Empowering Editors: How Automated Recommendations Support Editorial Article Curation

Anastasiia Klimashevskaia
anastasiia.klimashevskaia@uib.no
MediaFutures, University of Bergen
Bergen, Norway

Mehdi Elahi
mehdi.elahi@uib.no
MediaFutures, University of Bergen
Bergen, Norway

Dietmar Jannach
dietmar.jannach@aau.at
MediaFutures, University of
Klagenfurt
Klagenfurt, Austria

Christoph Trattner
christoph.trattner@uib.no
MediaFutures, University of Bergen
Bergen, Norway

Simen Buodd
simen.buodd@vg.no
VG, Schibsted Media
Oslo, Norway

ABSTRACT

The application of recommender systems in the news domain has experienced rapid growth in recent years. Various news outlets are proposing a full automation of a newspaper front page through automated recommendation. In this paper, however, we explore the synergy of editorial and algorithmic news curation by analyzing the front page of a real-world news platform, where news articles are either selected *automatically* by a recommendation algorithm or are selected *manually* by editors. An investigation of the interaction log data from an online newspaper revealed that while the editorial staff is focusing on content that is generally popular across large parts of the audience, the algorithmic curation can, in addition, provide small yet noteworthy personalization touches for individual readers. The results of the analysis demonstrate an example of a successful coexistence of editorial and algorithmic news curation.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

News Recommender Systems, Editorial Mission, Diversity

1 INTRODUCTION

In the current digital age, reading news articles online has become a part of our everyday routines. Traditionally, editors make choices on which news articles are relevant or important to include on

the front page or emphasize [3]. This process requires them to *manually* select the articles they consider to be the most interesting and important ones [29]. At the same time, overall news coverage and topic diversity should also be taken into account as part of editorial values [6, 11, 21, 29, 30].

However, with the enormous growth of news articles published every day, it is becoming a challenge for editors to go through the entire catalog of news articles and make choices on which articles to select. Thus, modern news platforms often leverage digital tools that can provide *automated* mechanisms, e.g., to extend the list of news articles selected by editors with additional articles to include. News Recommender Systems (NRS) can offer such services by analyzing the online click behavior of readers and offering them a personalized experience when navigating through the news.

Previously, various solutions were suggested by researchers to tackle the news recommendation problem [13, 26] — utilizing both collaborative filtering [8, 22, 34] and content-based approaches [16, 25], as well as hybridizing the two methods [7, 33]. Apart from classic challenges in recommendation such as cold start [18] and various undesired effects [6], news recommendation systems also face unique difficulties due to the dynamic nature of content and user preferences. News articles have short lifespans and time-dependent relevance, which necessitates timely delivery of relevant content [13, 24]. In addition, the unstructured format of news articles and the type of media used to deliver the news (short- and long-form article, quick headlines or images, video and audio formats), as well as the typical lack of user profiles due to anonymous browsing present additional challenges [2, 13]. These restrictions possibly make news recommendation extra challenging compared to other application domains. Last but not least, the specific setting in news recommendation also imposes certain technical restrictions — minimizing response time, scaling to handle large request volumes, and adapting to mobile device constraints [2]. Nevertheless, regardless of all the difficulties and challenges, utilizing automated recommendations can be beneficial for readers by enhancing their satisfaction and engagement [17], as well as for news platforms, by increasing key performance indicators (e.g., CTR) [20].

Various works explored NRS as a means to replace the work of newspaper editors in attempts to fully automate front page curation [13, 24]. In this paper, instead, we explore how editorial

efforts can be supported by an NRS, by conducting a comprehensive analysis of the interactions and the front page of a real-world commercial news platform. The news articles on the platform are delivered either based on choices made by editors or by an automated recommendation algorithm. Specifically, we compare the articles that were selected by editors and the algorithm in terms of several metrics that are commonly used to assess the popularity of the recommended content as well as its diversity, including the average recommendation popularity (ARP), Gini index, entropy, and miscalibration. We demonstrate that while the editorial staff can concentrate on more widely accepted content for wider demographics, even limited personalization through recommendation can still be beneficial and hence improve the content diversity and exposure of more niche content.

2 METHODS

We adopt an *observational* research approach in which we study the combination of editorial and algorithmic choices on a national online newspaper platform in terms of article diversity and popularity, as well as how well the curation matches user preferences in topical diversity and content popularity/mainstreamness. In the following section, we describe the observed setting, the collected dataset, and the used metrics.

Figure 1: Stylized Front Page.

2.1 Application Setting

Verdens Gang (widely known as VG)[1] is a national online newspaper in Norway, serving daily readers both with breaking and essential news (also often called *hard news* [27]) and with articles more focusing on sport, entertainment, and lifestyle (*soft news* [27]). While access to hard news is free of charge, parts of the soft news articles are behind a paywall requiring a premium subscription.

Most of VG's front page is curated and manually assembled by the editorial staff, who try to balance two primary goals in their article selection. First, they aim to maintain integrity and the journalistic mission to keep the population informed and updated from diverse standpoints. Secondly, they have a vested interest in maintaining the newspaper's revenue by placing some longer-form paywalled content on the front page, hoping to encourage readers to become subscribers. To support the editorial staff, VG employs a recommender algorithm that selects some of the paywalled soft-news content for individual users in a personalized way. For these purposes, the algorithm ranks a pool of articles based on how likely it is to entice the users to subscribe. As for personalization, the algorithm additionally takes into account factors such as (i) user demographics and previously expressed preference towards specific topics; (ii) articles that a particular user has previously seen but not interacted with repeatedly — such articles would be considered as "unsuccessful" for this particular user and would not be recommended again. In the end, the newspaper's front page contains premium articles selected both by editors and by the algorithm; see Figure 1. We emphasize that the proportion of free to paywalled content on the front page can change depending on the current news situation. We also note that the algorithm cannot select items that were already selected by an editor in order to avoid repetitions. As such, the pool of available items can be slightly smaller than the editors' pool of choices.

2.2 Dataset Characteristics

Our analysis is based on a dataset from VG, containing the front page user-item interaction data for a period of one month (June 2023), consisting of both impressions and clicks. The dataset also contained additional descriptive information for the news articles, such as whether the item was placed on the front page by the editors or by the algorithm, as well as the topic of the article. During preprocessing, users with less than 10 item impressions/clicks, and items with no clicks were excluded from the dataset to reduce potential noise. As a result, the preprocessed dataset consists of 272 premium articles and more than 50,000 users, spanning over 23 million rows in total.

It is worth noting that our analyses focus only on the premium soft-news articles. This has been conducted for a number of reasons: first, looking only at premium content means analyzing only logged-in user profiles with more reliable interaction data. Second, it is a fairer comparison between editorial and algorithmic curation, as the algorithm can select only premium content. To investigate topical diversity aspects, we consider article categories that are available for each news item. We limit our analyses to the four main categories (i.e., Consumer, General News, Sport and Celebrity) present in the collected dataset.

2.3 Metrics

To compare the characteristics of editorial and algorithmic choices, we investigated the popularity level and topic diversity of the premium articles selected with both approaches. Furthermore, we measured the extent of *miscalibration* in terms of these aspects [28]. Specifically, we used the following metrics:

- **Average Recommendation Popularity (ARP)** [35]. This metric quantifies the popularity of the items that are selected for recommendation. We define item popularity as a normalized number of clicks an item has received during the time of observation.
- **Gini Index** [5]. The Gini index is a measure to assess the inequality of a frequency distribution. In our setting, we measure how often each item is recommended on a front page. If the distribution is very uneven, this indicates that a recommendation might be focused on a small set of (usually popular) items. A higher Gini index value indicates a higher level of inequality — and in recommendation scenario, lower diversity.
- **Shannon Entropy** [12]. We adopt this metric as a measure of (category) diversity in the recommendations. With the entropy measure, we can determine how well-represented each news category is on average in the recommendations. Higher entropy values mean higher diversity and better topic coverage.
- **Miscalibration (MC)** [1, 14, 28]. The goal of calibration approaches is to ensure that the recommendations provided to a user match the *distribution* of past user preferences well. *Miscalibration*, in return, quantifies the discrepancy between the recommendation and user preference. To measure the extent of miscalibration we compared the probability distribution vectors describing user profiles with the distributions of the recommendations they received using Jensen-Shannon Divergence [19] as a distance metric between two distribution vectors. For item categories, we considered two separate cases with either article topics (MC-Div) or item popularities (MC-Pop), with item categories defined as highly, medium and less popular items[2].

In addition, we analyze the Click-Through-Rates (CTR) in an attempt to gauge user satisfaction and experience with both curation approaches. This metric is commonly used in real-life recommendation settings to evaluate recommendation accuracy and effectiveness, it is generally simple and quick to compute. However, we also acknowledge that CTR can have certain limitations and downsides in terms of reliability, which we discuss in further detail after presenting the results of the evaluation.

3 RESULTS

We report the summary of our results in Table 1. In terms of popularity metrics a notable difference can be observed between the algorithm-generated recommendation and manual news curation by the editors. According to the observation, the overall distribution

---

[2]The thresholds to separate the items in these popularity groups were set according to the Pareto rule and the majority previous research on popularity bias, which defines the top 20% of the most popular content as the most popular, and the lowest 20% as least popular, assigning the rest to the medium-popularity group [15].

of algorithm-generated recommendations is more spread out over the article pool compared to the editorial curation, which is demonstrated by the algorithm's Gini index value of 0.46, nearly half of the Gini index value for the editors with the value of 0.83. Similar trends are observed for the average recommendation popularity (ARP) metric, where the algorithm's recommendation achieved a value of 0.22, while the editor's news selection reached 0.37. However, the manual selection by editors surprisingly resulted in lower popularity miscalibration (with an MC-Pop value of 0.31), more closely matching the popularity tendencies of users compared to the algorithm's recommendation (with a value of 0.39). This suggests that the editorial approach may better align with the broad spectrum of user preferences in terms of news content popularity compared to the algorithmic approach. It also indicates a more general gravitation toward mainstream popular content, potentially being interesting for the majority of users. At the same time, the content recommended by the algorithm has an overall lower popularity, which can be interpreted as more personalized niche content for smaller demographics. However, both approaches combined on the front page have the potential to complement each other, providing enhanced utility to the whole reader base.

When discussing topic diversity, the recommendation algorithm interestingly exhibited an entropy level close to that of the editor's selection. Specifically, the recommendation algorithm achieved an entropy value of 1.09, while the editor's selection had a value of 1.02. This is a marginal difference, though it may still suggest that the algorithm could include slightly more diverse topics in its recommendations compared to the handpicked articles. Similarly, the miscalibration with respect to diversity (MC-Div) for the algorithm's recommendations is not very different from the manual news selection. Specifically, the MC-Div value for algorithm-generated recommendations is 0.28, while it is 0.29 for manual news selection by editors. Despite the similar values, these results may still suggest that the algorithmic approach might slightly better align with a broader spectrum of user preferences in terms of diversity. Overall, in regards to news topic diversity, both news selection approaches appear to be similar in matching user preferences towards various news topics, which is seen from diversity miscalibration results.

Last but not least, in terms of CTR results, again, both methods demonstrate fairly similar results, with the algorithm performing marginally better (0.03) than the editorial staff (0.02). This result may indicate that further personalized curation of news content has the potential to increase user engagement on the platform. However, as we provide a discussion in the following section, further investigation might be required to generalize better and make distinct conclusions on this behalf.

Last but not least, we can deduce news curation quality to some extent for both editorial and algorithmic picks from the CTR results and observe a marginal difference in favor of the algorithm. This can indicate that the general quality of recommendation does not diminish with the introduction of algorithmic curation, but can even undergo a slight improvement.

4 DISCUSSION

The overall results are promising and highlight the potential of an automated algorithmic approach to support editorial article

Table 1: Results of analyzing the front page of VG where news articles are selected either *manually* by the editors or recommended by an *automated* algorithm. The arrows indicate whether lower or higher values are considered desirable.

News Delivery	Personalized	Gini Index ↓	Content Popularity		Content Diversity		CTR ↑
			ARP ↓	MC-Pop ↓	Entropy ↑	MC-Div ↓	
Editorial Staff	No	0.83	0.37	**0.31**	1.02	0.29	0.02
Algorithm	Yes	**0.46**	**0.22**	0.39	**1.09**	**0.28**	**0.03**

curation. Several key aspects of the results may require further discussion. In this section, we explore some of these aspects.

Position Bias. Aiming to better understand the difference in observed metric results we would like to discuss possible limitations and factors that can have an impact on the results apart from just performance. One of such considerations might be the influence of position bias, as generally the very top of the front page is reserved for editorial curation since it might require a more cautious human touch. This can potentially give the content curated by the editors an advantage to gain more exposure, popularity, and consequently more clicks. We recall that within the row-based structure of the front page, typically first three rows are reserved for only "hard" free news for the most important current events. Position bias, while having an immense impact on the very first positions, generally tends to weaken rapidly further down the positions, having lesser of an effect [32]. Thus, while we acknowledge that position bias is present in our observed case, we believe that its influence on the results can be potentially negligible considering the positioning of the content we studied on the webpage.

Article Selection Pool. Another possible factor that require consideration could be potential variations in the article selection pool. For instance, there are cases when editors have early access to likely more popular content, thereby reducing the algorithm's opportunity to select these items. It would require further analysis of the overall content that can be available for recommendation from either side, as due to the particularities of news recommendation, such recommendations pool is expected to be very dynamic and constantly changing.

Click-Through-Rates. Measuring CTR has become one of the most common metrics in online evaluations of recommender models, it is simple, easy to calculate and can be monitored live through user activity logs. Some recommendation approaches proposed in the literature optimize the algorithms for predicted CTRs [9] as it appears to be more realistic than such classic information retrieval metrics as precision and recall. However, the CTR metric can be problematic in multiple ways — firstly, it was shown that higher CTR results do not necessarily mean increased profitability of a recommender [4], which warns against employing it as a main key performance indicator (KPI). Secondly, CTR results might be non-trivial to interpret, as it is directly connected to implicit user feedback, which can sometimes be falsely attributed to user satisfaction and recommendation relevancy [36]. Last but not least, CTR values are drastically affected by position bias, when sometimes the attention an item receives is connected not to its quality or relevance, but rather the good positioning on the front page [9]. Now, this particular drawback can be partially neglected in our

particular scenario as the contents of the front page on VG are very dynamic and change constantly.

Recommendation Diversity. While some of the research on news recommendation was previously raising concerns about algorithmic curation potentially causing filter bubbles and low topic diversity [10], other works demonstrate proof that algorithms are capable of recommending rather a diverse set of topics to newspaper readers and can help in avoiding filter bubbles [23, 31]. Addressing this challenge is out of the scope of our analysis, but at least our results demonstrate that in terms of topic diversity, an algorithmic recommendation does not necessarily differ significantly from human-performed news curation. Indeed, our findings (e.g., in terms of entropy metrics) indicate that both editors and the algorithm appear to utilize equally a diverse pool of news topics. More metrics shall be employed to investigate the diversity of news coverage further.

5 CONCLUSION

In this paper, we conducted an observational study to explore the synergy of editorial and algorithmic news curation. We examined the front page of a real-world news platform where the news articles are delivered both by an *automated* recommendation algorithm or they are *manually* selected by editors. The results of the experiment have revealed that recommendation generated by the automated algorithm can serve as an effective assistance to the editorial staff, complementing the front page selection with more personalized niche content with similar topic diversity for smaller demographics.

As a future work, it can be very insightful to conduct a qualitative analysis involving interviews with news editors to explore their methods and techniques for the process of content selection and curation. We are additionally interested in learning more about their reflection on how automated algorithms can be better integrated into this process and the extent to which this can facilitate the delivery of news content to the audience in an effective and responsible way.

6 ACKNOWLEDGMENTS

REFERENCES
[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 726–731.

[2] Jon Atle Gulla, Kevin C Almeroth, Mozhgan Tavakolifard, and Frank Hopfgartner. 2013. *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge.* ACM.

[3] Matt Carlson. 2018. Automating judgment? Algorithmic judgment, news knowledge, and journalistic professionalism. *New Media & Society* 20, 5 (2018), 1755–1772.

[4] Hung-Hsuan Chen, Chu-An Chung, Hsin-Chien Huang, and Wen Tsui. 2017. Common pitfalls in training and evaluating recommender systems. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 37–45.

[5] Robert Dorfman. 1979. A formula for the Gini coefficient. *The Review of Economics and Statistics* (1979), 146–149.

[6] Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Erik Knudsen, Helle Sjøvaag, Kristian Tolonen, Øyvind Holmstad, Igor Pipkin, Eivind Throndsen, Agnes Stenbom, et al. 2022. Towards responsible media recommendation. *AI and Ethics* (2022), 1–12.

[7] P Moreira Gabriel De Souza, Dietmar Jannach, and Adilson Marques Da Cunha. 2019. Contextual hybrid session-based news recommendation with recurrent neural networks. *IEEE Access* 7 (2019), 169185–169203.

[8] Florent Garcin, Kai Zhou, Boi Faltings, and Vincent Schickel. 2012. Personalized news recommendation based on collaborative filtering. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Vol. 1. IEEE, 437–441.

[9] Huifeng Guo, Jinkai Yu, Qing Liu, Ruiming Tang, and Yuzhou Zhang. 2019. PAL: a position-bias aware learning framework for CTR prediction in live recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 452–456.

[10] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. 2018. Burst of the filter bubble? Effects of personalization on the diversity of Google News. *Digital journalism* 6, 3 (2018), 330–343.

[11] Natali Helberger. 2021. On the democratic role of news recommenders. In *Algorithms, Automation, and News*. Routledge, 14–33.

[12] Lou Jost. 2006. Entropy and diversity. *Oikos* 113, 2 (2006), 363–375.

[13] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News recommender systems–Survey and roads ahead. *Information Processing & Management* 54, 6 (2018), 1203–1227.

[14] Anastasiia Klimashevskaia, Mehdi Elahi, Dietmar Jannach, Christoph Trattner, and Lars Skjærven. 2022. Mitigating popularity bias in recommendation: potential and limits of calibration approaches. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 82–90.

[15] Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. 2024. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction* (2024), 1–58.

[16] Michal Kompan and Mária Bieliková. 2010. Content-based news recommendation. In *E-Commerce and Web Technologies*. Springer, 61–72.

[17] Ting-Peng Liang, Hung-Jen Lai, and Yi-Cheng Ku. 2006. Personalized content recommendation and user satisfaction: Theoretical synthesis and empirical findings. *Journal of Management Information Systems* 23, 3 (2006), 45–70.

[18] Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. 2014. Facing the cold start problem in recommender systems. *Expert Systems with Applications* 41, 4 (2014), 2065–2073.

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

[20] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In *Proceedings of the 15th International Conference on Intelligent user interfaces*. 31–40.

[21] Feng Lu, Anca Dumitrache, and David Graus. 2020. Beyond optimizing for clicks: Incorporating editorial values in news recommendation. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 145–153.

[22] Zhongqi Lu, Zhicheng Dou, Jianxun Lian, Xing Xie, and Qiang Yang. 2015. Content-based collaborative filtering for news topic recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[23] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 2020. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. In *Digital Media, Political Polarization and Challenges to Democracy*. Routledge, 45–63.

[24] Özlem Özgöbek, Jon Atle Gulla, and R Cenk Erdur. 2014. A survey on challenges and methods in news recommendation. In *International Conference on Web Information Systems and Technologies*, Vol. 2. SCITEPRESS, 278–285.

[25] Owen Phelan, Kevin McCarthy, Mike Bennett, and Barry Smyth. 2011. Terms of a feather: Content-based news recommendation and discovery using twitter. In *European Conference on Information Retrieval*. Springer, 448–459.

[26] Shaina Raza and Chen Ding. 2022. News recommender system: a review of recent progress, challenges, and opportunities. *Artificial Intelligence Review* (2022), 1–52.

[27] Carsten Reinemann, James Stanyer, Sebastian Scherr, and Guido Legnante. 2012. Hard and soft news: A review of concepts, operationalizations and key findings. *Journalism* 13, 2 (2012), 221–239.

[28] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 154–162.

[29] Cagri Toraman and Fazli Can. 2015. A front-page news-selection algorithm based on topic modelling using raw text. *Journal of Information Science* 41, 5 (2015), 676–685.

[30] Christoph Trattner, Dietmar Jannach, Enrico Motta, Irene Costera Meijer, Nicholas Diakopoulos, Mehdi Elahi, Andreas L Opdahl, Bjørnar Tessem, Njål Borch, Morten Fjeld, et al. 2022. Responsible media technology and AI: challenges and research directions. *AI and Ethics* 2, 4 (2022), 585–594.

[31] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a mission: assessing diversity in news recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 173–183.

[32] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. 2018. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. 610–618.

[33] Hao Wen, Liping Fang, and Ling Guan. 2012. A hybrid approach for personalized recommendation of news on the Web. *Expert Systems with Applications* 39, 5 (2012), 5806–5814.

[34] Yingyuan Xiao, Pengqiang Ai, Ching-hsien Hsu, Hongya Wang, and Xu Jiao. 2015. Time-ordered collaborative filtering for news recommendation. *China Communications* 12, 12 (2015), 53–62.

[35] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. 2012. Challenging the long tail recommendation. *Proceedings of the VLDB Endowment* 5, 9 (2012), 896–907.

[36] Hua Zheng, Dong Wang, Qi Zhang, Hang Li, and Tinghao Yang. 2010. Do clicks measure recommendation relevancy? An empirical user study. In *Proceedings of the 4th ACM Conference on Recommender Systems*. 249–252.