What News Recommendation Research Did (But Mostly Didn't) Teach Us About Building A News Recommender

Karl Higley[1], Robin Burke[2], Michael D. Ekstrand[3] and Bart P. Knijnenburg[4]

[1]Department of Computer Science and Engineering, University of Minnesota
[2]Department of Information Science, University of Colorado, Boulder
[3]Department of Information Science, Drexel University
[4]School of Computing, Clemson University

Abstract

One of the goals of recommender systems research is to provide insights and methods that can be used by practitioners to build real-world systems that deliver high-quality recommendations to actual people grounded in their genuine interests and needs. We report on our experience trying to apply the news recommendation literature to build POPROX, a live platform for news recommendation research, and reflect on the extent to which the current state of research supports system-building efforts. Our experience highlights several unexpected challenges encountered in building personalization features that are commonly found in products from news aggregators and publishers, and shows how those difficulties are connected to surprising gaps in the literature. Finally, we offer a set of lessons learned from building a live system with a persistent user base and highlight opportunities to make future news recommendation research more applicable and impactful in practice.

Keywords

Recommender systems, News recommendation, Applications

1. Introduction

Everyone has a plan until they try to build a real system. — adapted from Mike Tyson [1].

Recommender systems is a strongly applied research field, which draws people from many disciplines (including machine learning, human-computer interaction, information retrieval, psychology, marketing, economics, and more) interested in a shared class of problems: helping people discover information, products, and other items that meet their needs in a personalized way. The community prides itself that the flagship RecSys conference attracts hundreds of industry researchers and practitioners each year, and encourages paper authors to pay attention to practical details like scalability, performance, and data quality.[1] In practice, however, recommender systems research often does not provide the full set of tools needed to build effective recommender systems applications, as has been noted previously [2, 3]. At the same time, reports from industry indicate that typical real-world recommender system deployments are quite simple and adopt few of the advanced techniques explored in the research literature [4].

Over the past two years, we have built a small-scale production recommender system, attempting as best we could to apply current best practices from the literature. We found gaps that go beyond the oft-discussed aspects of data management, deployment, and user experience. Many published results examine only one component of recommender systems (typically models), or single phases of the lifecycle of a user-recommender relationship (often after users have established interaction histories.) Published models are surprisingly difficult to apply to the kinds of data found in real-world datasets and practical recommendation problems. Many approaches for addressing issues like bias,

Beyond Algorithms: Reclaiming the Interdisciplinary Roots of Recommender Systems Workshop (BEYOND 2025), September 26th, 2025, co-located with the 19th ACM Recommender Systems Conference, Prague, Czech Republic.
khigley@umn.edu (K. Higley); robin.burke@colorado.edu (R. Burke); mdekstrand@drexel.edu (M. D. Ekstrand); bartk@clemson.edu (B. P. Knijnenburg)
0009-0002-6332-8997 (K. Higley); 0000-0001-5766-6434 (R. Burke); 0000-0003-2467-0108 (M. D. Ekstrand); 0000-0003-1341-0669 (B. P. Knijnenburg)

[1]These aspects have historically been emphasized by the Call For Contributions.

fairness, and diversity conflict with each other by intervening through the same system components (e.g. re-rankers) or pursuing conflicting goals, so they can not readily be used in conjunction with each other. In sum, it proves difficult to assemble a working system from the components that have received research attention. The result is that the substantial and growing body of recommender systems literature, taken as a whole, is likely having less impact on improving real people's experiences with the recommendations they receive in their daily lives than the field would like.

In this case study, we explore these disparities between research and practice, grounding the findings in our effort to build a news recommendation research platform. We briefly introduce the project and its goals, along with a short survey of existing news recommendation research. We then describe several specific challenges we faced (and still face) while creating and operating it — challenges that we expected the research literature to provide insights and techniques for, but found that it did not. We conclude with lessons and recommendations for fostering and producing research that is better able to deliver positive recommendation experiences for the real users of production recommender systems.

2. POPROX

The *Platform for OPen Research and Online eXperimentation* (POPROX)[2] is a community research infrastructure project funded by the U.S. National Science Foundation (NSF)[3]. Our aim is to build a live recommendation platform that can host research studies examining real-world interactions between users and recommender systems, enabling the kind of user-centered research that is commonplace in industry settings but difficult for academic researchers. We chose (and the NSF supported) news recommendation as the initial domain for POPROX because it is a recommendation domain of social import. With the number of US citizens receiving their news on digital platforms greater than 50% [5], concerns have rightly arisen about how the change from editorial curation of news to algorithmic curation impacts core democratic functions.

The POPROX platform has been live since January 2025, delivering personalized newsletters containing articles from the Associated Press (AP) to subscribers and supporting researchers investigating various aspects of recommendation. We chose a daily newsletter as the initial implementation of POPROX to be able to start studying real user responses quickly by leveraging the existing distribution and notification mechanisms of e-mail. It is also a relatively forgiving modality for recommendation generation, because researchers do not have to meet strict latency constraints in delivering recommendations and incorporating user feedback.

Since POPROX provides an existing pool of subscribers who are consented research subjects, researchers running experiments on the platform do not need to recruit experiment participants, secure their informed consent, or manage subscriptions. Although POPROX is a non-commercial research platform, maintaining a population of active subscribers exposes it to similar pressures as commercial products and industrial recommender systems. In particular, the "market logic" identified by Mitova et al. [6] makes "funnel thinking" that considers and addresses issues of reach, engagement, conversion, and retention [7] relevant for the platform.

3. News Recommendation Research

News recommendation has a 30+ year history, stretching back to *The Krakatoa Chronicle* [8], an early web-based personalized newspaper. In that time, it has been the focus of long-running workshops like INRA[4] and competitions such as the CLEF NewsREEL Challenge [9], the RecSys Challenge [10], and the MIND News Recommendation Competition[5]. We highlight and summarize a few aspects of

[2] https://poprox.ai
[3] https://www.nsf.gov/awardsearch/showAward?AWD_ID=2232551
[4] https://research.idi.ntnu.no/NewsTech/INRA/
[5] https://msnews.github.io/competition.html

the literature most relevant to our work on POPROX below, and direct readers to a number of survey papers that cover this research area more fully [11, 12, 13, 14, 15, 16, 17, 18].

3.1. Datasets

Available public news recommendation datasets in 2025 include MIND (English) [19], EB-NeRD (Danish) [20], Adressa (Norwegian) [21], plista (German) [22], NPR (Portuguese) [23], and IDEA (English) [24]. Other datasets commonly used in the news recommendation literature are not currently available to researchers, either because they are no longer provided (Globo [25], Yahoo! Webscope) or because the data is proprietary (MSN News, Google News, Yahoo News).

Although the data contained by public news recommendation datasets varies somewhat, it typically includes user and article identifiers, article categories or topics, named entities mentioned by articles, and textual features such as article headlines, abstracts, and body text. Some datasets have been extended into multi-modal formats with the addition of images (e.g. IM-MIND [26], VMIND [27]).

3.2. Modeling

Recent work in news recommendation modeling displays two major trends (which are not mutually exclusive and are sometimes combined): (1) text-based models applying techniques from natural language processing (e.g. BERT embeddings) to article headlines, abstracts, and/or body text, and (2) knowledge-aware models incorporating knowledge graphs and/or non-text attributes (e.g. categories or named entities). We refer readers to the modeling-oriented survey by Wu et al. [16] for detailed analysis of the model architectures and features used in news recommendation.

3.3. Evaluation

As is common throughout the field, most evaluation of news recommendation methods is performed offline using public or proprietary datasets, while online evaluation, A/B tests, and user studies are comparatively rare. Evaluation metrics are typically focused on predictive accuracy, with relatively few papers attempting to quantify "beyond accuracy" aspects like diversity, novelty, or serendipity. We refer readers to the evaluation-oriented survey by Karimi et al. [11] for a more detailed analysis of evaluation practices in the news recommendation literature. However, we highlight two noteworthy exceptions to these trends:

- CLEF NewsReel, a "living" evaluation lab for online and stream-based evaluation using plista's Open Recommendation Platform, where recommendations generated in response to live or replayed requests were required to adhere to realistic time constraints for real-world systems [9].
- Informfully, a news recommendation platform that includes a mobile app with which experimenters can conduct user studies on users they recruit and manage themselves [28]. Unlike POPROX, Informfully does not provide an associated pool of regular users; experimenters need to recruit their own research subjects.

3.4. Values

News recommendation presents a classic multistakeholder recommendation problem [29]: we expect that recommendation platforms will have journalistic objectives distinct from the goals that users might have relative to personalized content [30, 6]. This need is quite explicit for news organizations that have licensing requirements which include support for a statutorily-defined public interest [31, 32, 33]. Researchers have attempted to quantify and represent various journalistic objectives, especially news diversity, to be pursued in tandem with personalization [34, 35, 36].

As a tool for experimental evaluation of recommender systems and especially because of its survey capabilities, POPROX provides the opportunity to explore the consequences of algorithmic choices for journalistic values and users' experience of current events.

4. Practical Challenges

While the POPROX platform allows researchers to develop and deploy their own recommenders to support experiments, the system also needs to provide a satisfactory default news recommendation experience in the personalized newsletters it delivers to subscribers each day. This default experience serves two important purposes: supporting our efforts to recruit, engage, and retain a long-lived participant pool; and providing a solid base for researchers to use in building their own recommenders.

In this section, we outline some of the challenges we faced in building the default experience for POPROX and highlight where we were (and were not) able to rely on solutions from the research literature. We also describe our current solutions and their consequences for the platform.

4.1. Training A Recommendation Model

Key Issues In order to engage and retain users, the platform's default recommendation experience should present relevant news articles to POPROX subscribers. Relevance modeling is an important component of a recommendation system that does so, but since POPROX is a brand-new platform, we have not yet collected sufficient user behavior data to form a dataset that can be used for training or offline evaluation of recommendation models.

Moreover, our system both collects and uses data types that are not present in public datasets. Beyond article text, our AP news feed contains a variety of metadata, including named entities and topics, but the tags provided with Associated Press articles are substantially different from what is present in MIND and other available datasets. Our system also allows users to express explicit topic preferences that are not present in any public dataset as far as we know. We hoped to train a recommendation model that takes advantage of these data types to better understand which articles are relevant for which readers.

Relevant Literature While headlines are commonly used as inputs to news recommendation models, categorical features like named entities and subject/topic categories are provided in many public datasets. While some models do use combined textual and categorical features, we found few options for models that used all of these input types in combination and were suitable for deployment in a live system.

Models that incorporate a wide range of input data beyond headlines tend to be graph-based models that learn embeddings or weights directly for user and/or item IDs, instead of using mechanisms to compose user and item representations from history and content features. Models with article ID features are viable for offline datasets with static content pools, but do not provide a viable recommendation strategy in a live system that must select from fresh items each day. Models with user ID features are potentially workable in the context of a live system, but would require online learning approaches or daily retraining/fine-tuning to keep user interest profiles up to date, all of which are beyond the current capabilities of the POPROX system.

Our Approach In end, we chose to train the NRMS model [37] on the MIND dataset, relying on transfer learning from MIND to AP data, which has worked well enough to get us started. NRMS encodes news articles by embedding the words in their headlines, contextualizing the word embeddings with self-attention, and condensing them into a single article embedding with additive attention. Users are encoded similarly based on their clicked news articles by compressing article embeddings into a single user embedding with similar attention mechanisms. Candidate articles are then scored via dot products between user and article embeddings in a fashion similar to matrix factorization.

Consequences The NRMS user and news encoders solely use article headlines, so we are unable to make use of much of the article data that we receive in our AP news feed. Various work-arounds that we considered, such as augmenting headline text with metadata in training, were not workable with the MIND dataset because its metadata is different from what AP provides. Our experience with the recommended content suggests that the model's single-embedding user representation may be biased toward some topics over others, and not well-suited to representing interest in multiple distinct and

potentially non-overlapping topic areas. NRMS does not provide an obvious way to incorporate explicit user preference signals, so we have developed some workable-but-not-ideal approaches to providing such functionality in order to support user onboarding, which we describe below.

4.2. Providing User Preference Controls

Key Issues New users sign up for POPROX and expect to get something reasonable in their news feed right away, which presents a risk of low engagement, poor retention, and high churn in the subscriber base if we fail to meet that expectation. In our initial design work, we quickly determined that "something reasonable" meant allowing users to declare interests across news topics: some users want to see sports news every day, some want to avoid it altogether. Since POPROX is intended to be a platform for experimenting with a wide range of models and interfaces (including experiments with the user onboarding process), we had to represent preferences in a way that is not overly tied to any one specific model or experience. We also needed to have a way to elicit these topic preferences efficiently and to allow users to modify them over time.

Relevant Literature Industrial recommender systems from both publishers and news aggregators do commonly provide user interface controls that enable readers to explicitly declare their interests, but we have been unable to identify news recommendation research papers that incorporate declared user interests (like those collected during an onboarding process) as preference signals. We hypothesize that this may be related to the prevalence of implicit feedback and unavailability of such explicit preference features in public datasets. We also did not find good empirical evidence for the relative effectiveness of different preference elicitation strategies, in news or other recommendation domains.

We did find strictly attribute-based recommenders, where users can give explicit feedback on attributes, and implicit feedback (item clicks) is decomposed into attribute weight adjustments based on the attribute values of the clicked items [38], and there are examples of using query-specific attribute information for filtering recommendations [39]. None of these use cases involve the creation of explicit standing preferences to be integrated into the recommendation process.

We also found reranking approaches for improving the calibration of recommendations relative to categories of user interest. For example, the greedy reranking algorithm in [40] uses a topic distribution of the user's ratings to produce recommendations that are distributed in genre similarly to the user's profile. This technique could be extended to a topic distribution derived from explicit preferences.

Our Approach To elicit topic preferences, the POPROX interface allows subscribers to select from a set of 14 high-level topics that align with the sections provided at the top of the AP News website.[6] Subscribers indicate their interest in these topics using 5-point response scales ranging from "Not at all interested" to "Very interested". In the interest of providing users a measure of control over the articles they receive, we also allow subscribers to edit these preferences throughout their subscription period.

In order to make user preferences influence recommended content, we added a separate topic-based ranking pipeline whose output is merged with a click-based ranking pipeline before selecting recommended articles. The topic-based pipeline treats textual definitions of AP topics as headlines and shoehorns them into the NRMS article embedding space with the model's news encoder, then applies the remainder of the model as usual to produce an estimated topic-based relevance score.

In designing the topical ranker pipeline, we have drawn inspiration from the use of negative feedback in the NRNF model [41] since it is difficult to represent positive and negative signals together in the same user embedding. We therefore use separate embeddings and scorers to estimate interest and disinterest, and the resulting interest and disinterest scores are then combined via subtraction.

Consequences With an initial implementation of topic preferences in place, we now face the challenge of ensuring that they work as expected. Our topical ranker pipeline seems to work acceptably well

[6] https://apnews.com/

for users with narrow interests, but it is unclear how it could be extended. We would like to allow for more open-ended means of interest expression including named entities, locations, and others but we anticipate significant effort would be required to link those entities with descriptions or definitions (e.g. from an external knowledge base.) We are also not confident that a single-embedding user representation can adequately reflect a broader range of finer-grained interests.

Beyond the technical challenges, determining an appropriate blend of explicit preferences with implicit feedback is not straightforward. The recommender should honor user preferences to some extent and updating preferences should affect what recommendations are delivered in a noticeable way, but few offline accuracy metrics account for fidelity to explicit preferences. Furthermore, giving users more control over the news they receive may come at the cost of recommending informative content that, taken as a whole, adequately embodies journalistic values and fulfills the important roles of news providers beyond user engagement. For these reasons, we continue to rely in large part on "taste-testing" the recommendations but do not find this approach entirely satisfactory.

4.3. Combining Curation With Personalization

Key Issues News publishers and platforms take editorial stances not only in the news they cover and feature but also in the ways that news content is structured and presented for readers. Although recommendation technology expands the range of possibilities, the specific ways that personalization is deployed still reflects chosen resolutions of tensions between different logics and values, such as the tension between user engagement and the duty to inform. As a news recommendation research platform, POPROX is no different, and we are aware of the need to make informed, intentional decisions and to be explicit about the stances we take in designing the ways that articles are selected and displayed.

This presents significant practical challenges, since journalism and recommender systems have historically approached these issues in different ways. Newspapers and their digital equivalents have largely relied on the idea of sections, enabling curation across a range of diverse topics by providing different places to feature different kinds of news. Recommender systems research has heavily emphasized ranking a single recommendation list and mainly investigated ways to blend or balance multiple objectives when determining the order of items within. Finding appropriate ways to combine curation with personalization using these (and other) approaches remains an area of active investigation and exploration in both fields.

Relevant Literature We looked to several areas of the recommender systems and news recommendation literature for answers and approaches: multi-objective and multi-stakeholder approaches [29], grid/carousel interfaces [42, 43, 44], whole-page optimization [45], values in news recommendation [46, 47], and algorithmic auditing [48], among others. While each was informative and helpful in a general sense, we found few methods or results that were directly relevant to content-structuring approaches commonly taken by real-world news recommendation platforms and products, such as vertical sequencing of the ubiquitous "Top News" and "For You" sections (displaying curated and personalized content respectively) or the use of topical sections matching declared user interests.

Our Approach The current structure of POPROX newsletters contains a single ranked list of news articles selected for each user by our default recommender or by an experiment recommender (when an experiment is active on the platform). To the extent that we currently have an approach to combining editorial curation and recommender-based personalization, it is that we rely on the Associated Press to provide content that reflects their editorial standard and stances and apply a layer of personalization, resulting in a personalized selection of news from a curated content pool. We would like to move beyond this but still have many open questions about how to do so. Our AP news feed provides headline packages featuring the top ten stories for each of a range of topics (e.g. US News, Sports, Entertainment) but does not provide such a list of the overall top stories of the day. As a result, we are not able to fully rely on their editorial curation to determine what to feature as top news in POPROX newsletters, and would need to apply some form of (personalized or non-personalized) algorithmic selection.

Consequences On one hand, presenting only personalized content as we currently do could result in some degree of "filter bubble" issues [49, 50, 51, 52]. We expect these may be mitigated somewhat by recommending news articles from a fairly limited pool, since it is difficult for subscribers to delve too deeply into any single topic when the number of available articles per topic is low.

On the other hand, a potential future structure of the newsletter that incorporates multiple sections would become quite difficult to evaluate using standard accuracy-oriented offline evaluation techniques designed for single top-K lists. While the POPROX platform does provide a suite of tools for online and offline evaluation that includes some alternatives, we are hesitant to make changes to the newsletter that would deprive experimenters using the platform of familiar and useful tools without providing a workable substitute, which would present its own research, development, and validation challenges.

4.4. Assessing User Experiences And Satisfaction

Key Issues Online and offline evaluation methods based on content properties and behavioral tracking are necessary but not sufficient for understanding how recommendations are experienced by their recipients. User surveys therefore form an essential part of our platform, and provide several benefits:

- Because many newsletter consumers simply read the headlines without clicking through to the linked articles, users' satisfaction with the recommended news articles may not always be apparent from their interaction behavior. Surveys provide an additional signal of these "passive" users' satisfaction with their recommendations.
- Users may be interested in an article for several reasons, and a certain algorithm may be particularly good at fulfilling one specific type of user need (e.g. feel-good articles) or at catering to a balanced set of needs/interests. A survey can provide more contextual granularity to users' evaluations that can then be triangulated with their behavior [53, 54].
- Surveys can cover constructs related to the long-term effects of a recommender system, such as its ability to help users explore, understand, and develop their interests in a variety of news topics [55]. They can also include key user demographics and characteristics that can be used to evaluate the fairness of proposed recommendation algorithms and other interventions.

An important experimental advantage (but methodological challenge) of POPROX is the perpetual and longitudinal nature of our studies and therefore of our evaluations. Whereas most user-centered research studies in recommender systems involve short-lived interactions with a system, POPROX applies interventions to a user base that has ongoing interactions with our platform. This increases the realism and ecological validity of implemented studies and allows researchers to track the effects of interventions over time, but also complicates holistic evaluation because surveys must be administered perpetually (before, during, and after a study).

Relevant Literature To our knowledge, little to no published research has considered the longitudinal dimension of user experience in recommender systems, let alone the methodological question of how much time it takes for an intervention to "take hold" and how long an effect may linger post-intervention. Furthermore, while there exists a vast body of research on increasing survey participation in human subjects research [56], our platform much more closely resembles commercial settings in this regard, and much of the research in those settings is proprietary and unpublished.

Our Approach The POPROX platform sends users a short weekly survey, which rotates on a 5-week cycle through constructs related to users' perceptions, experience, and need fulfillment (which is at some occasions measured across the newsletters of the past week, while at other times we ask users to evaluate these constructs for a specific newsletter). On the fifth week of the cycle, the survey system rotates through a series of personal characteristics and demographics, which can be used to select/stratify a user sample (e.g. balance gender in a study, or target users of a certain age group), or

to contextualize evaluations (e.g. evaluate the effect of an algorithmic intervention across users with different personalities). To reduce survey length, we measure each construct with a single item taken from pre-validated scales.

Consequences While we have been able to rely on Qualtrics for core survey functionality, we had to devise our own solutions for creating a rotating schedule of periodic surveys involving nested cycles. Due to the voluntary, unpaid nature of users' participation in the POPROX platform, we have sought to balance user time and effort with our goal of collecting robust data on multiple user experience constructs. Despite our efforts to minimize their length, our initial deployment has shown very low engagement with weekly surveys. We are investigating whether we need to incentivize survey completion in order to improve the "conversion rate" of active readers into survey respondents.

5. Lessons for Future Research

Our work building POPROX and experience attempting to locate and apply research findings to support this effort leads us to several lessons for news recommendation as well as for the broader RecSys community. We expect many of these lessons may be unsurprising to people already working on production news recommenders, but we find them often missing from the literature, and therefore valuable to explicitly articulate for the community and for new researchers and practitioners looking to enter the field.

Treat using the available data as a first-order concern. Our challenges highlight several interrelated issues with the ways that data is used and discussed in research:

- Limiting modeling or evaluation to a subset of the attributes (or instances) in a dataset prevents the dataset's contents and capabilities from serving as a forcing function for developing adaptable models and systems. Accommodating data types beyond least-common-denominator attributes like clicks requires more flexibility than many of the models found in the literature provide.
- Rich item-level data is a natural substrate for facilitating user preferences and feedback, editorial control, business rules, and other functionality that provides human influence on system behavior. When little research is available on the use of a particular type of data, it is difficult to construct or extend recommender systems to provide user and stakeholder controls built on that data.
- Many papers do not provide clear and thorough details on data preparation for modeling and evaluation, as others have noted in the context of evaluation rigor [57, 58]. Thoroughly documenting data preparation decisions, including splitting, missing value imputation, feature engineering, etc. subjects data decisions to peer-review, allowing for feedback and community vetting of data practices. It also aids reproducibility and provides readers with worked examples of effectively using data, helping students and new researchers learn.

Given the importance of data to recommendation and evaluation, promoting more research on data seems promising as a path to impactful improvements in recommender systems that transcend domains and model families. There are some examples in the literature, such as work on feature engineering [59, 60], data minimization [61, 62] and the study of fairness research practicalities by Daniil et al. [63].

Model affordances matter. Real-world recommender systems must be developed with a degree of "mechanical sympathy" for how the models in use work, their strengths and weaknesses, and the interactions between those models and other system components. This is more straightforward with models that provide clear indications of how they can be used or extended to support common system functionality, and more difficult when those indications are lacking. In this regard, both domain-specific and domain-independent recommendation models often leave something to be desired.

On one hand, a domain-specific model may focus the proposed architecture solely on the features commonly available in public datasets from that domain. For example, many news recommendation models use only textual input features and require architectural modifications in order to accept categorical inputs. Others incorporate one specific categorical input (like topics) in a way that is difficult to extend to other categorical features. On the other hand, general purpose click-through rate prediction models accept a wide range of continuous and categorical features but provide few indications about how to construct appropriate input features to support desired system functionality (in general or in specific domains) even when that is possible and well-supported in practice.

We view these issues, at least in part, as a reflection of the field's emphasis on predictive accuracy over examining how models fit into the broader context of real-world systems, where models may either help or hinder practitioners in building systems that embody the requirements, values, and user experiences they aim to realize. We believe this represents an opportunity to investigate and improve the ways that recommendation models support common system design patterns and signal that support to recommender system developers.

Recommendation is more than modeling. Research and evaluation findings often focus on the scores produced by a model, or on rankings derived from those scores. However, in practice the quality of delivered recommendations depends on much more than accuracy, or even non-accuracy properties, of models and their outputs. A personalized news application may:

- Inquire about users' interests and disinterests to collect explicit preference signals
- Select candidate items from multiple sources centered on different news publishers, subjects, or geographical locations
- Exclude items from consideration based on user- or platform-defined criteria (e.g. "Hide all stories from [this source]")
- Estimate the likelihood of types of user engagement such as reading, saving, and sharing
- Recommend content across several modalities including articles, podcasts, and videos
- Satisfy list construction and ordering constraints imposed by product requirements (e.g. putting top news first, including certain percentages of local/national/international news)
- Balance multiple objectives for different stakeholders including readers, editors, journalists, and advertisers

Implementing this range of functionality requires a significant amount of software that is not frequently described in the literature or supported by recommender system frameworks or toolkits. To the best of our knowledge, there are only two academic or industrial frameworks (NVIDIA Merlin [3] and recent versions of LensKit [64]) that provide tools for building more complex multi-stage pipelines that combine models with other components, and neither offers specific support for news recommendation. As part of the POPROX platform, we aim to provide such a toolkit[7] for news recommendation researchers in order to support their efforts to build and evaluate news recommenders that deliver recommendations to users in the context of a live system.

Live systems present opportunities to apply a spectrum of evaluation methods. There are many well-known and frequently used evaluation approaches, including: accuracy and "beyond accuracy" offline evaluation metrics; online behavioral metrics and A/B testing; and user studies and surveys. We see opportunities to bridge the gaps between these methods by expanding both the system components included in evaluations and the ways that those components are evaluated (independently and together). We highlight three that are relevant for our system and that we believe are applicable in many contexts:

[7] https://github.com/CCRI-POPROX/poprox-recommender

- **Correctness testing:** The idea of evaluating whether a model or recommender meets particular functional requirements beyond statistical evaluation is relatively new to the field, but Michiels et al. [65] proposed the idea of recommender system test suites for specific behavior, which could be extended with domain- or application-specific behavioral evaluations and acceptance tests.
- **End-to-end offline evaluation:** Live systems with personalization features often use multi-stage recommenders, including retrieval and ranking models in conjunction with filters, re-rankers, and business logic. With appropriate system construction, these additional components can be included in offline evaluations to measure the accuracy and "beyond accuracy" impacts of design choices within and between these system components.
- **Distributional evaluation:** In real-world contexts, we are often concerned not only with aggregate performance and quality metrics but also with identifying which users and items are well-served or under-served in order to ensure that recommendations meet a minimum quality bar. For example, we have noticed biases in our system toward certain topics and away from others that we believe may originate from imbalances between topical categories in training data. Assessing such biases and disparities in utility across different categories and stakeholders calls for examining not only point estimates but also distributions, as suggested by Ekstrand et al. [66].

6. Conclusion

In this case study, we have described our experiences creating POPROX and some of the challenges we faced in doing so. Although news recommendation has been an active topic of research for several decades, we did not receive as much help from the research literature as we hoped. Part of this gap was expected, as we knew that news recommendation has unique characteristics among recommendation domains and that creating a long-running live recommender system as research infrastructure was not something that many research groups had attempted. Part of the gap was surprising though, since some features that are commonplace in commercial news recommendation products and platforms remain under-researched and have proved difficult to implement with published techniques.

We believe there are multiple reasons why these challenges exist. One is certainly the focus on the "horse race" of chasing an evaluation benchmark on a narrowly defined task. This kind of focus can lead to methods and models lacking functional characteristics that would make them suitable for deployment in real systems. We have also often encountered a push in recommender systems research (through reviewing, community expectations, etc.) to show that findings are generalizable: that they apply across multiple data sets, domains, and/or applications. Effectively building recommender systems that serve real users, however, requires deep and specific engagement not just with a domain in general, but with the particular characteristics of specific datasets, applications, and user communities.

While we recognize and have encountered significant systemic challenges and barriers, we nonetheless encourage more researchers to start, or get involved in, longer-term research infrastructure projects that build and operate live recommender systems. Engaging in a multi-year, interdisciplinary recommender systems effort provides a catalyst for integrating approaches from many areas and fields, which also serves to highlight what is missing. In the context of a long-lived system, different categories of issues surface that would not arise in a standalone research project over the course of a semester or a year.

In aggregate, these issues and gaps offer a different perspective on the extent to which RecSys research is making cumulative progress: while the field is making considerable advances on a number of important but narrowly defined recommendation problems, these advances do not yet "add up" to the knowledge base that is needed to build a real-world recommender system. By encouraging more researchers to engage in complex real-world projects, we hope to spur the RecSys community to become more aware of and attentive to research that fills the gaps that stand between the current state of the field and greater real-world impact.

Acknowledgments

This work is based on research supported by the National Science Foundation under Grant Nos. IIS 22-32551, 22-32552, 22-32555, and 24-09199. We are grateful to the rest of the POPROX team for their ongoing collaboration in this effort and many discussions that have informed this paper.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] R. Warner, Tall Man Tyrell Biggs thinks his height will add new dimension to fight, Orangeburg Times and Democrat (1987) 7.

[2] J. Beel, S. Dinesh, Real-world recommender systems for academia: The pain and gain in building, operating, and researching them., in: 5th Int. Workshop Bibliometric-Enhanced Inf. Retr. (BIR), 2017. URL: https://ceur-ws.org/Vol-1823/paper1.pdf.

[3] K. Higley, E. Oldridge, R. Ak, S. Rabhi, G. de Souza Pereira Moreira, Building and deploying a multi-stage recommender system with Merlin, in: Proceedings of the 16th ACM Conference on Recommender Systems, RecSys '22, Association for Computing Machinery, 2022, p. 632–635. doi:10.1145/3523227.3551468.

[4] J. Shah, M. González-Fierro, Adoption of recommendation systems: Observations, trends and leveling the playing field, in: Proceedings of the RecSys INTROSPECTIVES Workshop, 2024. URL: https://introspectives.github.io/2024/uploads/adoption_of_recommender_systems.pdf.

[5] C. St. Aubin, J. Liedke, News platform fact sheet, 2024. URL: https://www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet/.

[6] E. Mitova, S. Blassnig, E. Strikovic, A. Urman, C. de Vreese, F. Esser, When worlds collide: Journalistic, market, and tech logics in the adoption of news recommender systems, Journalism Studies 24 (2023) 1957–1976.

[7] H. Vandenbroucke, A. Smets, It's (not) all about that CTR: A multi-stakeholder perspective on news recommender metrics, in: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24, Association for Computing Machinery, 2024, p. 999–1003. doi:10.1145/3640457.3688183.

[8] T. Kamba, K. Bharat, M. C. Albers, The Krakatoa Chronicle: An interactive personalized newspaper on the web, in: Proceedings of the Fourth International Conference on World Wide Web, WWW4, Association for Computing Machinery, 1995, p. 159–170. doi:10.1145/3592626.3592638.

[9] F. Hopfgartner, T. Brodt, J. Seiler, B. Kille, A. Lommatzsch, M. Larson, R. Turrin, A. Serény, Benchmarking news recommendations: The CLEF NewsREEL use case, SIGIR Forum 49 (2016) 129–136. doi:10.1145/2888422.2888443.

[10] J. Kruse, K. Lindskow, S. Kalloori, M. Polignano, C. Pomo, A. Srivastava, A. Uppal, M. R. Andersen, J. Frellsen, RecSys Challenge 2024: Balancing accuracy and editorial values in news recommendations, in: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24, Association for Computing Machinery, 2024, p. 1195–1199. doi:10.1145/3640457.3687164.

[11] M. Karimi, D. Jannach, M. Jugovac, News recommender systems – survey and roads ahead, Information Processing & Management 54 (2018) 1203–1227. doi:https://doi.org/10.1016/j.ipm.2018.04.008.

[12] M. Li, L. Wang, A survey on personalized news recommendation technology, IEEE Access 7 (2019) 145861–145879. doi:10.1109/ACCESS.2019.2944927.

[13] C. Feng, M. Khan, A. U. Rahman, A. Ahmad, News recommendation systems - accomplishments, challenges & future directions, IEEE Access 8 (2020) 16702–16725. doi:10.1109/ACCESS.2020.2967792.

[14] S. Raza, C. Ding, News recommender system: a review of recent progress, challenges, and opportunities, Artificial Intelligence Review 55 (2022) 749–800.
[15] X. Meng, H. Huo, X. Zhang, W. Wang, J. Zhu, A survey of personalized news recommendation, Data Science and Engineering 8 (2023) 396–416.
[16] C. Wu, F. Wu, Y. Huang, X. Xie, Personalized news recommendation: Methods and challenges, ACM Trans. Inf. Syst. 41 (2023). doi:10.1145/3530257.
[17] A. Iana, M. Alam, H. Paulheim, A survey on knowledge-aware news recommender systems, Semantic Web 15 (2024) 21–82. doi:10.3233/SW-222991.
[18] C. Bauer, C. Bagchi, O. A. Hundogan, K. van Es, Where are the values? a systematic literature review on news recommender systems, ACM Trans. Recomm. Syst. 2 (2024). doi:10.1145/3654805.
[19] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, M. Zhou, MIND: A large-scale dataset for news recommendation, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3597–3606. doi:10.18653/v1/2020.acl-main.331.
[20] J. Kruse, K. Lindskow, S. Kalloori, M. Polignano, C. Pomo, A. Srivastava, A. Uppal, M. R. Andersen, J. Frellsen, EB-NeRD a large-scale dataset for news recommendation, in: Proceedings of the Recommender Systems Challenge 2024, RecSysChallenge '24, Association for Computing Machinery, 2024, p. 1–11. doi:10.1145/3687151.3687152.
[21] J. A. Gulla, L. Zhang, P. Liu, O. Özgöbek, X. Su, The adressa dataset for news recommendation, in: Proceedings of the International Conference on Web Intelligence, WI '17, Association for Computing Machinery, 2017, p. 1042–1048. doi:10.1145/3106426.3109436.
[22] B. Kille, F. Hopfgartner, T. Brodt, T. Heintz, The plista dataset, in: Proceedings of the 2013 International News Recommender Systems Workshop and Challenge, NRS '13, Association for Computing Machinery, 2013, p. 16–23. doi:10.1145/2516641.2516643.
[23] J. P. Lucas, J. F. G. da Silva, L. F. de Figueiredo, NPR: A news portal recommendations dataset., in: Proceedings of the RecSys NORMalize Workshop, 2023. URL: https://ceur-ws.org/Vol-3639/paper6.pdf.
[24] L. Heitz, N. Mattis, O. Inel, W. van Atteveldt, IDEA - informfully dataset with enhanced attributes, in: Proceedings of the RecSys NORMalize Workshop, 2024. URL: https://ceur-ws.org/Vol-3898/paper1.pdf.
[25] G. de Souza Pereira Moreira, F. Ferreira, A. M. da Cunha, News session-based recommendations using deep neural networks, in: Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems, DLRS 2018, Association for Computing Machinery, 2018, p. 15–23. doi:10.1145/3270323.3270328.
[26] J. Xun, S. Zhang, Z. Zhao, J. Zhu, Q. Zhang, J. Li, X. He, X. He, T.-S. Chua, F. Wu, Why do we click: Visual impression-aware news recommendation, in: Proceedings of the 29th ACM International Conference on Multimedia, MM '21, Association for Computing Machinery, 2021, p. 3881–3890. doi:10.1145/3474085.3475514.
[27] S. Han, W. Huang, X. Luan, Vlsnr: Vision-linguistics coordination time sequence-aware news recommendation, arXiv preprint arXiv:2210.02946 (2022).
[28] L. Heitz, J. A. Croci, M. Sachdeva, A. Bernstein, Informfully - research platform for reproducible user studies, in: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24, Association for Computing Machinery, 2024, p. 660–669. doi:10.1145/3640457.3688066.
[29] H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, L. Pizzato, Multistakeholder recommendation: Survey and research directions, User Modeling and User-Adapted Interaction 30 (2020) 127–158.
[30] L. A. Møller, Designing algorithmic editors: How newspapers embed and encode journalistic values into news recommender systems, Digital Journalism 12 (2024) 926–944.
[31] J. K. Sørensen, Public service media, diversity and algorithmic recommendation: Tensions between editorial principles and algorithms in European PSM organizations, in: Proceedings of 7th International Workshop on News Recommendation and Analytics (INRA 2019), volume 2554, 2019,

pp. 6–11.
[32] A. Grün, X. Neufeld, Transparently serving the public: Enhancing public service media values through exploration, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, 2023, p. 1045–1048. doi:10.1145/3604915.3610243.
[33] K. Tallerås, T. Colbjørnsen, K. Oterholm, H. Larsen, Cultural policies, social missions, algorithms and discretion: What should public service institutions recommend?, in: International Conference on Information, Springer, 2020, pp. 588–595.
[34] L. Heitz, J. A. Lischka, A. Birrer, B. Paudel, S. Tolmeijer, L. Laugwitz, A. Bernstein, Benefits of diverse news recommendations for democracy: A user study, Digital Journalism 10 (2022) 1710–1730.
[35] N. Helberger, On the democratic role of news recommenders, in: Algorithms, automation, and news, Routledge, 2021, pp. 14–33.
[36] S. Vrijenhoek, G. Bénédict, M. Gutierrez Granada, D. Odijk, M. De Rijke, Radio – rank-aware divergence metrics to measure normative diversity in news recommendations, in: Proceedings of the 16th ACM Conference on Recommender Systems, RecSys '22, Association for Computing Machinery, 2022, p. 208–219. doi:10.1145/3523227.3546780.
[37] C. Wu, F. Wu, S. Ge, T. Qi, Y. Huang, X. Xie, Neural news recommendation with multi-head self-attention, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6389–6394. doi:10.18653/v1/D19-1671.
[38] B. P. Knijnenburg, N. J. Reijmer, M. C. Willemsen, Each to his own: how different users call for different interaction methods in recommender systems, in: Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11, Association for Computing Machinery, 2011, p. 141–148. doi:10.1145/2043932.2043960.
[39] B. Loepp, K. Herrmanny, J. Ziegler, Blended recommending: Integrating interactive information filtering and algorithmic recommender techniques, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, Association for Computing Machinery, 2015, p. 975–984. doi:10.1145/2702123.2702496.
[40] H. Steck, Calibrated recommendations, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys '18, Association for Computing Machinery, 2018, p. 154–162. doi:10.1145/3240323.3240372.
[41] C. Wu, F. Wu, Y. Huang, X. Xie, Neural news recommendation with negative feedback, CCF Transactions on Pervasive Computing and Interaction 2 (2020) 178–188. doi:10.1007/s42486-020-00044-0.
[42] A. Raj, M. D. Ekstrand, Towards optimizing ranking in grid-layout for provider-side fairness, in: Proceedings of the 46th European Conference on Information Retrieval, volume 14612 of *LNCS*, Springer, 2024, pp. 90–105. doi:10.1007/978-3-031-56069-9_7.
[43] B. Rahdari, P. Brusilovsky, CARE: An infrastructure for evaluation of carousel-based recommender interfaces, in: Companion Proceedings of the 29th International Conference on Intelligent User Interfaces, IUI '24 Companion, Association for Computing Machinery, 2024, pp. 41–44. doi:10.1145/3640544.3645223.
[44] B. Loepp, J. Ziegler, How users ride the carousel: Exploring the design of multi-list recommender interfaces from a user perspective, RecSys '23, Association for Computing Machinery, 2023, p. 1090–1095. doi:10.1145/3604915.3610638.
[45] W. Ding, D. Govindaraj, S. V. N. Vishwanathan, Whole Page Optimization with Global Constraints, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, Association for Computing Machinery, 2019, pp. 3153–3161. doi:10.1145/3292500.3330675.
[46] S. Vrijenhoek, M. Kaya, N. Metoui, J. Möller, D. Odijk, N. Helberger, Recommenders with a mission: Assessing diversity in news recommendations, in: Proceedings of the 2021 Conference on Human

Information Interaction and Retrieval, Association for Computing Machinery, 2021, pp. 173–183. doi:10.1145/3406522.3446019.

[47] J. Stray, A. Halevy, P. Assar, D. Hadfield-Menell, C. Boutilier, A. Ashar, C. Bakalar, L. Beattie, M. Ekstrand, C. Leibowicz, C. Moon Sehat, S. Johansen, L. Kerlin, D. Vickrey, S. Singh, S. Vrijenhoek, A. Zhang, M. Andrus, N. Helberger, P. Proutskova, T. Mitra, N. Vasan, Building human values into recommender systems: An interdisciplinary synthesis, ACM Transactions on Recommender Systems 2 (2024) 20:1–57. doi:10.1145/3632297.

[48] E. Lurie, E. Mustafaraj, Opening Up the Black Box: Auditing Google's Top Stories Algorithm, in: The Thirty-Second International Florida Artificial Intelligence Research Society Conference (FLAIRS-32), AAAI, 2019. URL: https://aaai.org/papers/376-flairs-2019-18316/.

[49] M. van Alstyne, E. Brynjolfsson, Global village or cyber-balkans? Modeling and measuring the integration of electronic communities, Management Science 51 (2005) 851–868. doi:10.1287/mnsc.1050.0363.

[50] S. Flaxman, S. Goel, J. M. Rao, Filter bubbles, echo chambers, and online news consumption, Public opinion quarterly 80 (2016) 298–320.

[51] L. Michiels, J. Leysen, A. Smets, B. Goethals, What are filter bubbles really? A review of the conceptual and empirical work, in: Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '22 Adjunct, ACM, 2022, pp. 274–279. doi:10.1145/3511047.3538028.

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

[53] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, C. Newell, Explaining the user experience of recommender systems, User Modeling and User-Adapted Interaction 22 (2012) 441–504. doi:10.1007/s11257-011-9118-4.

[54] B. P. Knijnenburg, M. C. Willemsen, Evaluating Recommender Systems with User Experiments, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer US, 2015, pp. 309–352. doi:10.1007/978-1-4899-7637-6_9.

[55] B. P. Knijnenburg, S. Sivakumar, D. Wilkinson, Recommender systems for self-actualization, in: Proceedings of the 10th ACM Conference on Recommender Systems, RecSys '16, Association for Computing Machinery, 2016, p. 11–14. doi:10.1145/2959100.2959189.

[56] R. M. Groves, R. B. Cialdini, M. P. Couper, Understanding the Decision to Participate in a Survey, Public Opinion Quarterly 56 (1992) 475–495. doi:10.1086/269338.

[57] Z. Sun, D. Yu, H. Fang, J. Yang, X. Qu, J. Zhang, C. Geng, Are we evaluating rigorously? benchmarking recommendation for reproducible evaluation and fair comparison, in: Proceedings of the 14th ACM Conference on Recommender Systems, RecSys '20, Association for Computing Machinery, 2020, p. 23–32. doi:10.1145/3383313.3412489.

[58] E. Y. H. Tan, A Critical Study on MovieLens Dataset for Recommender Systems, B.Eng. FYP, Nanyang Technological University, Singapore, 2023. URL: https://hdl.handle.net/10356/171942. arXiv:10356/171942.

[59] B. Schifferer, G. Titericz, C. Deotte, C. Henkel, K. Onodera, J. Liu, B. Tunguz, E. Oldridge, G. De Souza Pereira Moreira, A. Erdem, GPU accelerated feature engineering and training for recommender systems, in: Proceedings of the Recommender Systems Challenge 2020, RecSysChallenge '20, Association for Computing Machinery, 2020, p. 16–23. doi:10.1145/3415959.3415996.

[60] T. Verdonck, B. Baesens, M. Óskarsdóttir, S. vanden Broucke, Special issue on feature engineering editorial, Machine Learning 113 (2024) 3917–3928. doi:10.1007/s10994-021-06042-2.

[61] A. J. Biega, P. Potash, H. Daumé, F. Diaz, M. Finck, Operationalizing the legal principle of data minimization for personalization, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, 2020, p. 399–408. doi:10.1145/3397271.3401034.

[62] N. Sonboli, S. Li, M. Elahi, A. Biega, The trade-off between data minimization and fairness in collaborative filtering, 2024. URL: https://arxiv.org/abs/2410.07182. arXiv:2410.07182.

[63] S. Daniil, M. Slokom, M. Cuper, C. Liem, J. van Ossenbruggen, L. Hollink, On the challenges

of studying bias in Recommender Systems: The effect of data characteristics and algorithm configuration, Information Retrieval Research 1 (2025) 3–27. doi:10.54195/irrj.19607.

[64] M. D. Ekstrand, Lenskit for python: Next-generation software for recommender systems experiments, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, Association for Computing Machinery, 2020, p. 2999–3006. doi:10.1145/3340531.3412778.

[65] L. Michiels, R. Verachtert, A. Ferraro, K. Falk, B. Goethals, A Framework and Toolkit for Testing the Correctness of Recommendation Algorithms, ACM Trans. Recomm. Syst. (2023). doi:10.1145/3591109.

[66] M. D. Ekstrand, B. Carterette, F. Diaz, Distributionally-informed recommender system evaluation, ACM Trans. Recomm. Syst. 2 (2024). doi:10.1145/3613455.