2. Potential and Challenges of Text-Based News Recommenders for News Companies

Recommender systems can be described as a personalization, that is, as "a form of user-to-system interactivity that uses a set of technological features to adapt the content, delivery, and arrangement of a communication to individual users' explicitly and/or implicitly determined preferences" (Thurman & Schifferes, 2012, p. 776). Based on this definition, we can first distinguish recommendations based on explicitly expressed user preferences from systems that draw on data implicitly (in reality, we often find hybrid forms; Spangher, 2015). These, in turn, fall in a continuum between user-data and content-data dependency. Each of these forms entails specific dilemmas for journalistic media that seek to personalize their content: Firstly, users have little motivation to provide explicit information about their preferences to improve recommendations (Thurman et al., 2019). Also, this information (such as interests in certain topics) quickly become obsolete (Kunert & Thurman, 2019, p. 762). Secondly, implicit recommendations are often "data hungry" (Head of Product at BBC News Online, 2016, cited in Kunert & Thurman, 2019, p. 777), i.e., they rely on the extensive collection and sharing of user data. Adams (2020) has pointed out that the "audience has been commodified and therefore instrumentalized" (p. 883)—a practice that threatens to undermine the authority of journalism as an institution committed to democratic norms and that is increasingly addressed by regulation authorities with restrictive legislation (such as the EU's General Data Protection Regulation; Eskens, 2019), even though its effectiveness in protecting consumers' data is debatable (Reviglio, 2020). Thirdly, the technologies that facilitate article recommendations are often provided by third parties such as the content aggregator, for example, Outbrain (Kunert & Thurman, 2019, p. 777). As a result, journalistic media are becoming increasingly dependent on platforms whose recommendation technologies are not transparent. The media themselves are in turn becoming wary of this practice of collecting user data and sharing it with third-party vendors (Kunert & Thurman, 2019, p. 777; von Nordheim & Fuchsloch, 2019, p. 254).

Given these problems, it seems an obvious and forward-looking choice for media to develop their own technologies that require little explicit participation by users and little disclosure of personal data, which rely instead on features of the news content. Today, rapid developments in the field of Natural Language Processing (or Natural Language Understanding) make it possible to compute text similarities based on complex language models. As an example of such technologies, we study article recommendations based on text similarities, operationalized by the BERT language model (Bidirectional Encoder Representations from Transformers), which was introduced by Google researchers (Devlin et al., 2018). The language model, pre-trained on Wikipedia and book texts, is used to compute document similarities (and other Natural Language Understanding tasks such as sentiment classification, natural language inference, and question answering) and achieved state-of-the-art accuracy (Wang et al., 2019). BERT is therefore an obvious algorithm for the development of new content-based recommender systems (Wang & Fu, 2020). Thanks to developments such as BERT, even small publishers (or service providers beyond the big advertising platforms, see Section 4.2.) can now create their own software for news recommendations. Just a few years ago, this level of independence involved huge development costs and was therefore only available to big media—Kunert and Thurman (2019) mention the *Financial Times* (p. 775), other examples include the *New York Times* (Spangher, 2015), or the *Washington Post* (Graff, 2015).

Journalistic media that seek to implement this form of text-based recommendation face the challenge that solely optimizing for text similarity satisfies neither the user's appetite for "news" nor the news media's normative aim to present their users' with a certain level of diversity. Such a similarity-based recommender thus needs to be calibrated against user satisfaction, e.g., a positive assessment of relevance and quality of the recommended article in accordance to personal needs (Bodó et al., 2019). User satisfaction, in turn, is assumed to translate into loyalty and trust, thus increasing the value of a news brand in the long term (Nelson & Kim, 2020).

3. User Satisfaction at the Intersection of Pleasurable Comfort and Valuable Diversity

On the one hand, similarity-based recommendations are likely to be evaluated positively, because users are familiar with the recommended topics, views, or facts. As the mere exposure effect (Zajonc, 1968) suggests, people tend to evaluate objects or people better only because they are more familiar with them (Bornstein, 1989). A similarity-based recommendation could thus encourage positive evaluations of the recommended article through repetition of topics, views, or facts and hence ease of processing. This should particularly hold true if

the original article that the similarity-based recommendation is based on is perceived of as being high quality.

On the other hand, it is obvious from the user's point of view that presenting more of the same makes "a perfectly boring, very foreseeable, very cold and technology-driven product, that doesn't feel like a proper journalistic product" (Bodó, 2019, p. 1068). Indeed, journalistic media are particularly regarded for their skill at providing users with a "reliable surprise" based on a wide range of high-quality content (Schoenbach, 2007), in other words, to let them encounter content that is pleasantly unexpected and new but without seeming accidental and arbitrary. Applied to recommendation engines, users might thus aim for serendipity in news recommendations.

"Serendipity" is a common design goal of recommendation engines not only in the context of news (Reviglio, 2019) and defined as the sweet spot, just the right degree, between novelty and unexpectedness (Maccatrozzo et al., 2017). This balance ensures that, although unexpected and new, the recommendation is still perceived as pleasant, enriching, and thus useful, which reflects in user satisfaction (Chen et al., 2019).

In contrast, optimizing recommendations solely in the direction of maximum content diversity is presumably not rated positively by users and could thus even work against the economic interests of publishers (Bernstein et al., 2020). However, a one-sided optimization in the direction of pleasure and convenience through very similar recommendations may in turn quickly lead to a limited range of content and counteract the societal role of journalistic media. Furthermore, narrowing down the selection of articles by only presenting the news a user likes is considered "the wrong path" (Bodó, 2019, p. 1065): "Our goal as a news organization is to inform people about what is happening and there are things that are not always fun" (Bodó, 2019, p. 1065). Thus, aiming for serendipity based on novelty as well as on unexpectedness might even translate into a more diverse news menu by challenging users' viewpoints from time to time.

Our first research question thus explores the relationship between the recommendation based on text similarity and user satisfaction with the recommendation (while controlling for the quality of the original article): How are text similarity, article quality, and overall recommendation satisfaction related (RQ1)? And how are the evaluations of the recommended article as (a) novel, and (b) unexpected and the overall level of recommendation satisfaction related (RQ2)?

The academic discussion of diversity in news is mostly limited to the supply side (Helberger et al., 2015). Still, there are indications that users have different expectations regarding the diversity of a news offering (Nielsen, 2016) and that they prefer different degrees of diversity (Bodó et al., 2019; Helberger et al., 2018). To perceive serendipity as enrichment, users need a certain "mental readiness" (Lutz et al., 2017, p. 1706) to encounter openly new, unexpected information that is

recommended to them. Individuals with a chronic need for cognitive closure (NfcC) generally prefer unambiguous situations and find ambiguity unpleasant (Webster & Kruglanski, 1997). Accordingly, they might benefit more from a recommended news item that is very similar to a previous, already known one.

A preference for algorithmic personalization (Thurman et al., 2019) and the share of algorithmically personalized news on overall news use (Schweiger, et al., 2019), in contrast, could increase satisfaction with article recommendations as both might reflect a higher acceptance of automated news recommendations. Similarly, a high technological affinity (Hampel et al., 2020) might lead to a more playful approach towards interactive online systems, again resulting in a greater mental readiness to encounter recommended news articles (McCay-Peet, 2013).

Differences in civic norms, such as in the duty to keep informed, general news interest, or trust, could influence satisfaction with article recommendations because they express an individually different motivation to engage with the recommended articles. This could lead to a very similar article being perceived as a welcome deepening of the topic. But it could also mean that the recommendation needs to be better, i.e., more tailored to already relatively specific needs and clear expectations of well-informed users who are strongly committed to being informed. Since serendipity does not contribute to satisfaction in the case of a highly purposeful use (Lutz et al., 2017), satisfaction with article recommendations solely based on text similarity might reach its limits here. Finally, sociodemographic characteristics such as age or education might also be influential (Möller et al., 2018).

Conceptualizing serendipity as the central variable of user satisfaction thus incorporates a "liberal-individualistic idea" of diversity (Helberger et al., 2018, p. 195), but it also takes into account the deliberative aspect of being exposed to a variety of different topics, facts, and points of view. Therefore, we assume that, as with diversity expectation, user satisfaction is not a "universal user trait" (Bodó, 2019, p. 208).

We, therefore, ask (RQ3): Which individual dispositions (preference for news personalization, relative share of algorithmically personalized news, technological affinity, NfcC, duty to keep informed, news interest, and trust) influence the relationship between text similarity, article quality, and recommendation satisfaction (RQ3a)? And what is the moderating role of individual dispositions regarding the relationship between recommendation satisfaction and evaluations of the recommendations as novel or unexpected (RQ3b)?

4. Research Design

In the reality of news companies, it is challenging to measure satisfaction with the recommendation beyond the actual click (Bodó, 2019) as additional user surveys are required. In communication research, it is in turn dif-

ficult to simulate realistic article recommendation and study recommendations as to the interaction between algorithms and the user (Loecherbach & Trilling, 2020). Data on user views of real recommendations are accordingly scarce in academia: most are derived from hypothetical instructions.

A unique feature of this study is the integration of a real recommendation engine for news articles into a survey. Even though the advantage of combining web tracking and survey data is clear (Bernstein et al., 2020; Loecherbach & Trilling, 2020), it is mainly the big tech platforms that have taken advantage of it so far (Stray, 2020). It is important to note that the design of the study is exploratory, it is a pilot study. Even though participants were randomly assigned to three different groups for their news recommendations (most similar article, least similar article, or article of random similarity), this is not a classic experimental study. Our analytical strategy aims at exploring and identifying relevant relationships between the different variables as a basis for further study, not at confirming hypotheses. For this reason, we have also retained the similarity score as a metric variable (and not a categorical variable identifying experimental groups).

4.1. Questionnaire Structure

Starting with questions about news usage, an interactive part follows in which the respondents freely search a database of actual news articles. Participants enter a search query that is of interest to them and related to politics, business, or culture in Germany and the world (hereafter depicted as "news"). The search query can consist of any number of terms. We used a search query as a starting point for the news browsing situation rather than a mock-up news webpage with a restricted set of articles as the latter may force participants to select articles on topics they are not interested in. By allowing users to freely select a topic of their choosing, participants are more likely to have a similar baseline of interest in the article on which the recommendation is then based. However, because users have to consciously decide on and type a search query, this overall level of interest in the article presented by the search query is likely to be somewhat higher than in a normal news browsing situation.

Respondents then select one of the multiple search results for further reading. To keep the time requirement reasonable, only articles with a minimum word count of 172 and a maximum of 736 are available. Immediately after reading, participants evaluate the quality of the self-selected article. Afterwards, another article is automatically recommended for further reading, randomly using one of three levels of text similarity (see Section 2). We instructed the participants: "The next click will take you to an article that might be of interest to you as well. This article is recommended to you based on the first article. Please read the article, just as you normally would do."

After reading, they again rate the recommended article. In addition, participants indicate their satisfaction with this recommendation. In order to avoid influencing the participants' response behaviour by preceding questions, for example, about their attitude towards personalization, these personal dispositions are surveyed after the interactive part.

4.2. Recommendation Engine

The recommendation engine was developed by the German start-up LakeTech, with whom we cooperated in this study. The start-up offers publishers the opportunity to integrate proprietary recommendation systems into their websites. The article recommendations are based on the content of the previously selected texts aiming to present similar texts. Similarity is calculated based on vector representations of each article—which are in turn based on the average of the vectors of each sentence (sentence embeddings), calculated with the pre-trained language model BERT developed by Google (Devlin et al., 2018). Thus, quantification of a statistical similarity between all articles is possible and represented as a similarity score (values between 0, *no similarity*, and 1, *identical*). For this study described here, three recommendation logics were implemented: (a) the most similar item is recommended; (b) the least similar; (c) a randomly drawn item. The three recommendation logics were randomly assigned to the participants (see Section 4.3).

To generate a representative news corpus, the URLs of relevant texts from ten different media (see Supplementary Files) were first saved via News API (2021) and scraped in the next step. These 194,167 German news texts from the year 2020 (published between 31 January 2020 and 1 January 2021) were then vectorized.

4.3. Measures

The main dependent variable is "satisfaction with the article recommendation," measured as agreement (5-point Likert scale) on items stating that: (a) the topic; (b) viewpoints; and (c) facts of the article are perceived as pleasant and enriching (e.g., "The second article was recommended to you based on the first article. We are interested in your evaluation of the second article compared to the first. Compared to the first article, I perceived the topic [viewpoints, facts] of the second article as pleasant and enriching"). Agreement on these three items is aggregated into a mean index (Cronbach's α = .89).

As possible independent variables related to the recommendation, we looked at text similarity, article quality, and evaluation of the recommendation. "Similarity score" is calculated using the vectorized articles (values between 0, *minimum*, and 1, *maximum similarity*). For each article in the corpus, the IDs and similarity scores of three other articles were stored as meta-data (the most similar, the least similar, and a randomly drawn article)

as a basis for the random assignment of recommended articles (as described in Section 4.2).

We operationalized the "evaluation of the recommendation" using the two dimensions of serendipity (see Section 3): novelty and unexpectedness. In analogy to the satisfaction measurement, we surveyed perception of how new and how unexpected the topics, viewpoints, and facts in the recommended article were (Haim et al., 2018) in comparison to the first article. Again, we aggregated agreement on these items into mean indices for "novelty" (Cronbach's α = .59) respectively "unexpectedness" (Cronbach's α = .73).

As we assume that recommendation satisfaction will be higher if the original article is rated as being of good quality, we also control for perceived "article quality." It is rated by the respondents, applying journalistic quality criteria previously used by Jungnickel (2011) using pairs of opposites (e.g., balanced, illustrative, comprehensible, trustworthy) on a 7-step scale and aggregated into a mean index (Cronbach's α = .89).

As possible independent variables relating to the individual dispositions of the users, we included the following: "Attitude towards news personalization" is measured with items applied by Bodó et al. (2019), Thurman et al. (2019), and Schweiger et al. (2019), in some cases with slight adjustments. The items relate both to perceived usefulness of news personalization, e.g., "when a news website highlights content that is particularly important to me," and to concerns about possibly (un)balanced ("I worry that personalized news will cause me to miss articles that contradict my views") or incomplete information ("I worry/fear that personalized news will cause me to miss important information") and privacy ("I worry that personalized news will make my privacy more vulnerable"). The mean index calculated from these six items shows good reliability (Cronbach's α = .75).

The "relative share of algorithmically personalized news" was calculated following the measurement of news usage proposed by Schweiger et al. (2019), it can take values from 0 (*no algorithmically personalized news used*) to 1 (*all used news are algorithmically personalized*). For technology affinity, we used three statements from the annual survey on technology attitudes among the German population (Hampel et al., 2020) aggregated into a sufficiently reliable mean index (Cronbach's α = .65).

"Duty to keep informed" was measured using four items proposed by McCombs and Poindexter (1983; α = .65). For "need for cognitive closure," we shortened the scale proposed by Schlink and Walther (2007) to five items as did Schweiger et al. (2019), but we replaced two of the items to provide a more specific reference to diversity (α = .62). All items were measured using five-point Likert scales and aggregated into mean indices. "News interest and news trust" are each single item measurements as used by Thurman et al. (2019). The full questionnaire is available in the Supplementary Files.

4.4. Sample

Findings are based on a sample representing all German-speaking internet users aged 18 and over. Participants were recruited by the online access panel provider Norstat, and cross-sampled according to education and age, as well as by place of residence (federal state), and gender. At the end of the ten-day field period in January 2021, 1,027 finished questionnaires resulted. After correcting for respondents who did not meet our pre-defined quality criteria (non-plausible answers, unrealistic response times, straightlining, incomplete cases), 588 valid cases remained for further analyses.

On average, participants are 48.2 years old. 51.5% identify with the male gender. 53.9% have a low level of formal education (no degree, secondary school diploma), 46.1 are higher educated (A-Levels, bachelor, master, doctorate). Accordingly, the sample offers a good representation of the German online population.

5. Results

To explore the relationships between the different variables as outlined in our research questions, we chose a regression analysis approach with recommendation satisfaction as the dependent variable. The predictors are included block-wise (forced entry), starting with sociodemographic variables (Model 1), evaluations of the recommendation and the original article (Model 2), other individual characteristics (Model 3), and selected interaction effects.

Regression assumptions were tested using plots (Luhmann, 2015). These plots (see the Supplementary Files) are used to verify the correct model specification (nonsystematic distribution, Lowess line parallel to the x-axis in the residuals vs. fitted diagram), to check the normal distribution of the residuals (in the Q-Q plot comparison with the diagonal), the homoscedasticity assumption (in the scale-location diagram unsystematic distribution of the residuals), and to diagnose outliers and influential values (with the residuals vs. leverage diagram using Cook's distance).

5.1. How Are Text Similarity, Article Quality, and Satisfaction With the Recommended Article Related?

Among the sociodemographic variables (Table 1, Model 1), only gender is initially influential. Those who identify themselves as male are a little more satisfied with the article recommendation (b = .17, p < .05). However, education and age do not correlate substantially with recommendation satisfaction. In total, sociodemographic characteristics explain only 1.6% of the variance (F(3,584) = 3.22, p = .02), opening up great explanatory potential for other predictors.

For this reason, all variables evaluating the recommendation or the article were included as the model's second block (Table 1, Model 2), increasing the explained

Table 1. Hierarchical regression predicting recommendation satisfaction.

	Model 1				Model 2				Model 3				Model 4			
	b	b 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	b	b 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	b	b 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	b	b 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]
Independent variables																
Intercept	2.71**	[2.57, 2.84]			2.40**	[2.09, 2.72]			2.61**	[2.28, 2.95]			2.64**	[2.31, 2.97]		
Block 1: Demographic variables																
Gender (1 = male)	.17*	[.01, .34]	.01	[−.01, .02]	.19**	[.05, .32]	.01	[−.00, .02]	.14*	[.01, .28]	.00	[−.00, .01]	.14*	[.00, .27]	.00	[−.00, .01]
Age	−.01*	[−.1, −.00]	.01	[−.01, .02]	−.00	[−.01, .00]	.00	[−.00, .01]	−.00	[−.01, .00]	.00	[−.00, .01]	−.00	[−.01, .00]	.00	[−.00, .01]
Education (1 = high)	−.14	[−.31, .02]	.00	[−.01, .02]	−.23**	[−.36, −.09]	.01	[−.00, .03]	−.24**	[−.38, −.11]	.01	[−.00, .03]	−.25**	[−.39, −.11]	.01	[−.00, .03]
Block 2: Recommendation measures																
Similarity score					.70**	[.28, 1.11]	.01	[−.00, .03]	.72**	[.31, 1.13]	.01	[−.00, .03]	.70**	[.29, 1.10]	.01	[−.00, .03]
Quality article 1					.07*	[.01, .13]	.01	[−.00, .02]	.03	[−.03, .09]	.00	[−.00, .01]	.03	[−.03, .09]	.00	[−.00, .00]
Novelty of recommendation					.67**	[.58, .75]	.28	[.22, .34]	.63**	[.55, .71]	.24	[.18, .29]	.63**	[.54, .71]	.24	[.18, .29]
Unexpectedness of recommendation					−.22**	[−.30, −.14]	.03	[.01, .06]	−.20**	[−.28, −.13]	.03	[.01, .05]	−.20**	[−.27, −.12]	.03	[.01, .05]

Table 1. (Cont.) Hierarchical regression predicting recommendation satisfaction.

	Model 1				Model 2				Model 3				Model 4			
	b	b 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	b	b 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	b	b 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	b	b 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]
Block 3: Individual dispositions																
Attitude towards personalization									.20**	[.11, .29]	.02	[.00, .04]	.20**	[.11, .29]	.02	[.00, .04]
Relative share of algorithmically personalized news									−.19	[−.49, .11]	.00	[−.00, .01]	−.19	[−.49, .10]	.00	[−.00, .01]
Duty to keep informed (DTKI)									−.06	[−.21, .08]	.00	[−.00, .00]	−.08	[−.23, .06]	.00	[−.00, .01]
Need for Cognitive Closure									.12*	[.02, .22]	.01	[−.00, .02]	.12*	[.02, .22]	.01	[−.00, .01]
News interest									.14*	[.02, .26]	.01	[−.00, .02]	.13*	[.01, .25]	.00	[−.00, .01]
News trust									.01	[−.07, .10]	.00	[−.00, .00]	.02	[−.06, .10]	.00	[−.00, .00]
Technical affinity									−.03	[−.11, .06]	.00	[−.00, .00]	−.02	[−.11, .06]	.00	[−.00, .00]
Block 4: Moderating effects																
Unexpectedness of recommendation * DTKI													−.18**	[−.30, −.06]	.01	[−.00, .02]
R^2		R^2 = .016*				R^2 = .351**				R^2 = .387**				R^2 = .396**		
95% CI		[.00, .04]				[.28, .40]				[.31, .43]				[.32, .44]		
ΔR^2						ΔR^2 = .335**				ΔR^2 = .036**				ΔR^2 = .009**		
95% CI						[.27, .40]				[.01, .06]				[−.00, .02]		

Notes: A significant b-weight indicates the semi-partial correlation is also significant; b represents unstandardized regression weights; sr^2 represents the semi-partial correlation squared; LL and UL indicate the lower and upper limits of a confidence interval, respectively; * p < .05; ** p < .01.

variance rises by 33.5 percentage points up to 35.1% for Model 2 ($\Delta R2$ = .335, p < .01; F(7,580) = 44.79, p = .00).

Among the predictors, text similarity has the largest impact (b = .7, p < .01) on recommendation satisfaction, with higher text similarity leading to better ratings. However, since the recommendation is based on similarity, this relationship should prove especially true if the first article is already rated as high quality. Although the evaluation of the first article itself only weakly contributes to recommendation satisfaction (b = .07, p < .05), Figure 1 might indicate a possible moderating effect. For those who already rate the first article better (mean + 1SD; all variables are mean-centred), the visualization suggests a stronger positive correlation between the text similarity used for recommending an article and recommendation satisfaction. A moderation analysis was run to determine whether the interaction between the evaluation of the first article and text similarity significantly predicts recommendation satisfaction. For this, the interacting variables were centred at their mean (using gscale from the jtools package; Long, 2021), then the linear model was fitted and plotted using the interactions package (Long, 2020). Since this interaction (b = .18, p = .4) does not become statistically significant ($\Delta R2$ = .39, F(15, 572) = 24,14, p < .01), the interaction term was not added to the final model as suggested by Hayes and Little (2018, p. 236).

5.2. How are the Evaluations of Article Recommendation and the Level of Recommendation Satisfaction Related?

Recommendation satisfaction is positively related to evaluating the recommended article as novel: If the recommended article presents new topics, perspectives, and/or facts, then the recommendation is perceived as more pleasant and enriching (b = .67, p < .01). By contrast, unexpected topics, viewpoints, and/or facts lead to the recommendation being experienced somewhat less as pleasant and enriching (b = −.22, p < .01)—even when, as in this model, all other factors such as text similarity and perceived quality of the first article are held constant. Apparently, the novelty and the unexpectedness of topics, facts, and/or viewpoints in a recommended article have a different impact on readers' recommendation satisfaction (see also Section 5.3, where we explore this relationship further).

By including the predictors related to article recommendation, formal education now also becomes significant. Users with a higher level of formal education are apparently less satisfied with the article recommendation (b = −.23, p < .01) holding all other predictors constant, which may be explained by their having higher or more specific content expectations (Helberger et al., 2018), and this will be further explored in the following sections.

5.3. Do Individual Dispositions Influence the Relationship Between Text Similarity, Article Quality, and Recommendation Satisfaction?

In a further analytical step, we consider individual dispositions as possible predictors (Table 1, Model 3), complementing the article and recommendation related measures and socio-demographics explored above. Contrary to our expectation, including individual dispositions improves model fit only minimally but still significantly ($\Delta R2$ = .036, p < .01) leading to Model 3
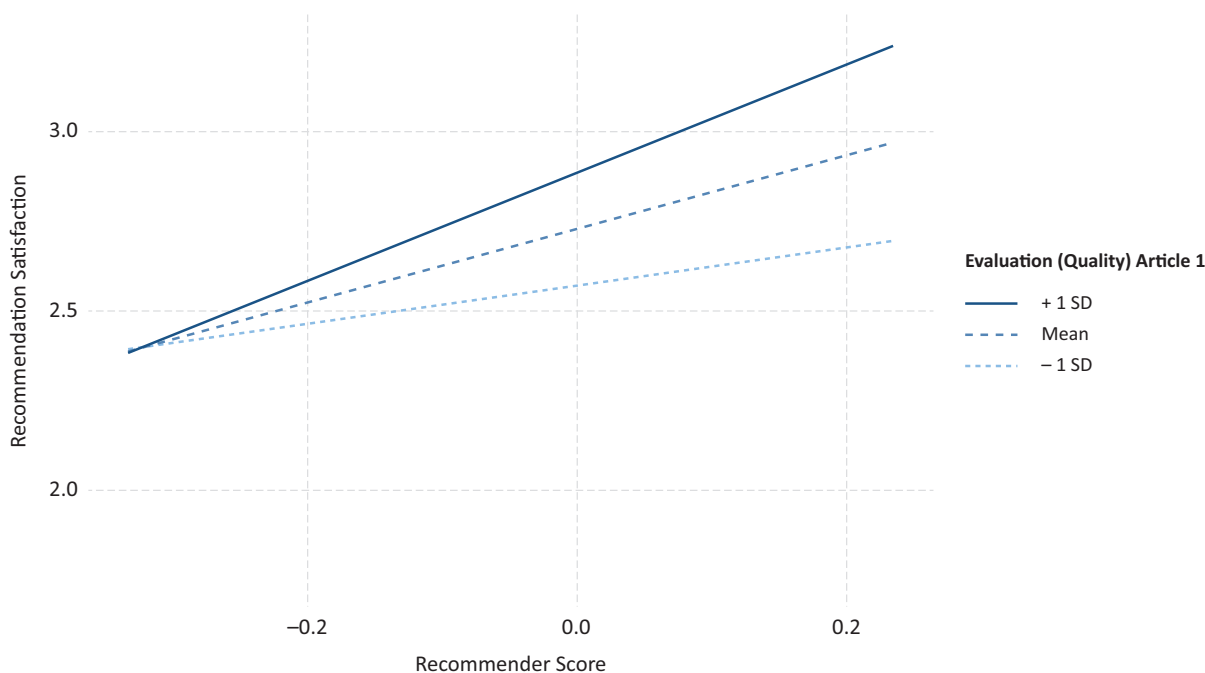

Figure 1. Moderating effect (n.s.) of quality article 1 on the conditional effect of text similarity on recommendation satisfaction.

(F(14,573) = 25.83, p = .00). Moreover, the inclusion of the individual variables does not lead to any significant change in the results already found.

Interest in news is positively but only moderately related to recommendation satisfaction (b = .14, p < .05). News trust (b = .01, n.s.) and a perceived duty to keep informed (b = −.06, n.s.), on the other hand, as well as a general affinity for technology (b = −.03, n.s.), show no correlation with recommendation satisfaction.

By contrast, the general attitude towards news personalization makes a more pronounced difference regarding the level of recommendation satisfaction. Those who report preferring news recommendations, as in the present case, also rate the recommendation as more pleasant and enriching (b = .2, p < .01), controlling for all other predictors. A further graphical exploration (Figure 2) reveals that for these news personalization endorsers, their positive evaluations of recommended articles are almost independent of the similarity to the original article. The (self-reported) personalization sceptics (mean −1SD), however, appear to prefer more similar article recommendations.

An additional moderation analysis did not show that opinion on personalization moderates the effect between text similarity and recommendation satisfaction was statistically significant (b = .27, p = .16; ΔR2 = .39, F(15, 572) = 24,28, p < .01). Again, the interaction term was dropped from the model, resulting in the simple effects only model (which is identical to Model 3). Similarly, the relative share of algorithmic news media in total news consumption becomes ineffective for recommendation satisfaction.

The significant, albeit weak, correlation of recommendation satisfaction with a NfcC (b = .12, p < .05) is surprising at first but plausible given the fact that the recommendation here is based on text similarity. People who avoid ambiguities and prefer closed-world views are more likely to perceive an article that matches their first, self-selected article, and thus the recommendation itself, as pleasant and enriching than those who enjoy challenging perspectives. Here, there might be a connection to the above finding that a differentiation of evaluation dimensions is apparently needed when measuring recommendation satisfaction, as shown by the opposing influence of articles that are evaluated as new compared to articles that are evaluated as unexpected.

The analysis of RQ3 shows that rating the recommended item as novel versus unexpected has an opposite effect on recommendation satisfaction: Recommendations of articles rated as novel are evaluated more positively, while at the same time recommendations of articles rated as unexpected are evaluated negatively. We explore the interrelations of these three measures further by analysing how individual dispositions might interact with rating the recommended article as unexpected and as novel, respectively.

Including the corresponding interaction terms again only leads to minimal further variance explanation (Model 4; ΔR2 = .009, p < .01; F(15,572) = 25.00, p = .00). In the direct comparison of the novel vs. unexpected dimension, it is noticeable that none of the interactions with novelty contributes significantly to the model, also underlined by a visual exploration (see the Supplementary Files). Again, the moderation analysis did not find any of the assumed moderating effects.

However, for the correlation between the recommendation satisfaction and its unexpectedness, the impact of different individual dispositions can be visually detected,
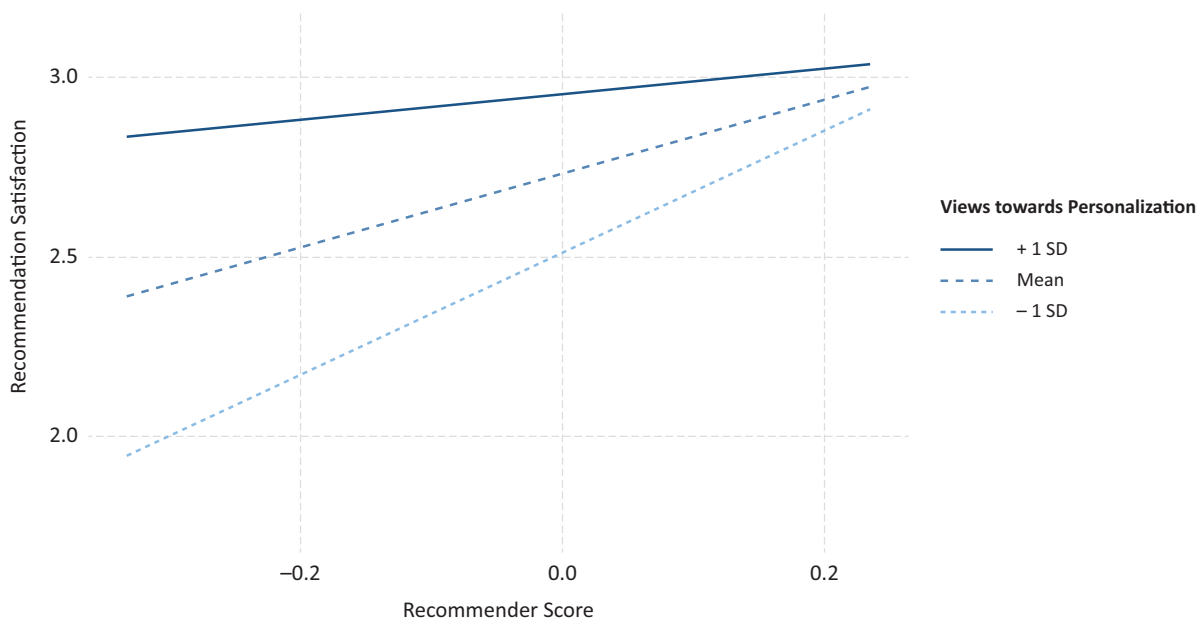

Figure 2. Moderating effect (n.s.) of view towards personalization on the conditional effect of text similarity on recommendation satisfaction.

each showing different strengths in their correlations. Accordingly, a moderation analysis was conducted to determine whether the interaction between the personal characteristics and the rating of the recommended article as unexpected significantly predicts satisfaction. Results show that duty to keep informed (DTKI) moderated the effect between unexpectedness and recommendation satisfaction significantly, $F(15,572) = 25.00$, $p < .001$.

To investigate this effect in more detail, a Johnson-Neyman diagram was plotted (Figure 3). For average (i.e., around the mean) values of DTKI, there is no significant moderation effect. For below-average values of the DTKI (from about .8SD below the mean), on the other hand, we see a positive effect. Thus, for people with a low DTKI, recommendations perceived as pleasant and enriching are also more likely to be perceived as unexpected. For people with a greater sense of duty to inform themselves, on the other hand, there is a negative effect, i.e., here the recommendations not perceived as pleasant and enriching are more likely to be perceived as "not unexpected," in other words, as expected and unsurprising.

This refers to the challenge of finding the right balance between being novel and unexpected, while
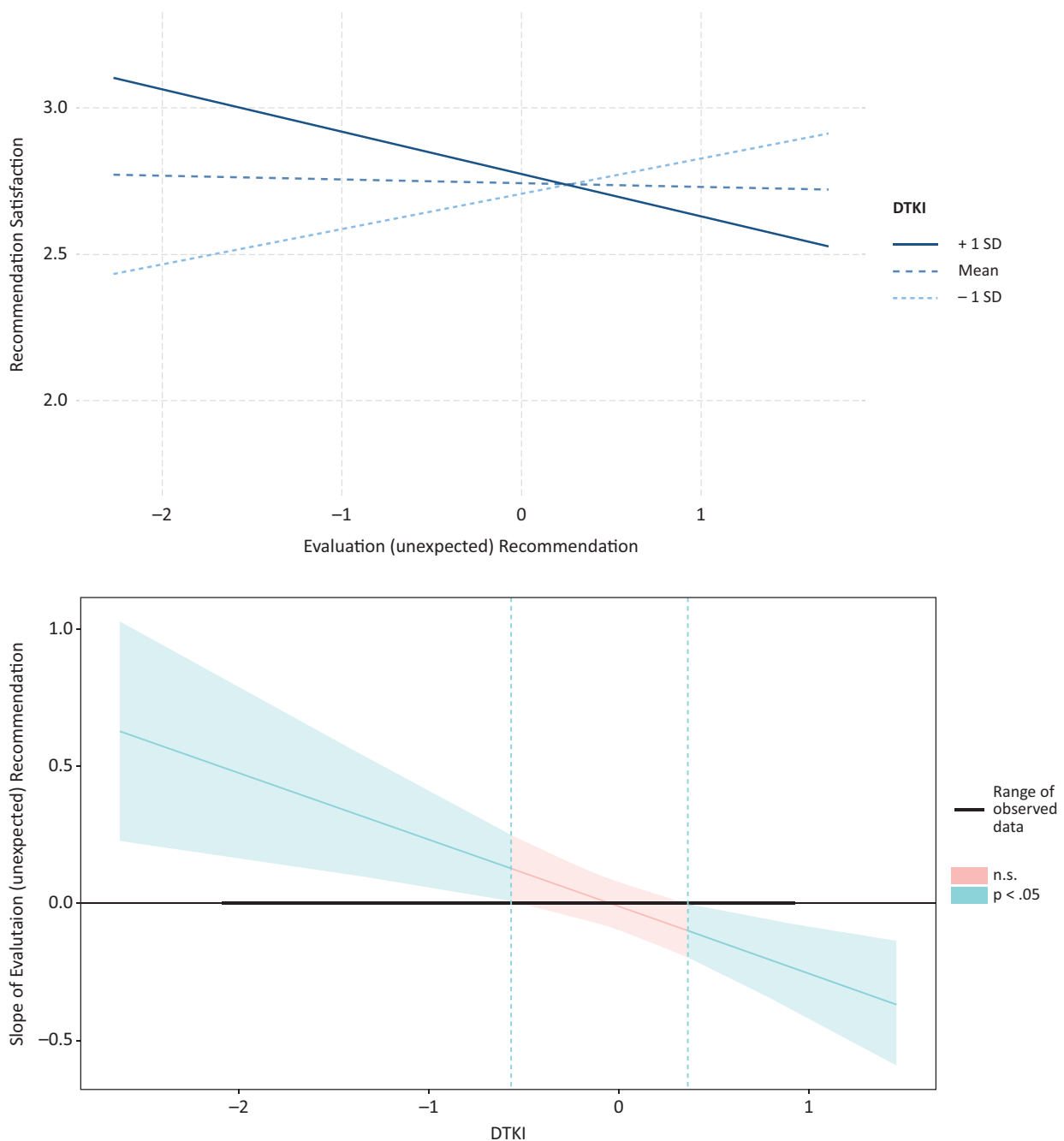

Figure 3. Moderating effect of duty to keep informed on the conditional effect of evaluation as unexpected on recommendation satisfaction.

still being perceived as pleasant and enriching in the overall experience and thus contributing to satisfaction. Optimizing toward pleasant and enriching through novelty thereby represents the simpler strategy of satisfying users through more familiar ways of recommendation. Optimizing satisfaction through unexpected items is much more challenging and obviously also depends on how strong the civic norm to keep informed is. Nevertheless, from a normative point of view, this strategy has more potential to create diversity in article recommendations.

6. Discussion

In summary, the calibration of the recommender, i.e., the degree of similarity between the original and the recommended article, turns out to be the strongest predictor of the entire model. The stronger the recommendation is based on article similarity, the more pleasant and enriching it is perceived to be (RQ1). Moreover, this satisfaction strongly depends on whether the recommended article is evaluated as novel (RQ2a). By contrast, if the topic, facts, and/or viewpoints of the recommended article are perceived as unexpected, this decreases satisfaction with the recommendation (RQ2b). This would confirm previous research on the preference of news users for "reliable surprises" (Schoenbach, 2007), news media should aim to recommend—and produce—news content that adds novel positions or facts, but still falls within the expectations of users for the topic.

In all models, the more educated rate article recommendations as less pleasant and enriching (RQ3a). Yet, higher news interest leads to slightly higher recommendation satisfaction. Users with a high NfcC are also more likely to be satisfied, which stands to reason given that the recommendation is based on text similarity. For general attitudes toward news personalization, a significant simple effect emerges—endorsing algorithmic personalization leads to greater recommendation satisfaction. Even if the moderation analysis is not significant, it can still be deduced based on the visual analysis that, in particular, people with a greater scepticism towards such news recommendations are satisfied precisely with a more similar recommendation, thus preferring less diverse recommendations (RQ3b). This finding deserves further exploration in the future.

For the fascinating opposing relationship between perceiving the recommended article as novel or unexpected and recommendation satisfaction, a significant moderating effect of the civic information norm emerges. For people with a strong sense of duty to inform themselves about current events, rating the recommended article as unexpected nevertheless goes along with recommendation satisfaction, i.e., perceiving the article recommendation as pleasant and enriching. For those users, news media should aim to recommend even more diverse news and thus fulfil their democratic role in the best possible way. By contrast, people who are less concerned about informing themselves about current events are also less satisfied with article recommendations on unexpected topics, facts, or points of view. Here, news media should aim to recommend a comparatively homogeneous news diet to avoid alienating them during these current times when more and more citizens are losing the connection that news media provides them to the public sphere.

7. Conclusion

So, do our results indicate that there is a possible calibration of the recommender that satisfies all user groups equally? This seems not to be the case. It seems certain that news organizations apparently cannot go far wrong with a text-based recommendation algorithm, as indicated by the strength of the predictor text similarity on recommendation satisfaction. At the same time, we find several individual dispositions for which this relationship is less strong. For example, higher educated people (a key target group of many news organizations) are generally less satisfied with the article recommendation if it is based on text similarity. And if we avoid only focusing on the "liberal-individualist" goal of satisfaction (Helberger et al., 2018), and also take into account unexpected and thus potentially challenging content (which is important from a deliberative point of view), a trade-off becomes apparent. The moderator effect of the duty to keep informed make clear that a single, standardized recommender solution will be difficult to achieve. This is especially true if normative goals beyond user satisfaction are to be met.

This leaves media organizations with the data dilemma in that content-based algorithms alone can hardly meet the individual requirements of different target groups. A mixed strategy of implicit and content-based recommendations could remedy this, as Loecherbach and Trilling (2020) also argue. However, the fact that we have already been able to identify different user segments with a standardized survey should also encourage media organizations to meet user expectations of personalized recommendations sufficiently well with comparatively simple means. For example, on-site surveys segmenting one's own target group on the basis of social science concepts such as the duty to keep informed, personalization preference, or the NfcC used here could already be informative enough to increase recommendation satisfaction in the future and thus contribute to greater trust and customer loyalty.

This brings us to the limitations of our study. To achieve a similar level of interest in the recommended articles for all participants, we asked participants to formulate an active search query in the first step. Only in the subsequent step did they receive an article by automated recommendation. However, the instruction to enter a real information need may have had an effect on the participants' expectations regarding the second, recommended article. Within the context of a goal-oriented

search, users tend to hold a fairly specific set of expectations regarding the characteristics of the article (Lutz et al., 2017). This clear set of expectations may have carried over to the second article. Purposeful searches are also possible on news websites, but open browsing to pass time, or at least quite undirected behaviours in which people just update themselves with current events, are even more prevalent. We, therefore, assume that our design influenced the findings especially in terms of the negative correlation between unexpected topics, viewpoints, and facts and article recommendation satisfaction. In a further study, our exploratory findings need to be investigated in a confirmatory design and under systematic variation of the instructions (search task vs. browsing).

Furthermore, though individual dispositions such as the NfcC or news interest are considered to be relatively stable (and our items aimed to identify the more general, not situational attitudes), there is the possibility that the topics selected by our participants had an impact on these attitudes. A more "emotional" topic such as the Covid-19 pandemic might have temporarily increased the NfcC, whereas a "safer" or very familiar topic may have decreased it. Future studies might consider including more items on the user interest in the selected topic and the level of attention while reading it, to control for these possible priming effects.

Even if the inclusion of individual dispositions already means a shift away from short-term engagement metrics, our study was only able to provide a snapshot of the interrelations between user characteristics and article recommendation evaluation. Especially against the background of building brand loyalty through satisfaction (Nelson & Kim, 2020), the mid- and long-term development of these interrelations need greater attention in the future as does the question of which other individual dispositions might be relevant. Informational self-efficacy, for example, contributes to a higher mental readiness to value serendipity (Lutz et al., 2017). Serendipity could be a valuable link between the sometimes challenging diversity in news and a pleasant user experience. Self-efficacy could also increase the sense of control and agency in news use which might in the long term contribute to the willingness to actively engage with the personalization settings and thus to implicitly or explicitly provide personal data (Monzer et al., 2020).

Despite these limitations, the major advantage of our study lies in the practical relevance and transferability of the recommendation algorithm used. With BERT, we not only simulated a realistic content-based personalization based on genuine articles (an approach that is used cost-effectively by smaller news organizations) but also embedded it in a survey interface that enables authentic recommendations tailored to user interests. Here, typical news portal features such as headlines and images were omitted, as was the media brand, which has undoubtedly reduced the ecological validity of our research design. However, this allowed us to avoid

confounding in our exploratory setting. Further studies should aim to gradually include these parameters as well. The simplicity of the design, however, made it possible to achieve a sample that is representative of the German internet population.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

All raw data and scripts necessary to reproduce the results can be found on https://osf.io/kh5va/?view_only=aff4561db93d434cbb14d7b2b3f2de0c. All attachments are available in this repository.

References

Adams, P. C. (2020). Agreeing to surveillance: Digital news privacy policies. *Journalism & Mass Communication Quarterly*, *97*(4), 868–889. https://doi.org/10.1177/1077699020934197

Bernstein, A., Vreese, C. d., Helberger, N., Schulz, W., Zweig, K., Baden, C., Beam, M. A., Hauer, M. P., Heitz, L., Jürgens, P., Katzenbach, C., Kille, B., Klimkiewicz, B., Loosen, W., Moeller, J., Radanovic, G., Shani, G., Tintarev, N., Tolmeijer, S., . . . Zueger, T. (2020). *Diversity in news recommendations*. ArXiv. http://arxiv.org/pdf/2005.09495v1

Bodó, B. (2019). 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. https://doi.org/10.1080/21670811.2019.1624185

Bodó, B., Helberger, N., Eskens, S., & Möller, J. (2019). Interested in diversity. *Digital Journalism*, *7*(2), 206–229. https://doi.org/10.1080/21670811.2018.1521292

Bornstein, R. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin*, *106*, 265–289.

Chen, L., Yang, Y., Wang, N., Yang, K., & Yuan, Q. (2019). How serendipity improves user satisfaction with recommendations? A large-scale user evaluation. In L. Liu & R. White (Eds.), *The world wide web conference on—WWW '19* (pp. 240–250). ACM. https://doi.org/10.1145/3308558.3313469

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. ArXiv. http://arxiv.org/pdf/1810.04805v2

Eskens, S. (2019). A right to reset your user pro-

file and more: GDPR-rights for personalized news consumers. *International Data Privacy Law*, *9*(3), 153–172. https://doi.org/10.1093/idpl/ipz007

Graff, R. (2015, June 3). How the *Washington Post* used data and natural language processing to get people to read more news. *Knight Lab*. https://knightlab.northwestern.edu/2015/06/03/how-the-washington-posts-clavis-tool-helps-to-make-news-personal

Haim, M., Graefe, A., & Brosius, H.-B. (2018). Burst of the filter bubble? *Digital Journalism*, *6*(3), 330–343. https://doi.org/10.1080/21670811.2017.1338145

Hampel, J., Zwick, M., & Störk-Biber, C. (2020). *Technik Radar 2020: Was die Deutschen über Technik denken* [Technology radar 2020: What Germans think about technology]. Acatech; Koerber-Stiftung. https://www.koerber-stiftung.de/fileadmin/user_upload/koerber-stiftung/redaktion/technikradar/pdf/2020/TechnikRadar-2020_Langfassung.pdf

Hayes, A. F., & Little, T. D. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. Methodology in the social sciences*. The Guilford Press.

Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191–207. https://doi.org/10.1080/1369118X.2016.1271900

Helberger, N., Kleinen-von Königslöw, K., & van der Noll, R. (2015). Regulating the new information intermediaries as gatekeepers of information diversity. *Info*, *17*(6), 50–71. https://doi.org/10.1108/info-05-2015-0034

Jungnickel, K. (2011). Nachrichtenqualität aus Nutzersicht. Ein Vergleich zwischen Leserurteilen und wissenschaftlich-normativen Qualitätsansprüchen [News quality from the perspective of the users. A comparison between user evaluations and scientific-normative quality requirements]. *M&K Medien & Kommunikationswissenschaft*, *59*(3), 360–378. https://doi.org/10.5771/1615-634x-2011-3-360

Kunert, J., & Thurman, N. (2019). The form of content personalisation at mainstream, transatlantic news outlets: 2010–2016. *Journalism Practice*, *13*(7), 759–780. https://doi.org/10.1080/17512786.2019.1567271

Loecherbach, F., & Trilling, D. (2020). 3bij3: Developing a framework for researching recommender systems and their effects. *Computational Communication Research*, *2*(1), 53–79. https://doi.org/10.5117/CCR2020.1.003.LOEC

Long, J. A. (2020). Interactions: Comprehensive, user-friendly toolkit for probing interactions (1.1.3) [Computer software]. https://cran.r-project.org/package=interactions

Long, J. A. (2021). Jtools: Analysis and presentation of social scientific data (2.1.3) [Computer software]. https://cran.r-project.org/package=jtools

Luhmann, M. (2015). *R für Einsteiger: Einführung in die Statistik-Software für die Sozialwissenschaften* [R for beginners: An introduction to the statistics software for social scientists]. Beltz.

Lutz, C., Pieter Hoffmann, C., & Meckel, M. (2017). Online serendipity: A contextual differentiation of antecedents and outcomes. *Journal of the Association for Information Science and Technology*, *68*(7), 1698–1710.

Maccatrozzo, V., Terstall, M., Aroyo, L., & Schreiber, G. (2017). SIRUP: Serendipity in recommendations via user perceptions. In G. A. Papadopoulos, T. Kuflik, F. Chen, C. Duarte, & W.-T. Fu (Eds.), *Proceedings of the 22nd international conference on intelligent user interfaces—IUI '17* (pp. 35–44). ACM. https://doi.org/10.1145/3025171.3025185

McCay-Peet, L. (2013). *Investigating work-related serendipity, what influences it, and how it may be facilitated in digital environments.* [Doctoral dissertation, Dalhousie University]. DalSpaceg. http://dalspace.library.dal.ca/handle/10222/42727

McCombs, M., & Poindexter, P. (1983). The duty to keep informed: News exposure and civic obligation. *Journal of Communication*, *33*(2), 88–96. https://doi.org/10.1111/j.1460-2466.1983.tb02391.x

Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, *21*(7), 959–977. https://doi.org/10.1080/1369118X.2018.1444076

Monzer, C., Moeller, J., Helberger, N., & Eskens, S. (2020). User perspectives on the news personalisation process: Agency, trust, and utility as building blocks. *Digital Journalism*, *8*(9), 1142–1162. https://doi.org/10.1080/21670811

Nelson, J. L., & Kim, S. J. (2020). Improve trust, increase loyalty? Analyzing the relationship between news credibility and consumption. *Journalism Practice*, *15*(3), 348–365. https://doi.org/10.1080/17512786.2020.1719874

News API. (2021). *Documentation*. https://newsapi.org/docs

Nielsen, R. K. (2016). People want personalised recommendations (even as they worry about the consequences). In N. Newman, R. Fletcher, D. A. L. Levy, & R. K. Nielsen (Eds.), *Reuters institute digital news report 2016* (pp. 112–114). Reuters Institute.

Nordheim, G. v., & Fuchsloch, S. (2019). Der gemeinwohlorientierte Intermediär [The common good orientated intermediary]. In J. Krone & A. Gebesmair (Eds.), *Reihe Medienstrukturen: Band 14, Zur Ökonomie gemeinwohlorientierter Medien: Massenkommunikation in Deutschland, Österreich und der Schweiz* [Series media structures: Volume 14, On the economics of public interest media: Mass communication in Germany, Austria and Switzerland] (pp. 245–260). Nomos.

Reviglio, U. (2019). Serendipity as an emerging design

principle of the infosphere: Challenges and opportunities. *Ethics and Information Technology*, *21*(2), 151–166. https://doi.org/10.1007/s10676-018-9496-y

Reviglio, U. (2020). Towards a right not to be deceived? An interdisciplinary analysis of media personalization in the light of the GDPR. In I. O. Pappas, P. Mikalef, Y. K. Dwivedi, L. Jaccheri, J. Krogstie, & M. Mäntymäki (Eds.), *Digital transformation for a sustainable society in the 21st century* (pp. 47–59). Springer.

Schlink, S., & Walther, E. (2007). Kurz und gut: Eine deutsche Kurzskala zur Erfassung des Bedürfnisses nach kognitiver Geschlossenheit [Short and good: A German short scale for assessing the need for cognitive closure]. *Zeitschrift Für Sozialpsychologie*, *38*(3), 153–161. https://doi.org/10.1024/0044-3514.38.3.153

Schoenbach, K. (2007). "The own in the foreign": Reliable surprise—An important function of the media*? Media, Culture & Society*, *29*(2), 344–353. https://doi.org/10.1177/0163443707074269

Schweiger, W., Weber, P., Prochazka, F., & Brückner, L. (2019). *Algorithmisch personalisierte Nachrichtenkanäle: Begriffe, Nutzung, Wirkung* [Algorithmic personalized news channels: Terminology, usage, effects]. Springer.

Spangher, A. (2015, August 11). Building the next *New York Times* recommendation engine. *The New York Times*. https://open.blogs.nytimes.com/2015/08/11/building-the-next-new-york-times-recommendation-engine

Stray, J. (2020). Aligning AI optimization to community well-being. *International Journal of Community Well-Being*, *3*(4), 443–463. https://doi.org/10.1007/s42413-020-00086-3

Thurman, N., Moeller, J., Helberger, N., & Trilling, D. (2019). My friends, editors, algorithms, and I. *Digital Journalism*, *7*(4), 447–469. https://doi.org/10.1080/21670811.2018.1493936

Thurman, N., & Schifferes, S. (2012). The future of personalization at news websites. *Journalism Studies*, *13*(5/6), 775–790. https://doi.org/10.1080/1461670X.2012.664341

Wang, T., & Fu, Y. (2020). Item-based collaborative filtering with BERT. In S. Malmasi, S. Kallumadi, N. Ueffing, O. Rokhlenko, E. Agichtein, & I. Guy (Eds.), *Proceedings of the 3rd workshop on e-Commerce and NLP* (pp. 54–58). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.ecnlp-1.8

Wang, W., Bi, B., Yan, M., Wu, C., Bao, Z., Xia, J., Peng, L., & Si, L. (2019). *StructBERT: Incorporating language structures into pre-training for deep language understanding*. ArXiv. http://arxiv.org/pdf/1908.04577v3

Webster, D. M., & Kruglanski, A. W. (1997). Cognitive and social consequences of the need for cognitive closure. *European Review of Social Psychology*, *8*(1), 133–173. https://doi.org/10.1080/14792779643000100

Zajonc, R. B. (1968). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*, 117–123.

About the Authors

Mareike Wieland is a research assistant and doctoral student in journalism and communication studies at the University of Hamburg. Her research focuses on the use, perception, and processing of political information in automated news environments.

Gerret von Nordheim is a postdoctoral researcher in the field of communication science at the University of Hamburg. His research focuses on intermedia effects in the hybrid media system, especially between journalism, social media, and populist actors. He is specialized in the field of computational methods.

Katharina Kleinen-von Königslöw is a professor of journalism and communication studies, especially digital communication and sustainability at the University of Hamburg. Her research focuses on the impact of digitalisation and technological innovations on political communication and, in particular, the role of social network platforms and their use by citizens and political actors.